On the transition to a sustainable economy

Boomsma, Mirthe

DOI:
10.26116/center-lis-2113

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

CentER

On the transition to a sustainable economy: Field experimental evidence on behavioral interventions

MIRRON ADRIANA BOOMSMA

Tilburg University

On the transition to a sustainable economy: Field experimental evidence on behavioral interventions

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op woensdag 8 september 2021 om 16.30 uur

door

Mirron Adriana Boomsma,
geboren te 's-Gravenhage

Promotor:
Prof. dr. D.P. van Soest (Tilburg University)

Copromotor:
Dr. B.A. Vollaard (Tilburg University)

Leden Promotiecommissie:
Prof. dr. A.J. Dur (Erasmus University Rotterdam)
Prof. dr. J.A. List (The University of Chicago)
Prof. dr. S. Rosenkranz (Utrecht University)
Prof. dr. S. Suetens (Tilburg University)
Dr. J.A. Bouma (Planbureau voor de Leefomgeving)

This thesis was made possible through financial support from PBL, The Netherlands Environmental Assessment Agency. Deze thesis is mede mogelijk gemaakt door een financiële bijdrage van het Planbureau voor de Leefomgeving (PBL).

Abstract

This dissertation studies the impact of behavioral interventions on waste sorting and energy conservation, two domains where sustained environmental conservation has the potential to substantially reduce social costs. The interventions are evaluated by means of field experiments. The first essay investigates the relative impact of behavioral interventions versus neoclassical interventions. It finds that interventions that draw on extrinsic motivations have an immediate and sizable effect on waste sorting behavior, but also that the average treatment effects attenuate steeply over time. In contrast, the essay finds equally sizeable yet long-lasting effects of a treatment designed to increase households' intrinsic motivation to sort waste. The second essay analyzes the effect of social learning interventions. It considers two interventions, one aimed at leveraging social learning via role models and a second one via feedback on the prevalence of organic waste sorting in the household's direct vicinity. The essay finds that both interventions increase waste sorting in the short run, but only the social feedback's impact is long-lasting. The third essay analyzes residential energy consumption, and how real-time disaggregated consumption feedback corrects consumer mistakes in this domain. The essay finds this feedback, provided by way of in-home displays, to reduce household energy consumption. The savings are largest on gas consumption, and the evidence suggests the effect to reflect reductions in space heating. The three essays are preceded by an introductory chapter that introduces the topic of study and the field experimental methodology. The dissertation closes with a concluding chapter that reviews its contribution to the literature on behavioral interventions and its implications for Dutch environmental policy.

Acknowledgements

It is strange to reflect on a PhD written for a large part during a worldwide pandemic. The days on campus with familiar faces seem a long time ago, with most interactions currently taking place by means of Zoom or simply by phone. Yet thinking back to the over 6 years I spent at Tilburg University, working towards this PhD, I cannot feel differently than tremendously grateful for the people that I have met, and the amount I was able to learn. Before I move forward, I would like to direct some words of gratitude to those that supported me during this journey.

To start off my words of thanks, I would like to thank PBL (the Netherlands Environment Assessment Agency) for the fruitful collaboration and the generous funding that made my PhD research in Tilburg possible.

Then, of course, a word of thanks to my first supervisor. Daan, you enabled me to work independently, but always made time for me when I needed direction. Thank you for the insightful discussions regarding field experimental design and analysis, and all the comments and feedback you have provided me with over the years. I am very grateful for the opportunity to step into the field, and to learn as much as I did.

There are more researchers I would like to thank. I am greatly indebted to Ben, Jetske, and Kees. Ben, thank you for your insightful comments, and helping me become a better writer and empirical researcher. Jetske, your kind words of advice have been extremely valuable to me during this trajectory. Thank you. And Kees, thank you for your perseverance and your support. You were key in making sure that our display research-project survived until the very end, and our talks enriched my understanding of energy research. I also thank all committee members, for their valuable comments that improved this doctoral thesis and increased its societal relevance.

This thesis bundles three field experiments, and therefore many people 'from the field' fulfilled important roles in ensuring that these projects came to a successful outcome. Regarding the waste management experiments, I direct a great amount of thanks to everyone involved in the Afval in de hoogbouw project. Especially, Gijs and Addie for project management, and Marn, for sharing his insights and his patience handling my email requests. In addition, special thanks to Cees, for our conversations about psychology. Regarding the display project, I am grateful to the whole EDA consortium, and to Ernestine and Erik of Quintens for managing the project. Also the University of Groningen and Energysense were important partners, that helped us with the project website and the energy data infrastructure.

During my PhD, I got the opportunity to go on a research visit at the University of Chicago. I would like to express my gratitude for receiving the Aart de Zeeuw scholarship, for John allowing me to visit his group, and the fantastic learning experience that was my research visit. I am also grateful to Winnie, who provided me with a great place to stay during my time there.

I also thank my cohort of the Research Master for their companionship over the years. Especially, thank you Dorothee, Laura, Santiago, Oliver, Thijs, and Sophie – our little group, shirking in practice – for the coffees, lunches, Columbian food, econometrics consulting and so much more. You made Tilburg the home it was. Sophie, you were the best office mate one could have during the PhD. But Albert, my pandemic office mate, you are a close second. Thank you for helping me out with practical matters at the office, that greatly reduced stress on my part during busy times.

Thank you to other friends, and family. Thanks to the girls of Novem Invictae, especially Caroline, Lisanne and Hanna. Your (video) phone calls, texts, talks over dinners and even flowers have provided me with great support. But also thanks to Harold and Nicoline. Our diners together reminded me to not take myself so seriously, and provided a much welcomed distraction during stressful times. A warm thank you to my family, who have always provided me with their unconditional support. I feel lucky having been brought up by parents who always believe in chances and opportunities. I would not have made it this far without you.

Finally, I would like to thank Marcel. Your idea to invest in a good coffee machine has been one of the best decisions made during my PhD – especially since working at home became the norm. On more serious terms, thank you, so much, for your love and support. You have been by my side from the end of the Research Master until the very end of the PhD. Surrounded by your love, this whole trajectory has been a lot easier than it would have been without. Being with you truly is a grand adventure, and an adventure that continues, now that this PhD adventure ends.

Mirthe Boomsma
Eindhoven, June 13, 2021

Contents

List of Tables

Chapter 4

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 1

Introduction

Behavioral economics has broadened the economist's view of human decision making. Actual behavior may deviate from that of 'homo economicus' because of bounded self-interest, bounded rationality and bounded self-control (Mullainathan & Thaler, 2001). This broader perspective provides economists with new instruments to alter behavior, with better insight into how traditional policy instruments (such as taxes) would function in actual markets, and allows for better analysis of the welfare implications of policy changes (Chetty, 2015). Behavioral economics also offers another rationale for policy intervention, next to traditional market failures such as externalities: the existence of behavioral biases which prevent the efficient outcome to materialize (Shogren & Taylor, 2008). People may know what is in their best interest, but they may not be able to commit to their preferred course of action, resulting in a welfare loss.

The insights of behavioral economics are of great relevance to the field of environmental economics, where inefficient outcomes are often associated with substantial environmental and thus social costs. For instance, the behavioral economics insight that households are not always selfish is fruitful ground for policies that appeal to social preferences, which can be harnessed to induce households to incorporate the environmental costs of their decisions. In addition, the insight that households may not always make rational use of all available information highlights the potential of interventions that help households make decisions that are in line with their preferences. When behavioral biases induce households to over-consume environmental resources, mitigating these private behavioral failures also lowers the harm from environmental externalities.

In this thesis, I analyze the effects of behavioral interventions on environmental conservation behaviors relevant for policy. Can these interventions influence household behavior? And if so, is the behavioral change persistent over time?

1.1 Topic of study

I examine the effectiveness of behavioral interventions in two domains in which sustained environmental conservation has the potential to substantially reduce social cost: residential waste sorting and residential energy conservation. If households fail to sort their waste into the reusable and non-reusable (residual) waste flows, valuable scarce resources are squandered. Non-sorted or contaminated waste flows also increase environmental pollution. Polluted waste flows may leak into the ground (or into ground water) when land-filled or processed to compost, but pollutants can also be emitted to the air in the form of methane and carbon-dioxide in case of waste incineration (Bijleveld et al., 2021). Residential energy consumption is associated especially with greenhouse gas emissions, with the amount of emissions depending on the energy mix employed (Andor et al., 2020).[1]

The behavioral interventions studied in this thesis capitalize on the bounded self-interest and bounded rationality of households to induce better environmental outcomes. Specifically, in the domain of residential waste sorting I study the effects of (i) appeals to extrinsic and intrinsic motives, and (ii) social learning. In the domain of energy conservation, I analyze the impact of an information treatment in the form of feedback on a household's energy consumption.

1.2 Methodology

I provide evidence for the effectiveness of these behavioral interventions by means of field experiments. Field experiments use random treatment assignment in a setting that captures important characteristics of the real world (List & Reiley, 2008). Collecting data in the field as opposed to the laboratory is expected to make the results similar to those in the policy context and

[1] In addition to the environmental externality, electricity consumption also yields a congestion externality. Supply always needs to meet demand, and hence the volatility of demand makes electricity consumption quite expensive – and especially so during peak hours. In addition, the sustainability of the energy grid is an important policy matter due to the shift to wind and sun energy carriers combined with a growing energy demand (Ministry of Economic Affairs of the Netherlands, 2016).

thus to score high on ecological validity (Roe & Just, 2009). For instance, both the energy conservation and waste sorting experiments discussed in this thesis, take place within the household home. The context may be especially relevant as habits may play an important role in the environmental behaviors that I analyze. Habits are ingrained behaviors that take place in a stable context (Verplanken & Aarts, 1999), and field experiments allow me to research such behaviors in their natural environment. Field experiments have the real-world characteristic in common with observational data. The advantage of using the experimental approach is the control over the way households are (randomly) assigned to treatment. This randomization feature of field experiments enables me to provide credible estimates of the causal effects of the researched interventions. These causal effects explain what would happen when these interventions are implemented, compared to the alternative situation (also referred to as the counterfactual) in which the policy would not have been implemented.

To examine how experiments can help with the identification of causal estimates, it is useful to analyze the the fundamental identification problem in causal inference (Rubin, 1974; Holland, 1986).[2][3] The treatment effect (τ) on individual i can be defined as the change in outcome when she would receive treatment, compared to the situation where she would not have received treatment. Let us use $T \in \{0, 1\}$ to denote an individual's treatment status, with $T = 1$ if the individual is assigned to treatment, and with $T = 0$ if the individual is assigned to control. Using y_{ij} to denote the potential outcome of individual i if her treatment status would be $T = j$, $j \in \{0, 1\}$, the treatment effect for individual i is $\tau = y_{i1} - y_{i0}$. Here, the term potential outcome refers to an outcome that may or may not be observed, as observing a certain outcome depends on the state of the world— in our case, the treatment individual i is assigned to. Unfortunately, researchers cannot estimate τ directly, as they do not observe how treated individuals' outcomes would have been had they not received treatment, nor how control group individuals' would have fared had they received the treatment. Researchers therefore typically estimate β instead (Angrist & Pischke, 2009);

$$
\begin{aligned}
\beta &= E[y_{i1}|T = 1] - E[y_{i0}|T = 0] \\
&= \underbrace{E[y_{i1} - y_{i0}|T = 1]}_{\text{TOT}} + \Big(\underbrace{E[y_{i0}|T = 1] - E[y_{i0}|T = 0]}_{\text{selection bias}}\Big).
\end{aligned}
$$

[2]This discussion builds on Duflo et al. (2007); Czibor et al. (2019) that analyze the causal identification strategy used in field experiments.

[3]For simplicity, we assume in this example that treatment compliance is perfect.

The researchers typically estimate (some variant of) the first line, and thus estimate β as the difference in mean behavior among treatment households (those i for who $T = 1$) and control group households (those i for who $T = 0$). We can rewrite this equation. By subtracting and also adding $E[y_{i0}|T = 1]$, we obtain the two new terms shown in the second line. This rewriting allows us to analyze under what circumstances $\beta = E[\tau]$, the average treatment effect (ATE), and when the parameter denotes another causal parameter of interest, such as the treatment effect on the treated. The first term, $E[y_{i1} - y_{i0}|T = 1]$ equals the average treatment effect on the treated (TOT), while the second term, $E[y_{i0}|T = 1] - E[y_{i0}|T = 0]$ denotes the differences in outcomes due to selection bias. This bias arises when households that ended up in the control group are not identical to those who ended up in the treatment group. To convince readers that β equals the causal treatment effect on the treated, the researcher should thus provide a compelling argument of why selection bias is of minor importance, or absent. In other words, why the potential control outcome is the same for the treated as for the non-treated households. Yet, as the potential outcome for treatment households in the control state of the world ($E[y_{i0}|T = 1]$) cannot be observed, the reader is unable to empirically verify the plausibility of this assumption.

Random treatment assignment, with sufficiently large sample sizes, solves the selection problem. In this case, "flipping a coin" to decide who ends up receiving the treatment results in the treatment and control group being very similar (if not identical) in terms of the (distribution of) all possible characteristics – observed, and also unobserved. When successful, randomization ensures that the average outcome observed in the control group is in expectation equal to the average outcome of the treatment group households, in case they would have not been assigned to treatment ($E[y_{i0}|T = 1] = E[y_{i0}|T = 0]$). That means that the selection bias term is zero and the treatment control difference can directly be interpreted as the causal treatment effect on the treated. When the researcher additionally assumes that the potential outcomes of an individual is independent of the treatment status of other individuals, the so-called SUTVA assumption ("Stable Unit Treatment Value Assumption"; (Angrist et al., 1996)) and that no individual can select themselves into or out of the treatment during data collection, β can more generally be interpreted as the average treatment effect (ATE). In this case, $E[y_{i1}|T = 1] - E[y_{i0}|T = 0] = E[y_{i1} - y_{i0}]$ and thus the estimate equals the expected difference in potential outcomes for the experimental population (Duflo et al., 2007; Czibor et al., 2019).[4]

1.3 Thesis outline

The remainder of this thesis is as follows. The second chapter analyzes the effects of interventions aimed at either providing extrinsic incentives or at strengthening intrinsic motives. Especially extrinsic incentives, such as convenience and price incentives, have spurred a vast literature in the domain of residential waste sorting. Convenience is generally found to be conducive to higher rates of waste sorting (Hornik et al., 1995; Ferrara & Missios, 2012; Bucciol et al., 2015), but it is not always easy to facilitate. This is especially true in urban settings where curbside collection may not be feasible due to space constraints. Price incentives generally have been found to be effective (e.g. Fullerton & Kinnaman, 1996; Usui, 2008; Dijkgraaf & Gradus, 2004; Bel & Gradus, 2016; Bucciol et al., 2015), but the results vary substantially across studies and with recycling program characteristics. For instance, Bucciol et al. (2015) argue that curbside collection increases waste sorting by 15%, an effect that doubles when accompanied with a waste pricing scheme, while Bel and Gradus (2016) do not find that the existence of a curbside collection scheme influences waste price elasticity. Consequently, the role of price incentives and convenience within this context is still a matter of debate.

Moreover, as price incentives may crowd out intrinsic motivation to sort waste (Frey & Jegen, 2001), residual waste taxation raises concerns about possible contamination of the recyclable waste flows, and increased dumping and waste burning practices to evade the pricing scheme (Fullerton & Kinnaman, 1995). This fear, but also the limited political feasibility of imposing taxes, has led to persistent barriers in the uptake of direct price incentives.[5] Behavioral interventions, especially those that appeal to non-selfish motives, may not suffer from these limitations, and may be a fruitful alternative, or complement. The empirical literature on behavioral interventions within

[4] There are other reasons to favor experiments. For instance, as treatment assignment does not depend on (a complex function of) covariates, the results might be easier to replicate. Another reason could be that the more simple analysis methods of experiments might be more credible, compared to more complex observational methods in which researchers have more degrees of freedom in their choice of analysis. See for a discussion of these and more arguments, Czibor et al. (2019).

[5] Note that another difficulty in the implementation of waste taxes could be the costs associated with monitoring waste production, especially for weight-based pricing schemes, see (Kinnaman, 2006).

the domain of waste sorting is relatively undeveloped, however (Briguglio, 2016). There is thus little evidence of their effect relative to traditional policy instruments, which will be the topic of investigation of this chapter.

The third chapter examines the potential of social learning to improve waste sorting. Social learning refers to the process by which new behaviors can be acquired by observing and imitating others (Bandura & Walters, 1963; Bandura, 1977). This may help households deduce valuable information about the act and the (public) benefits of sorting which can elevate some of the informational constraints found in the literature (Hornik et al., 1995; Gamba & Oskamp, 1994; Vollaard & van Soest, 2020)[6]. However, the peer information might also appeal to reciprocal or conformity preferences (Goldstein et al., 2008; Schultz et al., 2007), or induce 'conditional cooperation' when households respond to the contributions of others by also choosing to contribute Fischbacher et al. (2001); see also Ostrom (2009). Applications of social learning, especially social comparisons (see for example Allcott, 2011b; Allcott & Rogers, 2014), have been shown to be promising tools to foster behavioral change. However, there are only few applications in the domain of waste sorting (e.g. Linder et al., 2018; Milford et al., 2015; Nomura et al., 2011). Chapter three analyzes two applications of social learning, social feedback and social modeling, using straightforward paper-based interventions in the waste sorting domain.

The fourth chapter discusses the effect of an intervention aimed at stimulating energy conservation. Within this context, households face a private monetary incentive to engage in resource conservation, in contrast to the two earlier chapters addressing sorting of household waste. A large literature has focused on the energy price incentive, examining effective energy pricing such as dynamic pricing to smooth energy consumption in and outside of peak hours and so ease the load for energy production (for an overview, see Price, 2014). Yet, pricing does not seem to be the sole solution to lower private energy costs and lower environmental costs. Demand is relatively price inelastic without consumption feedback (Jessoe & Rapson, 2014), and households face several barriers that limit their uptake of energy-saving investments, even when these are attractively priced (e.g. Allcott & Greenstone, 2012; Gillingham & Palmer, 2014). An important reason seems to be bounded rationality. In particular, households apply heuristics when they respond to energy prices (Ito, 2014), pay less attention to energy costs when these are

[6]Agents may attribute more weight to imperfect signals provided by their peers compared to their own private information Banerjee (1992); Bikhchandani et al. (1992); see also Asch (1956).

not salient (Allcott, 2011a; Sexton, 2015), and have trouble identifying the large contributors to their energy bill (Attari et al., 2010). Chapter four contributes to the existing work of consumption feedback interventions in the energy domain designed to limit behavioral barriers (Allcott & Rogers, 2014; Jessoe & Rapson, 2014; Ito et al., 2018).[7] The chapter analyzes the effect of real-time electricity and gas consumption feedback on energy conservation.

The fifth and final chapter provides some concluding remarks. The first section discusses how this thesis contributes to the literature on the impact of behavioral interventions on household habits. The second section analyzes the relevance of the interventions studied to Dutch environmental policy. It applies insights of the recent literature on scaling (Al-Ubaydli et al., 2021) and discusses the generalizability and scalability of the interventions studied to the Dutch policy context. The chapter concludes with a brief discussion of the role field experiments may play in evidence-based policy making.

1.4 Preview of results

The findings of the three field experiments discussed in this thesis can be summarized as follows. I resort to the plural 'we' when discussing them rather than the singular 'I', as these three chapters are the result of collaborative research efforts. In the second chapter, we find that an appeal to a household's intrinsic motives to sort waste induces persistently higher levels of organic waste sorting. In contrast, we fail to find persistent results for a treatment offering extrinsic incentives (a reward) or for a treatment that harnesses reciprocity. An accompanying survey reveals that this treatment has been effective in stimulating positive attitudes of households towards at-home organic waste sorting.

In the third chapter, we find that providing neighborhood-level feedback on the amount of organic waste collected induced households to increase their waste sorting in both the short as well as in the longer run. Distributing flyers that depicted the behavior of a 'social model' only had a short-term impact, and we do not find any evidence of adding this leaflet to the neighborhood-level feedback letter increases the latter's effectiveness. An analysis of underlying behavioral determinants reveals a change in household attitudes in response to the treatments, but we also find evidence for households to perceive both higher benefits to sorting, and a higher-self efficacy to sort their organic waste.

[7] See the meta-reviews by Abrahamse et al. (2005); Darby (2006); Fischer (2008); Ehrhardt-martinez et al. (2010); Delmas et al. (2013); Karlin et al. (2015).

In the fourth chapter, we find feedback by way of an in-home display to reduce household gas consumption by almost 7 percent, and electricity consumption by about 2 percent. The relatively high gas savings on cold days suggest that reductions in space heating explain most of the gas savings. We find evidence that these changes are due to households becoming better informed of the size of their gas bill. As a result, our findings suggest that real-time disaggregated consumption feedback may be an effective tool to mitigate consumer mistakes in the energy domain.

1.5 References

Abrahamse, W., Steg, L., Vlek, C., & Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, *25*(3), 273–291.

Allcott, H. (2011a). Consumers' perceptions and misperceptions of energy costs. *American Economic Review: Papers & Proceedings*, *101*(3), 98–104.

Allcott, H. (2011b). Social norms and energy conservation. *Journal of Public Economics*, *95*(9-10), 1082–1095.

Allcott, H., & Greenstone, M. (2012). Is there an energy efficiency gap? *Journal of Economic Perspectives*, *26*(1), 3–28.

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037.

Al-Ubaydli, O., Lee, M. S., List, J. A., Mackevicius, C. L., & Suskind, D. (2021). How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy*, *5*(1), 2–49.

Andor, M. A., Gerster, A., Peters, J., & Schmidt, C. M. (2020). Social norms and energy conservation beyond the us. *Journal of Environmental Economics and Management*, *103*, 102351.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics : An empiricist's companion.* Princeton University Press.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Attari, S. Z., DeKay, M. L., Davidson, C. I., & Bruine de Bruin, W. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences*, *107*(37), 16054–16059.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-hall.

Bandura, A., & Walters, R. H. (1963). *Social learning and personality development.* New York: Holt, Rinehart and Winston.

Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, *107*(3), 797–817.

Bel, G., & Gradus, R. (2016). Effects of unit-based pricing on household waste collection demand: A meta-regression analysis. *Resource and Energy Economics*, *44*, 169–182.

Bijleveld, M., Beeftink, M., Bruinsma, M., & Uijttewaal, M. (2021). *Klimaatimpact van afvalverwerkroutes in Nederland: CO_2-kentallen voor recyclen en verbranden voor 13 afvalstromen* (Tech. Rep. No. 20.190400.163). Delft: CE Delft.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Briguglio, M. (2016). Household cooperation in waste management: Initial conditions and intervention. *Journal of Economic Surveys*, *30*(3), 497–525.

Bucciol, A., Montinari, N., & Piovesan, M. (2015). Do not trash the incentive! Monetary incentives and waste sorting. *Scandinavian Journal of Economics*, *117*(4), 1204–1229.

Chetty, R. (2015). Behavioral economics and public policy: A pragmatic perspective. *American Economic Review: Papers & Proceedings*, *105*(5), 1–33.

Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432.

Darby, S. (2006). *The effectiveness of feedback on energy consumption. A review for DEFRA of the literature on metering, billing and direct displays* (Tech. Rep.). Environmental Change Institute.

Delmas, M. A., Fischlein, M., & Asensio, O. I. (2013). Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012. *Energy Policy*, *61*, 729–739.

Dijkgraaf, E., & Gradus, R. H. (2004). Cost savings in unit-based pricing of household waste. The case of The Netherlands. *Resource and Energy Economics*, *26*(4), 353–371.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61 using randomization in development economics research: A toolkit. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of development economics* (Vol. 4, pp. 3895–3962). Elsevier B.V.

Ehrhardt-martinez, K., Donnelly, K. A., & Laitner, J. A. (2010). *Advanced metering initiatives and residential feedback programs: A meta-review for household electricity-saving opportunities* (Report No. E105). Washington D.C: ACEEE.

Ferrara, I., & Missios, P. (2012). A cross-country study of household waste prevention and recycling: Assessing the effectiveness of policy instruments. *Land Economics*, *88*(4), 710–744.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Fischer, C. (2008). Feedback on household electricity consumption: A tool for saving energy? *Energy Efficiency*, *1*(1), 79–104.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.

Fullerton, D., & Kinnaman, T. C. (1995). Garbage, recycling, and illicit burning or dumping. *Journal of Environmental Economics and Management*, *29*, 78–91.

Fullerton, D., & Kinnaman, T. C. (1996). Household responses to pricing garbage by the bag. *American Economic Review*, *86*(4), 971–984.

Gamba, R. J., & Oskamp, S. (1994). Factors influencing community residents' participation in commingled curbside recycling programs. *Environment and Behavior*, *26*(5), 587–612.

Gillingham, K., & Palmer, K. (2014). Bridging the energy efficiency gap: Policy insights from economic theory and empirical evidence. *Review of Environmental Economics and Policy*, *8*(1), 18–38.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hornik, J., Cherian, J., Madansky, M., & Narayana, C. (1995). Determinants of recycling behavior: A synthesis of research results. *Journal of Socio-Economics*, *24*(1), 105–127.

Ito, K. (2014). Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing. *American Economic Review*, *104*(2), 537–563.

Ito, K., Ida, T., & Tanaka, M. (2018). Moral suasion and economic incentives: Field experimental evidence from energy demand. *American Economic Journal: Economic Policy*, *10*(1), 240–267.

Jessoe, K., & Rapson, D. (2014). Knowledge is (less) power: Experimental evidence from residential energy use. *American Economic Review*, *104*(4), 1417–1438.

Karlin, B., Zinger, J. F., & Ford, R. (2015). The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin*, *141*(6), 1205–1227.

Kinnaman, T. C. (2006). Policy watch: Examining the justification for residential recycling. *Journal of Economic Perspectives*, *20*(4), 219–232.

Linder, N., Lindahl, T., & Borgström, S. (2018). Using behavioural insights to promote food waste recycling in urban households - Evidence from a longitudinal field experiment. *Frontiers in Psychology*, *9*(352), 1–13.

List, J. A., & Reiley, D. (2008). Field experiments. In S. N. Durlauf & L. E. Blume (Eds.), *The new palgrave dictionary of economics* (2nd ed.). London: Palgrave Macmillan.

Milford, A. B., Øvrum, A., & Helgesen, H. (2015). *Nudges to increase recycling and reduce waste* (Discussion Paper No. 2015–01). NILF Norwegian Agricultural Economics Research Institute.

Ministry of Economic Affairs of the Netherlands. (2016). *Energy report: Transition to sustainable energy* (Tech. Rep.). Ministry of Economic Affairs.

Mullainathan, S., & Thaler, R. H. (2001). Behavioral economics. In *International encyclopedia of the social & behavioral sciences* (1st ed., pp. 1094–1100). Pergamon Press.

Nomura, H., John, P. C., & Cotterill, S. (2011). The use of feedback to enhance environmental outcomes: a randomised controlled trial of a food waste scheme. *Local Environment*, *16*(7), 637–653.

Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. *Science*, *325*(5939), 419–422.

Price, M. K. (2014). Using field experiments to address environmental externalities and resource scarcity: Major lessons learned and new directions for future research. *Oxford Review of Economic Policy*, *30*(4), 621–638.

Roe, B. E., & Just, D. R. (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics*, *91*(5), 1266–1271.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007, May). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–434.

Sexton, S. (2015). Automatic bill payment and salience effects: Evidence from electricity consumption. *Review of Economics and Statistics*, *97*(2), 229–241.

Shogren, J. F., & Taylor, L. O. (2008). On behavioral-environmental economics. *Review of Environmental Economics and Policy*, *2*(1), 26–44.

Usui, T. (2008). Estimating the effect of unit-based pricing in the presence of sample selection bias under japanese recycling law. *Ecological Economics*, *66*(2-3), 282–288.

Verplanken, B., & Aarts, H. (1999). Habit, attitude, and planned behaviour: Is habit an empty construct or an interesting case of goal-directed automaticity? *European Review of Social Psychology*, *10*(1), 101–134.

Vollaard, B., & van Soest, D. (2020). *Breaking habits.*

Chapter 2

On the relative effectiveness of neoclassical and behavioral interventions

2.1 Introduction

A sizeable share of the pollution generated by consumers and households is caused by everyday activities that are better described as habits than as the outcome of continuous utility (re-)optimization (Carrus et al., 2008). Commuters' transportation choice, residential energy and water usage, and household waste recycling are examples of recurring behaviors of which the desirability, or optimality, is only infrequently (re-)evaluated by the individuals undertaking these activities. From an environmental policy perspective, the notion that polluting and resource-intensive activities might be habit-driven implies that even temporary policies can have long-lasting effects – if they can break old polluting habits and introduce greener ones (Becker &

This chapter is based on joint work with Cees Midden and Daan van Soest, under the working title *'On the relative effectiveness of neoclassical and behavioral interventions in fostering environmental change'*. We thank the municipality of Amsterdam, and then especially Esther Somers, Stef le Fevre, Rini de Jong and Jos de Bruin, for their help and support in implementing the project. We gratefully acknowledge financial support by the Netherlands' Ministry of Infrastructure and Water Management (IenW), the Department of Waterways and Public Works (RWS), the Dutch Waste Management Association (VA), the Dutch Association of Waste Management Firms (NVRD), and the Association of Dutch Municipalities (VNG). We also thank Odette van de Riet, Jessanne Mastop, Reint Jan Renes, Jorn Horstman and Kevin de Goede, and especially Gijs Langeveld and Addie Weenk, for their help and support in setting up the experiment, and Tom Melis for research assistance.

Murphy, 1988).

In this paper we experimentally test how to improve improve a specific type of environmental behavior – increasing households' propensity to sort waste, and then especially organic waste sorting. Studying waste sorting is important from an environmental policy perspective, because recycling helps reduce resource depletion, as well as the emission of greenhouse gases.[1] More importantly, however, residential waste sorting is very interesting from an academic perspective because in-house (organic) waste sorting, unlike residential water or energy conservation, is an activity that is driven by especially behavior (as opposed to technology) that are largely habit-driven, and with little (if not zero) private returns. "Technologies" like recycling bins and kitchen top receptacles can facilitate waste sorting, but by themselves they can not recycle resources.[2] Because behavioral change is indispensable for improved waste sorting, improving waste sorting provides a strong challenge for any intervention – as behavior is notoriously difficult to change (Brandon et al., 2017). But it also holds the promise that *if* behavior is changed, it may have changed (semi-)permanently, as waste sorting is strongly habit-driven.

We test the relative effectiveness of three different interventions by implementing a Randomized Controlled Trial (RCT) among 1090 households (or 2254 individuals) living in apartment buildings in downtown Amsterdam, the Netherlands. The first intervention was targeted at strengthening people's attitudes towards waste sorting by providing them with detailed information on why organic waste sorting is important, together with an explicit appeal for the recipient household to engage in organic waste sorting. The second intervention consisted of promising households a reward for proper waste sorting over a three-month period. The third intervention aimed at improving waste sorting by harnessing reciprocity. Independent of their past recycling behavior, households in this treatment arm received a gift, which was accompanied by a (written) appeal to either start or intensify in-house waste sorting. The three treatments differ in the extent to which they target the recipient household's intrinsic or extrinsic motivation

[1]For instance, organic waste sorting helps reduce the emission of greenhouse gases because the separate treatment of organic waste improves the technical efficiency of residual waste incineration and because organic waste is an energy source of and by itself (Bijleveld et al., 2021). It helps reduce resource scarcity because organic waste can be reused to produce new products such as organic fertilizer (van Dijk & Hultermans, 2020).

[2]This is in stark contrast to domains of energy and water conservation, where energy-efficient refrigerators or low-flow shower heads can provide savings without a necessitating a change in day-to-day usage.

to sort waste. The first intervention, which we will refer to as the persuasive appeal treatment, focuses on improving the recipient's intrinsic motivation to recycle. As such, it can be classified as a behavioral intervention (or "nudge"), as it aims to alter people's behavior without forbidding any options or significantly changing their economic incentives (Thaler & Sunstein, 2008). The second intervention, which we will refer to as the reward treatment, aims to induce increased waste sorting by improving the activity's private cost-benefit ratio. As such it can be viewed as (a specific form of) a neoclassical policy intervention aimed at providing extrinsic incentives to sort waste (Lazear, 2000). The third intervention, which we will refer to as the reciprocity treatment, combines elements of both extrinsic and intrinsic incentives as it aims to harness reciprocity to improve waste sorting outcomes by giving an unconditional gift (Fehr et al., 1993; Rabin, 1993). As the three interventions' budgets were roughly the same, this paper can be viewed as a test regarding the relative cost-effectiveness of neoclassical versus more behavioral interventions in changing behavior.

We monitor the frequency with which individual households dispose of their organic waste over a period of 14 months; from at least nine months before the start of the interventions until three of five months after their start. We find that all three interventions positively affect waste sorting behavior in the short run. On average, all three interventions increase the frequency with which households make use of the collective organic waste collection facilities by about 20% in the first month after the treatment has been implemented. This increase is due to a combination of previously non-sorting households starting to use the waste sorting facilities, and of already sorting households intensifying their waste sorting activity. We find that the persuasive appeal treatment is especially effective in inducing non-sorter households to start sorting, while the gift and reward treatments' impact is predominantly via intensifying the activity of sorter households. We also find, however, that the effect is short-lived for both the reward and the reciprocity treatment. In these two treatments, the frequency of organic waste disposals falls back to the level as observed in the control group within 8 weeks after the start of the intervention. This is in stark contrast to the long-lasting impact of the persuasive appeal treatment, as we observe that the initial improvement in waste sorting activity does not appreciably depreciate over the five-month post-intervention period for which we have data.

We also conducted a number of surveys to explore the underlying mechanisms of these differential impacts. Consistent with the observed long-lasting impact of the persuasive appeal treatment on waste sorting, we find that this treatment significantly improved our respondents' attitude towards the desirability of in-home waste sorting up to six months after the intervention. We document that the persuasive appeal treatment increased households' sense of personal responsibility to sort waste, consistent with a model of morally-motivated behavior (Brekke et al., 2003). We fail to detect any changes in the perceived personal responsibility to sort waste in response to either the reciprocity or the reward treatment.

Our study contributes to two strands in the literature. First and foremost, we contribute to the literature on the effectiveness of behavioral interventions, typically referred to as 'nudges' (Thaler & Sunstein, 2008), within the context of environmental behavior. Conveying the message that a change in behavior can help protect the environment, whether done so in the form of a moral appeal or by communicating a social norm, has been found to be effective in reducing households' consumption of both water and electricity (see for example Allcott, 2011; Allcott & Rogers, 2014; Ayres et al., 2012; Ferraro et al., 2011; Ferraro & Price, 2013; Brent et al., 2020), and also in encouraging households to sort waste (Linder et al., 2018; Abrahamse & Steg, 2013; Osbaldiston & Schott, 2012; Czajkowski et al., 2017; Nomura et al., 2011; Alacevich et al., 2020).

Our study is among the few studies that document a behavioral intervention not just having an immediate but also a longer-lasting (or maybe even permanent) impact on behavior– see Brandon et al. (2017), but also see Bernedo et al. (2014) and Kesternich et al. (2019) for notable exceptions. The fact that we observe this longer-run effect in the context of waste sorting is especially interesting because any change in environmental outcomes is caused predominantly by a change in behavior. This is in stark contrast to, for example, water and electricity usage, where the adoption of technically more efficient appliances can result in savings without requiring any concomitant change in behavior (see for example Stern, 2000; Karlin et al., 2015; Brandon et al., 2017).

The second strand of literature our paper contributes to is the one on intrinsic versus extrinsic motivation, and then especially on the relative importance of these factors as drivers of environmental behavior. Both intrinsic motivation and extrinsic incentives matter for performance outcomes in many domains in life – from school achievements to job performance (Cerasoli et al., 2014).

While the literature on the *interaction* between intrinsic and extrinsic motivation (especially the issue of 'crowding out') is quite substantial (for overviews, see Frey, 1997; Frey & Jegen, 2001; Deci et al., 1999, 2017; Gneezy et al., 2011; Meier, 2007), little is known about these motivations' relative effectiveness in steering behavior. Kuvaas et al. (2016) correlate measures of intrinsic motivation and external incentives on a series of job performance indicators, and find that the correlation between intrinsic motivation and work performance was three times stronger than for extrinsic motivation (see also Kuvaas et al., 2017). And Wrzesniewski et al. (2014) find that the stronger a recruit's intrinsic motivation to attend West Point, the higher the probability of receiving early promotion during the five years of mandatory service, and the higher the probability of remaining in the military beyond the mandatory five year period. The opposite was found for recruits with stronger extrinsic motives to attend West Point. These insights seem to apply to residential waste management too. Cecere et al. (2014) analyze the outcomes of a large European-wide survey on waste prevention, and find that households that attach more weight to cost-benefit considerations and reputational concerns are less likely to care about food waste prevention.

These correlational studies suggest (but do not prove) that, all else equal, strengthening people's intrinsic motivation would be more effective than offering extrinsic incentives. Causal evidence for this conjecture is, however, scant, especially in the realm of public good provision in which private monetary benefits to conserve are absent. A notable exception is Vollaard and van Soest (2020), who implemented an RCT to test the impact of three interventions aimed at improving residential waste sorting. Two of these interventions were aimed at increasing households' intrinsic motivation. They consisted of sending out a moral appeal letter about the societal importance of recycling (i.e., an injunctive social norm intervention; see Ferraro & Price, 2013), and of a campaign to communicate the local prevalence of proper waste sorting (i.e., a descriptive social norm intervention; see Goldstein et al., 2008). The impacts of these two interventions on residential recycling rates were then compared to that of a third one: a short but intensive campaign to monitor and enforce proper waste sorting. Vollaard and van Soest (2020) find that the regulatory crackdown resulted in significant and substantial increase in the various recyclable waste flows (paper, plastics and organic waste) and a concomitant decrease in the total amount of residual waste collected. The instantaneous and long-lasting effect of the enforcement activity is in sharp contrast to the lack of impact, in both the short and the longer run, of the injunctive and descriptive social norm interventions. Extrinsic incentives thus outperformed the intrinsic motivation interventions, but the amount of money spent on the enforcement activity was an order of magnitude larger than the budget for the other two interventions. Our paper complements the work by Vollaard and van Soest

(2020) in that we test the effectiveness of intrinsic and extrinsic motivation interventions using equal budgets, providing better insight into not just interventions' absolute but also their relative effectiveness.

We are not the first to study waste sorting. There is a series of papers developing theories on the drivers of recycling behavior, from neoclassical determinants (Fullerton & Kinnaman, 1996) to behavioral determinants such as as social norms, environmental attitudes and self-image concerns (Viscusi et al., 2011; Abbott et al., 2013; Brekke et al., 2003; Nyborg et al., 2006; Brekke et al., 2010). We use the insights provided by these studies to design our three treatments that differ in the manner they appeal to these two categories of motives. There is also a fairly large empirical literature analyzing the importance of neoclassical attributes such as benefits and costs of sorting (e.g., the presence of waste pricing schemes, and the convenience of sorting). While convenience is identified as an important factor (Hornik et al., 1995; Ferrara & Missios, 2012; Bucciol et al., 2015), the role of prices seems more complex, with results depending on the structural characteristics of the recycling program studied (Bel & Gradus, 2016; Bucciol et al., 2015). The literature analyzing the effect of behavioral interventions on waste sorting behavior is much less well-developed (Briguglio, 2016). In a town in Greater Manchester (United Kingdom), Nomura et al. (2011) randomized the provision of injunctively-framed feedback on residents' street food waste recycling rate and found that providing such information resulted a three percent increase in waste sorting in the three months post treatment. Milford et al. (2015) analyzed the impact of a similar intervention – providing information on a household's individual waste sorting compared to that of average household in their district – among 6000 households in a city in the south-east of Norway. Relaying this information, in the form of sending out two letters nine months apart, resulted in a two percentage point increase in recycling up to (at least) five months after sending out the second letter. Linder et al. (2018) tested the impact on recycling rates of a so-called community-based marketing campaign in a Stockholm suburb, aimed at informing households about their fellow residents' views on waste recycling. They observed higher recycling rates in the treatment group up to eight months after the information leaflet had been sent out. We complement the insights provided by these papers by testing the relative effectiveness of behavioral and neoclassical interventions in inducing organic waste sorting.

The setup of this paper is as follows. In section 2 we present our experimental design. In section 3 we analyze which household characteristics are correlated with the decision to sort organic waste, and in section 4 we present the treatment effects in both the short and long(er) run. Section 5 discusses the survey evidence regarding the mechanism via which treatments affect behavior, and section 6 concludes.

2.2 Experimental design

2.2.1 Setting

We implemented a Randomized Controlled Trial (RCT) in apartment buildings in the city center of Amsterdam. Recycling rates in urban areas are notoriously low compared to less urbanized or rural areas (Callan & Thomas, 2006; Halvorsen, 2008). Amsterdam is no exception, with recycling rates of about 16% in 2017, compared to the Dutch average of 50%.[3] Waste sorting is driven by a variety of factors, including convenience and feasibility (Briguglio, 2016). In-house waste sorting typically requires the usage of a multitude of bins, each intended for the temporary storage of a different waste stream (such as paper, plastic, organic materials and glass) within the home, or outdoors on a balcony or in the garden. Effective waste sorting rates tend to be lower in multi-family dwellings than in single-family houses because per-capita living surfaces tend to be smaller in the latter, and sizeable outdoor spaces (such as gardens or large balconies) are more likely to be lacking (Hage et al., 2009; Abbott et al., 2013). Recycling rates thus tend to be low in highly urbanized areas, and in the case of Amsterdam there are also no financial incentives for households to engage in waste sorting as the municipality charges a flat fee for household waste collection.

The city of Amsterdam wanted to learn what types of incentives are most effective in stimulating organic waste sorting. Recovery of organic waste – predominantly food waste and leftovers in the case of apartment buildings – is important because of two reasons. First, because of the environmental benefits associated with the separate collection of organic material – in the form of re-use (as compost, but also for the production of energy and biogas; Hogg et al., 2002), and in the form of savings in the amount of energy used to incinerate residual waste (as the amount of gas needed for proper incineration tends to be higher the larger the share of (especially humid) organic waste). Second, because of the financial cost savings recycling yields for the municipality – these environmental benefits translate, to a considerable degree, into lower waste management costs.

[3]Calculated on the basis of data collected by the Netherlands' Bureau of Statistics (https://opendata.cbs.nl). The numbers exclude recycling of bulkier household waste (like chairs, couches and mattresses).

We partnered with the municipality of Amsterdam to experimentally test the effectiveness of three interventions aimed at improving organic waste recycling by households living in apartment buildings. The RCT's pilot area is home to 2254 individuals living in 1090 separate households. Prior to the start of the field experiment households in the pilot area had the opportunity to sort paper and glass (as is the case virtually everywhere in the Netherlands), but not organic waste. Ten collective organic waste disposal containers were scheduled to be placed in the pilot area in November 2016, and another four in June 2017.[4]

To ensure that usage of the facilities was restricted to just the residents of the selected apartment buildings, each container was equipped with an electronic access card system. The card readers registered both the user's identity and date of usage, which allows us to record the frequency with which each household made use of the collective organic waste sorting facilities. Our main variable of interest, households' organic waste sorting behavior, is thus measured by the frequency with which household members made use of the outdoor organic waste containers.

At the moment at which the first ten organic waste containers were fully functional (in the fourth week of November 2016), the municipality sent out a letter to inform all residents in the apartment buildings about the placement of the new organic waste containers. The letter was accompanied by an access card as well as by a leaflet that contained information about the location and usage of the organic waste containers. As the main type of organic waste produced by households living in apartment buildings is kitchen waste – consisting of food preparation waste and leftovers – each household was presented with a kitchen-top organic waste receptacle as well as with a set of 100 biodegradable bags to line the receptacle with. Households were also informed that they could replenish their supply of biodegradable plastic bags for free in a nearby shop. Our data collection started on November 28, 2016 and ended on January 28, 2018, a period of 61 weeks. For an overview of the timeline of this study, see Figure 2.1.

[4]At the same time at which the four new organic containers would be placed, one residual waste container was scheduled to be removed. The reason for this container phase-in and phase-out is because the municipality wanted to test how walking distance affects usage. We will get back to this in Section 2.2.4.

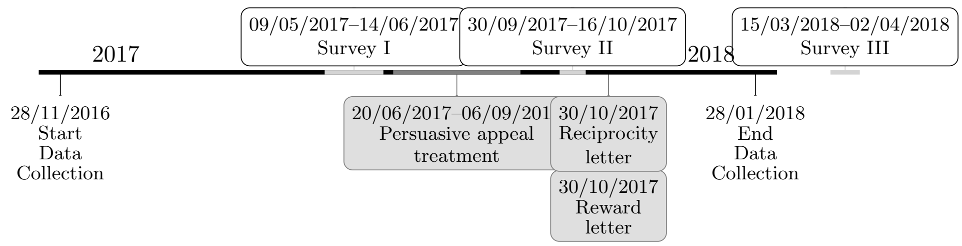


Figure 2.1: Timeline of the Randomized Controlled Trial.

2.2.2 The interventions

We aimed to test the relative effectiveness of three interventions on the organic waste sorting behavior. The three interventions differed in the extent to which they appealed to households' intrinsic or extrinsic motivations to sort waste. The first intervention consisted of sending out two letters, ten weeks apart, that were designed to strengthen households' attitudes towards waste sorting, by emphasizing the positive environmental consequences of organic waste recycling. The second intervention aimed to harness reciprocity by offering households an unconditional gift, combined with an appeal to engage in organic waste recycling. The third intervention consisted of informing households that they would be entitled to receive an in-kind reward if, in the next three months, they would actively engage in organic waste recycling. The main variable of interest is the frequency with which households made use of the apartment buildings' collective organic waste collection facilities.[5]

We will describe the three interventions in more detail in the following three subsections. Before doing so, we would like to stress the temporary nature of our interventions. Depending on how one evaluates the implementation process, the length of the interventions were three months (in case of the reward treatment), ten weeks (in case of the persuasive appeal treatment), or one day (the reciprocity treatment). We conjecture that temporary incentives can have long(er)-run impacts because of three reasons. First, if, once acquired, waste sorting is a habit (as conjectured by, for example, Carrus et al., 2008), temporary incentives may give rise to a permanent change in behavior (Becker & Murphy, 1988). Second, organic waste sorting may

[5]Originally the plan was to not only record the frequency with which households use organic waste containers, but also the weight of each individual disposal. Unfortunately, at the time that we designed the experiment weighing mechanisms were both unreliable and very expensive, and hence we decided to only focus on frequency and not on weight. We will get back to this in section 2.4.3.

be viewed as an experience good (Nelson, 1970). Especially inconvenience and hygienic consequences may be attributes of organic waste sorting that are difficult to properly assess ex ante (Lange et al., 2014; Briguglio, 2016). If the intervention is able to induce households to start sorting, they may downward adjust their assessment of these disadvantages of organic waste sorting, making it more likely that the activity will be sustained. Third, it may also be the case that the marginal costs of implementing the activity fall as the household gains more experience with the activity, in the form of learning by doing (Foster & Rosenzweig, 1995). Whether and to what extent the three temporary interventions will be able to have long(er)-term impacts thus depends on whether they are able to induce a short-run improvement in waste sorting, and also whether the induced effect lasts sufficiently long for the short-term change in behavior to become long-lasting.[6]

In addition, we did not only aim to assess the relative effectiveness of the three interventions in improving waste sorting, but also whether their effectiveness varies with household types. In other words, we were not only interested in the average impact of the three interventions in the short- and in the longer run, but also whether interventions were especially effective in inducing non-sorters to start sorting, or in inducing sorter-households to intensify their waste sorting. This distinction may be of key importance for whether any short-run improvements are likely to have long-lasting effects, as the scope for affecting behavior via the three mechanisms mentioned above – habit formation, experience good considerations and learning by doing effects – is likely to be larger among previously non-sorting households than among households that were already engaged in the activity.

We will describe the three treatments in more detail in the following three subsections.

[6]Note that purchasing waste sorting technologies (like bins and kitchen top receptacles) can make in-house waste sorting easier to sustain, but also that human input remains indispensable to have positive sorting rates – household members still need to do the waste sorting themselves, and they also need to bring the bin's contents to the collective organic waste container. This is in sharp contrast to, for example, energy savings, which can be achieved by changing behavior (switching off lights), purchasing technologies (like an energy-efficient refrigerator), or both. Because human effort is indispensable for in-house waste sorting, investing in waste sorting technologies can facilitate long-term waste sorting, but it cannot be, by itself, a mechanism that causes temporary change in waste sorting ending up becoming permanent.

The persuasive appeal treatment

The persuasive appeal treatment consisted of sending out two letters, about ten weeks apart, that provided information on the positive environmental benefits associated with waste recycling. The first letter was delivered on June 20, 2017. It contained information about the societal relevance and environmental impact of organic waste sorting, and requested households to sort their organic waste. The importance of recycling behavior was stressed by mentioning how much biogas and compost had been produced since the start of organic waste recycling in the area (i.e., since late November 2016; see Figure 2.1), and how much CO_2 had been saved as a result. The letter also contained a pictogram explaining the recycling process and providing examples of the products that can be manufactured from recycled organic waste. This first letter also contained an invitation to visit the municipality's waste treatment facility, so that households could see for themselves how useful and beneficial organic waste sorting is. The second letter was delivered to the treatment households on September 6, 2017. This letter was aimed at re-emphasizing the information already provided in the first letter. In addition, the second letter was accompanied by a tangible result of organic waste recycling – a small bar of soap made of orange peelings.

The intervention was thus aimed at directly changing people's attitudes towards organic waste sorting. As such, it can be characterized as a persuasive information treatment, aimed at changing behavior by stimulating households to think about the issue at hand (Petty & Briñol, 2008; Dellavigna & Gentzkow, 2010). The letters provided information that is likely to be new to many of the recipients – both in terms of the new products and services that can be produced with recycled organic waste (such as energy in the form of biogas, and personal hygiene products like soap), as well as the size of the environmental benefits. This new information may change people's perception (or beliefs) of the activity's environmental benefits (O'Keefe, 2015), which in turn may affect the decision to engage (more intensively) in waste sorting (Brekke et al., 2003).

The intervention consisted of sending out two letters, about ten weeks apart. We sent out the second letter to reinforce the message of the first, to induce households to really take in and process the information. Compared to more spontaneous forms of information processing (aimed at increasing the salience of the issue; Bowles & Polanía-Reyes, 2012), thoughtful information processing is more likely to be retained in memory, and hence is likely to be more effective in the long(er) run (Petty et al., 2002; Petty & Briñol, 2008). By providing households with information about the environmental impacts of the organic waste sorting activity in their own neighborhood, we aimed to increase treatment households' sense of personal responsibility to sort (Brekke et al., 2003). Not just the information itself is relevant, but also the reliability of the sender and the strength of the provided evidence (Petty et al., 1981). The invitation to visit the municipality's waste treatment facility – even though it was cancelled because of a lack of interest – was extended to also enhance (or ensure) the credibility of the provided information.

The reciprocity treatment

On October 30, 2017, households assigned to the reciprocity treatment received a gift – a high-end vegetable cutting board, made of bamboo. The gift was delivered in their letter boxes, and was accompanied by a letter from the municipality. In that letter the municipality expressed the hope that the gift would be perceived as a token of appreciation by those households that already actively sorted their organic waste, and as an encouragement to start organic waste sorting by all others. This treatment was intended to harness reciprocity – presenting households with a gift may instill feelings of gratitude and possibly indebtedness (Fehr et al., 1993; Rabin, 1993), inducing them to either start or intensify their waste sorting activities. The gift may also be viewed to signal a high level of trust of the municipality regarding the households' intentions to implement the desired behavior, which the recipient may be reluctant to betray (Sherry, 1983).

The role of reciprocity in economic behavior has been studied in a wide range of contexts from labor market interactions (Akerlof, 1982) to charitable donations (Falk, 2007). A substantial share of humanity is endowed with reciprocal preferences (including conditional cooperation; see Fischbacher et al., 2001), and offering up-front and unconditional gifts has been found effective in stimulating pro-social behaviors such as donations to charities (Falk, 2007) and to nature conservation (Alpizar et al., 2008). Whether and to what extent one-time gifts result in (semi-) permanent (as opposed to just transitory) effects, is an open question (cf. Charness et al., 2004; Gneezy & List, 2006). As argued at the beginning of Section 2.2.2, in the case of waste sorting temporary interventions may have long-lasting consequences because waste sorting is habitual, because it is an experience good, and/or because learning by doing results in lower costs of implementing the activity. In case of the the reciprocity treatment there may even be a fourth mechanism at play. If used, the cutting board may be perceived as a constant reminder of the desirability of organic waste sorting, and hence as a cue to sort waste (see Laibson, 2001).[7]

The reward treatment

The reward treatment consisted of sending a letter, delivered on October 30, 2017 (the same day as the letter of the reciprocity treatment; see the timeline in Figure 2.1). The letter announced that the recipient household would be entitled to receive an in-kind reward – the nature of which was yet to be disclosed – in the last week of January 2018, conditional on the household actively engaging in organic waste sorting in the next three months. The intention was to use the same high-end bamboo cutting board of the reciprocity treatment as reward in the reward treatment. Due to unforeseen developments the cutting boards were no longer available, and the households in the reward treatment ended up receiving a luxury packet with high-end soap bars (made of recycled organic materials, especially coffee residue) instead. Not disclosing the nature of the reward beforehand undoubtedly affected the reward's effectiveness (negatively if the treatment households dislike uncertainty, or positively if the actual reward ended up less valuable than expected on beforehand). But it also means that our study's integrity was not compromised by the the fact that we were not able to hand out the intended reward.

The reward treatment is expected to be effective in improving waste sorting because the prospect of being able to earn a reward improves the activity's expected cost-benefit ratio. This may be especially effective among households with a low intrinsic motivation to sort waste. Compared to a pay-per-throw scheme (or a piece rate payment), rewards are expected to be an efficient remuneration scheme. Binary reward schemes can be effective with less than perfect monitoring (Lazear, 2000), and for the same expected payment they tend to extract higher effort than piece rate compensation schemes (Brown, 1992; Parent, 1999). And one-time rewards may have long(er)-lasting effects in this context because of the habitual nature of waste

[7] Note that in the reciprocity treatment households received a cutting board, while households in the persuasive appeal treatment received a piece of soap made of recycled organic materials. Both the value and the way an item is offered, affect how it is perceived. While in the reciprocity treatment the cutting board was offered with an explicit reference to the donor's gratitude for the household's willingness to consider (improving their) organic waste sorting, in the persuasive appeal treatment the soap tablet was explicitly presented as "an example of what useful items can be produced with recycled organic waste". Because of that framing, and its near-negligible financial value, we expect the tablet to have reinforced the message that sorting is important as opposed to have invoked reciprocity.

sorting, its experience good characteristics or its learning by doing aspects (as discussed before), but also because the receipt of a reward may generate feelings of gratitude and appreciation, resulting in rewards having an impact beyond the period over which performance is rewarded (Bulte et al., 2021).

2.2.3 Timeline of the RCT

The timeline of this study is summarized in Figure 2.1. The frequency of usage of the collective waste sorting facilities was recorded from November 28, 2016, to January 28, 2018. The first letter of the persuasive appeal treatment was delivered on June 20, 2017, and the second letter was delivered about ten weeks later, on September 6, 2017. That means that we have seven months of pre-intervention data and five months of post-intervention data for the persuasive appeal treatment.

The letters for the reciprocity and reward treatments were both delivered on October 30, 2017. The duration of the pre- and post-intervention periods for the reward and reciprocity treatments were thus eleven and three months, respectively. We will refer to the two periods in which treatments took place as intervention period 1 (persuasive appeal treatment) and intervention period 2 (reciprocity and reward treatment). All observations before the start of intervention period 1 will be referred to the baseline period.[8]

In addition, we fielded three surveys. Survey I was fielded in May-June 2017, Survey II in September-October 2017, and Survey III in March-April of 2018. The surveys were intended to provide additional insight into the various interventions' underlying mechanisms – did the treatments strengthen attitudes towards waste sorting, did they induce people to revise their assessment of the inconvenience of waste sorting, etc.? Because of the timing of the surveys and the interventions, we can assess possible impacts of the persuasive appeal treatment in the short run (in Survey II) and in the longer run (in Survey III), while for the reciprocity and reward treatments we can only assess the short-run impacts (in Survey III).

[8]The three treatments are thus not implemented in the exact same period. This difference in timing may result in an unfair comparison of the treatments' relative effectiveness if the treatment effect is likely to be season-dependent. For example, if in some periods there is substantially more organic waste production than in others, the potential effectiveness of any treatment is higher too. While this undoubtedly would be an issue for an RCT using households living in single family dwellings (because of the strong seasonality of garden refuse), organic waste sorting by the control group households does not vary much between the different intervention periods; see also Figure 2.4.

2.2.4 Randomization

We used stratified-randomization to allocate our 1090 households into the four treatment groups. The randomization was stratified on households characteristics that have been documented to be strongly correlated with waste sorting (Briguglio, 2016) – distance to the nearest organic waste container, household size and composition (single adults with and without children, non-single adults with children, non-single adults without children) and observed sorting behavior in the baseline period. Larger families are expected to produce more (organic) waste, but they may also be more prone to engage in waste sorting, possibly because waste sorting is subject to increasing returns to scale. Family composition may matter too, as the propensity to sort seems to increase with age (Vining & Ebreo, 1990), while the presence of children may strengthen pro-environmental preferences among adult family members (Dupont, 2004). Finally, the larger the walking distance, the larger the expected inconvenience of disposing waste at the waste facilities (Rousta et al., 2015). We also stratified our sample for those 179 households whose walking distance to the nearest organic waste container would be affected by the phase-in of the four additional organic waste containers in June 2017, and also for the 39 households whose walking distance to the nearest residual waste container would be increased, in July 2017, because of the removal of one residual waste container.[9]

The last characteristic we stratified on was observed sorting behavior in the baseline period. We counted the number of times households made use of the collective organic waste collection facilities in March 2017, and classified them into three different groups: those with zero usage of the organic waste collection facilities, and those with above-median and below-median usage (conditional on having used the organic waste containers at least once). Using sorting behavior in the baseline period as a stratification variable is important because of two reasons. First, if waste sorting is indeed a habit, baseline sorting behavior is likely to be predictive of waste sorting later on. Stratification on baseline behavior would then enable us to estimate possible treatment effects with higher precision. Second, it allows us to test whether the interventions' main impact, if there is one, is via inducing non-sorting households to to start sorting, or rather via inducing sorter-households to step up their game.

[9]Note that we stratify on the changes in walking distances to mitigate the dependent variable's error variance – not because we view the change in walking distance as an additional treatment. Even though the change in distance is exogenous, the number of affected households is too small for a properly powered test.

Table 2.1 presents the characteristics of each of our four treatment groups (in columns (1)-(4)) as well as those of the overall experimental population (in column (5)). On average households consisted of two individuals, and males and females were equally represented across households. There were relatively few households with young children in our sample (about 3 percent), and about 28 percent of the households had one or more household members above the age of 65. Apartments either had a balcony or a garden, but never both. Not surprisingly, 98 percent of the households in our sample had a balcony, and hence only 2 percent had a garden. More than 90 percent of the apartments were serviced by an elevator, and hardly any households had access to a separate storage space. The average walking distance, at baseline, to the outdoor organic and residual waste containers was about 100 and 75 meters, respectively.

Finally, regarding baseline waste deposit behavior, a little over half of households made use of the outdoor organic waste collection facilities at least once in the baseline period (and hence the other half did not even use them once). On average our sample of households used the organic waste containers about 0.7 days per week (or about once in every 1.5 weeks). Defining 'regular organic sorters' as those households that made use of the collective organic waste sorting facilities at least once in every 1.5 weeks, we find that 28 percent of the households in our sample fitted this definition.

Variable	N	(1) Control Mean/SE	N	(2) P. Appeal Treatment Mean/SE	N	(3) Reciprocity Treatment Mean/SE	N	(4) Reward Treatment Mean/SE	N	(5) Total Mean/SE	(6) Difference (1)-(2)	(7) Difference (1)-(3)	(8) Difference (1)-(4)	(9) Normalized difference (1)-(2)	(10) Normalized difference (1)-(3)	(11) Normalized difference (1)-(4)
Nr of persons	465	2.049 (0.052)	197	2.091 (0.079)	190	2.042 (0.080)	192	2.115 (0.086)	1044	2.068 (0.035)	-0.042	0.007	-0.065	-0.038	0.007	-0.057
Nr of elderly	465	0.286 (0.027)	197	0.254 (0.039)	190	0.253 (0.042)	192	0.307 (0.041)	1044	0.278 (0.018)	0.032	0.033	-0.021	0.057	0.058	-0.037
1 person hh	465	0.398 (0.023)	197	0.381 (0.035)	190	0.405 (0.036)	192	0.385 (0.035)	1044	0.394 (0.015)	0.017	-0.007	0.012	0.035	-0.015	0.025
2 person hh	465	0.320 (0.022)	197	0.305 (0.033)	190	0.316 (0.034)	192	0.318 (0.034)	1044	0.316 (0.014)	0.016	0.005	0.003	0.034	0.010	0.006
>2 hh members	465	0.282 (0.021)	197	0.315 (0.033)	190	0.279 (0.033)	192	0.297 (0.033)	1044	0.290 (0.014)	-0.033	0.003	-0.015	-0.073	0.006	-0.034
Elderly hh	465	0.224 (0.019)	197	0.198 (0.028)	190	0.184 (0.028)	192	0.250 (0.031)	1044	0.216 (0.013)	0.026	0.039	-0.026	0.062	0.096	-0.062
hh with kids <3 years	465	0.028 (0.008)	197	0.036 (0.013)	190	0.032 (0.013)	192	0.016 (0.009)	1044	0.028 (0.005)	-0.008	-0.004	0.012	-0.044	-0.022	0.080
Floor	454	2.993 (0.102)	187	2.888 (0.149)	184	3.103 (0.161)	188	2.793 (0.152)	1013	2.957 (0.067)	0.106	-0.110	0.201	0.049	-0.050	0.093
Elevator Access	488	0.932 (0.011)	202	0.911 (0.020)	200	0.935 (0.017)	200	0.945 (0.016)	1090	0.931 (0.008)	0.021	-0.003	-0.013	0.082	-0.010	-0.052
Home value above median (300k Euro)	488	0.424 (0.022)	202	0.470 (0.035)	200	0.510 (0.035)	200	0.510 (0.035)	1090	0.464 (0.015)	-0.046	-0.086**	-0.086**	-0.093	-0.172	-0.172
Has balcony	488	0.977 (0.007)	202	0.990 (0.007)	200	0.985 (0.009)	200	0.965 (0.013)	1090	0.979 (0.004)	-0.013	-0.008	0.012	-0.093	-0.053	0.078
Storage	486	0.002 (0.002)	201	0.005 (0.005)	200	0.000 (0.000)	199	0.000 (0.000)	1086	0.002 (0.001)	-0.003	0.002	0.002	-0.054	0.054	0.054
Baseline walking distance in mtr organic	488	96.234 (2.562)	202	102.982 (3.949)	200	102.168 (3.874)	200	103.264 (4.238)	1090	99.863 (1.721)	-6.748	-5.933	-7.030	-0.119	-0.106	-0.122
Baseline walking distance in mtr residual	488	74.103 (2.376)	202	77.675 (3.783)	200	75.183 (3.780)	200	75.980 (4.024)	1090	75.308 (1.626)	-3.572	-1.080	-1.876	-0.068	-0.020	-0.035
Number of organic disposal days per week	488	0.441 (0.033)	202	0.406 (0.052)	200	0.495 (0.063)	200	0.430 (0.055)	1090	0.442 (0.024)	0.035	-0.054	0.011	0.047	-0.069	0.014
Deposited organic waste at least once at baseline	488	0.533 (0.023)	202	0.545 (0.035)	200	0.485 (0.035)	200	0.550 (0.035)	1090	0.529 (0.015)	-0.012	0.048	-0.017	-0.024	0.096	-0.034
Regular organic waste disposal at baseline	488	0.273 (0.020)	202	0.287 (0.032)	200	0.285 (0.032)	200	0.285 (0.032)	1090	0.280 (0.014)	-0.015	-0.012	-0.012	-0.033	-0.028	-0.028

Notes: Columns (6)-(8) present the absolute difference, for each variable, between each of three intervention groups, and the control group. The stars indicate the significance level of the associated t-test. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Columns (9)-(11) present the normalized difference. The indicator for garden is not included in table, this garden and balcony are mutually exclusive as households without balcony have a garden and vice versa. This table was adapted from a table made using the `iebaltab` command of the `ietoolkit` package (DIME Analytics, 2019).

Table 2.1: Descriptive statistics of the household sample, and treatment balance tests.

Regarding balance, columns (6)-(8) of Table 2.1 present the differences in means between each of the three intervention groups and the control group. According to standard t-tests, only two of these differences are statistically significant – the shares of households living in an apartment with an above-median tax value in the reciprocity and reward treatments are significantly higher than that in the control group (both at $p = 0.04$). Quantitatively the differences are small (less than 9 percentage points), and this is confirmed by the size of the normalized differences, as presented in columns (9)-(11). The largest normalized differences (in absolute values) are 0.172 (for the shares of the above-median tax dwellings in the reward and reciprocity treatments), and hence the absolute values of these (and all other) normalized differences are well below the conventional cut-off value of 0.25 standard deviations (Abadie & Imbens, 2011; Imbens & Wooldridge, 2009; Imbens & Rubin, 2015).

In addition, we also ran tests on all the bicomparisons – not just treatment versus control for each treatment and for each variable (the 51 tests shown in columns (6)-(8) and (9)-(11) of Table 2.1), but also of each treatment group versus each of the other two treatment groups (another set of 51 tests; not shown here, but available upon request). Only one of those additional 51 bicomparisons turned out to be significantly different from zero. Households in the persuasive appeal group are significantly more likely to have a balcony compared to the households in the reward treatment (at $p = 0.09$), and hence they are also less likely to have a garden. In short, just three of the in total 102 bicomparisons were statistically significantly different from zero, and the size of the (normalized) differences were quite small. We thus conclude that our randomization procedure has been successful in creating treatment groups that are very similar in terms of their observable characteristics – and hence this probably holds for their unobservable characteristics too.

2.2.5 Power

We thus used stratified random assignment to allocate the 1090 households in our study to four treatment groups (persuasive appeal, reward, reciprocity, and control). Because of both financial and statistical reasons each of the three treatment groups was envisaged to consist of 200 households; the control group thus consisted of almost 500 households. Because the control group is used in multiple tests (against each of the three intervention groups), statistical power is increased if the control mean is estimated with more precision; see List et al. (2011).[10]

Statistical power – the probability of being able to detect a treatment effect if there is one – does not only depend on the number of observations; it also depends on the model used to estimate the treatment effects. As will be explained in more detail in Section 2.4.2, we use a panel fixed-effects model to estimate the treatment effects. Following Bertrand et al. (2004), standard errors are clustered at the unit of randomization – the household – to control for serial correlation. For each test we have about 700 cross-sectional units (200 households in treatment, and almost 500 in control), for which we have waste sorting information over a period of 61 weeks. That means that we have close to 43,000 observations for each individual test, but the statistical power of the test crucially depends on the serial correlation in individual households' waste sorting decisions – oftentimes in a complex and non-linear fashion (Burlig et al., 2020).[11]

We used STATA package `pcpanel` to derive the Minimum Detectable Effect (MDE) for the persuasive appeal treatment, and also for the reward and reciprocity treatments. MDEs differ because of the differences in the number of pre- and post-intervention periods. We used data on observed waste sorting behavior of all households in the period between November 28, 2016 (the start of the registration of access card usage) and April 2, 2017 (the data available at the time of randomization).We calculated the standard deviation of the idiosyncratic error component (0.479, having controlled for household fixed effects and for week fixed effects) as well as the serial correlation in the idiosyncratic error term (0.107, assuming an AR(1) process). The number of cross-sectional units was 700, of which 200 (or 28%) were allocated to the treatment group. The targeted significance level was 0.05 (two-sided), and the targeted probability of being able to reject the null of no difference (if there is, in fact, a non-zero treatment effect)

[10] With three interventions (with equal budgets), the power of the between-group test is maximized if the number of units in the control group is $\sqrt{3}$ times the number of units in each of the three treatment groups. Because of budgetary considerations our control group is 2.5 as large as each of the three intervention groups (rather than the optimal 1.73 times).

[11] As shown by Burlig et al. (2020), the relationship between power and the level of serial correlation is complex. On the one hand, higher levels of serial correlation facilitate detecting differences between pre- and post-intervention observations. On the other hand, higher levels of serial correlation imply (steeply) declining added benefits of having a longer time series for each cross-sectional unit, because each extra observation yields less new information. So the relationship between power and serial correlation is ex-ante ambiguous. And the same holds for, for example the relationship between power and the length of the panel.

was 80 percent. With 41 pre-intervention periods and 20 post-intervention periods, the Minimum Detectable Effect (MDE) for the persuasive appeal treatment is an increase in the organic waste facilities' usage frequency of 0.034 days per week. And with 49 pre-intervention periods and 12 post-intervention periods, the MDE for the reciprocity and reward treatments is a 0.040 increase in number of days per week the collective organic waste sorting facilities are used. These translate into Minimum Detectable Effect Sizes of 0.071 and 0.084 standard deviations for the persuasive appeal and the reciprocity/reward treatments, respectively.

2.3 Analysis of the decision to sort organic waste

Before estimating the impact of the three types of interventions on waste sorting behavior, we analyze which household characteristics are correlated with the decision to sort organic waste. This analysis is important for two reasons. First, it can provide insight into the behavioral mechanisms underlying waste sorting – once one is engaged in the activity, is the behavior likely to be permanent, or is the decision to sort much more erratic? And second, we hypothesize that the three interventions may differ in the extent to which they are effective in inducing non-sorting households to start sorting, and maybe also in the extent to which they are able to induce sorter-households to intensify their waste sorting behavior. This section aims to identify what types of households are likely to become sorters or non-sorters.

Figure 2.2 provides insight into households' propensity to start sorting. It shows, for each week in the baseline period, the number of households that make their first-ever visit to the organic waste collection facilities in that specific week. Overall, 53% of the sample accessed a container at least once in the baseline period, and the vast majority of those did so, for the first time, in the first two weeks after the facilities became available in the last week of November, 2016. That suggests that by and large households either used the facilities in the first few weeks after they became operational, or they did not use them at all.

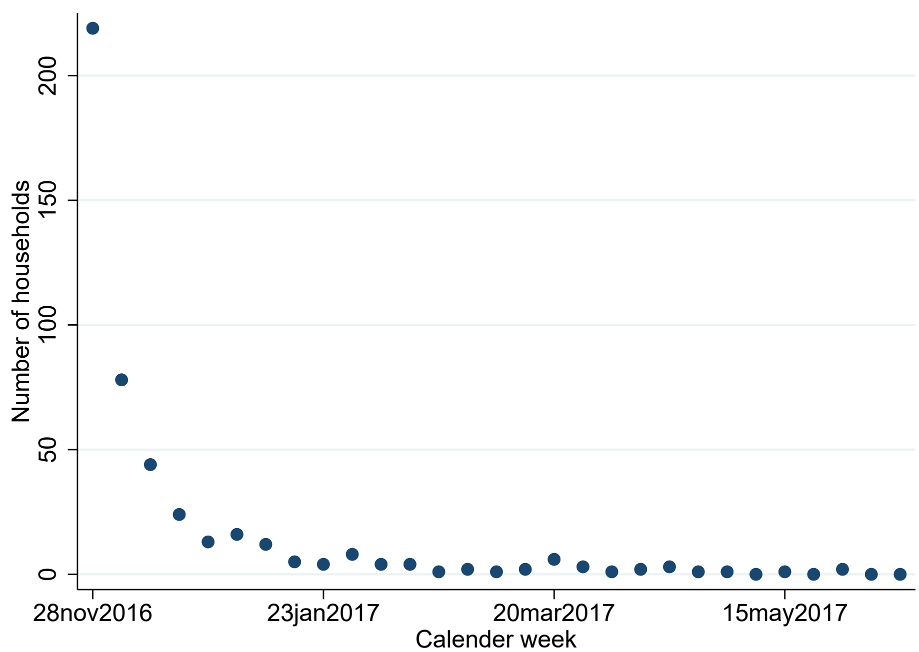


Figure 2.2: Number of households making their first organic waste disposal by calendar week in the baseline period.

In the next subsection we take a closer look at the persistence of organic waste sorting behaviors, and in the second subsection we analyze what household characteristics are correlated with organic waste sorting activity.

2.3.1 The persistence in organic waste sorting behavior

To assess the persistence of waste sorting behavior, we analyze the conditional probability of a household having made use, at least once, of the organic waste collection facilities, in three different months: the first month of the baseline period, the month before the first intervention period, and the month before the second intervention period. Figure 2.3 demonstrates that the household's decision to (not) sort waste is quite persistent. As shown in the bottom arm of the tree diagram, 73% of households who made use of the organic waste sorting facilities in the first month of the baseline period, also used them in the month before the first intervention takes place, five months later. And again 78% of those households that used the facilities, at least once, in the month before the start of the first intervention period, also used them at least once in the month before the start of the second intervention period.[12] And as shown in the top arm of the tree diagram, households that had not used the organic waste collection facilities in the first month at baseline, had a 90 percent probability to remain inactive in the month before the first intervention period 1. And 97% of those households that had not used the facilities in the month before the first intervention period, also did not use them in the month before the second intervention.

[12]Obviously we excluded all those households in this second step of the Markov chain that had been assigned to the persuasive appeal treatment.

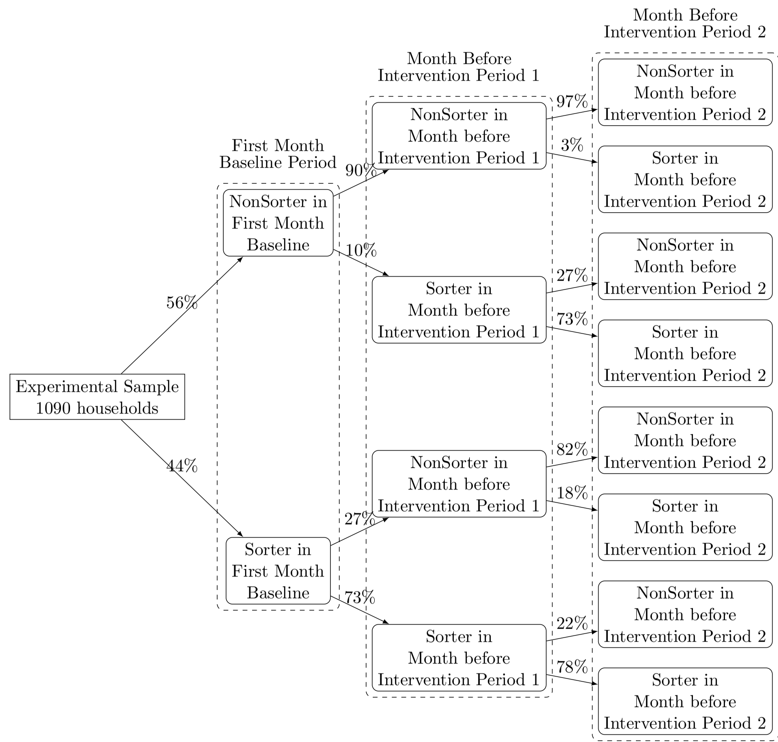


Figure 2.3: Markov diagram of the shares of households sorting or not-sorting in each of the months.

The percentages denote the conditional probability of becoming a sorter (a household that disposes organic waste at least once in a specific month) in the next period. The percentages are calculated based on observed organic disposal behavior using the full experimental sample, except for the evaluation of the changes from the month before the first to that before the second intervention period, where the households of the persuasive appeal treatment are omitted.

2.3.2 Identifying the types of households that are likely to engage in organic waste sorting

We now turn to exploring what characteristics correlate with the decision to (not) sort waste, for each of the three periods we focused on in the previous subsection – the first month of the baseline period, the last month before the start of the first intervention period, and in the last month before the start of the second intervention period. We aim to uncover what types of households made use of the facilities as soon as they became available, and also the characteristics of those households that continued to sort in the longer run.

The remainder of this subsection is set up as follows. In section 2.3.2 we present our estimation strategy, and in section 2.3.2 we present the results.

Estimation strategy

Our estimation strategy is as follows. Let us use $y_i = 1$ to denote if household i made use of the collective organic waste sorting facilities at least once in the month under consideration, and $y_i = 0$ otherwise. We then estimate the following model:

$$y_i^* = \alpha + \beta X_i + \epsilon_i. \tag{2.1}$$

In this equation y_i^* is a latent variable; if the estimated value is positive ($y_i^* > 0$), the household is predicted to have made use (at least once) of the outdoor organic waste sorting facilities in the time period under consideration. Next, X_i is the vector of household characteristics at baseline (i.e., prior to the installation of the organic waste sorting facilities). Characteristics include household size and composition (family size, and whether or not children below the age of 3 and/ or people over 65 were present in the household) as well as a set of apartment characteristics (the floor the apartment was located on, the presence of a balcony (as opposed to a garden), whether the tax value of the apartment was above or below median, as well as the walking distances, in meters, to the nearest organic and residual waste containers). Finally, ϵ_i is the error term. We estimate his regression using the Huber/White/sandwich estimator, to correct for heteroscedasticity in the error term.

Regression results

Table 2.2 presents the regression results of equation (2.1). Column (1) of this Table identifies the characteristics of the households that made use of the organic waste sorting facilities at least once in the first month of the baseline period. Column (2) shows the estimation results explaining the sorting decision in just the last month before the start of the first intervention period (i.e., in the last month of the baseline period), and column (3) does the same but then for last month before the start of the second intervention period. Comparing regression coefficients across the three columns of Table 2.2 sheds light on whether specific characteristics induce households to make use of the facilities at once in the very short run (column (1)), in the short run (column (2)), or also in the slightly longer run (column (3)). Note that the number of observations in columns (1) and (2) is 968 (rather than 1090) because of missing data; the number of observations in column (3) is even lower because of the exclusion of the households of the persuasive appeal treatment.

As shown in column (1) of Table 2.2, we find that in the very short run one-person households are about 18 percentage points less likely to use the organic waste sorting facilities than multi-person households. Regarding family composition, we do not find evidence for households with small children to be more or less likely to take up sorting, whereas households with one or more members of 65 or older, are 12 percentage points more likely to engage in sorting. The probability to sort increases by about 1.5 percentage point if, at baseline, the organic (residual) waste container is 10 meters closer (farther away).

Comparing the results in columns (2) and (3) of Table 2.2 to those in column (1), we find that with time, the share of households sorting does not fall substantially over time, and also that the composition of sorter households does not appreciably change either. As shown at the bottom of Table 2.2 about 45% of the households make use of the organic waste sorting facilities at least once in the first month after the facilities became available. This share drops to 38% in the last month before the start of the first intervention period, five months later, and to about 34%, another two months later. Overall, this suggests that the share of sorting households is fairly constant over time, and certainly so after the novelty of having organic waste sorting facilities in the neighborhood wears off. And regarding the characteristics of the households that continue to sort in the longer run, we see that in columns (2) and (3) they are by and large the same as in column (1), with the exception of household size (number of household members and presence of small children).

Table 2.2: Factors correlated with households' usage of the outdoor organic waste collection facilities.

	(1) Waste sorting (y/n) First Month Baseline	(2) Waste sorting (y/n) Month before Intervention Period 1	(3) Waste sorting (y/n) Month before Intervention Period 2
1-Person household	-0.180***	-0.100***	-0.0491
	(0.037)	(0.037)	(0.040)
>2 Household members	-0.0474	-0.0250	0.00242
	(0.044)	(0.043)	(0.047)
Elderly household	0.123***	0.123***	0.0990**
	(0.040)	(0.040)	(0.044)
Household with kids <3 years	-0.0172	-0.188**	-0.180**
	(0.097)	(0.080)	(0.088)
Floor	-0.00415	0.00164	-0.000144
	(0.008)	(0.008)	(0.008)
Elevator Access	-0.0232	-0.0203	-0.0606
	(0.065)	(0.065)	(0.075)
Home value above median (300k Euro)	0.0464	0.0323	0.0177
	(0.033)	(0.033)	(0.036)
Balcony	0.121	0.114	0.0336
	(0.101)	(0.101)	(0.113)
Distance to nearest organic container in mtr	-0.00168***	-0.00126***	-0.00112**
	(0.000)	(0.000)	(0.000)
Distance to nearest residual container in mtr	0.00141***	0.00115**	0.00112**
	(0.000)	(0.000)	(0.001)
Dep. Var. Sample Mean	0.447	0.379	0.344
Observations	968	968	786
Excl. p. appeal treatment	N	N	Y
Model	Probit	Probit	Probit
Container Facilities Change Controls	N	N	Y

Standard errors in parentheses

$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The presented coefficients are average marginal effects of a probit model estimated with robust standard errors. We code a household as sorting waste in a certain time period if we have observed the household to have disposed organic waste at least once during this period. Finally, when we analyze the decision to sort in the month before Intervention Period 2 (Column (3)), we exclude households that have started the persuasive appeal treatment in Intervention Period 1. The nearest container distance variables measure the distances at the start of the experiment. As 218 households have experienced a shock in nearest container distance in the period leading up to Intervention Period 2, we estimate effects of the container distance measures controlling for these shocks. We code these controls as indicators that equal one when a household has experienced a shock in nearest organic or residual container distance respectively.

2.4 Estimating the impact of the three interventions on the frequency of waste sorting

Having determined the correlates of organic waste sorting, we now turn to analyzing the effectiveness of the three interventions in increasing households' waste sorting activities. We measure the frequency of use as the number of distinct days per week an household used at least one of the organic waste containers. That means that this variable takes on values between 0 and 7, and if we report an average frequency of usage of, say, 0.5, this means that households use their cards, on average, one day every two weeks.[13]

In this section we first present graphical evidence on the (relative) impact of the three interventions on the frequency with which households make use of the organic waste sorting facilities, and then we turn to a parametric analysis.

2.4.1 Graphical evidence on the impact of the three interventions

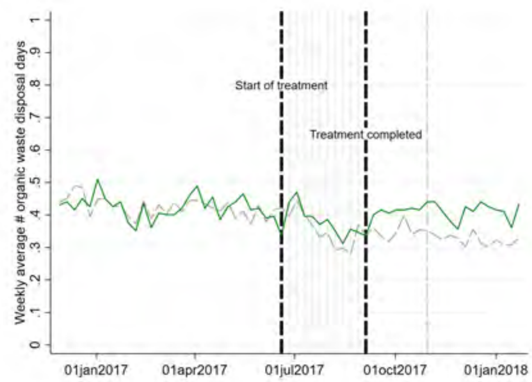
Panels (a)-(c) of Figure 2.4 present the frequency with which the outdoor organic waste sorting facilities were used in, respectively, the persuasive appeal, reciprocity and reward treatments – as well as in the control group. The frequencies of use of the average household in each of the three intervention groups are depicted using solid lines; the frequency of use by the average household in the control group is depicted by the light-grey dashed lines. The vertical dashed lines in each of the three panels mark the end of the baseline period and/or the start of the intervention period of the treatment under consideration. In case of the reciprocity and reward treatments (shown in panels (b) and (c) of Figure 2.4, respectively), the two coincide – the baseline period ends and the intervention period starts on the day at

[13] Alternatively, we could have used the sheer number of times households used their card in a specific week. In 92 percent of the cases households swiped their access card just once on a specific day, suggesting that it is not too important which of the two frequency measures we use. However, the mean frequency of usage using the sheer number is inflated by implausibly high numbers of times some households used their cards on a specific day – the highest number of times a card was swiped on a specific day was 35. Using one's card 35 times on one day may reflect highly dedicated waste sorting; it is more likely that the user's card malfunctioned. Because the two frequency measures are identical for 92 percent of the cases and the 'number of distinct days per week' measure is more conservative than 'the sheer number of times per week', we use the former in the rest of the analysis.

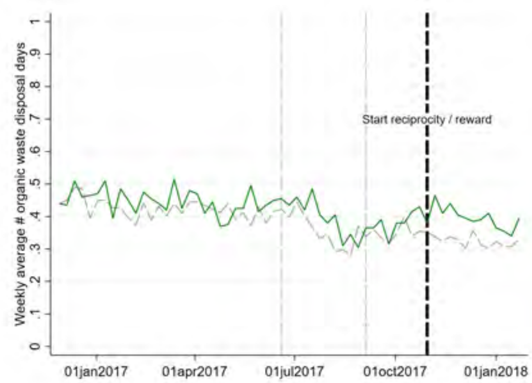
which the letters were delivered to the treatment households (October 30, 2017). The persuasive appeal treatment, however, consisted of two contact moments, an initial letter on June 20, 2017, and a follow-up letter about ten weeks later, on September 6, 2017.

Comparing the average sorting frequencies between the treatment groups and the control group in the baseline period, the temporal patterns are very similar in all four groups, and also the levels of usage in the three treatment groups are quite similar to those in the control group – albeit less so for the households that were yet to receive the reward treatment (see panel (c) of Figure 2.4). In the baseline period the average frequency of use was 0.45 days per week (or once every 2.2 weeks). The frequency dropped substantially during the Summer period (starting mid July 2017), and did not fully bounce back to the pre-summer level in the Fall (as evidenced from the time pattern of the control group).

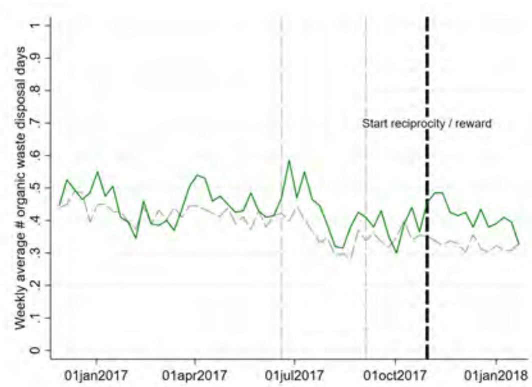
Comparing usage frequencies in the intervention and control groups, panels (a)-(c) of Figure 2.4 suggests that all three interventions were effective in increasing organic waste sorting frequencies compared to the control group. Interestingly, the first letter that was sent out as part of the persuasive appeal treatment did not have any impact on waste sorting, while the second letter's impact is quite considerable. Whether this difference is caused by the content of the second letter having been more effective than that of the first or by the fact that the timing of the first letter was, in retrospect, unfortunate (as it was sent out just before the start of the Summer holidays), is an open question.



(a) Persuasive appeal treatment versus the control group.



(b) Reciprocity treatment versus the control group.



(c) Reward treatment versus the control group.

Figure 2.4: Average frequency of organic waste facilities usage in each of the three intervention groups, compared to the control group (dashed line).

Regarding the relative effectiveness of the three interventions, Figure 2.4 suggests that the short-term impacts of the three treatments are quite similar – at least when taking the arrival of the second letter as the starting point of the persuasive appeal treatment. Yet the panels also suggest that the longer-run dynamics are quite different. Whereas the effect of the persuasive appeal treatment seems to be fairly stable over its five-month (post-) intervention period, the impacts of the reciprocity and reward treatments seem to decline over time. In the next subsections we analyze these patterns in more detail.

2.4.2 Parametric analysis

We use difference-in-difference regression models to estimate the short- and long(er)-run effectiveness of our three interventions in inducing waste sorting. We present our estimation strategy in Section 2.4.2, and the regression results in Section 2.4.2.

Estimation strategy

We use three different regression specifications. The standard intention-to-treat (ITT) model, to be estimated for each treatment j separately ($j \in \{Appeal, Rew, Recip\}$), is:

$$y_{it} = \alpha_i + \gamma_t + \sum_{p=\{1,2\}} \beta^{jp} \, T_i^j \, I_t^{jp} + \epsilon_{it}. \tag{2.2}$$

Here, y_{it} is the number of distinct days in week t household i made use of the outdoor organic waste collection facilities. Next, α_i and γ_t are respectively the household and week fixed effects, and T_i^j is a dummy variable that takes on a value of one if household i was assigned to treatment j under consideration, and zero otherwise. We estimate the average treatment effect for two different subperiods. We do so because we have data up to five months after the completion of the persuasive appeal treatment, whereas we only have three months of post-intervention data for the reward and reciprocity treatments. We construct two indicator functions, I_t^{j1} and I_t^{j2}. I_t^{j1} equals 1 for all weeks t in months 1-3 after the completion of intervention j (and zero otherwise), and I_t^{j1} equals 1 for all t in months 4 and 5 since the completion of that treatment (and zero otherwise).[14] Our key parameter of interest is β^{j1}, $j \in \{Appeal, Rew, Recip\}$, as this coefficient captures the average size of the treatment effect over the three-month period since the completion

[14]Note that $I_t^{j2} = 0$ for all t for the reward and reciprocity treatments, but not for the persuasive appeal treatment.

of the treatment rollout. We are also interested in $\beta^{Appeal2}$, as it captures the longer-run impact of the persuasive appeal treatment. Finally, ϵ_{it} is the error term, which is clustered at the unit of randomization – i.e., the household level – to account for serial correlation (Bertrand et al., 2004).

We estimate equation (2.2) using OLS. Our dependent variable can take on values between 0 and 7, and hence, strictly speaking, we are not allowed to use OLS. As a robustness check we re-estimate the treatment effects using a negative binomial model.[15] Following Allison and Waterman (2002), we include both household and week fixed effects by directly estimating household- and week-specific intercepts. The estimated treatment effects are reported as incidence ratios, and hence they reflect the percentage difference in the dependent variable between the treatment and control households in the post-intervention subperiod of interest, and in the control group.

We also estimate two other versions of equation (2.2). The first alternative specification aims to provide better insight into the dynamics of the treatment effects, by estimating them separately for each of the post-intervention months. Equation (2.3) is identical to equation (2.2), except that it estimates the average treatment effects of each of the three interventions for each of the post-intervention months for which we have data:

$$y_{it} = \alpha_i + \gamma_t + \sum_m \beta^{jm} \, T_i^j \, I_t^{jm} + \epsilon_{it}. \tag{2.3}$$

Using m to index each month in the intervention period, I_t^m equals 1 if week t falls in the m^{th} month after the completion of the intervention, and zero otherwise. Note that equation (2.3) is estimated using $m \in \{1, .., 5\}$ for the persuasive appeal treatment, and with $m \in \{1, 2, 3\}$ for the reciprocity and reward treatments.

The second modification we implement aims to provide insight into possible heterogeneous treatment effects. If effective, does a treatment improve average waste sorting because it induces intensification of waste sorting among households that were already actively sorting their waste in the baseline period, or because it induces previously non-sorting households to

[15] We use a negative binomial model rather than a Poisson model because the latter assumes that the dependent variable's variance is equal to its mean. In this study the assumption is not likely to be met because of the fairly large number of non-sorting households (i.e., those with zero organic waste disposal per week). The presence of a large number of zeros is likely to give rise to overdispersion – the variance in the dependent variable is larger than the mean. The negative binomial does not require equal mean and variance, and hence the negative binomial model is the preferred approach (Cameron & Trivedi, 2013).

start sorting? We test for heterogeneous impacts of the treatments by allowing the treatment effects to differ among the (at baseline) sorter households and the non-sorting households. To this end, we re-estimate equation (2.3) on sample splits of households based on their sorting activity in the month before treatment.

Finally, two remarks are in order regarding our empirical specification. First, note that equations (2.2) and (2.3) yield intention-to-treat (ITT) estimates, whereas it would also be very relevant to estimate the treatment on the treated (TOT). All three treatments were implemented by means of letters, but unfortunately we have no information on whether the recipients actually read them. In case of the reciprocity treatment, we can be quite sure that the ITT is, in fact, equal to the TOT. The gift, the high-end cutting board, was small enough to fit in the letterboxes, and it is quite likely that the recipient households were curious enough to open the package. For the other two treatments it is less likely that all households actually opened the letter(s), so the TOT is expected to differ from the ITT. Unfortunately we are unable to estimate the TOT because we have no information on which households opened the letters. Second, our treatment estimates are unbiased if treatment spillovers are absent. While our RCT's sample size too small for us to implement a two-stage randomization design to test whether spillovers are indeed present (Hudgens & Halloran, 2008; Baird et al., 2018), we were able to do so in a companion RCT (Boomsma & van Soest, 2020), in a different city but with similar interventions and much larger sample sizes. As we did not find any evidence of spillovers in that companion study, we are quite confident that spillovers are not likely to be substantial in the present study either. In any case, if we are incorrect and spillovers were non-negligible, our results would be lower-bound estimates of the true treatment effects.

Results

The regression results of equation (2.2) are presented in Columns (1)-(3) of Table 2.3. Column (1) of this table shows that the persuasive appeal treatment's impact is about 0.07 extra disposal days in the first three months after the end of the persuasive appeal treatment ($p = 0.126$), and 0.10 in months 4 and 5 ($p = 0.034$). Both impact estimates are well above our RCT's Minimum Detectable Effect (MDE). The impact thus increases over time, and only becomes statistically significant at conventional significance levels in the second subperiod. Compared to the average number of usage days of about 0.35 in the control group (see the bottom panel of the table), the 0.07 extra days per week in the first three months imply a 20 percent increase in organic waste sorting compared to the control group, and the 0.10 extra days in months 4 and 5 translate into a more than 30 percent increase compared to the control group – or an increase in the frequency from one day in every 3.3 weeks to one day in every 2.3 weeks.

Columns (2) and (3) of Table 2.3 show that the reciprocity and reward treatments had a very similar effect on waste sorting behavior; the frequency of use of the organic waste facilities increased by between 0.045 and 0.048 extra days per week (compared to an MDE of 0.04). This constitutes a percentage change of about 14%, or an increase from using the facilities one day every 3.3 weeks to one day every 2.6 weeks. This effect is significantly different from zero (at $p < 0.10$) for the reward treatment, and it only just fails to be significant ($p = 0.11$) for the reciprocity treatment. When combining all four treatment groups in one single regression (see column (4) in Table 2.3), we obtain essentially the same results as in columns (1)-(3).

Columns (1)-(4) of Table 2.3 are obtained using OLS while, strictly speaking, they should have been estimated using a negative binomial model. Column (5) shows the results of this robustness check. The coefficients presented are incidence ratios, and hence they capture the percentage change in behavior compared to that in the control group. The treatment effects of the first three months of the persuasive appeal, reciprocity and reward treatments are estimated to reflect a 17, 15 and 16 percent increase with respect to the mean usage frequency of the control group. Moreover, the effects during the fourth and fifth month of the persuasive appeal treatment are estimated to reflect a 32 percent increase. The results are very similar to the earlier findings of the linear model. We therefore stick to our linear modeling strategy in the remainder of this paper.

Table 2.3: The average impact of the three treatments on the number of distinct organic waste container usage days per week.

Coefficients reported as:	Linear Marginal Effects				Incidence Ratios
	(1) Disposal days	(2) Disposal days	(3) Disposal days	(4) Disposal days	(5) Disposal days
Persuasive Appeal Treatment Month 1-3	0.0678			0.0597	1.167
	(0.0443)			(0.0427)	(0.122)
Persuasive Appeal Treatment Month 4-5	0.1000**			0.103**	1.321**
	(0.0471)			(0.0461)	(0.150)
Reciprocity Treatment Month 1-3		0.0452+		0.0464*	1.148*
		(0.0276)		(0.0272)	(0.0810)
Reward Treatment Month 1-3			0.0484*	0.0495*	1.159**
			(0.0280)	(0.0276)	(0.0816)
Constant	0.436***	0.439***	0.440***	0.439***	0.857***
	(0.0248)	(0.0243)	(0.0248)	(0.0194)	(0.0389)
Control mean month 1-3 (P.appeal)	0.345			0.345	0.345
Control mean month 4-5 (P.appeal)	0.318			0.318	0.318
Control mean month 1-3 (Recipr/Reward)		0.323	0.323	0.323	0.323
R^2	0.008	0.009	0.009	0.007	
Number of hh (N)	690	688	688	1090	1090
Observations (N × T)	42090	41968	41968	66490	66490

Dependent variable is the number of organic disposal days per week. Standard errors, clustered at the household level, in parentheses. All specifications include household- and week- fixed effects. The negative binomial is estimated using unconditional fixed effects (that is, household identifiers are included as dummy variables in the estimation). $^{+}p < 0.125$, $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The effect sizes are considerable (at least in percentage terms) and not very different between the three interventions – at least not in the first three months after the implementation. But the post-intervention period averages may hide important temporal patterns. In Table 2.4 we present the regression results when estimating the average treatment effects for each of the post-intervention months; see equation (2.3).

Table 2.4: The dynamics of the treatment effects, for each month in the post-intervention period.

	Persuasive Appeal	Reciprocity	Reward
	(1)	(2)	(3)
	Disposal days	Disposal days	Disposal days
Treatment × Month 1	0.0718	0.0697^{**}	0.0792^{**}
	(0.0493)	(0.0311)	(0.0315)
Treatment × Month 2	0.0639	0.0481	0.0450
	(0.0473)	(0.0316)	(0.0322)
Treatment × Month 3	0.0676	0.0178	0.0210
	(0.0458)	(0.0323)	(0.0331)
Treatment × Month 4	0.109^{**}		
	(0.0486)		
Treatment × Month 5	0.0906^{*}		
	(0.0482)		
Constant	0.436^{***}	0.439^{***}	0.440^{***}
	(0.0248)	(0.0243)	(0.0248)
R^2	0.008	0.009	0.010
Number of hh (N)	690	688	688
Observations (N × T)	42090	41968	41968

Dependent variable is the number of organic disposal days per week. Standard errors, clustered at the household level, in parentheses. $^{+}p < 0.125$, $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Column (1) of Table 2.4 shows that indeed the immediate impact of the persuasive appeal treatment is sizeable (albeit that the standard error is quite considerable too, resulting in a $p-$value of 0.146 in month 1), but also that the impact actually increases over the interaction period. This is in marked contrast to the treatment effects for the reciprocity and reward treatments (shown in columns (2) and (3) of Table 2.4, respectively). These are statistically significantly different from zero in the first post-treatment month, but then they actually drop sharply over time.[16]

We thus document that the initial impact of the three interventions is roughly similar, but also that the persuasive appeal treatment still remains effective up to five months after completion of the intervention. Because the budgets were roughly the same for each of the three interventions, the persuasive appeal treatment is thus found to be the most cost-effective one. Note, however, that this conclusion is reinforced when taking into account that the fixed costs of designing this intervention were relatively high while the variable costs were quite low, whereas the opposite is true for the reciprocity and reward interventions. When scaled up, the persuasive appeal treatment is thus likely to become even more cost-effective.

So why is it that the persuasive appeal treatment is more effective than the reciprocity and reward treatments? One possible reason is that gifts and rewards are simply less effective than interventions aimed at strengthening people's attitudes toward waste sorting. But another reason may be that these treatments operate on different types of people that make up differential shares of the population. For example, it may be the case that the one treatment predominantly encouraged non-sorting households to start sorting their organic waste, while the other may have induced households that already engaged in sorting, to intensify sorting. We test this using a simplified

[16]These results clearly indicate that the reciprocity treatment's impact is quite transitory at best. However, we cannot be so sure about the reward treatment. The results presented in Table 2.4 indicate that the impact of the prospect of a reward extinguishes before the completion of the evaluation period. If households are rational (as posited by neoclassical economics, the actual receipt of the reward (by those that did well) should not affect their behavior in the post-evaluation period. However, as suggested in section 2.2.2, the receipt of the reward may instill feelings of gratitude, possibly inducing households that received the reward, to continue to sort better. We cannot test this because the municipality shut down the access card registration system because of privacy considerations. so we do not have any information on households' behavior after (not) having received the reward. However, because the behavior in the third evaluation month of the reward treatment is not significantly different from that of the control group in that same month, the reward is very similar as a gift to those households that are sorting well. We fail to find a long-lasting effect of the gift (also for that subgroup of sorter households), and hence the long-run consequences of the reward treatment are unlikely to be substantive.

version of equation (2.3) in which months two and three are pooled, and also months four and five. We estimate these treatment effects by sample splits based on pre-treatment sorting behavior (sorters versus nonsorters), defining 'sorters' as those households that made use of the organic waste containers at least once in the last month prior to the start of the treatment.

Table 2.5 presents the average treatment effects for the two subgroups for the three different subperiods. The results are quite striking. Columns (1) and (2) of Table 2.5 show that the persuasive appeal treatment was especially effective in changing the behavior of those households that had not made use of the collective organic waste sorting facilities in the month prior to the start of the intervention. In other words, the persuasive appeal treatment induced non-sorting households to start sorting, and it was by and large ineffective in intensifying the waste sorting activities of households that were already actively sorting their waste in the month prior to the start of the intervention.

The patterns in the reciprocity and reward interventions are markedly different; see columns (3)-(6) in Table 2.5. As expected on the basis of Table 2.4, the treatment impacts were very short-lived (as they are only found to be effective in changing behavior in the first month after receipt of the letter). But columns (3)-(6) in Table 2.5 also reveal that they were only able to change the behavior of those households that were already actively sorting their organic waste. In other words, they were able to induce sorter-households to intensify their sorting behavior (albeit only very temporarily), and they were by and large ineffective in inducing non-sorter households to start sorting.

Table 2.5: Impact estimates for households used the organic waste containers at least once in the month before the start of the relevant treatment, and those that did not.

	Persuasive Appeal		Reciprocity		Reward	
	(1) Sorter	(2) Non-Sorter	(3) Sorter	(4) Non-Sorter	(5) Sorter	(6) Non-Sorter
Treatment × Month 1	0.0484	0.0831	0.138*	0.0388	0.154*	0.0314
	(0.0898)	(0.0566)	(0.0767)	(0.0270)	(0.0781)	(0.0206)
Treatment × Month 2-3	0.0266	0.0871+	0.0796	0.0115	0.0650	0.0150
	(0.0759)	(0.0542)	(0.0762)	(0.0237)	(0.0745)	(0.0227)
Treatment × Month 4-5	0.100	0.0963*				
	(0.0867)	(0.0533)				
Constant	0.870***	0.171***	0.904***	0.222***	0.899***	0.213***
	(0.0538)	(0.0221)	(0.0542)	(0.0240)	(0.0556)	(0.0241)
R^2	0.020	0.021	0.012	0.045	0.013	0.043
Number of hh (N)	262	428	219	469	228	460
Observations (N × T)	15982	26108	13359	28609	13908	28060

Standard errors, clustered at the household level, in parentheses. $^{+}p < 0.125$, $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

2.4.3 Weight versus frequency

We thus documented that the three treatments were able to – at least temporarily – increase the frequency with which households made use of the organic waste sorting facilities. However, from an environmental perspective, it is not the frequency per se that is of interest, but the total amount of organic waste deposited. Intuitively, an increase in the frequency with which households make use of the organic waste sorting facilities is likely to also result in an increase in the amount of organic collected. This is obviously true for non-sorting households that were induced to start sorting. But it is also plausible that the increased frequency of use by the sorter households reflects an intensification of organic waste sorting. This is the case if either the sorter household maintained the same recycling process (in which case the increased frequency reflects a proportional increase in the waste deposited), or they invested in the purchase of (likely larger) in-house waste receptacles (implying that the observed percentage increase in the frequency of use is an underestimate of the actual increase in weight).

As already stated in footnote 5, at the time of the setup of the RCT the technology to weigh individual deposits was both very expensive as well as not sufficiently accurate for the municipality to be willing to purchase a per-deposit weighing technology for each organic waste container. However, in a companion study we were able to test the impact of behavioral interventions not just on the frequency with which organic waste was deposited, but also the average weight of the deposits; see Boomsma and van Soest (2020). The companion RCT was implemented in predominantly apartment buildings in the city of Schiedam (also located in the highly urbanized western part of the Netherlands). We were able to measure the weight consequences of the behavioral interventions in Schiedam because of a much larger sample size – 4000 households making use of 42 collective organic waste containers (as opposed to the current study's 1090 households making use of 10 collective organic waste containers). The larger sample size allowed us to implement a cluster-randomized design with treatments being assigned at the level of households making use of the same container (rather than individual-level assignment). That means that we were able to compare the weights of containers assigned to different treatments. We did not find any evidence for the average weight of a deposit being *reduced* as a result of any of the three behavioral interventions we implemented; if anything, the average weight seemed to have been increased. That means that in that companion RCT, the percentage change in the frequency of use was a lower-bound estimate of the percentage increase in the weight collected. Although these results were obtained from an RCT in a different municipality and with different behavioral interventions, they provide suggestive evidence that behavioral treatments can affect the frequency of use without decreasing the per-deposit weight, and hence that the percentage changes in the frequency of waste sorting in this study are likely to be lower-bound estimates of the treatments' impacts in terms of weight.

2.5 Treatment effects on households' perceived desirability and feasibility of waste sorting

We thus find that the persuasive appeal treatment had a long-lasting effect on organic waste sorting behavior, and that the impacts of the reciprocity and reward treatments were rather short-lived. In this section we aim to identify the underlying mechanisms – did the persuasive appeal treatment indeed result in a non-transitory increase in the motivation to sort (or in the attitude towards organic waste sorting), and is there no appreciable impact of the reciprocity and reward treatments on motivation? Or worse, is there evidence of especially the reward treatment actually reducing the motivation to sort – as predicted by motivational crowding theory (Deci et al., 1999; Frey & Jegen, 2001; Gneezy et al., 2011)? After all, a non-negligible share of the households in our RCT seem to have been intrinsically motivated to sort waste, as almost 30 percent of them could be labelled as 'sorter households' in the baseline period.

To shed light on these issues we conducted a series of surveys; see also Section 2.2.3 and Figure 2.1. The first survey was implemented in May-June 2017, and two follow-up surveys were fielded during periods of a little over two weeks in September-October 2017 and in March-April 2018. The timing of the follow-up surveys was such that for the persuasive appeal treatment, we can measure both the short- and the longer-run impact of the intervention on households' motivation to sort, as the roll-out of the intervention itself was completed in early September 2017; see Figure 2.1. For the reward and reciprocity treatments, the final survey was implemented four months after the start of the interventions, and hence one month after the end of the reward treatment's evaluation period. All three surveys had a similar setup, and included questions eliciting households' attitudes towards waste sorting, their beliefs about the necessity or usefulness of organic waste sorting, their self-efficacy, their perception of how (in)convenient waste sorting is, and whether they perceived waste sorting to be a social norm.

All 1090 households in our RCT were invited to participate in each of the three surveys. However, not all households accepted the invitation. The number of responding households declined from 286 for the first survey, to 212 for the second, and to 199 for the third. Although response rates are decent (more than 25% of the households responded to the first survey, and still about 18% responded to the third), identifying the treatment effects poses a challenge for three reasons. First, the total number of respondents is still very limited especially for the second and the third survey. Second, households that participated in one survey did not necessarily accept to also participate in the next. More specifically, of the households that participated in the first survey, about 60 percent participated in the second survey, and about the same percentage responded to the third. We thus have more statistical power when analyzing the cross-section of all households that responded to a particular survey (either the second, or the third) than when analyzing a panel excluding all survey responses of households that failed to participate in the previous survey. And third, we are severely hampered by (our survey company's interpretation of) the European Union's General Data Protection Regulation (GDPR) that came into effect in April 2016. According to our survey firm, informed consent was needed to be allowed to match survey respondents to the waste sorting frequency data and to the municipality registers. Because very few households gave consent, we are unable to properly control for biases in survey responses due to potentially treatment-induced non-random selection into the survey.[17] We can, however, use the data of the households that gave explicit consent to test for treatment-induced selection into the survey.

Our empirical strategy is as follows. We first test whether treatment status affects the propensity to participate in the survey. This is an (albeit weak) test of the internal validity of our survey analysis. Second, we explore whether treatment status affects survey responses by means of simple cross-sectional analyses.

2.5.1 Testing for treatment-induced survey selection bias

Selection into a survey is unlikely to be random. Some households are more likely to be willing to participate than others – because some have more time to do so than others, or because the survey's topic is more appealing to some than to others. That means that the survey responses may not be representative of the opinions, motivations and preferences held by the participants in the RCT. However, if the decision to participate in the survey is independent of treatment status, the estimates of the treatment effects on survey responses are unbiased for the subgroup of respondents (Gerber & Green, 2012). In other words, the analysis is still internally valid, albeit that any statement about the treatment impacts is only valid for the subset of households that were willing to participate in the survey.

As already stated in the introduction to this section, the analysis of selection into the survey is complicated because of (our survey company's interpretation of) the GDPR. Although we have access to the municipality's register for all 1090 households in the RCT, we can only assess the drivers of selection into the survey for those households that consented to us linking the survey data to the register and waste deposit data – about half of the respondents in both the second and the third survey.

The results of the selection tests are presented in Table 2.6. Tests are based on linear probability regression models estimated using OLS, with the dependent variable being whether or not the household participated in either the second or the third survey.[18] Columns (1a)-(1c) of this table present the

[17] One way to correct for selection biases is to run a Heckman selection model, which tests whether households that are more likely to participate in a survey tend to have more positive or more negative responses on individual survey questions; see Heckman (1979).

[18] We use OLS and not probit because of the issues associated with properly calculating the marginal effects in non-linear models; see Ai and Norton (2003). Conclusions regarding

results of three different specifications of the analysis of the factors that drive selection into the second survey. At the time of the second survey the only intervention that had been implemented was the persuasive appeal treatment, and hence the treatment indicator distinguishes those households that had received the persuasive appeal treatment from those households that had not received any intervention yet – all households from the control group that participated in the survey, as well as those from the reciprocity and reward treatment groups. At the time of the third survey all three interventions had been completed In columns (2a)–(2c) of Table 2.6 we test whether having received the persuasive appeal treatment affected selection into the third survey (as compared to the households in the control group), and in columns (3a)-(3c) we do the same for the combined reward and reciprocity treatments.[19] Finally, note that the a-column of each series of analyses (i.e., columns (1a), (2a) and (3a)) includes all possible covariates, the b-column only includes the covariates that showed up significantly in the associated a-column (as well as the treatment indicator), and the c-column presents the results of the set of significant coefficients interacted with the treatment indicator.

the outcomes of all models without interaction effects are robust to re-estimating them using probit, and hence the OLS regressions are likely to provide reliable estimates for the models with interactions terms as well.

[19]The latter analysis thus tests for the *pooled* effect of treatment-induced selection into the survey, and we do so for two reasons. First, the number of respondents is too low to test for differential impacts. Second, we also believe that pooling the two treatments is not likely to severely bias the results, because the treatments seemed to have been fairly similar in terms of their impact on sorting behavior.

Table 2.6: Factors explaining households' propensity to participate in either the second of the third survey.

	Survey II respondent Persuasive Appeal (T) vs. Control			Survey III respondent Persuasive Appeal (T) vs. Control			Survey III respondent Reciprocity/Reward (T) vs. Control		
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)	(3a)	(3b)	(3c)
1 person hh	-0.027			-0.050*			0.006		
	(0.023)			(0.026)			(0.023)		
>2 hh members	-0.007			-0.064			0.026		
	(0.040)			(0.042)			(0.038)		
Children	-0.016			0.052			-0.002		
	(0.039)			(0.043)			(0.037)		
Elderly hh	0.053*	0.042*	0.052*	0.039	0.026	0.031	0.058**	0.038	0.031
	(0.027)	(0.024)	(0.027)	(0.030)	(0.028)	(0.032)	(0.028)	(0.025)	(0.032)
Floor	-0.004			-0.001			-0.000		
	(0.004)			(0.005)			(0.005)		
Above Median Home value ($\geq$ 300k Eur)	-0.010			0.013			-0.003		
	(0.019)			(0.023)			(0.020)		
Baseline Organic waste distance above median	-0.006			0.007			0.013		
	(0.018)			(0.023)			(0.019)		
Disposited organic waste at least once at baseline	0.043**	0.045**	0.037*	0.032	0.042*	0.056**	0.045**	0.055***	0.056**
	(0.018)	(0.018)	(0.019)	(0.022)	(0.022)	(0.026)	(0.020)	(0.020)	(0.026)
Regular organic waste disposal at baseline	0.149***	0.149***	0.154***	0.127***	0.114***	0.107**	0.130***	0.120***	0.107***
	(0.030)	(0.029)	(0.032)	(0.036)	(0.034)	(0.041)	(0.032)	(0.032)	(0.041)
Treatment	0.004	-0.000	-0.008	0.019	0.010	0.032	0.011	0.015	0.006
	(0.023)	(0.022)	(0.017)	(0.025)	(0.024)	(0.026)	(0.019)	(0.019)	(0.018)
Treatment × Elderly hh			-0.058			-0.014			0.015
			(0.063)			(0.066)			(0.052)
Treatment × Disposited organic waste at least once at baseline			0.048			-0.046			-0.004
			(0.047)			(0.049)			(0.040)
Treatment × Regular organic waste disposal at baseline			-0.025			0.023			0.030
			(0.076)			(0.073)			(0.064)
Constant	0.053**	0.019**	0.020**	0.037	0.019^{+}	0.012	-0.008	0.008	0.012
	(0.026)	(0.009)	(0.010)	(0.032)	(0.012)	(0.012)	(0.029)	(0.012)	(0.012)
Observations	968	1044	1044	613	662	662	786	847	847

Results obtained using OLS regressions. Robust standard errors in parentheses. $^{+}p < 0.125$, $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The results are quite homogeneous across the two survey waves and the two (groups of) interventions. Households that engaged in sorting at baseline were more likely to select into the second and third survey, and the same holds – albeit to a lesser extent – for households with one or more members over 65. None of the other covariates show up significantly, including the treatment indicators. We thus do not find any evidence of the treatments affecting households' propensity to select into the surveys – not in terms of levels (see the a- and b-columns), but also not when interacted with the characteristics that showed up significantly (the baseline sorting behavior and the age indicators; see the c-columns).

We thus do not find any evidence for the treatments having affected the composition of the set of respondents. Any treatment differences with respect to respondents' motivations or opinions are therefore unlikely to be biased. However, while any differences found are valid for the group of households that selected into the survey (on average older and more prone to sort waste), they are not necessarily representative of the overall opinions and motivations of our experimental population.

2.5.2 Treatment effects on the factors affecting waste sorting

The surveys were designed to gain insight into the (potential) mechanisms by which the interventions affected waste sorting behavior. The surveys contained slightly more than thirty statements on factors affecting waste sorting behavior. These statements can loosely be categorized into motivation (e.g., the perceived desirability of waste sorting), self-efficacy and perceived (in)convenience. Ex ante, the chances of finding treatment-induced changes in any of these categories are not very high, as households that engaged in organic waste sorting in the baseline period are overrepresented in our survey. These households are thus likely to already have a positive attitude towards waste sorting in general, and towards organic waste sorting in particular, and the same holds for how they score on the other categories. Indeed, of the over thirty statements included in the second survey, we only found a significant treatment effect on one statement, and a borderline significant treatment effect on another, similar statement. These statements are, respectively, whether the household thinks that it is desirable if it sorts its waste in general, and also, more specifically, its organic waste. Table 2.7 presents the probit regression results for these two statements, and for both surveys. Shortly after the persuasive appeal treatment had been completed (i.e., when filling out the second survey), households in the persuasive appeal treatment group are 14.4 percentage points more likely to fully agree

Table 2.7: Treatment effects of the persuasive appeal treatment on the second and third survey.

	Survey II		Survey III	
	(1) Desirable to sort waste?	(2) Desirable to sort organic waste?	(3) Desirable to sort waste?	(4) Desirable to sort organic waste?
Persuasive Appeal Treatment	0.144**	0.111^{+}	0.115*	0.070
	(0.065)	(0.071)	(0.068)	(0.078)
Dispositied organic waste at least once at baseline	0.208	0.430***	0.186	0.259^{+}
	(0.148)	(0.161)	(0.163)	(0.160)
Regular organic waste disposal at baseline	0.204***	0.222***	0.036	0.246***
	(0.072)	(0.079)	(0.079)	(0.087)
Above Median Home value ($\geq$ 300k Eur)	0.007	0.018	0.000	0.065
	(0.060)	(0.064)	(0.072)	(0.076)
Baseline Organic waste distance above median			0.033	0.061
			(0.069)	(0.073)
Observations	169	168	114	114

Coefficients are average marginal effects from a probit regression. Robust standard errors in parentheses. $^{+}p < 0.125$, $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

with the first statement ($p = 0.026$), and 11.1 percentage points more likely to fully agree with the second ($p = 0.117$); see columns (1) and (2) of Table 2.7. Of course, these outcomes may be false positives, because they are the only two questions with (borderline) significant treatment differences out of a set of 30 questions. But the likelihood of these being false positives is reduced by the fact that we still find a significant difference for the first statement in the third survey, five months after the implementation of the second. Households in the persuasive appeal treatment group are still 11.5 percentage points more likely to fully agree with the first statement, and this difference remains significant at the 10 percent level ($p = 0.091$); see column (3) of Table 2.7. The persistence of the treatment impact provides suggestive evidence that indeed the persuasive appeal treatment has been effective in establishing a long-lasting increase in support for (organic) waste sorting – consistent the lack of treatment decay in terms of actual (organic) waste sorting behavior.[20] For completeness we also present the third survey's regression results for the second statement; see column (4) of Table 2.7. While the coefficient remains positive, it is not significantly different from zero ($p = 0.340$).

[20] We are unable to test this formally using Multiple Hypothesis Testing because, to the best of our knowledge, no procedure is available that allows researchers to test the null of no effect on two outcome variables (in our case, on household attitude to sort in the short-term and longer-term) using multiplicity-corrected test statistics.

Consistent with the lack of a long-lasting impact of the reward and reciprocity treatments on actual organic waste sorting behavior, we do not find any significant differences for any of the factors affecting the propensity to sort waste between households in the pooled reciprocity and reward treatment group and those in the control group (not shown here; available upon request). The only significant treatment effect that we find is a 14.1 percentage point increase in the likelihood of fully agreeing with the statement that recycling is necessary to economize on natural resources ($p = 0.050$). We do not attach much value to this outcome for two reasons. Scoring better on this variable is, of course, by no means a guarantee that households' own motivation to sort waste has been improved, and we also cannot rule out that this result is, in fact, a false positive.

Overall, we thus do not find convincing evidence that the reciprocity and reward treatments affected any of the determinants of waste sorting behavior – self-efficacy, motivation, and perceptions of the (in-)convenience of sorting. We also find no evidence of the persuasive appeal treatment affecting these determinants, except for changes in the perceived desirability of in-home waste sorting by the household itself. Therefore the persuasive appeal treatment did not just affect households' perceived benefits of waste sorting or the desirability of household participation in general; it also strengthened their appreciation of the relevance of their own actions in this domain. This result provides support for the role of personal responsibility in driving waste sorting behavior, in line with the theory of Brekke et al. (2003).

2.6 Conclusion

Increasing households' propensity to sort waste poses a challenge because of the private (monetary and/or non-monetary) benefits of waste sorting tend to be small, while the costs can be substantial – for example, think of the inconveniences of in-house waste sorting in general, and of sorting organic waste in particular. The challenge is even greater because waste sorting is typically viewed as habitual behavior; inducing humans to shed old habits and acquire new ones is notoriously difficult.

In this paper, we presented the results of three interventions aimed at stimulating households to sort (organic) waste. We implemented a Randomized Controlled Trial among 1090 households living in apartment buildings in downtown Amsterdam. We observe the frequency with which households make use of newly installed collective organic waste sorting facilities. We find considerable support for the hypothesis that waste sorting is habitual.

Hardly any of the households that did not make use of the organic waste facilities in the first month after they became available, started using them at a later date. Similarly, we find that the characteristics of the households that were still engaged in waste sorting nine months after they became available, are very similar to those that took up waste sorting in the period immediately after the facilities were put in place.

The interventions tested in this study were a persuasive appeal treatment aimed at strengthening households' positive attitude toward waste sorting, a treatment in which households were promised an in-kind reward conditional on active waste sorting, and treatment aimed at harnessing reciprocity – offering a gift in the hope of households being willing to reciprocate in the form of improved waste sorting. Our interventions thus differ in whether they aim to strengthen households' intrinsic or extrinsic motivations to sort waste, and hence whether they can be classified as more behavioral policy instruments (the persuasive appeal treatment) or as more neoclassical policy instruments (the reward treatment), with the reciprocity treatment being somewhere in the middle. Because the treatments' financial budgets were roughly the same, we view our RCT as an experimental test of the relative effectiveness of our neoclassical and behavioral interventions – in the short run and in the longer run, and also whether they are more effective in inducing non-sorting households to start sorting organic waste, or rather in increasing the activity's intensity among households that are already engaged in the activity.

We find that the direct impact of each of the three interventions is positive and very similar. We document an increase in the frequency of waste sorting of over 20% in the first month after the start of the interventions (albeit that in case of the persuasive appeal treatment behavior was too noisy for this impact to be statistically significant). We provide tentative evidence that the observed increase in the frequency of organic waste sorting is likely to be a lower-bound estimate of the increase in organic waste collected. While the direct impact is roughly the same in all three treatments, we find marked differences in both the permanence of the impact, as well as in whether they were especially effective in inducing new waste sorting behavior or in intensifying existing behaviors. While the persuasive appeal treatment's impact strengthens over time and is especially effective in encouraging non-sorting households to start sorting, the reciprocity and reward treatments' main impact is via a highly temporary intensification of waste sorting by households that were already engaged in the activity. Results from two survey waves, implemented about one month and six months after the completion of the persuasive appeal treatment, provide corroborative evidence of the persuasive appeal treatment's long-lasting impact.

We thus find that with similar budgets, the behavioral intervention – the persuasive appeal treatment – is more effective in inducing pro-environmental behavior than the neoclassical intervention – offering a conditional reward. We document this result in an environmental domain and in a context that are not very conducive to pro-environmental behavior – organic waste sorting in apartment buildings. Whether these insights spillover to other environmental behaviors that can be characterized as habits – such as commuters' modal transportation choice – is an open question.

2.7 References

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, *29*(1), 1–11.

Abbott, A., Nandeibam, S., & O'Shea, L. (2013). Recycling: Social norms and warm-glow revisited. *Ecological Economics*, *90*, 10–18.

Abrahamse, W., & Steg, L. (2013). Social influence approaches to encourage resource conservation: A meta-analysis. *Global Environmental Change*, *23*(6), 1773–1785.

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*(1), 123–129.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, *97*(4), 543–569.

Alacevich, C., Bonev, P., & Soderberg, M. (2020). *Pro-environmental interventions and behavioral spillovers: Evidence from organic waste sorting in Sweden.*

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9-10), 1082–1095.

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037.

Allison, P. D., & Waterman, R. P. (2002). Fixed-effects negative binomial regression models. *Sociological Methodology*, *32*, 247–265.

Alpizar, F., Carlsson, F., & Johansson-Stenman, O. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, *92*(5-6), 1047–1060.

Ayres, I., Raseman, S., & Shih, A. (2012). Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *Journal of Law, Economics, and Organization*, *29*(5), 992–1022.

Baird, S., Bohren, J. A., McIntosh, C., & Özler, B. (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, *100*(5), 844–860.

Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy*, *96*(4), 675–700.

Bel, G., & Gradus, R. (2016). Effects of unit-based pricing on household waste collection demand: A meta-regression analysis. *Resource and Energy Economics*, *44*, 169–182.

Bernedo, M., Ferraro, P. J., & Price, M. (2014). The persistent impacts of norm-based messaging and their implications for water conservation. *Journal of Consumer Policy*, *37*, 437–452.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, *119*(1), 249–275.

Bijleveld, M., Beeftink, M., Bruinsma, M., & Uijttewaal, M. (2021). *Klimaatimpact van afvalverwerkroutes in Nederland: CO_2-kentallen voor recyclen en verbranden voor 13 afvalstromen* (Tech. Rep. No. 20.190400.163). Delft: CE Delft.

Boomsma, M. A., & van Soest, D. P. (2020). *What do my neighbours do? leveraging social learning to stimulate organic waste sorting.*

Bowles, S., & Polanía-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature*, *50*(2), 368–425.

Brandon, A., Ferraro, P. J., List, J. A., Metcalfe, R. D., Price, M. K., & Rundhammer, F. (2017). *Do the effects of social nudges persist? Theory and evidence from 38 natural field experiments* (NBER Working Paper Series No. 23277). National Bureau of Economic Research.

Brekke, K. A., Kipperberg, G., & Nyborg, K. (2010). Social interaction in responsibility ascription: The case of household recycling. *Land Economics*, *86*(4), 766–784.

Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics*, *87*(9-10), 1967–1983.

Brent, D. A., Lott, C., Taylor, M., Cook, J., Rollins, K., & Stoddard, S. (2020). What causes heterogeneous responses to social comparison messages for water conservation? *Environmental and Resource Economics*, *77*(3), 503–537.

Briguglio, M. (2016). Household cooperation in waste management: Initial conditions and intervention. *Journal of Economic Surveys*, *30*(3), 497–525.

Brown, C. (1992). Wage levels and method of pay. *RAND Journal of Economics*, *23*(3), 366–375.

Bucciol, A., Montinari, N., & Piovesan, M. (2015). Do not trash the incentive! Monetary incentives and waste sorting. *Scandinavian Journal of Economics*, *117*(4), 1204–1229.

Bulte, E., List, J. A., & van Soest, D. (2021). Incentive spillovers in the workplace: Evidence from two field experiments. *Journal of Economic Behavior & Organization*, *184*, 137–149.

Burlig, F., Preonas, L., & Woerman, M. (2020). Panel data and experimental design. *Journal of Development Economics*, *144*, 102458.

Callan, S. J., & Thomas, J. M. (2006). Analyzing demand for disposal and recycling services: A systems approach. *Eastern Economic Journal*, *32*(2), 221–240.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge: Cambridge University Press.

Carrus, G., Passafaro, P., & Bonnes, M. (2008). Emotions, habits and rational choices in ecological behaviours: The case of recycling and use of public transportation. *Journal of Environmental Psychology*, *28*(1), 51–62.

Cecere, G., Mancinelli, S., & Mazzanti, M. (2014). Waste prevention and social preferences: The role of intrinsic and extrinsic motivations. *Ecological Economics*, *107*, 163–176.

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980–1008.

Charness, G., Frechette, G. R., & Kagel, J. H. (2004). How robust is laboratory gift exchange? *Experimental Economics*, *7*(2), 189–205.

Czajkowski, M., Hanley, N., & Nyborg, K. (2017). Social norms, morals and self-interest as determinants of pro-environment behaviours: The case of household recycling. *Environmental and Resource Economics*, *66*(4), 647–670.

Deci, E. L., Olafsen, A. H., & Ryan, R. (2017). Self-determination theory in work organizations: The state of a science. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 19–43.

Deci, E. L., Ryan, R. M., & Koestner, R. (1999). A meta-analytic review of eperiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627–668.

Dellavigna, S., & Gentzkow, M. (2010). Persuasion: Empirical evidence. *Annual Review of Economics*, *2*, 643–669.

DIME Analytics. (2019). *IETOOLKIT: Stata module providing commands specially developed for impact evaluations.*

Dupont, D. P. (2004). Do children matter? An examination of gender differences in environmental valuation. *Ecological Economics*, *49*(3), 273–286.

Falk, A. (2007). Gift exchange in the field. *Econometrica*, *75*(5), 1501–1511.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of*

Economics, *108*(2), 437–459.

Ferrara, I., & Missios, P. (2012). A cross-country study of household waste prevention and recycling: Assessing the effectiveness of policy instruments. *Land Economics*, *88*(4), 710–744.

Ferraro, P. J., Miranda, J. J., & Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: Evidence from a randomized environmental policy experiment. *American Economic Review: Papers & Proceedings*, *101*(3), 318–322.

Ferraro, P. J., & Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, *95*(1), 64–73.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Foster, A. D., & Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, *103*(6), 1176–1209.

Frey, B. S. (1997). *Not just for the money: An economic theory of personal motivation.* Cheltenham, UK and Brookfield, US: Edward Elgar Publishing.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.

Fullerton, D., & Kinnaman, T. C. (1996). Household responses to pricing garbage by the bag. *American Economic Review*, *86*(4), 971–984.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: design, analysis, and interpretation.* New York: W.W. Norton.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, *74*(5), 1365–1384.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 1–21.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482.

Hage, O., Söderholm, P., & Berglund, C. (2009). Norms and economic motivation in household recycling: Empirical evidence from Sweden. *Resources, Conservation and Recycling*, *53*(3), 155–165.

Halvorsen, B. (2008). Effects of norms and opportunity cost of time on household recycling. *Land Economics*, *84*(3), 501–516.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Hogg, D., Favoino, E., Nielsen, N., Thompson, J., Wood, K., Penschke, A., … Papageorgiou, S. (2002). *Economic analysis of options for managing biodegradable municipal waste* (Final Report to the European Commission). Eunomia Research & Consulting, Scuola Agraria del Parco di Monza, HDRA Consultants, ZREU and LDK ECO on behalf of ECOTEC Research & Consulting.

Hornik, J., Cherian, J., Madansky, M., & Narayana, C. (1995). Determinants of recycling behavior: A synthesis of research results. *Journal of Socio-Economics*, *24*(1), 105–127.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference: For statistics, social, and biomedical sciences an introduction.* Cambridge: Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Karlin, B., Zinger, J. F., & Ford, R. (2015). The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin*, *141*(6), 1205–1227.

Kesternich, M., Römer, D., & Flues, F. (2019). The power of active choice: Field experimental evidence on repeated contribution decisions to a carbon offsetting program. *European Economic Review*, *114*, 76–91.

Kuvaas, B., Buch, R., Gagné, M., Dysvik, A., & Forest, J. (2016). Do you get what you pay for? Sales incentives and implications for motivation and changes in turnover intention and work effort. *Motivation and Emotion*, *40*(5), 667–680.

Kuvaas, B., Buch, R., Weibel, A., Dysvik, A., & Nerstad, C. G. (2017). Do intrinsic and extrinsic motivation relate differently to employee outcomes? *Journal of Economic Psychology*, *61*, 244–258.

Laibson, D. (2001). A cue-theory of consumption. *Quarterly Journal of Economics*, *116*(1), 81–119.

Lange, F., Brückner, C., Kröger, B., Beller, J., & Eggert, F. (2014). Wasting ways: Perceived distance to the recycling facilities predicts proenvironmental behavior. *Resources, Conservation and Recycling*, *92*, 246–254.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, *90*(5), 1346–1361.

Linder, N., Lindahl, T., & Borgström, S. (2018). Using behavioural insights to promote food waste recycling in urban households - Evidence from a longitudinal field experiment. *Frontiers in Psychology*, *9*(352), 1–13.

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, *14*(4), 439–457.

Meier, S. (2007). A survey of economic theories and field evidence on pro-social behavior. In B. S. Frey & A. Stutzer (Eds.), *Economics and psychology: A promising new cross-disciplinary field* (pp. 51–88). Cambridge, Massachusetts: MIT Press.

Milford, A. B., Øvrum, A., & Helgesen, H. (2015). *Nudges to increase recycling and reduce waste* (Discussion Paper No. 2015–01). NILF Norwegian Agricultural Economics Research Institute.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311–329.

Nomura, H., John, P. C., & Cotterill, S. (2011). The use of feedback to enhance environmental outcomes: a randomised controlled trial of a food waste scheme. *Local Environment*, *16*(7), 637–653.

Nyborg, K., Howarth, R. B., & Brekke, K. A. (2006). Green consumers and public policy: On socially contingent moral motivation. *Resource and Energy Economics*, *28*(4), 351–366.

O'Keefe, D. J. (2015). *Persuasion: Theory and research.* Thousand Oaks, London, New Delhi: Sage Publications.

Osbaldiston, R., & Schott, J. P. (2012). Environmental sustainability and behavioral science: Meta-analysis of proenvironmental behavior experiments. *Environment and Behavior*, *44*(2), 257–299.

Parent, D. (1999). Methods of pay and earnings: A longitudinal analysis. *Industrial and Labor Relations Review*, *53*(1), 71–86.

Petty, R. E., & Briñol, P. (2008). Persuasion: From single to multiple to metacognitive processes. *Perspectives on Psychological Science*, *3*(2), 137–147.

Petty, R. E., Briñol, P., & Tormala, Z. L. (2002). Thought confidence as a determinant of persuasion: The self-validation hypothesis. *Journal of Personality and Social Psychology*, *82*(5), 722–741.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, *41*(5), 847–855.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *83*(5), 1281–1302.

Rousta, K., Bolton, K., Lundin, M., & Dahlén, L. (2015). Quantitative assessment of distance to collection point and improved sorting information on source separation of household waste. *Waste Management*, *40*, 22–30.

Sherry, J. F. (1983). Gift giving in anthropological perspective. *Journal of Consumer Research*, *10*(2), 157–168.

Stern, P. C. (2000). New environmental theories: Toward a coherent theory of environmentally significant behavior. *Journal of Social Issues*, *56*(3), 407–424.

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

van Dijk, E., & Hultermans, R. (2020). *Handreiking kwaliteitsverbetering gft-afval via verwerkingscontracten* (Tech. Rep.). RHDHV.

Vining, J., & Ebreo, A. (1990). What makes a recycler? *Environment and Behavior*, *22*(1), 55–73.

Viscusi, W. K., Huber, J., & Bell, J. (2011). Promoting recycling: Private values, social norms, and economic incentives. *American Economic Review: Papers & Proceedings*, *101*(3), 65–70.

Vollaard, B., & van Soest, D. (2020). *Breaking habits.*

Wrzesniewski, A., Schwartz, B., Cong, X., Kane, M., Omar, A., & Kolditz, T. (2014). Multiple types of motives don't multiply the motivation of west point cadets. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(30), 10990–10995.

Chapter 3

What do my neighbors do? Leveraging social learning

3.1 Introduction

Combatting climate change and natural resource depletion are among the greatest challenges of our times. Recovery and subsequent reuse of waste materials can contribute to solving both issues – because materials recycling reduces the need for extracting new resources, and because typically less energy is required to produce goods from recycled materials than from virgin resources (Ferrara & Missios, 2005; Dijkgraaf & Gradus, 2004). Residential waste sorting and recycling are thus key ingredients of any policy aimed at fostering the circular economy and at mitigating global warming (Hogg et al., 2002).

Despite the considerable environmental – and hence societal – benefits of residential waste recycling, the bulk of recyclable materials in household waste ends up being incinerated or land-filled (Bartl, 2014). One important

This chapter is based on joint work with Daan van Soest, under the working title *'What do my neighbors do? Leveraging social learning to stimulate organic waste sorting'*. We gratefully acknowledge the very fruitful collaboration with the municipality of Schiedam (and then especially Jeffrey van Steenes and George Derksen) and with the municipality's waste collection firm. We also gratefully acknowledge the financial support by the Netherlands' Ministry of Infrastructure and Water Management (IenW), the Department of Waterways and Public Works (Rijkswaterstaat), and the Associations of Dutch Waste Collection Firms (VA), of the Dutch Waste Management Sector (NVRD), of Dutch Municipalities (VNG). We also thank Odette van de Riet, Jessanne Mastop, Reint Jan Renes, Jorn Horstman, de Goede, and especially Addie Weenk, Cees Midden, and Gijs Langeveld, for their help and support in setting up the experiment, as well as Tom Melis for his excellent research assistance.

reason for this is that waste sorting and materials recovery are valuable to society, but costly for the household. Residential waste recycling requires a series of actions for each of the separate waste flows, from the temporary storage of waste materials on the premises to their actual disposal. Costs include the space taken up by the various waste flow receptacles for temporary storage as well as the time and effort needed for sorting and disposing – and then there may be hygienic considerations too (Briguglio, 2016). Many municipalities in developed countries facilitate residential waste sorting by offering all sorts of services, such as placing bottle banks and waste paper containers in neighborhoods, and organizing the curbside collection of plastics and organic waste. But because most municipalities charge flat waste handling fees, households typically do not face any financial incentives to sort their waste (Matheson, 2019; Chamizo-Gonzalez et al., 2016).

Waste collection charges that vary with the amount of non-sorted waste collected may provide proper incentives for households to recycle (e.g. Fullerton & Kinnaman, 1996; Usui, 2008; Dijkgraaf & Gradus, 2004; Bel & Gradus, 2016). The effectiveness of waste-based taxation schemes is, however, predicated on the regulator's ability to monitor and deter evasive behavior, as they may result in recyclable waste flows being contaminated by non-recyclable materials, and also in illegal dumping or burning (Fullerton & Kinnaman, 1995). Because monitoring and enforcement typically do not enjoy much popular support in the domain of residential waste sorting, there is a policy demand for softer interventions that can be used to stimulate recycling – as a substitute for regulatory approaches, or maybe as a complement.[1]

In this paper we present the results of a Randomized Controlled Trial, implemented in a mid-sized city in the Netherlands, aimed at improving sorting of organic waste. Using a sample of close to 4000 households, we test the separate as well as the combined impact of two interventions that aim to improve recycling by providing information about fellow residents' engagement in the waste sorting activity. One intervention consists of providing information on the waste sorting behavior of local role models – residents in the neighborhood who actively sort their waste and who are willing to help convince their fellow residents of the desirability and feasibility of in-home waste sorting by (i) demonstrating that they themselves are actively engaged in the activity, and (ii) showing how they organized the waste sort-

[1] Policies requiring monitoring and enforcement may pose a political risk, but this does not mean that they cannot be effective; see for example Vollaard and van Soest (2020) for evidence to the contrary. For studies on policy interventions to prevent illegal dumping, see for example Dur and Vollaard (2015, 2019).

ing process in their homes. The other intervention consists of providing information on the waste sorting behavior of a household's direct peers, in the form of providing feedback on the average amount of organic waste collected in the nearest collective organic waste container. We will refer to these interventions as the social modelling and social feedback treatment, respectively.

Our interventions aim to harness elements of social learning to improve organic waste sorting by households. Social learning is a process by which new behaviors can be acquired by observing and imitating others (Bandura & Walters, 1963; Bandura, 1977). We conjecture that (enabling) social learning is particularly relevant in the context of (organic) waste sorting for four reasons. First, lack of information on how to sort ranks high among the reasons households state why they fail to engage in waste recycling schemes (Hornik et al., 1995; Gamba & Oskamp, 1994; Vollaard & van Soest, 2020). Seeing how others sort their waste may then be an efficient way for the household to acquire the relevant information. Second, the fact that other households are observed to engage in waste sorting may provide a signal about the (societal) relevance of the activity that is more persuasive than individually acquired information. For example, Banerjee (1992) and Bikhchandani et al. (1992) show that agents may be more inclined to base their decision on imperfect signals transmitted by others than on their own private information; see also Asch (1956). Third, people tend to infer what "normal" behavior is from the behavior of others (Goldstein et al., 2008). Because people have a preference for conforming their behavior to what is perceived to be normal (Schultz et al., 2007), information on the pro-social behavior of others (either by personal observation, or via explicit communication) may induce the recipient to start acting pro-socially too (see for example Allcott, 2011; Allcott & Rogers, 2014; Ferraro et al., 2011; Ferraro & Price, 2013; Giaccherini et al., 2019). Fourth, the perceived benefits of contributing to a public good may depend on how many others are willing to do so as well. Ostrom (2009) identified "conditional cooperation" as a key preference that helps support collective action, and Fischbacher et al. (2001) documented its prevalence in humans (see Vringer et al. (2017) for a field test).

We find that both the social modelling and the social feedback treatment are effective in improving organic waste sorting – as measured by increases in the average frequency with which the collective organic waste sorting facilities are used, as well as by increases in the share of households using them. We also find that both treatments seem to be especially effective in improving organic waste sorting by households that were already engaged in the activity before the start of intervention (albeit that the evidence for a heterogeneous impact is much weaker for the social modelling treatment than for the social feedback treatment). We document these positive impacts during the treatment implementation period, and we also find that the impacts of both treatments remain positive and significant, on average, over the nine month post-implementation. However, while we find no evidence of the social feedback treatment's impact deteriorating over time, the social modelling treatment's effect sizes do decrease over time. Last but not least, we fail to document any added impact of combining the two interventions (above and beyond the impact of just the social feedback intervention).

The design of our Randomized Controlled Trial is fairly complex, but it provides high-powered tests for a variety of key aspects of the behavioral interventions. For example, it allows us to explicitly test for possible spillover effects from treatment to control households – for which we find no evidence – and it also allows us to explicitly test whether indeed an increase in the *frequency* of organic waste disposals actually translates into an increase in the *weight* of organic waste collected. More specifically, we document that the percentage increases in the frequency with which the organic waste facilities are used provide a lower-bound estimate of the percentage increase in the amount of organic waste collected.

In addition to collecting information on households' usage of the collective organic waste sorting facilities, we also gathered survey evidence providing insights into the mechanisms via which the treatments affected waste sorting behavior. We find that, on average, having been exposed to one of the three treatments significantly strengthened the household's attitude towards waste sorting. We also find that the treatments increased the perceived benefits of organic waste sorting, a result that seems to be driven by especially higher confidence in the waste recycling process. However, we also find (somewhat weaker) evidence that self-efficacy was improved too. Households for instance reported to find it easier to remember the sorting rules and reported to have better insight in sorting performance. The survey outcomes thus document an improvement in the (perceived) cost-benefit ratio of organic waste sorting, making it more likely that the improved organic waste behaviors are likely to be sustained in the future.

The insights obtained from this study are important for two reasons. First, for all those waste flows and/or waste collection situations where ex-post waste sorting (i.e., sorting not at home but at the waste treatment facility) is either not feasible or too costly, our Randomized Controlled Trial provides evidence on whether (and to what extent) recycling rates can be improved by means of fairly simple information-provision interventions. In terms of the contribution to the literature on behavioral policies (or nudges, see Thaler & Sunstein, 2008), this is important because unlike domains like energy and water savings, engaging in waste sorting is a pure public good without any (direct) private financial benefits. And because of the fact that we have data up to nine months after the completion of the interventions, we are able to assess not just the interventions' short-run impact but also their effectiveness in the longer run. Second, and related, waste sorting behavior is typically thought of as a behavior that is driven by habits (Knussen & Yule, 2008; Carrus et al., 2008). Changing them thus poses a major challenge (as getting rid of past non-sorting habits is difficult), but also that, if the behavior is changed in the short run, we may expect waste sorting behavior to continue long after the interventions themselves have been discontinued. And because many other environmental behaviors are also habit-driven (Carrus et al., 2008; Gsottbauer & van den Bergh, 2011), the insights of this paper may spill over to other environmental policy domains as well.

The literature on residential waste sorting is quite rich and has covered a large variety of topics. These include analyzing the importance of economic incentives for recycling (see for example Fullerton and Kinnaman (1995, 1996); Fullerton and Wolverton (2000); Kinnaman and Yokoo (2011); for overviews see Dijkgraaf and Gradus (2015) and Bel and Gradus (2016)), but also the role of non-financial drivers therein (Aadland and Caplan (2003) and Kinnaman (2006); for an overview see Briguglio (2016)). Theory that can explain the decision to sort – even in the absence of financial returns – has been developed by Brekke et al. (2003, 2010) and also by Bénabou and Tirole (2006), highlighting the relevance of, among others, self-image and reputational concerns; for empirical tests see for example Viscusi et al. (2011); Czajkowski et al. (2019); Hage et al. (2009); Kirakozian and Charlier (2015); Nyborg et al. (2006). The literature using Randomized Controlled Trials to assess the effectiveness of different (behavioral) policy interventions on recycling is, however, much less well-developed (Briguglio, 2016). Linder et al. (2018) tested the impact on recycling rates of a so-called community-based marketing campaign in one of the suburbs of Stockholm. They found that informing households about their fellow residents' views on waste recycling resulted in a significant increase in recycling rates up to eight months after the information leaflet had been sent out. Nomura et al. (2011) randomized the provision of information on street-level food waste recycling rates in Greater Manchester, and found that providing such information resulted a three percent increase in waste sorting up to three months post treatment. Milford et al. (2015) analyzed the impact of a similar intervention: providing households in a city in the south-east of Norway with information on how their waste sorting performance compared to that of the average household in their district. The information was relayed by means of two letters, sent out nine months apart. Milford et al. (2015) found that relaying this information resulted in a two percentage point increase in recycling up to (at least) five months after sending out the second letter. Linder et al. (2018), Milford et al. (2015) and Nomura et al. (2011) all document positive effects with little decay. We contribute to the insights they provided by testing the impact of interventions harnessing social learning over a substantially longer time horizon, allowing us to better test whether indeed the behavioral impacts are long-lasting.

The remainder of this paper is set up as follows. Section 3.2 introduces the study context as well as our experimental design. Our identification strategy is presented in Section 3.3, and the results are presented in Section 3.4. The survey results, aimed at probing the treatments' underlying mechanisms, are presented in Section 3.5. Section 3.6 concludes.

3.2 Design and Methods

3.2.1 Background and experimental setting

We implemented our Randomized Controlled Trial (RCT) in a neighborhood in Schiedam, a city of about 80,000 inhabitants that is part of the Rotterdam–The Hague metropolitan area in the highly urbanized Western part of the Netherlands. As is the case in most municipalities in the Netherlands, increasing organic waste recycling rates – and then especially of cooking waste and food leftovers – ranks high on Schiedam municipality's policy agenda. Schiedam's standard system of organic household waste collection is via the use of wheelie bins, to be placed at the curbside on organic waste collection days. The wheelie bins are quite sizable and require outdoor storing space with easy access from the storage location to the roadside. That means the wheelie bin system is not well-suited for a considerable share of households (in Schiedam, but also elsewhere in the Netherlands). This includes households living in multi-family dwellings (such as low- and high-rise apartment buildings). But it also includes all households living in single-family homes with insufficient outdoor storage space for the wheelie bin, and/or without easy access from the storage location to the wheelie bin collection point at the curbside. At the time we implemented the RCT all households in the municipality of Schiedam had access to waste sorting facilities for some waste fractions (such as paper and glass), but only those households that had been provided with wheelie bins were able to separately sort their organic waste.[2]

To facilitate organic waste sorting by households living in dwellings not serviced by the wheelie bin system, collective organic waste collection facilities were introduced in our pilot neighborhood in the Fall of 2017. The neighborhood covers about one square kilometer and is home to about 4,000 households (or 9,000 individuals), most of whom live in multi-family dwellings. Forty-four outdoor collective organic waste containers were placed in the neighborhood's public spaces; see Figure 3.A1 in Appendix 3.A.1 for a picture of one of these containers. Usage of these facilities was intended for all those households that were not serviced by individual curbside organic waste collection using wheelie bins. Access to the collective organic waste containers was restricted by means of a card system; opening the lid required swiping a card that was made available only to those households that could not be serviced by wheelie bins.

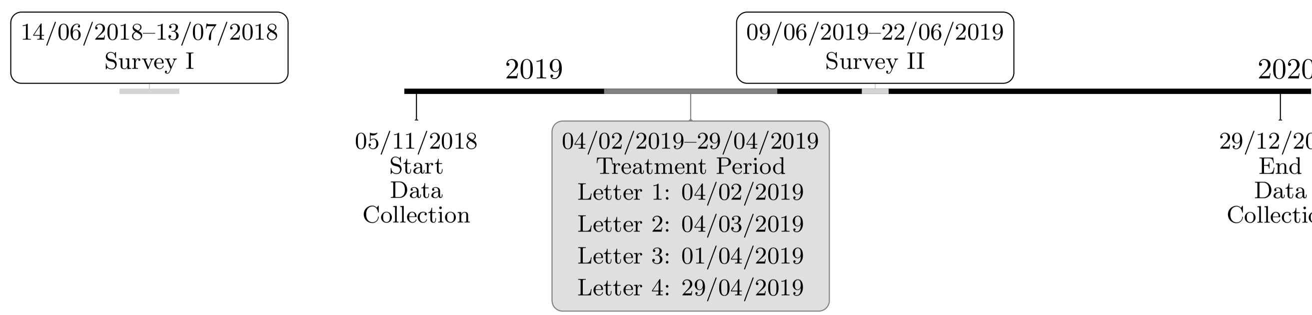
The municipality realized that while the availability of collective waste collection facilities is a necessary condition for the targeted households to sort their organic waste, engagement in the activity is not automatic. Organic waste recycling rates – and then especially those for cooking waste and food leftovers – are typically fairly low even among households serviced by way of wheelie bins, and they are even lower among households in multi-family dwellings. Organic waste sorting is typically viewed as one of the most inconvenient types of recycling activities – because the temporary in-house storage requires space and may be perceived as unhygienic, and because walking distances to the collective outdoor organic waste containers can be considerable (Sidique et al., 2010; Lange et al., 2014). The municipality therefore combined the introduction of the collective organic waste containers with an information campaign aimed at emphasizing the environmental and societal benefits of organic waste sorting. Each household

[2]The situation in Schiedam is very much representative of the overall situation in the Netherlands. A substantial share of the Dutch population engages in recycling, and then especially of paper and glass. Paper waste is often collected at the curbside by sports clubs (as a supplementary source of revenue for these club in addition to their membership fees), and bottle banks are conveniently located next to supermarkets so that households can dispose of their glass waste when shopping for groceries. Metals, textiles and household chemical waste can be deposited at the municipality's waste treatment site. Organic waste is typically collected only from households that can be serviced by means of wheelie bins; the availability of (collective) organic waste sorting facilities for households living in multi-family dwellings or in bin-inaccessible single-family dwellings is still novel in the Netherlands. The same holds, to a large extent, for plastics. The waste sorting system in Schiedam is thus representative of that in the Netherlands.

in the research area received a package that contained information on how to sort waste (e.g. what types of waste are suitable for organic waste recycling) in the form of a leaflet, a cheat-sheet and a sticker for quick reference. Households were also offered a receptacle to be placed on the kitchen top, to facilitate the temporary storage of cooking waste and food leftovers to be disposed of later in the outdoor collective organic waste collection facilities. Each receptacle was accompanied by a roll of compostable bags to line the receptacle with.

3.2.2 Treatment description

Despite all the efforts described in the previous Subsection, use of the collective organic waste facilities was very modest at best, as evidenced by our baseline data. In the baseline period (from November 5, 2018, to February 4, 2019), less than 10 percent of the households made regularly use of the organic waste collection facilities. We teamed up with the municipality of Schiedam to test whether (and how much) organic waste sorting can be increased by means of three interventions aimed at harnessing elements of social learning; an intervention presenting fellow residents of the neighborhood as a role models (the social modelling treatment), an intervention aimed at providing feedback on the actual sorting behavior of one's peers (the social feedback treatment), as well as the combination of the two (the combined treatment). The time line of the experiment is depicted in Figure 3.1. The two core treatments (that are implemented in isolation as well as combined), are presented in more detail in the next two subsections; the added benefits of also testing the combined treatment, are discussed in the third subsection.



The black line visualises the organic waste data collection period. Both the Social modelling & Social Feedback Treatment took place in the Treatment Period.

Figure 3.1: Timeline of the Randomized Controlled Trial.

The social modelling treatment

The social modelling treatment was designed and implemented as follows. We aimed to identify four residents of the neighborhood, two male and two female and of different age classes, (i) who were (likely to be) engaged in organic waste sorting, and (ii) were likely to be well-known in the neighborhood, for example because they were active in the local community center etc. Age and gender mattered for the selection, because the persuasiveness of a message tends to vary with its recipient's sense of shared identity with its sender (Petty et al., 1981; Bandura, 1988; Goldstein et al., 2008). We asked the municipality to use their contacts in the neighborhood, to suggest four names; the municipality was well-positioned for this task because of the close connections the municipality's community workers had with the neighborhood. The four individuals suggested were subsequently contacted, again via the intermediation of the neighborhood's community workers, to ask them whether they would be willing to help set up an information campaign in the neighborhood to improve organic waste recycling rates. All four agreed. They were then asked to participate in a photo shoot in which they demonstrated the steps they take when sorting their organic waste, and they were also interviewed about their motivations to sort waste. We then created four comic-like types of leaflets, one for each model. Figure 3.A2 in Appendix 3.A.1 presents one of those leaflets. The social modelling households received all four leaflets, one at the time and accompanied by a letter, at four-week intervals. The first leaflet was sent out on February 4, 2019, and the fourth on April 29, 2019.

The essence of each of the four leaflets is the role model informing the recipient of her motives to engage in (organic) waste sorting, as well as demonstrating how she has organized her in-house waste sorting process to make it as convenient as possible. Both aspects, but especially the latter, are key aspects of social learning as defined by Bandura and Walters (1963) and Bandura (1977). By watching the actions of the role model the observer learns how an activity is done best. The information on the social model's behavior and her in-house organic waste sorting process, may enhance the recipient's perception of the attainability of the task, or how well he is able to implement waste sorting himself (e.g. Bettinger et al., 2005; Oster & Thornton, 2012; Eble & Hu, 2020). It may so enhance the recipient household's self-efficacy in the domain of in-house waste sorting, as defined by Bandura (1982). But the observer is also led to infer that, *because* the role model engages in the activity, she is of the opinion that organic waste sorting is an important societal cause to contribute to. The fact that we added explicit information on the role model's motivations to sort thus simply reinforced the conclusions the observer would have drawn himself by just observing the model's behavior.[3] The social modelling information may so induce households to update their beliefs regarding the (environmental) benefits associated with the activity.

While social learning is the key mechanism via which the leaflets are expected to induce behavioral change, other mechanisms may be at play as well. The information that a specific individual thinks that waste sorting is important may also be perceived as an appeal to moral preferences (Brekke et al., 2003, 2010) or even as the communication of an (injunctive) social norm (Goldstein et al., 2008).[4]

The social feedback treatment

The social feedback intervention consisted of informing households of the average amount of organic waste that was collected and deposited by all households in their direct vicinity.[5] We first identified which of the newly installed collective organic waste containers was closest to each treatment household ; for each social feedback household, its 'households in the direct vicinity' were then defined as all those households for which that container was also the nearest. Households were then sent a series of four letters, at monthly intervals, that informed them of the amount of organic waste that was deposited by the – on average 100 – households in their direct vicinity. We also added an injunctive norm in the form of a clear objective (an average amount of organic waste sorted, per household per month, of 1 kg) as well as a yardstick (in the form of a star system) for households to

[3]The main reason why we explicitly added the social model's motivations to the instructions in the leaflet is that instructions without motivation can be construed as quite paternalistic.

[4]As argued in a recent paper by DellaVigna and Linos (2021), large-scale field experiments typically do not allow for a clean estimate of a single possible mechanism. The main reason for that is that implementation of large-scale field experiments requires the cooperation of an implementing partner (typically a firm, or a government body). These partners are typically more interested in identifying an intervention that generates an impact than in obtaining insight into the exact mechanism via which the impact is generated. Because of that reason, treatments in large-scale field experiments are better described as package interventions (moving various variables at the same time) than as clean experiments in which just one variable differs between each two treatment groups. Our study is not an exception.

[5]The implicit assumption is that if households engage in organic waste sorting, they most likely use the container that is closest to their home. We find that indeed in 82% of the cases households use the container nearest to their home.

Average organic weight per household in the household-container cluster per week	Number of stars awarded
<150 gram	*
150-350 gram	**
350-675 gram	***
675-1000 gram	****
>1 kilo	*****

Table 3.1: Threshold weights of the weekly amount of organic waste deposited as used to compute the number of stars in the social feedback treatment.

easily evaluate the organic waste sorting performance of the other households in their direct vicinity. Every week the amount of organic waste deposited in each container was weighed. If that weight, divided by the number of households in the vicinity and averaged over the previous four weeks, was below 150 grams, the feedback letter sent to all households in the container's vicinity would depict one star. Two stars would be awarded if the average weight was between 150 and 350 grams; to receive five stars, the average weight of organic waste deposited per household over the past four weeks needed to be a 1000 grams, or more. Table 3.1 shows the cutoff weights per household that are used to determine the star scores. An example of a social feedback letter is presented in Figure 3.A3 in Appendix 3.A.1. The timing of the letters coincided with those of the social modelling treatment; the first was thus sent on February 4, 2019, and the fourth and last on April 29, 2019.

As was the case with the social modelling intervention, the social feedback intervention also consisted of several elements. Two of the social learning elements that may be present in the social modelling treatment, may also be at work in the social feedback treatment. The information that a certain amount of organic waste has been deposited by households in one's direct vicinity may induce treatment households to update their beliefs about the feasibility of organic waste sorting; if others can do it, they themselves are likely to be able to do it too (see for example Bramoullé & Kranton, 2007). And the fact that other households in one's direct vicinity engage in organic waste sorting may induce households to update their beliefs about the societal relevance of organic waste sorting: if sorting happens in one's direct vicinity, this is a signal that at least some neighbors find recycling a worthy cause to contribute to.

Of course, other mechanisms may be at work too. Being provided with information that at least some of the neighbors engaged in waste sorting may have been perceived as a signal of what responsible behavior looks like, thus providing an appeal to moral preferences and/ or self-image considerations (Nyborg et al., 2006). Information about the average sorting behavior of all neighbors in one's direct vicinity may also have been perceived as a descriptive social norm the treatment household may want to conform to (Goldstein et al., 2008); we added the star system as an injunctive norm to prevent possible boomerang effects that might be caused by the very low pre-intervention organic waste sorting rates (Schultz et al., 2007). The star system also provided a collective waste sorting challenge households may have been sensitive to (Gómez-Miñambres, 2012; Ashraf et al., 2014).[6]

The combined treatment

The above description of the social feedback and the social modelling treatments presents the possible mechanisms via which they may affect household behavior. The fact that households are informed that other households engage in organic waste sorting (regarding the role model's behavior, or because of the information on the weight collected in the container closest to them over the previous month) may induce them to update their beliefs about (i) the feasibility of organic waste sorting (if others can do it, they themselves are likely to be able to do it too) and (ii) the social relevance of organic waste sorting (if others do it, this is a signal that at least some neighbors find recycling a worthy cause to contribute to). These mechanisms may thus be at play in both the social modelling and social feedback treatments — albeit possibly to different degrees. The main difference between the two is that the social modelling intervention explicitly shows, in the form of the comic, the steps taken by the role model of how to sort waste. Demonstrating these steps, as well as the tips and tricks associated with that, can help increase the recipient household's self-efficacy – a mechanism that cannot be present in the social feedback treatment. The social feedback letters did not provide tips and tricks on how to organize one's (organic) waste sorting process, and one cannot derive such information from observing others

[6]Households may have perceived the star scores as a challenge to reach a specific collective waste sorting target (1 kg per household per week, for all households making use of that container). This mechanism may well have been in place, but it is unlikely that this has resulted in a competition between neighborhoods, as households were not informed of the (star score) performance of other neighborhoods. For evidence on the impact of harnessing between-neighborhood competition, see Nomura et al. (2011).

depositing their organic waste in the outdoor collective container. We thus decided to test not just the effectiveness of the social feedback and the social modelling treatments, but also of the combination thereof. The treatment thus consisted of adding the social modelling flyer (the comic) to the social feedback package.

3.2.3 Experimental design

We thus implemented three different treatments – the social modelling treatment, the social feedback treatment, a treatment that combined both types of interventions – to test whether the two core treatments are mutually reinforcing when implemented jointly. Waste sorting outcomes in each of these treatment groups were compared against those of the fourth group that did not receive any information – the control group. In terms of outcome variables, we have information on the daily waste sorting behavior of 3987 households (consisting of, in total, 8990 individuals), as well as on the weight of organic waste deposited, per week, in 42 container-units[7] (from here on, 'containers'). Frequency of use of the collective organic waste sorting facilities is available in the form of the number of times households used their card to get access to the collective organic waste containers. The weekly weight data are available because we asked the municipality to weigh each organic waste container's content before it was emptied. Usage data are thus available at the household level, the weight data only at the container level.

From an environmental (and cost-effectiveness) perspective, the key treatment impacts of interest are those on the amount of organic waste collected (i.e., the change in weight). Behaviorally, however, we are also interested in the impact of the treatments on the frequency with which households made use of the facilities – for example because we are interested in establishing whether the treatments are effective in inducing non-sorting households to start sorting (the extensive margin), or whether they are especially effective in intensifying waste sorting of those households that were already engaged in the activity (the intensive margin). If frequency and aggregate weight are perfectly correlated – that is, if the treatments affected the frequency

[7] As stated before, 44 containers were installed in the research area. One of those containers was used by just seven households, and one organic waste collection location housed two containers. We excluded the former container as well as its seven users because the number of users is too low to be able to weigh the amount of organic waste deposited therein with sufficient precision. Regarding the latter two containers, households in the vicinity were equally likely to use the one or the other, and hence we treat the two as just one single container-unit.

of use but not the average weight per deposit – we can design the RCT to maximize our chances of detecting treatment effects on the frequency of use. If we cannot rule out that the treatments affected the weight per deposit, the optimal design is slightly more complicated.

The assumption of the per-deposit weight being independent of the treatment may be valid, because households are likely to postpone making the next trip to the collective waste sorting facilities until their in-house temporary storage receptacle is full. So for a receptacle of a given size, the (unobserved) total weight deposited by a household would be proportional to the (observable) number of times it made use of the outdoor organic waste collection facilities. Yet frequency and weight may not be perfectly correlated for two reasons. First, for households that already engaged in some organic waste sorting at baseline, the weight per deposit may change if the treatment induced them to purchase a larger bin for their temporary in-house organic waste storage. The number of trips to the outdoor organic waste container may then have remained constant or may even have decreased, while the amount of waste deposited (and hence the weight) was actually larger. Second, for households that previously did not sort but were induced by the treatment to start sorting, changes in the average frequency would only be a good indicator of the changes in total weight collected if the average weight of the waste they deposit happened to be equal to the average weight deposited by those households that already engaged in the activity.

We thus cannot assume that the average weight per deposit is unaffected by the treatments, and hence we face the following trade-off regarding the design of our RCT. On the one hand we can maximize the statistical power to detect changes in behavior by implementing our (stratified) randomization at the level of individual households. With about 4000 households in the RCT and with four treatment arms, we would expect to then have very good balance on all characteristics (observable and unobservable) that affect waste sorting. We would then also expect to be able to obtain precise estimates of the treatment effects on the frequency with which the average household made use of the organic waste collection facilities – assuming that there were no treatment spillovers between neighbors (see below). And we would also have maximum power for any test on heterogeneous treatment effects – including those on whether a treatment was especially effective in moving the intensive or the extensive margin of waste sorting. With randomized assignment at the individual level, however, the share of deposits made by households in each of the four treatment arms would be the same for each container, and hence it would be impossible to estimate the treatment impact in terms of weight collected.

On the other hand we could maximize our chances of being able to estimate our treatments' weight impacts by allocating all households in the vicinity of a specific container to one and the same treatment. Assuming that households are most likely to make use of the nearest organic waste container, we could create 42 household-container clusters, each containing all households for which a specific container was most proximate to, and subsequently randomize each household-container cluster to one of the four treatment arms. While a full cluster-randomized design would maximize our chances of being able to estimate the treatment impacts on the amount of weight collected, estimating the treatment impacts on households' waste sorting behavior would be more hazardous because households within a household-container cluster are likely to be more similar to each other than households in different household-container clusters. Dwellings in the same location are more similar in terms of characteristics like size and sales price than those in different locations, and they may also attract occupants that are more similar to each other (in terms of preferences, but also in terms of disposable income) than to those that end up living in different types of dwellings elsewhere in the neighborhood. This would imply that outcomes of households in the same cluster are likely to be correlated, and hence, for the same sample size, statistical power would be (much) lower with a cluster-randomized design than with individual randomization. Because we are interested in the treatment impacts on both weights and frequency of use, we opted for a mixed design, combining a cluster-randomized approach for the estimate of one treatment effect and a within-location individual-level randomization to estimate the impact of the other two treatments. Our experimental design is summarized in Figure 3.2; details about the actual randomization process, including stratification, is relegated to Section 3.2.4.

We decided to use cluster-level randomization to estimate the impact of the social feedback treatment. The main reason for this is that social feedback on the average weight collected needs to be implemented at the container level, and hence estimating the impact of this treatment necessitates the implementation of a cluster-randomized design. Of the 42 household-container clusters, we allocated 11 to just the control group, and 10 to just the social feedback treatment; see Figure 3.2.[8] The impact of the social

[8]Figure 3.2 summarizes the design, with the number of containers allocated to one or more treatments in the first stage of the randomization and the numbers of households to be allocated to each individual treatment in the second stage. On average there are about 95 households per household-container cluster (about 4000 households making use of 42

feedback treatment can then be estimated by means of a simple two-way difference-in-difference estimator using these two groups of households, with standard errors clustered at the unit of randomization – at the household-container cluster level (Bertrand et al., 2004). We will refer to this as Test I; see the bottom part of Figure 3.2. A more detailed discussion of the actual empirical strategy is deferred to Section 3.3.

Next, to estimate the impact of social modelling, we applied individual randomization to the households in 11 household-container clusters. Half of the households in each of these household-container clusters were to receive the social modelling treatment, and the other half ended up in the control group. With individual randomization we can again implement a simple two-way difference-in-difference estimator with standard errors clustered at the unit of randomization – the individual household. This test is referred to as Test II in Figure 3.2.

The third treatment combines all elements of both the social feedback and the social modelling intervention. To estimate whether there is any interaction effect between the two constituent interventions, all of the households in the remaining 10 household-container clusters were to receive the social feedback treatment, and a randomly selected half of the households in each household-container cluster were to also receive the social modelling treatment. Treatment assignment is again based on individual randomization, and estimation is therefore by means of a two-way difference-in-difference model with standard errors clustered at the level of individual households. Because all households in this test receive the social feedback intervention, the coefficient on the social modelling treatment indicator captures the treatments' interaction effect. In Figure 3.2 we refer to this test as Test III.

containers); the exact number of households ending up in a treatment is the result of the actual randomization, to be discussed in Section 3.3. This also explains why the number of household-container clusters to be allocated to the pure control and pure feedback groups is not exactly the same; the average number of households per household-container cluster is slightly lower in the Pure Control household-container clusters than in the Pure Feedback household-container clusters.

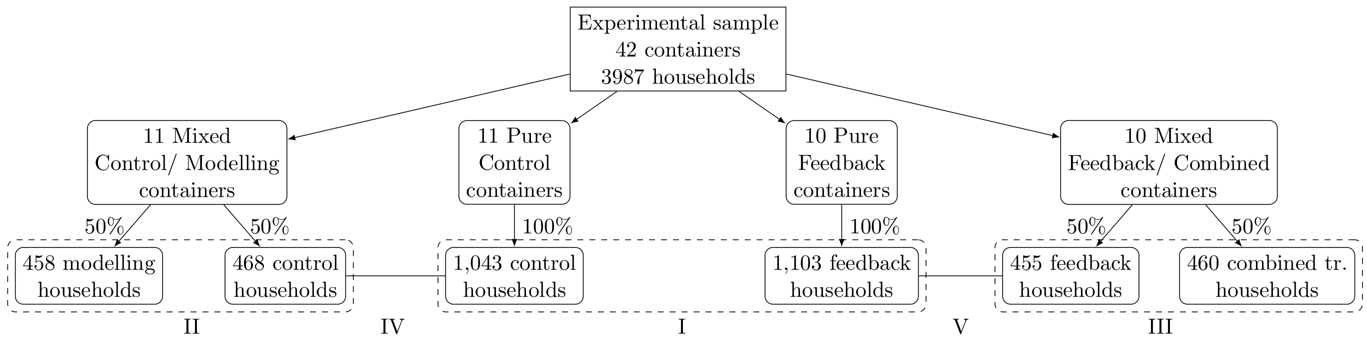


Figure 3.2: Visual representation of our two-stage randomization strategy.

Note that the first level presents the allocation of containers (and hence household-container clusters) to container types, and the second level the allocation of households within the household-container clusters to treatments. The dashed boxes at the second level indicate the subsamples used to estimate the key behavioral treatment effects, the uninterrupted lines indicate the subsamples used to test for possible spillover effects.

Our design serves three purposes. First, the two-stage randomization process allows for properly powered tests of each of the three main treatment effects on household waste sorting behavior – the frequency with which they use the collective organic waste sorting facilities. Despite the fact that relatively few household-container clusters are used for Tests II and III (11 and 10, respectively; see Figure 3.2), these tests are adequately powered because randomization is at the individual level, and takes place within specific household-container clusters. The latter implies that the randomization is stratified on location, and hence implicitly on a fairly large number of other characteristics, observed and unobserved, such as dwelling type, sales price and possibly even disposable incomes. That households within a household-container cluster are more similar than between clusters thus enhances statistical power for Tests II and III, but it also implies less statistical power, all else equal, for Test I (the impact of the social feedback intervention), because of its cluster-randomized design. To compensate for this, we assigned relatively many household-container clusters to Test I (21 in total; all those households in the household-container clusters that have been labelled Pure Control and Pure Feedback in Figure 3.2).

By adjusting the number of household-container clusters assigned to each of the three tests depending on the type of randomization, we thus ensure that we have adequate statistical power for all three. We calculated the Minimum Detectable Effects (MDEs) using the STATA package `pc_simulate`, developed by Burlig et al. (2020). Power analyses of panel data are complicated because the statistical power of the test crucially depends on the serial correlation in individual households' waste sorting decisions – oftentimes in a complex and non-linear fashion (Burlig et al., 2020). On the one hand, higher levels of serial correlation facilitate detecting differences between pre- and post-intervention observations. On the other hand higher levels of serial correlation imply (steeply) declining added benefits of having a longer time series for each cross-sectional unit, because each extra observation yields less new information. Burlig et al. (2020)'s STATA package `pc_simulate` allows the researcher to derive the MDE for difference-in-difference models in which within-unit and/or between-unit correlation in the error terms is controlled for using clustered standard errors. We find that with 80% power, the MDE for Test I is equal to an increase in the frequency of use of the collective organic waste sorting facilities of 0.021 extra days per week, and for Tests II and III it is 0.034 extra days per week.

Second, this design also allows us to test for the presence of spillovers (Baird et al., 2018). The claim that Tests I-III provide reliable estimates of the actual treatment effects is predicated on two assumptions. For test I, we assume that there is partial inference, there may be spillovers between neighbors located in the same container-cluster in the social feedback treatment, since these households receive a letter informing them of the sorting efforts of their neighbors, but spillovers between container-clusters are absent. Under this assumption, the difference in behavior between social feedback and pure control clusters equals the total effect of treatment (Hudgens & Halloran, 2008). This effect embeds both the direct and indirect effect of the treatment, which may materialize when households react to a new feedback letter communicating changes in the sorting behavior of their neighbors. For test II and III we invoke the Stable Unit Value Assumption (SUTVA); see (Rubin, 1978). Both these assumptions imply that spillovers from the treated to control group households to which they are compared are absent. While Test I is likely to yield unbiased impact estimates of the social feedback treatment, this is not necessarily the case for the impact estimates of the social modelling and the combined treatments (Tests II and III, respectively). As neighbors can be randomized into different treatment arms, the effectiveness of the treatment may be underestimated if households in the treatment group (the social modelling treatment for Test II, and the combined treatment for Test III) affect the behavior, directly or indirectly, of nearby households that had been assigned to the other treatment arm (the control group in case of Test II, and the households that only received the social feedback in case of Test III). Our two-stage randomization design allows us to explicitly test for the presence of such spillovers (Hudgens & Halloran, 2008). More specifically, we can test for possible spillovers of the social modelling treatment by comparing the organic waste deposit frequencies of the control group households in the Mixed Control/ Modelling household-container clusters to those of the households in the Pure Control household-container clusters; see Test IV in Figure 3.2. Similarly, we can also test for the impact of neighbors' having received the combined treatment on those that only received the social feedback treatment – by running tests on the differences in behavior between the households in the Mixed Feedback/ Combined household-container clusters that just received the social feedback intervention and the social feedback households in the Pure Feedback household-container clusters; see Test V in Figure 3.2.

Third, this design does not only allow us to estimate the treatment impacts on the frequency of organic waste sorting; it also allows us to determine whether the estimated percentage changes in the frequency of use is an overestimate or an underestimate of the percentage change in organic waste collected. The first randomization step provides us with 42 containers that are allocated to four groups that differ in the shares of nearby households that are treated by a specific intervention – zero, fifty or one hundred percent. The differences in the shares of treated households making use of the various household-container clusters yield the variation needed to test whether the average weight of organic waste deposits is affected by each treatment, and hence whether the observed change in frequency of use is an under- or an overestimate of the treatments' impact on the amount of organic waste collected.

3.2.4 Data, randomization procedure and treatment balance

Measuring organic waste sorting activity

We construct two measures of organic waste sorting activity. The digital log books of the collective containers' access card readers provide us with information on the number of times each household accessed the outdoor organic waste collection facilities. It is safe to assume that if a household accessed an organic waste container, it was to deposit (some of their) organic waste.[9] Our first measure of organic waste sorting activity is the number of unique days in a week a household used their card to get access to an organic waste container. This variable, which we will refer to as "the number of disposal days", can take on a value between 0 and 7. If, for example, we report the average frequency with which households used the organic waste collection facilities is 0.2 days per week, this means that the average household deposits its organic waste once every 5 weeks.[10]

Our second measure of organic waste sorting activity captures whether or not a household made use of the organic waste sorting facilities in a week.

[9]Ideally, of course, (i) all of a household's organic waste ended up in the collective organic waste containers, and (ii) no non-organic materials were deposited in them (no glass, paper, textiles, and also no residual waste items). We have no information on the total amount of organic waste produced by each household, so we cannot measure what share was recycled. However, we were able to monitor to what extent the organic waste fraction was polluted by non-organic materials. Pollution rates were very low, and hence we can safely assume that if households made use of the organic waste collection facilities, it was to dispose near-unpolluted organic waste.

[10]In 84% of the cases in which households used their card to gain access to an organic waste container on a specific day, they just used it once. However, there are also instances in our data set in which the system registered that a household swiped their card multiple times on the same day – up to 59 times. While swiping one's card 59 times on a day may reflect highly dedicated waste sorting, it is more likely that the system malfunctioned, and the lid failed to open the first time the card was used. Information on the number of distinct days in a week the facilities were used is thus more likely to properly capture the intensity of waste sorting behavior than the total number of times the card was used in a week.

Changes in this variable reflects the extensive margin of organic waste sorting, and hence the comparison with the treatment impact on this variable and that on the number of disposal days gives insight into whether the treatment affected the intensive margin of waste sorting, the extensive margin, or both.

Our two main dependent variables are thus households' frequency of use of the collective organic waste sorting facilities, as well as the share of households making use of them. These are important behavioral variables, but ultimately the success or failure of the interventions depends on whether or not they were able to increase the weight of organic waste collected. We were unable to weigh individual deposits, but we have access to the total weight of organic waste deposited in each container, for a large number of weeks. The total amount of weight deposited is thus our third key outcome variable of interest.

Stratification procedure and treatment balance

Our two-stage randomization process, as explained in Section 3.2.3 and summarized in Figure 3.2, was implemented as follows. In the first stage, 11 of the 42 household-container clusters were to be assigned to the just control group, 10 were to be assigned to the just the social feedback group, 11 were to be selected whose households were subsequently to be split, 50-50, to either social modelling or control, and 10 were to be selected whose households would subsequently be split, 50-50, to either just social feedback or the combined social feedback and social modelling intervention. We will refer to these four groups of household-container clusters as Pure Control, Pure Feedback, Mixed Control/ Modelling and Mixed Feedback/ Combined, respectively.

Randomized assignment of the 42 household-container clusters to each of the four groups was stratified on (i) whether the average value of the homes of the household-container cluster was above or below the median, (ii) whether or not the number of households in the household-container clusters was above or below median, and (iii) whether the average weight of the organic waste collected by each container over the last four weeks preceding the randomization was above or below the median. Because we stratified on both the number of households in a household-container cluster and on the average weight collected, we also implicitly stratified on the average intensity with which households engage in organic waste sorting. We also stratified on the average tax value of the homes because it is likely to capture a variety of dwelling and occupant characteristics (e.g., type of dwelling, size, disposable income) which may affect households' ability to or perceived relevance of waste sorting (Briguglio, 2016).

The randomization process for the social feedback intervention (Test I) consisted of just one stage – because of the cluster-randomized approach. A second randomization stage was implemented for all households in the Mixed Control/ Modelling and Mixed Feedback/ Combined household-container clusters (Test II and Test III, respectively). The second-stage treatment allocation of households within each household-container cluster was stratified on characteristics that have been documented to be strongly correlated with waste sorting (Briguglio, 2016): (i) a median split indicator for distance to the nearest organic waste container, (ii) family size (whether or not the number of individuals in a household was 3 or more), and (iii) whether the household made use of the collective organic waste container at least once in the time period between November 12, 2018 and January 9, 2019 – the two-month period between the start of the data collection and the moment at which the randomization was executed. As explained in Section 3.2.3, our design implies that the randomization for Tests II and III was also stratified on location, as randomization is within household-container clusters. And because dwellings (and hence occupants) are more likely to be similar within a geographic area than between, that means that we expect to have good balance on quite a number of other variables too – think of dwelling type, size, and tax value, and hence also family size, composition and disposable income.

Table 3.2 provides the descriptive statistics of the full household sample (in column (1)), as well as those of the households that ended up in the treatment groups that are used for our main tests (i.e., Tests I – III); see columns (2)–(7). As shown in column (1), the average household in our sample consists of 2.25 persons, with a by and large equal representation of single-person, two-person and more than two person households. At least one household member is above 65 in about 21 percent of the households, and just above nine percent of the households have one or more children below the age of three. The average home surface is about 100 square meters, and almost 70 percent of the households in the sample lives in a multi-family dwelling. The average (tax) value of a home is 141,000 euros, and about 56 percent of the dwellings are owner-occupied. The walking distance to the nearest organic waste container is between 15 and 100 meters, with an overage of about 70 meters (or about a one-minute walk).

Regarding sorting behavior in the baseline period (see 3.1), the average household made use of the collective residual waste containers on about 1.8 separate days per week, but it used the collective organic waste containers less than 0.12 days per week (i.e., about once every eight weeks); see the bottom rows of column (1) of Table 3.2. The average frequency of use of the organic waste containers was thus quite low, but this average hides substantial variation in organic waste sorting behavior. Slightly less than 17 percent of the households made use of the organic waste sorting facilities at least once during the baseline; 83 percent did not use them at all. The share of households that made regular use of the organic waste sorting facilities was even smaller. Defining "regular sorters" as those households that made use of the organic waste containers on at least two different days every 3 weeks in the baseline period, we find that they make up about 7 percent of the sample. These data suggest that simply offering households the possibility to sort their organic waste does not automatically result in active engagement in organic waste sorting.

Next, columns (2)-(7) of Table 3.2 present the means and standard deviations of the households in our key treatment groups – those that are used for Tests I, II and III. Columns (2) and (3) present the means and standard deviations of those households that have been been assigned to the pure control group and to the pure social feedback group, respectively. Together they form the sample of 2146 households (1103 and 1043 respectively) that will be used for Test I. Columns (4) and (5) present the means and standard deviations of the 926 households that are to be used in Test II, on the impact of social modelling. All households in this sample belong to the Mixed Control/ Modelling household-container clusters (see Figure 3.2, with 458 households receiving the social modelling treatment and 468 households ending up in the control group. The characteristics of the two subsamples employed in Test III (460 of the households in those household-container clusters that were assigned to the combined treatment, and 455 to just the social feedback treatment) are presented in columns (6) and (7).

Comparing columns (2)–(7) of Table 3.2, the means and standard deviations are very similar in the six subsamples. This observation is confirmed by columns (8)-(10) that provide the absolute differences between the characteristics of the relevant pairs of treatment groups, as well as the outcomes of a series of *t*-tests on these differences. We fail to reject the null hypothesis of no difference for most *t*-tests. Of the 45 tests implemented, we reject the null of no difference for just five at the 5 percent level, or lower. In absolute values the differences are quite small, but they are statistically significant because of the fairly large number of observations. An alternative test for balance is on the basis of normalized differences, which present scale-free comparisons (Imbens & Rubin, 2015; Imbens & Wooldridge, 2009; Abadie & Imbens, 2011). The normalized differences are presented in columns (11)-

(13) of Table 3.2. Only one normalized difference is larger than the critical value of 0.25 that is typically used in this literature, and that is the difference in mean distances to the nearest collective organic waste container between the Pure Control and the Pure Feedback household-contained clusters. While the absolute difference in distance of 9.6m is not significant according to the t-test provided in column (8) of Table 3.2, the normalized difference is close to 0.5 standard deviations. Overall we conclude that the randomization has been successful in creating six groups of households that very similar in all observable characteristics, and hence they are very likely to be similar in all unobservable characteristics as well.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Type of household-container cluster		Pure Control	Pure Feedback	Mixed Control/ Modelling		Mixed Feedback/ Combined		Absolute	Absolute	Absolute	Normalized	Normalized	Normalized
Treatment assigned	*Total* Mean/SE	*Control* Mean/SE	*Feedback* Mean/SE	*Control* Mean/SE	*Modelling* Mean/SE	*Feedback* Mean/SE	*Combined* Mean/SE	difference (2)-(3)	difference (4)-(5)	difference (6)-(7)	difference (2)-(3)	difference (4)-(5)	difference (6)-(7)
Average number of household members	2.255	2.267	2.300	2.186	2.151	2.253	2.296	-0.034	0.035	-0.043	-0.025	0.028	-0.028
	[0.038]	[0.072]	[0.080]	[0.099]	[0.096]	[0.063]	[0.032]						
Share of one person households	0.355	0.354	0.346	0.382	0.393	0.374	0.298	0.007	-0.011	0.076**	0.016	-0.022	0.160
	[0.014]	[0.022]	[0.030]	[0.038]	[0.028]	[0.030]	[0.018]						
Share of two person households	0.311	0.306	0.306	0.310	0.301	0.288	0.367	-0.001	0.009	-0.079**	-0.001	0.018	-0.169
	[0.009]	[0.010]	[0.016]	[0.038]	[0.020]	[0.025]	[0.023]						
Share of households with three or more members	0.334	0.340	0.347	0.308	0.306	0.338	0.335	-0.007	0.002	0.004	-0.014	0.004	0.008
	[0.012]	[0.022]	[0.026]	[0.030]	[0.029]	[0.019]	[0.017]						
Share of households with at least one member over 65 older	0.207	0.182	0.184	0.239	0.240	0.240	0.220	-0.002	-0.001	0.020	-0.005	-0.002	0.048
	[0.012]	[0.018]	[0.014]	[0.031]	[0.039]	[0.020]	[0.025]						
Share of households with at least one child younger than 3 yrs	0.093	0.099	0.110	0.066	0.087	0.084	0.083	-0.011	-0.021	0.001	-0.031	-0.072	0.003
	[0.006]	[0.011]	[0.009]	[0.008]	[0.022]	[0.014]	[0.008]						
Average living surface of the dwelling (in m2)	99.197	99.024	102.361	97.813	94.196	94.821	102.642	-3.337	3.617	-7.821***	-0.084	0.077	-0.217
	[2.684]	[4.957]	[6.164]	[7.835]	[5.638]	[2.600]	[3.205]						
Average tax value of the dwelling (in euro's)	1.41e+05	1.35e+05	1.45e+05	1.49e+05	1.44e+05	1.33e+05	1.43e+05	-1.03e+04	5554.761	-9672.016***	-0.178	0.073	-0.187
	[4720.107]	[5271.923]	[9684.654]	[17482.566]	[14435.424]	[4949.878]	[6205.745]						
Share of owner-occupied dwellings	0.564	0.583	0.561	0.536	0.507	0.567	0.611	0.022	0.030	-0.044	0.044	0.060	-0.089
	[0.038]	[0.051]	[0.083]	[0.118]	[0.106]	[0.044]	[0.036]						
Share of households living in multi-family dwellings	0.693	0.674	0.724	0.716	0.723	0.695	0.607	-0.050	-0.007	0.088***	-0.109	-0.016	0.185
	[0.028]	[0.061]	[0.053]	[0.066]	[0.065]	[0.051]	[0.052]						
Average distance to nearest organic waste container (in m)	72.450	67.561	77.162	75.159	75.339	69.555	69.469	-9.601	-0.180	0.086	-0.477	-0.009	0.009
	[2.980]	[6.610]	[6.448]	[6.267]	[6.146]	[2.779]	[2.745]						
Average number of unique residual waste disposal days per week	1.781	1.799	1.763	1.736	1.688	1.846	1.860	0.036	0.047	-0.014	0.030	0.042	-0.012
	[0.028]	[0.047]	[0.071]	[0.057]	[0.058]	[0.060]	[0.037]						
Average number of unique organic waste disposal days per week	0.117	0.118	0.100	0.099	0.132	0.139	0.134	0.018	-0.033	0.005	0.042	-0.085	0.012
	[0.012]	[0.025]	[0.022]	[0.032]	[0.036]	[0.026]	[0.009]						
Share of households having used the organic waste facilities	0.166	0.154	0.154	0.160	0.162	0.189	0.204	0.000	-0.001	-0.015	0.001	-0.004	-0.039
at least once during the baseline period	[0.011]	[0.014]	[0.014]	[0.040]	[0.040]	[0.021]	[0.013]						
Share of households that regularly used	0.066	0.066	0.053	0.066	0.076	0.075	0.083	0.014	-0.010	-0.008	0.058	-0.040	-0.029
the organic waste facilities during the baseline period	[0.008]	[0.015]	[0.012]	[0.023]	[0.022]	[0.014]	[0.011]						
Number of households	3,987	1,043	1,103	468	458	455	460						

Table constructed using the iebaltab command of the ietoolkit package (DIME Analytics, 2019).
The t-tests are conducted on the treatment coefficients obtained by regressing each covariate on the treatment indicator. Standard errors are clustered at the household-container level for Test I and at the household level for Tests II and III.
***, **, and * respectively indicate significance at the 1, 5, and 10 percent critical level.

Table 3.2: Summary statistics of the sample characteristics and balance tests for the treatment groups to be used to estimate the treatment effects (Tests I – III).

3.3 Empirical Strategy

As already announced in Section 3.2.3, we estimate our treatment effects using a standard two-way difference in difference model:

$$y_{ict} = \alpha_i + \gamma_t^{RS} + \gamma_t^{INS} + \sum_{p \in \{T,P\}} \beta^p \; I_t^p \; T_i + \epsilon_{ict}. \tag{3.1}$$

In this equation, y_{ict} reflects the organic waste sorting behavior of household i in household-container cluster c in week t. Household fixed effects are captured by α_i, and to improve precision we allow the week fixed effects to differ between those households that were already engaged in regular organic waste sorting in the baseline period, i.e. those households that used the facilities at least on two separate days every three weeks (γ_t^{RS}; with RS referring to "regularly-sorting households"), and those households that did not regularly use the waste sorting facilities in the baseline period (γ_t^{INS}; with INS referring to "irregularly-sorting and non-sorter households"). We use two different types of week fixed effects, because seasonal influences are less likely to affect sorting activities for the latter type. Variable T_i reflects the treatment household i in household-container cluster c was assigned to; we estimate equation (3.1) separately for each of the three treatments. We estimate the average treatment effects for two subperiods – the period in which the treatments are implemented ($p = T$, between February 4, 2019, and April 29, 2019), and the post-intervention period ($p = P$, between April 30 and December 29, 2019). Indicator function I_t^p is equal to one in all weeks t either during or after the treatment implementation period (that is, if $p = T$ or $p = P$, respectively), and zero otherwise. Hence β^T is the average treatment effect during the intervention period, and β^P captures the average treatment effect in the post-intervention period. The last term in equation (3.1) is ϵ_{ict}, the error term. We follow Abadie et al. (2017) and cluster the standard errors at the level of the unit at which treatments were assigned – at the household-container cluster level for Test I, and at the level of individual households for Tests II and III (see also Section 3.2.3).

In our core anlyses we use OLS to estimate equation (3.1) even though y_{ict} is either a count variable (with integer numbers between 0 and 7, for the number of separate days on which a household used the organic waste facilities in week t) or a binary variable (capturing whether or not the household used the facilities in week t, yes or no). We do so for both technical reasons as well as for ease of interpretation, but we also probe the robustness of the OLS results using negative binomial and probit models too.

Equation (3.1) is our key model, but we estimate two other specifications as well. The first alternative specification is designed to provide insight into possible heterogeneous treatment effects. If effective, does a treatment improve average waste sorting because it induces intensification of waste sorting among households that were already actively sorting their waste in the baseline period, or because it induces previously non-sorting households to start sorting? We construct an indicator variable I^S_{ic} which takes on value one if the household i in household-container cluster c has used the organic waste sorting facilities at least once in the baseline period, and zero otherwise. We specify the following regression model:

$$y_{ict} = \alpha_i + \gamma_t^{RS} + \gamma_t^{INS} + \sum_{p \in \{T,P\}} (\beta^p \ + \ \beta^{pS} \ I^S_{ic}) \ I^p_t \ T_i + \epsilon_{ict}. \tag{3.2}$$

Here, β^p captures the treatment's impact on those households that never used the waste sorting facilities in the baseline period, and $\beta^p + \beta^{pS}$ then captures the treatment's impact on those household that were already engaged in waste sorting – at least to some extent – in the baseline period.

The second alternative specification is designed to provide better insight into the dynamics of the treatment effects, by estimating them separately for each each month – in the intervention period itself, but also in the post-intervention period. Equation (3.3) is identical to equation (3.1), except that it estimates the average treatment effects of each of the two treatments for each of the treatment months for which we have data:

$$y_{ict} = \alpha_i + \gamma_t^{RS} + \gamma_t^{INS} + \sum_{m=-4}^{11} \beta^m \ I^m_t \ T_i + \epsilon_{ict}. \tag{3.3}$$

Here, I^m_t is an indicator function that equals one if week t falls in the m^{th} month since the start of the intervention (with $m \in \{-4, ..., 11\}$), with negative numbers identifying pre-treatment months), and zero otherwise. All other variables are defined as specified in equation (3.1). We explicitly estimate pre-treatment differences (as well as post-treatment effects) to obtain insight into the extent to which the treatment and control groups have similar pre-treatment organic waste sorting patterns (and hence whether there is support for the implicit parallel trend assumption). In both equations (3.2) and (3.3) standard errors are clustered either at the household level or at the household-container level, depending on the treatment examined.

3.4 Results

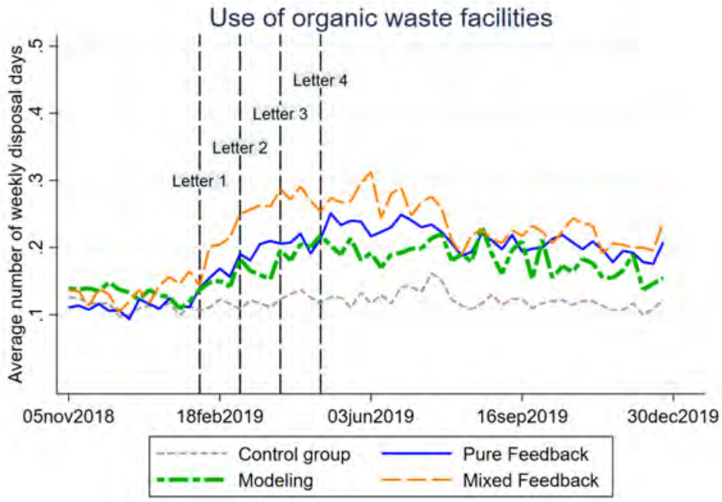
Our two key behavioral outcome variables are the number of unique days per week households made use of the collective organic waste collection facilities, and the number of unique households accessing the facilities per week. Figure 3.3 presents the time patterns of each of these two variables, in the top and bottom panel respectively, for the three treatment groups and the control group. The graphs need to be interpreted with caution, as they are based on household-level treatment assignment, and hence possible treatment spillovers are not accounted for. Yet the time patterns provide suggestive evidence that all three treatments were effective in increasing the intensity of organic waste sorting. Before the start of the interventions organic waste sorting behavior was fairly similar in all four groups, and as of the start of the interventions there is a marked increase in the overall frequency of usage in the three treatment groups (top panel), as well as in the number of unique households using the facilities per week (bottom panel). Finally, the two graphs also suggest that the impacts of the three treatments seem to be fairly permanent, as the difference between each of the three intervention groups and the control group does not appreciably decline from the fourth post-intervention month onward.

In the remainder of this section, we present the results of our statistical analyses. Section 3.4.1 presents our estimates of the average treatment effects, Section 3.4.2 tests for the presence of treatment spillovers, Section 3.4.3 explores whether some types of households respond more strongly to the treatments than others, Section 3.4.4 tests whether the size of the treatment effects change over time and Section 2.4.3 tests whether or not the estimated changes in the frequency of use of the organic waste collection facilities are likely to be an over- or an underestimate of the treatment impacts on the amounts of organic weight collected.

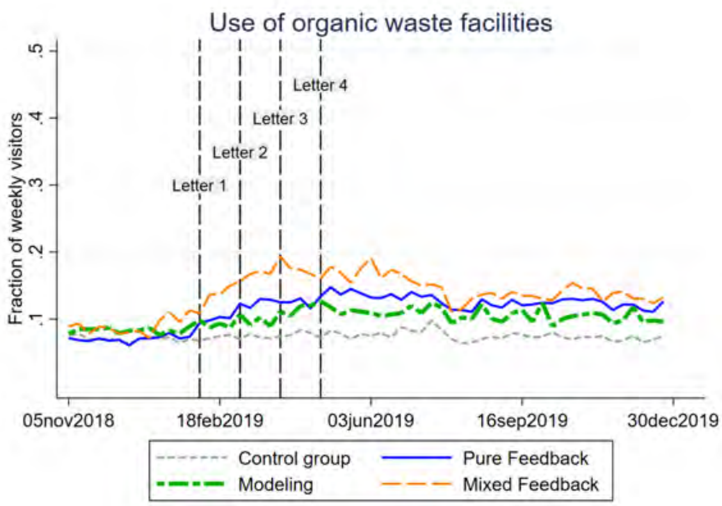
3.4.1 Main treatment estimates

We start by presenting the results of our three main tests, Tests I-III, estimated using equation (3.1). The results are presented in Table 3.3. The odd-numbered columns present the treatment effects, during the intervention period and in the post-intervention period, on the average number of unique days households use the organic waste facilities per week; the even-numbered columns present the effects for the share of unique households making use of the facilities per week.

The impacts of the social feedback treatment (Test I) are presented in



(a) The average number of households' unique organic waste disposal days per week.



(b) Share of unique households having used the organic waste collection facilities per week.

Figure 3.3: Temporal pattern of the average organic waste sorting performance, measured by the number of unique usage days per week (panel (a)) or by the share of unique user households per week (panel (b)), in each of the four treatment groups.

columns (1) and (2) of Table 3.3. We find that this treatment had a positive impact on both measures of organic waste sorting in both sub-periods; all estimated coefficients are significant at the 1 percent level, or better. Moreover, the effects are sizeable; the social feedback treatment increased the frequency of use of the organic waste collection facilities by 0.058 days per week during the intervention period (or a 46% increase compared to the control group's mean frequency of 0.126 visits per week; see column (1)) and there was a 3.7 percentage point increase in the average number of unique households that made use of the organic waste sorting facilities (or a 48% increase compared to the control group's mean percentage of 7.6% unique sorting households per week; see column (2)). The effect sizes for the two dependent variables are thus fairly similar in percentage terms, and hence the social feedback seems to have been equally effective in increasing the intensive as well as the extensive margins of organic waste sorting. Moreover, columns (1) and (2) also show that the effect does not seem to decrease over time, even when the feedback information is discontinued; the average impacts are, if anything, larger in the nine-month post-intervention period than during the three-month intervention period (although this difference is not significant; see the results of the appropriate F-test at the bottom of Table 3.3).

The results of the social modelling treatment (Test II) are presented in columns (3) and (4) of Table 3.3. We find that this intervention also affected both measures of organic waste sorting activity – during the intervention period, and also in the post-intervention period. During the intervention period, the treatment resulted in a 35% increase in the frequency of use and a 30% increase in the number of distinct users. And while we do not find that the treatment effect increased over time, we also do not find much evidence of effect size attenuation – the average effects are very similar in the last nine months for which we have data to the average effects during our three-month intervention period. Overall, the qualitative impacts of the social modelling treatment are thus similar to those of the social feedback treatment: the intervention seems to be equally effective in increasing both the intensive and extensive margin of sorting behavior as evidenced by the similarity in effects on both the number of disposal days and the number of weekly container visitors, and we observe no significant reduction in the size of the treatment effect between the intervention and post-intervention periods.[11] However, the sizes of the impact estimates of the social modelling

[11] The fact that we find significant effects for the social modelling treatment, implemented "old school style" via leaflets delivered via mailboxes, is in a stark contrast to the

treatment seems to be smaller than those of the social feedback treatment.

As shown in columns (5) and (6) of Table 3.3, we fail to find evidence for the social modelling and social feedback interventions to be mutually reinforcing (Test III). The estimated impacts on the frequency of use of the organic waste collection facilities are very close to zero, both during the intervention and after the intervention has been completed. The estimated impact on the share of unique households having used the facilities is somewhat larger, but the coefficients are not measured with sufficient precision for them to be significantly different from zero.

The results in Table 3.3 are obtained using OLS regressions. The estimation procedure therefore does not take into account that our dependent variable is either a count variable with values between 0 and 7 (in case of the impact estimates on the frequency of use, as measured by the number of unique days a household made use of the collective organic waste sorting facilities), or a binary variable (in case of the impact estimates on whether or not a household made use of the facilities at least once in a week). We test the robustness of our results by re-estimating equation (3.1) using negative binomial [12] and probit models; for more details, see Appendix 3.A.2. We find that the results presented in Table 3.3 are robust, both qualitatively and also quantitatively, to using these alternative regression models; see Table 3.A2. As a second robustness check, we test whether our results are robust to correcting for multiple hypothesis testing. The test results are presented in Appendix 3.A.2. We find that all but one of the results of the social feedback intervention (as measured by the frequency of use of the organic waste sorting facilities or the share of households using them, both during the intervention period and thereafter) are robust to correcting for multiple hypothesis testing (MHT). The p-values are $p = 0.032$, or better, for the ones that remain significant after the MHT correction, and $p = 0.155$ for the one that does not survive the MHT correction. The social modelling treatment's impacts on the frequency of sorting remain borderline significant

lack of an effect of an intervention, aimed at reducing food waste, implemented via social media; see Young et al. (2017). Whether this difference in outcomes is due to (i) the delivery method (Facebook versus leaflets, (ii) the method of harnessing social influence (with self-selected role models posting hints how to reduce food waste versus our hand-picked models) or (iii) the targeted behavior (reducing food waste versus stimulating organic waste sorting) is an open question.

[12] We use a negative binomial model rather than a Poisson model because the latter assumes that the dependent variable's variance is equal to its mean. The number of unique visits per week is, however, likely to be overdispersed because of the fairly large number of non-sorter households. The negative binomial model is the preferred approach because it does not require equal mean and variance (Cameron & Trivedi, 2013).

within and post intervention (with $p = 0.091$ and $p = 0.107$, respectively), but the treatment's impact on the share of households using the facilities (about $p = 0.174$ for each of the two subperiods) now fail to be significant.

	Social feedback Pure Feedback vs. Pure Control clusters		Social modelling within Mixed Control/ Modelling clusters		Social modelling on top of feedback within Mixed Feedback/ Combined clusters	
	(1) Disposal days	(2) Decision to sort	(3) Disposal days	(4) Decision to sort	(5) Disposal days	(6) Decision to sort
Social feedback × Intervention period	0.0579***	0.0366***				
	(0.0104)	(0.00595)				
Social feedback × Post intervention	0.0893***	0.0511***				
	(0.0255)	(0.0138)				
Social modelling × Intervention period			0.0365**	0.0193***		
			(0.0157)	(0.00694)		
Social modelling × Post intervention			0.0353*	0.0191**		
			(0.0187)	(0.00868)		
Combined treatment × Intervention period					0.00674	0.0103
					(0.0244)	(0.0128)
Combined treatment × Post intervention					-0.000747	0.00262
					(0.0278)	(0.0136)
R^2	0.0203	0.0218	0.0140	0.0136	0.0247	0.0303
Reference group mean (Intervention period)	0.126	0.0763	0.104	0.0707	0.243	0.149
Reference group mean (Post intervention)	0.125	0.0752	0.116	0.0752	0.244	0.144
Number of observations (N × T)	128760	128760	55560	55560	54900	54900
Number of households (N)	2146	2146	926	926	915	915
Week fixed effects	Y	Y	Y	Y	Y	Y
Household fixed effects	Y	Y	Y	Y	Y	Y
F-statistic $\left[\beta^T = \beta^P\right]$	2.577	1.820	0.00715	0.000539	0.143	0.677
p-value $\left[\beta^T = \beta^P\right]$	0.124	0.192	0.933	0.981	0.705	0.411

Standard errors, clustered at the household-container level in (1) and (2) and at the household level in (3)-(6), in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.3: Treatment impact estimates for the three main interventions (Tests I-III).

3.4.2 Treatment spillovers

The treatment effect estimates of Test II and III presented in Section 3.4.1 are predicated on the assumption that there are no within-cluster spillovers from households receiving a specific treatment to those that did not directly receive it. So the question is whether there are significant differences, after the start of the interventions, in the (trends in) behavior of control group households that are very likely to have neighbors who received the social modelling intervention (i.e., the control households in the Mixed Control/ Modelling household-container clusters), and of control group households whose neighbors also did not receive any treatment (i.e., those in the Pure Control household-container clusters). And similarly, are there significant differences between social feedback households that are likely to have neighbors who received the combined intervention (i.e., the social feedback households from the Mixed Feedback/ Combined household-container clusters), and social feedback households whose neighbors also just received the social feedback treatment (i.e., those from the Pure Feedback household-container clusters)?

As explained in Section 3.2.3, our design allows us to explicitly test for spillovers; see Tests IV and V in Figure 3.2. Before presenting the test results, we first verify the absence of major imbalances at baseline that may affect treatment outcomes and therefore bias our spillover estimates. The results of these balance tests are presented in Table 3.A1 in 3.A.1, and indicate that the randomisation was also successful in creating balanced treatment arms for the spillover tests. We only find one significant difference between the two groups within the sample of households used for Test IV, and also just one between those used for Test V; with values of 0.101 and 0.139, both associated normalized differences are well below the typical imbalance threshold of 0.25.

The results of the spillover tests are presented in Table 3.4. Columns (1) and (2) of this table report the results of Test IV on the presence of spillovers from the social modelling treatment to the control group. The results are clear-cut; there is no evidence for any spillover effects of the social modelling treatment on non-treated households in the vicinity. The lack of positive spillovers suggests that whatever induced treatment households to better sort their waste did not have any direct or indirect effects on the behavior of non-treated households (think of strengthening the social norm of organic waste sorting – e.g., via direct communication, or via learning by watching others – or by affecting personal norms).

Results are less straightforward, however, for the spillover test of the combined treatment – Test V. As shown in columns (3) and (4) of Table 3.4, social feedback households living in areas in which (a random) half of the households received the combined treatment (i.e., the social feedback households from the Mixed Feedback/ Combined household-container clusters) use the organic waste sorting facilities more frequently during the intervention period than the social feedback households living in areas in which *all* households received the social feedback intervention (i.e., those from the Pure Feedback household-container clusters). During the intervention period the frequency of usage by the former is about 20% higher than that of the latter (0.038 days per week more often, compared to a baseline rate of 0.188), and also the share of households making use of the facilities is 15% higher (a 0.018 percentage points increase compared to a baseline user share of 0.118).

The results for Test V are puzzling. On the one hand we find that there is no significant difference between households that received the combined treatment and those that just received the social feedback treatment; see columns (5) and (6) of Table 3.3. This suggests that our social modelling information does not enhance the effectiveness of the social feedback information. On the other hand, columns (3) and (4) of Table 3.4 show that social feedback households located in household-container clusters of which half of their neighbors received the combined intervention engaged in more sorting than those households of which all neighbors received just the social feedback treatment.

	Spillover effect social modeling Mixed Control/ Modelling vs. Pure Control clusters *Only control households*		Spillover effect modelling on top of feedback Mixed Feedback/ Combined vs. Pure Feedback clusters *Only pure feedback households*	
	(1) Disposal days	(2) Decision to sort	(3) Disposal days	(4) Decision to sort
Mixed Control/ Modelling cluster × Intervention period	-0.00371 (0.00633)	-0.00386 (0.00334)		
Mixed Control/ Modelling cluster × Post intervention	0.0101 (0.0152)	0.00182 (0.00898)		
Mixed Feedback/ Combined cluster × Intervention period			0.0377** (0.0168)	0.0181* (0.00906)
Mixed Feedback/ Combined cluster × Post intervention			0.00875 (0.0320)	-0.000305 (0.0176)
R^2	0.0157	0.0157	0.0203	0.0225
Reference group mean (Intervention period)	0.126	0.0763	0.188	0.118
Reference group mean (Post intervention)	0.125	0.0752	0.212	0.127
Number of observations (N × T)	90660	90660	93480	93480
Number of households (N)	1511	1511	1558	1558
Week fixed effects	Y	Y	Y	Y
Household fixed effects	Y	Y	Y	Y

Standard errors, clustered at the household-container level, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.4: Estimates of the presence of spillovers effects between treated and non-treated households for the social modelling treatment (Test IV) and the combined treatment (Test V), respectively.

We see two possible explanations. The first is that the combined intervention does create spillover effects to social feedback neighbors, but that our previous analysis is not able to detect these effects because of its focus on average effects may hide temporal differences. If the impact of the combined treatment is stronger than that of just the social feedback treatment, we would expect to observe a stronger short-term impact of the combined treatment, with that of the social feedback treatment catching up later. The second possible mechanism is that the results are driven by small differences in baseline sorting behavior. From Table 3.4 we know that the households in the Pure Feedback household-container cluster were very similar to those in the Mixed Feedback/ Combined household-container cluster. But the estimate of the spillover effect may be affected by small differences in group organic waste sorting performance in the month prior to sending out the first letter.

To provide insight into the validity of the first explanation, Figure 3.4 presents the temporal pattern of organic waste sorting of all households in the Mixed Feedback/ Combined household-container clusters, as well as that of all households in the Pure Feedback household-container clusters. This figure clearly suggest that if there are spillovers from the combined to the social feedback treatment group, these spillovers must have been instantaneous and perfect; there is no evidence for the households in the combined treatment group taking the lead and those in the social feedback group following suit. We therefore conclude that it is unlikely that the difference is caused by spillovers.

Figure 3.4 does, however, also suggest that small differences at baseline may indeed have resulted in biased estimates of the treatment effects. Compared to the households in the Pure Feedback treatment group, average waste sorting was slightly higher among the social feedback households in the Mixed Feedback/ Combined clusters at baseline, and the first letter of the social feedback treatment seems to have had a stronger effect. We also observe, however, that this difference in waste sorting intensity does not last, as the level of waste sorting in the Mixed Feedback/ Combined household-container clusters decays in the first months post-treatment to a level more similar to that observed in the Pure Feedback groups. Figure 3.4 therefore seems to support the hypothesis that small baseline differences may have been key in driving the difference in social feedback performance between these two groups.

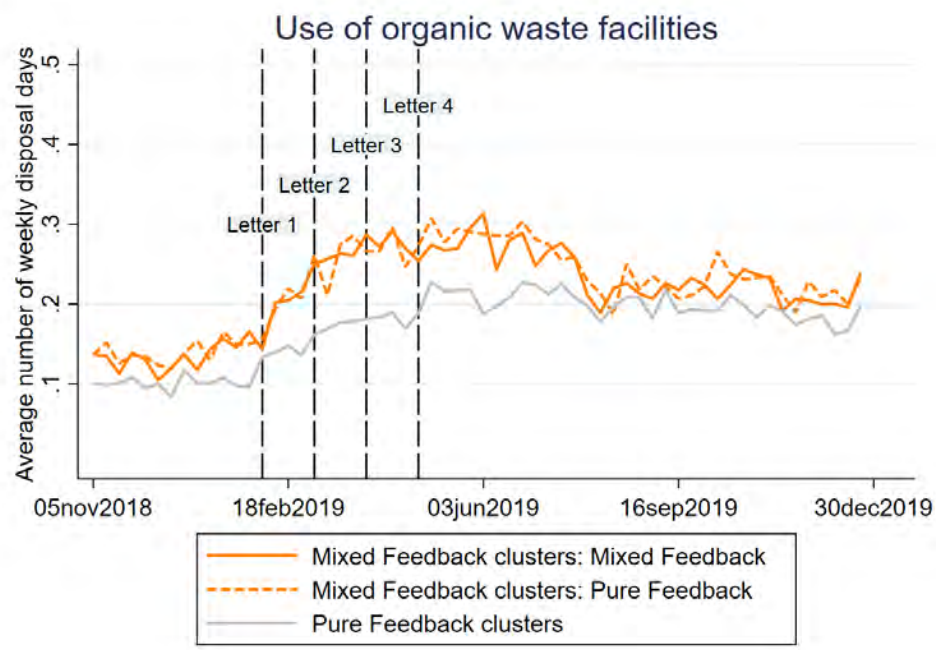


Figure 3.4: The average number of unique organic waste disposal days per week of each of the two types of social feedback households (in the Pure Feedback and in the Mixed Feedback/ Combined household-container clusters) and of the combined households (in the Mixed Feedback/ Combined household-container clusters).

To statistically test the relevance of small differences at baseline, we use an alternative regression strategy than the difference-in-difference approach presented above – one that explicitly controls for any baseline differences in household-container cluster waste sorting performance. More specifically, we estimate the following ANCOVA model:

$$y_{ict} = \alpha\ \bar{y}_{ic0} + \phi_{ic} + \gamma_t^{RS} + \gamma_t^{INS} + \sum_{p \in \{T,P\}} \eta^p\ I_t^p\ S_{ic} + \epsilon_{ict}. \tag{3.4}$$

The essence of ANCOVA is that the outcome variable of interest, y_{ict}, is regressed on its average level before the introduction of the treatment, $\bar{y}_{ic0}$, and possibly also on a set of baseline covariates. In our specification, we control for household-container clusters' waste sorting behavior in the last month before the start of the intervention. We do so by including a vector of dummy variables, denoted by ϕ_{ic}, that capture (intervals of) the average amount of organic waste deposited by the household-container cluster in the month prior to the receipt of the first letter. The treatment period indicator (I_t^p) and the time fixed effects, separate for household that regularly sorted at baseline (γ_t^S) and for those that did not (γ_t^{INS}), are defined as before; see equation (3.1). Defining dummy variable $S_{ic} = 1$ if household i in cluster c is a social feedback household in a Mixed Feedback/ Combined household-container cluster and 0 otherwise, coefficient η^p captures the extent of a possible spillover from the combined treatment group households on their neighbors who only received the social feedback intervention in treatment period p. We conclude that spillovers are absent if we cannot reject the null of $\eta^p = 0$ for $p = \{T, P\}$. Estimation is via OLS, and the standard errors are clustered at the household-container cluster level.

The results of this analysis are presented in Table 3.5. We first verify the extent to which our ANCOVA analysis is able to replicate the results of the difference-in-difference approach presented in Table 3.4. Columns (1) and (2) of Table 3.5 present the ANCOVA regression results, for respectively the frequency of use and user shares, without controlling for the baseline waste sorting performance dummies. The results are very similar to those obtained with the two-way difference-in-difference approach, as shown in columns (3) and (4) of Table 3.4. Replacing the difference-in-difference strategy by an ANCOVA strategy does not affect outcomes; the coefficient capturing possible spillovers is positive and highly significant.

So are the difference in waste sorting performance between the Pure Feedback households and the feedback households in the Mixed Feedback/ Combined household-container clusters the result of small differences in waste sorting performance at baseline? When we control for baseline neighborhood waste sorting performance (see columns (3) and (4) of Table 3.5), the spillover coefficients become insignificant, with p-values of 0.358 or higher.

We thus conclude that there is evidence of treatment spillovers between treated and non-treated households for neither the social modelling treatment nor the combined treatment, and hence that the treatment estimates presented in Table 3.3 are unbiased.[13]

[13]The letters received as part of the social learning treatments may thus have induced households to update their beliefs about (i) the feasibility of organic waste sorting (if others can do it, they themselves are likely to be able to do it too) and (ii) the social relevance of organic waste sorting (if others do it, this is a signal that at least some neighbors find recycling a worthy cause to contribute to). Similar considerations may have been evoked by observing increased sorting behavior by neighboring households – when noticing that more people make use of the facilities, or when observing more people carrying waste bags in the streets. "Direct observation" is unlikely to be an important channel for social learning, however – if only because we do not find any evidence for spillovers between treated and non-treated households. We therefore conclude that treatment effects occur via our letter interventions – not via households observing a (substantial) increase in the number of sorting households, and subsequently follow their example.

	Spillover effect combined treatment to just feedback Mixed Feedback/ Combined vs. Pure Feedback clusters *Just feedback households only*			
	(1)	(2)	(3)	(4)
	Disposal days	Decision to sort	Disposal days	Decision to sort
Mixed Feedback/ Combined cluster × Intervention period	0.0376**	0.0184*	0.0199	0.0119
	(0.0167)	(0.00913)	(0.0246)	(0.0126)
Mixed Feedback/ Combined cluster × Post intervention	0.00860	-0.0000626	-0.00908	-0.00657
	(0.0319)	(0.0175)	(0.0381)	(0.0210)
R^2	0.394	0.336	0.395	0.337
Reference group mean (Intervention period)	0.188	0.118	0.188	0.118
Reference group mean (Post intervention)	0.212	0.127	0.212	0.127
Number of observations (N × T)	73226	73226	73226	73226
Number of households (N)	1558	1558	1558	1558
Week fixed effects	Y	Y	Y	Y
Neighborhood baseline performance strata	N	N	Y	Y

Standard errors, clustered at the household-container level, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.5: Results of the spillover test between combined treatment and social feedback households (Test V) using ANCOVA.

3.4.3 Treatment heterogeneity

We thus established that both social feedback and social modelling were effective in increasing organic waste sorting behaviors; see Table 3.3. We also documented that that there is no added benefit from combining the two interventions (compared to just implementing social feedback). We now turn to answering two – related – questions. First, were the treatments effective because they induced households that did not previously engage in organic waste sorting to start separating their organic waste? Or were the treatments effective because they were especially successful in inducing households that already participated in organic waste sorting, to step up their game? And second, does the star score received affect the size of the treatment response? After all, policy makers who want to use the social feedback mechanism can choose how to set the thresholds for the various star scores. Does more positive feedback (in the form of awarding a higher star score for a given performance) encourage households to do even better, or does it make them more complacent instead?

We aim to answer these two questions in the following two subsections. Before doing so, we pool all data into three groups – the control group (pooling the control households in the Pure Control and in the Mixed Control/ Modelling household-container clusters), the social feedback group (pooling the households in the Pure Feedback and in the Mixed Feedback/ Combined household-container clusters), and the social modelling group (consisting of only the social modelling households in the Mixed Control/ Modelling household-container clusters). We pool these groups of households because this increases the power of our analysis; we are allowed to do so because we found evidence of neither treatment spillovers nor of any added impact of social modelling above and beyond the impact of the social feedback intervention.

Estimating the treatment effects on sorter and non-sorter households

To test whether the treatment effects differ between those households that were already engaged in some organic waste sorting prior to the introduction of the treatment, we estimate equation (3.2). This specification allows the treatment effect to differ depending on households' baseline sorting status – having used the organic waste collection facilities at least once during the baseline period, or not. For brevity, we will refer to these as 'sorter households' and 'non-sorter households', respectively.

	Social feedback Pooled Feedback vs. Pure Control clusters		Social modelling within Mixed Control/ Modelling clusters	
	(1) Disposal days	(2) Decision to sort	(3) Disposal days	(4) Decision to sort
Pooled feedback × Intervention period	0.0498***	0.0333***		
	(0.00849)	(0.00532)		
Pooled feedback × Post intervention	0.0646***	0.0431***		
	(0.0190)	(0.0102)		
Pooled feedback × Intervention period × Sorter household in baseline period	0.152***	0.0782***		
	(0.0311)	(0.0191)		
Pooled feedback × Post intervention × Sorter household in baseline period	0.170***	0.0501**		
	(0.0411)	(0.0222)		
Social modelling × Intervention period			0.0175*	0.0128**
			(0.00998)	(0.00584)
Social modelling × Post intervention			0.0240*	0.0175**
			(0.0129)	(0.00717)
Social modelling × Intervention period × Sorter household in baseline period			0.121	0.0413
			(0.0769)	(0.0300)
Social modelling × Post intervention × Sorter household in baseline period			0.0726	0.0106
			(0.0823)	(0.0388)
R^2	0.0221	0.0240	0.0148	0.0139
Reference group mean (Intervention period)	0.126	0.0763	0.104	0.0707
Reference group mean (Post intervention)	0.125	0.0752	0.116	0.0752
Number of observations (N × T)	183660	183660	55560	55560
Number of households (N)	3061	3061	926	926
Week fixed effects	Y	Y	Y	Y
Household fixed effects	Y	Y	Y	Y

Standard errors, clustered at the household-container level in (1) and (2) and at the household level in (3) and (4), in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.6: Testing for treatment heterogeneity between sorter and non-sorter households .

Table 3.6 presents the results of our regression analyses using equation (3.2). Columns (1) and (2) of this table show that the social feedback intervention has been effective in increasing organic waste sorting among both sorter and non-sorter households, but especially so for the sorter households. For example, whereas the non-sorter households in the social feedback treatment increase the frequency with which they use the facilities by 0.05 extra days per week during the intervention period, the sorter households increase their frequency of use by (0.0498 + 0.152 =) 0.20 days per week; see column (1) of Table 3.6.

We thus find heterogeneous impacts for the social feedback treatment, but the evidence for the social modelling treatment is less compelling. As shown in columns (3) and (4) of Table 3.6, the estimates indicate a stronger response among sorter households than among non-sorter households, but the coefficients are estimated with insufficient precision for these differences to be statistically significant. We interpret these outcomes as absence of evidence for the existence of heterogeneous treatment effects in the social modelling intervention – not evidence of absence.

The impact of star score feedback on organic waste sorting behavior

The second type of heterogeneity we test for, is the impact of the number of stars received in the social feedback letters. If the amount of sorting is such that a container just fails to receive an extra star, does that stimulate or rather discourage sorting compared to the case in which the star cutoff had been set slightly more leniently? That is, does it matter whether thresholds are set such that they are either just above or just below the current sorting level?

To answer this question we resort to regression discontinuity analysis (Lee & Lemieux, 2010). More specifically, we compare the social feedback treatment's impact on households assigned to household-container clusters with per-household average container weights just below the threshold that merits an extra star, to that on households with per-household average container weights just above that threshold. Regression discontinuity analysis is predicated on the assumption that clusters similar in average household organic waste weight deposited are likely to be similar in other dimensions as well, and thus that there are no fundamental differences between households located in household-container clusters just above or just below a star cutoff level.

Figure 3.5 presents the amount of organic waste deposited, averaged

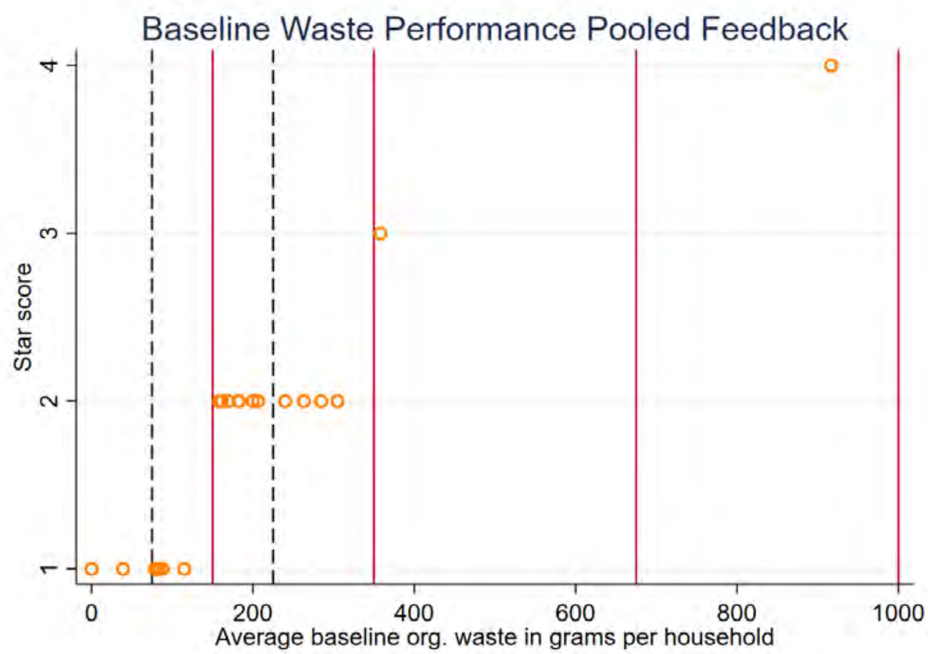


Figure 3.5: The distribution of the amount of organic waste collected in the containers of the 42 associated household-container clusters in the last month before the start of the social feedback intervention.

Solid vertical lines indicate the threshold weights for the number of stars. All containers received at least one star; extra stars are earned when crossing 150 grams (**), 350 grams (***), 675 grams (****) and 1000 grams (*****). The dashed lines indicate the interval of 75-225 grams per household, which is the interval of household-container clusters selected for the regression discontinuity analysis.

over all households in a household-container cluster and measured in grams, in the final pre-treatment month. The uninterrupted vertical lines in this figure indicate the threshold levels for the second, third, fourth and fifth star (with levels 150, 350, 675 and 1000 grams; see also Table 3.1). As is clear from Figure 3.5, the bulk of the containers received either one or two stars. We thus focus our analysis on all households in those household-container clusters of which the container's average per-household weight was just above or just below the second-star cutoff. More specifically, we focus on the impact on the behavior of the households in the 12 household-container clusters of which the container's average per-household weight was less than 75 grams away from the 150 gram threshold (i.e., between 75 and 225 grams; see the dashed vertical lines in Figure 3.5). This interval contains more than 500 households whose container received 1 star in the first social feedback letter, and more than 700 households whose container received 2 stars in that first letter. The regression discontinuity model is specified as follows:

$$y_{ict} \;=\; \alpha_i + \gamma_t^{RS} + \gamma_t^{INS} + \sum_{p \in \{T,P\}} \delta^p \; I_t^p \; I_c(w_c \;>\; 150) + \epsilon_{ict}, \quad (3.5)$$

where $I_c(w_c > 150)$ is an indicator function that takes on value 1 if container c's average amount of organic weight collected over the last pre-intervention month is between 150 and 225 grams per household in the household-container cluster, and zero otherwise. We cluster the standard errors at the household-container cluster level. The other regressors are defined similarly as in equation (3.1). The coefficients of interest are δ^T and δ^P, which capture the difference in the frequency of organic waste sorting caused by the first feedback letter containing two stars rather than one in, respectively, the intervention period and the post-intervention period.

	Interval (75-225 grams)	
	(1)	(2)
	Nr of disposal days	Decision to sort
Intervention Period × Two Stars in First Letter	0.0202	0.0159*
	(0.0155)	(0.00759)
Post-Intervention × Two Stars in First Letter	0.0430	0.0171
	(0.0378)	(0.0203)
R^2	0.0184	0.0221
Reference group mean (Intervention Period)	0.137	0.0885
Reference group mean (Post Intervention)	0.156	0.0993
Observations (N x T)	76260	76260
Households (N)	1271	1271
Week fixed effects	Y	Y
Household fixed effects	Y	Y

Standard errors, clustered at the household-container level, in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.7: The impact of the receipt of an extra star on household organic waste sorting.

Table 3.7 presents the results. The evidence for the surprise of receiving a two-star letter at the start of treatment affecting sorting behavior – either in the short run, or also in the longer run – is weak at best. The impact on the share of households making use of the organic waste sorting facilities in the intervention period is positive and significant. But the size of the effect is small (less than 1.6 percentage points), and only marginally significant ($p = 0.060$). None of the other coefficients are significant at the 10 percent level, or better.[14]

We thus conclude that baseline cluster performance as measured by being just above or below the second star-threshold at baseline is not associated with significantly higher sorting behavior after treatment.

3.4.4 Treatment dynamics

We thus found that both the social feedback and the social modelling treatments were effective in improving waste sorting behavior, both during the intervention period as well as in the post-intervention period. Estimating the average effect over each of these sub-periods may, however, hide temporal patterns in treatment response. To analyze the treatment dynamics, we estimate equation (3.3). Figure 3.6 presents the estimated treatment effects for each month – the four months prior to the start of the intervention, and the twelve months thereafter (including month 0, the month in which the first letters were sent). The coefficients capture the difference in sorting behavior between the treatment and control households, compared to the difference observed in the month preceding treatment (month -1). Panel A of Figure 3.6 presents the treatment effects on the number of distinct days households made use of the organic waste sorting facilities per week, and Panel B the share of households having used them at least once during a week.

As is clear from Figure 3.6, pre-treatment differences are relatively small and statistically insignificant. Regarding the treatment impacts, the effects of the social feedback treatment are sizeable both during the intervention as well as after the interventions have been completed; Panels A and B suggest that from the fourth month since the completion of the intervention (i.e., from month 6 onwards), both the frequency of use and the share of users

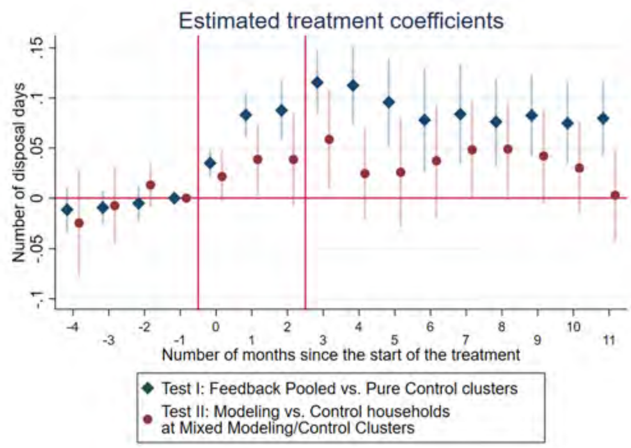
[14] As a robustness check, we enlarge the number of household-container clusters from 12 to 16 by estimating equation (3.5) using all households in household-clusters within 150 grams of the two-star threshold (i.e., those with per-household weights collected between 0 and 300 grams). The results – not shown here, available upon request – are both qualitatively and quantitatively very similar to the ones presented in Table 3.7.

stabilized at a level that is about 0.05 days per week (or about two-thirds) higher than that of the control group. In contrast, the effects of the social modelling intervention are smaller and they also tend to attenuate over time.

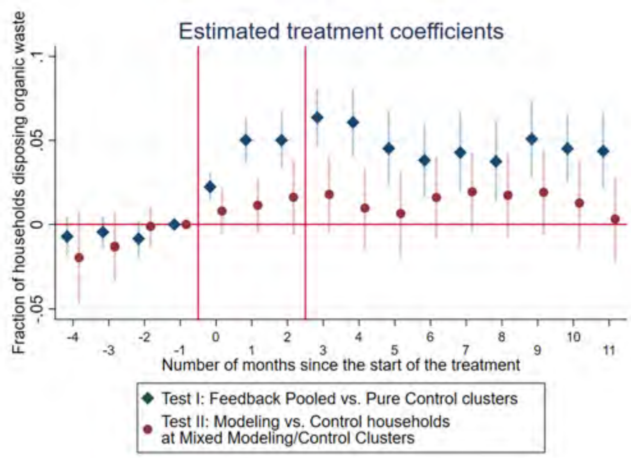
We thus conclude that the social feedback treatment has not only been successful in motivating households to better sort their organic waste in the short run, but also in the long(-er) term – and maybe even permanently.[15] The social modelling treatment's impact are, however, found to be transitory at best.[16]

[15]This permanent change in behavior may be due to habit formation (Becker & Murphy, 1988), or due to waste sorting having experience good characteristics (because of the difficulty of properly assessing the costs and benefits of waste sorting ex ante; (Nelson, 1970; Foster & Rosenzweig, 1995), or both. Unfortunately, our RCT does not allow us to differentiate between these different channels, but based on the extant literature both are plausible candidate causes for the short-term interventions having permanent effects.

[16]We also formally tested whether the social feedback intervention is significantly more effective than the social modelling treatment – in terms of the frequency of sorting as well as in terms of the share of households using the facilities, during the intervention period as well as thereafter. We were able to reject the null of no difference for all four comparisons at p-values of 0.105, or better.



(a) Frequency of use (days per week)



(b) User fraction

Figure 3.6: Treatment estimates per month, for the social feedback and the social modelling interventions

The estimates are the coefficients obtained using estimating regression (3.3) using the number of separate usage days per week (panel (a)) or the fraction of households making use of the facilities in a week (panel (b)) as dependent variable. The indicator for final month of the baseline period is the omitted category. A month is defined here as four weeks, with the exception of months -4 (1 week) and 11 (3 weeks). The period in which the interventions were implemented (the 'intervention period') is demarcated by the vertical lines.

3.4.5 The relationship between usage frequency and weight

We documented that the social modelling and social feedback treatments resulted in a significant increase in the frequency with which households made use of the organic waste sorting facilities. From an environmental perspective this increased frequency is beneficial only if it translates into an increase in the total weight of organic waste sorted. After all, for the environment it is not the frequency of visits per se that matters, but the amount of waste collected (in terms of improved composting, reuse as organic fertilizer, and fuel and CO_2 emission savings when incinerating the cleaned-up residual waste flow). So did our treatments increase the total amount of weight collected?

Obviously, an increase in the frequency of use of the collective organic waste collection facilities results in an increase in the amount of organic waste collected if the weight per deposit does not decrease. A sufficient condition for this is that the percentage reduction in weight per visit is smaller than the concomitant percentage increase in the visiting frequency. Average weight per visit may go down if a treatment induces households to wait less long before they go out and dispose of the organic waste collected. It may also go down if the treatment induced non-sorting households to start sorting, but at lower intensity than households that already sorted prior to the treatment implementation. Did our treatments result in a reduction in the average weight of organic waste deposits?

We empirically verify the impact of an increased waste sorting frequency on the average weight per deposit as follows. We regress the weight of container c in week t, W_{ct}, on (i) the total number of disposals made at container c in week t ($NumbDeposits_{ct}$, as measured by the aggregate number of unique disposal days registered at container c in week t), and (ii) a series of dummy variables capturing the accompanying household-container cluster's type (Pure Control, Pure Feedback, Mixed Control/ Modelling, or Mixed Feedback/ Combined), interacted with dummies capturing either the intervention or the post-intervention period:

$$W_{ct} = \alpha NumbDeposits_{ct} \quad + \sum_{p\in\{T,P\}} \sum_{k\in K} \beta_k^p \; Z_{kc} \; I_t^p + \gamma_c + \mu_t + \epsilon_{ct}. \quad (3.6)$$

In this model, coefficient α measures how the total weight collected in container c in week t varies with the number of deposits made in that week.[17] Next, Z_{kc} is a dummy variable capturing whether or not container c had

[17] That is, we still retain the definition that a household made one deposit if the number

been assigned to a specific container type k, with $k \in K$, and K ={Pure Feedback, Mixed Control/ Modelling, Mixed Feedback/ Combined}). As before, I_t^p is an indicator function that is equal to one in the weeks during or after the treatment implementation period (that is, for $p = T$ and $p = P$, respectively), and zero otherwise. Having interacted Z_{kc} and I_t^p, coefficient β_k^p then captures whether, having controlled for the number of deposits having been made, the total amount of weight collected in containers of type k is higher or lower than that in the Pure Control containers in period p). Finally, γ_c and μ_t capture the container and week fixed effects; standard errors (ϵ_{ct}) are clustered at the container level.

Note that this method is, in essence, a simple production function approach. The container fixed effects capture any time-invariant factors affecting the weight of organic waste collected in the container, and the week fixed effects correct for possible seasonal influences. The weight of organic waste collected in a container is assumed to vary with the total number of deposits the container receives. This total number of deposits varies with the treatment status of the households to which the container is nearest to, as all three treatments have been documented to increase the frequency of organic waste deposits; see Section 3.4.1). The approach assumes that, independent of whatever caused the number of deposits to differ between containers, an extra deposit results in the container's weight to increase by α kilos. The coefficients on the container's type therefore capture whether the observed container weight is larger or smaller than predicted by the number of disposals under this assumption, controlling for container and week fixed effects. If none of the β-coefficients are negative and significantly different from zero, we can conclude that any change in the frequency of use did not result in a lower average weight per deposit.

The results of this analysis are presented in Table 3.8. The coefficient on the number of visits shows that Pure Control households that made use of the organic waste container facilities dropped off, on average, 1.16 kilos of organic waste per visit. Next, and importantly, all coefficients on the container-type dummies have a positive sign – in the intervention period, and also in the post-intervention period. While not all coefficients are significantly different from zero, none of them is negative either. For example, in case of the Mixed Control/ Modelling containers, the total weight collected,

of times it swiped its card to get access to the container, is strictly positive. Having swiped their card twice or more on the same day is thus coded as one deposit for that day. As already explained in footnote 10, we do so because swiping the card more than once very likely to be the result of a system malfunction, in the form of the lid not opening immediately at the first swipe.

averaged over the intervention period is 6.65 kilos higher ($p = 0.052$) than one would expect if the weight per deposit would have been equal to that in the pure control group. That means that weight per deposit in the intervention period is thus higher for these mixed-usage containers than for the containers that are exclusively (or at least predominantly) used by households in the control group. And the same conclusion holds for the other types of household-container clusters, and also for the post-intervention period.

This analysis thus shows that the average weight per deposit is not lower in any of the (mixed) treatment containers than in the pure control containers. That means that the percentage change we documented in the change in frequency with which households made use of the organic waste container is a lower-bound estimate of the percentage change in weight collected.

	Weight collected per container in a week
Number of unique organic waste disposal days in week	1.164***
	(0.177)
Intervention period × Pure Feedback container	8.698*
	(5.068)
Intervention period × Mixed Control/Modelling container	6.647*
	(3.311)
Intervention period × Mixed Feedback/Combined container	4.414
	(4.208)
Post intervention period × Pure Feedback container	8.807**
	(4.321)
Post-intervention period × Mixed Control/Modelling container	5.536
	(4.208)
Post-intervention period × Mixed Feedback/Combined container	1.370
	(4.334)
Constant	5.213
	(4.023)
R^2	0.220
Pure Control container mean (Intervention period)	26.80
Pure Control container mean (Post-intervention period)	29.17
Number of containers (N)	40
Number of observations (N × T)	1104
Container fixed effects	Y
Week fixed effects	Y

Standard errors, clustered at the container level, in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.8: Analysis of the weekly amount of weight collected in a container.

3.5 Identification of the underlying mechanisms

In Section 3.4 we documented that both the social modelling and the social feedback treatments significantly affected waste sorting. We also find that in case of the social feedback treatment the impacts are long-lasting. These results do, however, raise the question via which mechanism these treatments affected waste sorting. Is it because the treatments increased the households' motivation to sort? Did the exposure to the intervention improve the extent to which households felt capable of properly sorting their waste, as they became more knowledgeable on what type of waste needs to be disposed of how? And were the effects long-lasting because (a more intensive) engagement in waste sorting resulted in mitigating concerns regarding how cumbersome (or inconvenient) organic waste sorting is?

Based on Briguglio (2016) we selected a number of factors that are thought to be important determinants of organic waste sorting behavior, ranging from households' attitudes towards waste sorting to their perceived self-efficacy in waste sorting.

These (categories of) variables can be considered potential mediator variables via which the treatments have improved organic waste sorting. We developed a survey instrument to elicit households' scores on different aspects of each mediator; an index was then constructed for each mediator by aggregating the survey answers to all questions within the mediator category under consideration (for details, see below).

We administered the survey instrument twice, with a one-year interval. The first survey was implemented in June and July 2018, roughly 5 months before the start of the baseline period, while the second survey took place in June and July 2019, about one month after the intervention period ended (see the time line in Figure 2.1). The surveys were conducted in person, with the survey team knocking on the doors of pre-selected households. We did not have budget to contact all 4000 households in our study. For the baseline survey we contacted, in total, 1324 households, of which 242 (or 18%) responded, and for the second survey we had budget to visit 1093 households. With relatively few respondents, statistical power is increased substantially if the regression model can be estimated as an ANCOVA model (regressing endline variables on baseline values, treatment status and other controls) rather than as a simple cross-section; see McKenzie (2012). To increase our chances of being able to estimate ANCOVA models, we revisited the baseline respondents for the second survey. We also revisited baseline non-respondents, but only a subsample of those – if they were unwilling or unable to participate in the first survey, they may not be willing or able to participate in the second survey either. We thus spent the rest of the remaining budget for the second survey on trying to recruit households that had not been contacted for the baseline survey. Of the 1093 households that were visited for the second survey, 222 (or just over 20%) ended up filling out the survey, of which 105 also had participated in the baseline survey.

We face two challenges. First, we need to choose whether we wish to analyze the second survey data using ANCOVA (i.e., using the baseline value of the dependent variables as a control), or whether we analyze the data as a cross section (using just the data of the second survey). With an ANCOVA we have 105 responding households to work with; with a cross section, we have 222 households. Second, as is the case with any survey, the sample of households that answered the survey is not necessarily representative of the population of households in the RCT; there may be non-random selection into the survey. We decided to opt for the cross-sectional analysis because the number of households having filled out both surveys is too low to be able to properly correct for non-random survey participation.

The set-up of this section is as follows. Section 3.5.1 presents our identification strategy to uncover possible avenues via which the treatments affected waste sorting activity. Here, selection into the survey plays a key role, and Section 3.5.2 documents whether non-random selection was indeed an issue in our study. Section 3.5.3 then presents the results of the analysis.

3.5.1 Mechanism identification strategy

In the endline survey we asked a series of questions to gauge five mediator categories relevant for waste sorting behavior: households' attitudes towards waste sorting, the extent to which they view waste sorting to be a personal norm, their beliefs about the benefits of waste sorting, their beliefs about the (in-)convenience of waste sorting, and their perceived self-efficacy in waste sorting. Each mediator category consisted of between 3 and 11 questions (or survey items), and households could answer using a five-point Likert scale. We normalized households' answers to each question by subtracting the item's mean score and subsequently dividing this difference by the item's standard deviation. We then created an index for each mediator category by taking the average of all normalized items in that category.

Determining which of the mediator categories had been affected by the treatments – if any – is not straightforward. While allocation to the treatment arms is random, the decision to answer the survey (when invited to do so), is not. If the households that filled out the endline survey were a representative sample of the population of households that participated in our RCT, we could simply regress each of the indices on treatment status using OLS, to uncover the impact of treatment (β). We would then estimate the following model:

$$y_i = \beta T_i + \gamma X_i + \epsilon_i. \tag{3.7}$$

Here, y_i is household i's score on the mediator index under consideration, T_i denotes the treatment household i was assigned to, and X_i is a vector of covariates that includes a constant term as well as those household characteristics that are potentially correlated with outcome variable y_i. Finally, ϵ_i is the error term that is normally distributed with mean 0 and variance σ^2.

Our parameter of interest, β, is an unbiased estimate of the treatment effect on y_i if selection into the survey is (quasi-) random. Let us use binary variable p_i to denote whether or not household i participated in the endline survey ($p_i = 1$), or not ($p_i = 0$). We can model the decision to select into the survey as follows:

$$p_i^* = \mu T_i + \delta X_i + \nu Z_i + \varphi_i. \tag{3.8}$$

In this equation p_i^* is a latent variable such that $p_i = 1$ if and only if $p_i^* > 0$. Next, X_i is the vector of household characteristics that may affect moderator values (see equation (3.7)), and Z_i is a vector of household characteristics that affect the decision to participate in the survey, but not the survey outcome variable y_i. Finally, φ_i is the error term. Household i's survey responses are observed (i.e., $p_i = 1$ in (3.8) and y_i is non-missing in (3.7)) if and only if $p_i^* > 0$. Equation (3.8) can be estimated using probit. As shown by Heckman (1976, 1979), not taking into account sample selection gives rise to an omitted variable problem in the outcome equation (3.7). A necessary condition for the OLS estimate of β in (3.7) to be unbiased, given selection as shown in equation (3.8), is that $\rho \equiv corr(\varphi_i, \epsilon_i) = 0$; see also Vella (1998). If $\rho \neq 0$, an unbiased estimate of β can be obtained by jointly estimating (3.7) and (3.8) using Maximum Likelihood (Puhani, 2000).[18]

The key test for selection bias is thus whether $\rho \neq 0$. The reliability of this test depends crucially on the availability of instrumental variables that are strongly correlated with the decision to participate without being correlated with the survey outcome variable of interest (i.e., those variables in Z_i); see Wolfolds and Siegel (2019)). We will address this issue in the next subsection.

[18]We use the Maximum Likelihood estimator (and not Heckman's two-step estimator) because it has better small sample properties; see Puhani (2000).

3.5.2 Selection into the survey

The vector of covariates of selection model (3.8) must thus include (i) all covariates that are expected to affect households' responses to the survey questions making up the various mediator categories (i.e. X_i in equation (3.7)), as well as (ii) additional covariates Z_i that strongly affect the decision to participate (p_i^* in equation (3.8)), but that are unlikely to be correlated with the answers to those survey questions (y_i in equation (3.7)). We selected these two sets of variables as follows.

Potential candidates to be included in X_i and Z_i can be selected from the set of available baseline characteristics (see Table 3.2). As explained before, information on most of these variables are obtained from Schiedam municipality's register data; those on sorting behavior in the baseline period have been collected as part of the RCT. In addition to these variables we possibly also need to control for whether households participating in the endline survey also participated in the baseline survey. Households that participated in both surveys are likely to have stronger pro-sorting opinions and behaviors than others. We explicitly need to control for this because for the endline survey we deliberately oversampled households that had already participated in the baseline (in our attempt to obtain panel survey data rather than a cross-section).

Regarding the choice of instrumental variables in the Heckman selection model (those to be included in Z_i), we have two candidate variables: whether a household was not just invited to participate in the endline survey but for the baseline survey as well, and whether the household was at home when the interviewers passed by for the endline survey. These two variables are unlikely to have a direct influence on the mediator variables of interest, and hence may serve as potential instruments in the Heckman selection model. The invitation to participate in the baseline survey was extended randomly to households in selected blocks deemed representative of the study area, and their propensity to accept the invitation to participate in the second survey may be higher or lower than for households that have not been contacted before. Also, whether or not an adult member of the household was at home when the interviewers came by is another variable that is likely to affect households' probability to participate in the survey without an obvious correlation with the mediator variables of interest.

The regression results of model (3.8) are presented in Table 3.9. All variables in the top panel of this table (i.e., above the dashed horizontal line) are candidates to be included in X_i; those in the bottom panel are our potential instruments (Z_i). Column (1) presents the regression results from the full model. Three of the variables in the top panel are found to be significantly correlated with the decision to participate, and may also affect households' appreciation of the various aspects of the waste sorting process. Households that were already engaged in waste sorting at least to some extent in the baseline period are more likely to participate in the survey, and – having controlled for baseline sorting behavior – so are households living farther away from the organic waste sorting facilities. The third variable is whether the household had participated in the baseline survey, 12 months earlier; those who participated in the baseline survey were also more likely to participate in the endline survey.

As shown in column (1) at the bottom part of Table 3.9, our two potential instruments significantly affect the probability of participating in the endline survey. The coefficients of both variables are positive and significantly different from zero at the 5% level; the coefficient of having been at home at the time of the endline survey is particularly large. The result of the test whether our two instruments are relevant is presented at the very bottom of Table 3.9. The value of the appropriate F-test is 120 ($p < 0.01$), and hence we conclude that our two instruments are indeed strongly correlated with the decision to participate in the endline survey.

The number of variables included in column (1) of Table 3.9 is large, and a large share of them turns out to not significantly different from zero. Multicollinearity may cause variables not to show up significantly, and hence we reran equation (3.8) in search of a more parsimonious model. We therefore kept the treatment indicator and sorting variables (the latter were important characteristics relating to treatment effect heterogeneity), but removed those variables of which the average marginal effect was smaller than its standard error. As shown in column (2) of Table 3.9, omitting them resulted in one of the previously insignificant variables to become significant: households with larger homes are now also found to be more likely to participate in the survey. However, we also find that the coefficient of the variable does not appreciatively differ between columns (1) and (2). This also holds for the coefficient on treatment status, which remains small and not statistically different from zero ($p = 0.623$). We thus find no evidence of treatment-induced selection into the survey.

	(1)	(2)
Received any of the three treatments	0.0116	0.0101
	(0.0205)	(0.0205)
Used organic waste facilities at least once during baseline	0.0935***	0.0950***
	(0.0307)	(0.0306)
Regular sorter during baseline	0.0363	0.0355
	(0.0398)	(0.0398)
Distance to nearest organic waste container (in m.)	0.00167***	0.00164***
	(0.000482)	(0.000474)
Share of households living in multi-family dwellings	-0.0176	
	(0.0260)	
Household with at least one member over 65	-0.0347	-0.0336
	(0.0228)	(0.0226)
One person household	0.0225	
	(0.0234)	
Household with three or more members	-0.0172	
	(0.0272)	
Surface of the dwelling (ln)	0.0506	0.0445*
	(0.0416)	(0.0266)
Household with children	0.0306	0.00999
	(0.0290)	(0.0214)
Tax value of the home (in ln)	-0.0183	
	(0.0486)	
Owner-occupied dwelling	0.0129	
	(0.0217)	
Participated in baseline survey	0.210***	0.209***
	(0.0304)	(0.0305)
Was invited to participate in the baseline survey	0.0539**	0.0519**
	(0.0232)	(0.0231)
Was at home when contacted for the endline survey	0.472***	0.468***
	(0.0309)	(0.0303)
Number of households (N)	1091	1093
Relevance test: χ^2-statistic	120.315	122.206
Relevance test: p-value	0.000	0.000

Robust standard errors in parentheses

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

Dependent variable is whether the household filled out the survey. Estimation using probit; the coefficients are average marginal effects. Variables included as instruments in the Heckman model, are “Was Contacted for the Baseline Survey”, and “Was at Home when Contacted for the Endline Survey”. Relevance tests capture whether the instruments have sufficient predictive power in predicting survey participation.

Table 3.9: Household characteristics correlated with the propensity to participate in the survey.

3.5.3 Identification of the treatment mechanisms

We now turn to estimating whether the treatments had any impact on (the categories of) variables that are thought to be important for households' decision to engage in organic waste sorting. We distinguish five types of intermediate variables – those related to households' motivation to sort: their attitudes towards organic waste sorting, their personal norms regarding waste sorting, their beliefs regarding the benefits of waste sorting, the extent to which they feel empowered to properly sort waste (self-efficacy), and their perception of how (in)convenient it is to sort organic waste.

To probe different aspects of our mediators, each mediator category consisted of between 3 to 11 survey questions, which were subsequently aggregated into mediator indices as follows. Answers to each of the questions within a mediator category were normalized by subtracting the question's mean and dividing by its standard deviation ('I do not know' answers were coded as missing). To construct an index for each mediator category, we took the mean of the normalized values of the answers to all the questions (or items) that make up a specific mediator category.

As explained in Section 3.5.1, the parameters of equations (3.7) and (3.8) are estimated jointly, using Maximum Likelihood. Table 3.10 presents the parameter estimates of equation (3.7) for each of the five mediator categories, as well as the test of whether or not selection bias is an issue (i.e., whether $\rho \neq 0$, or not). We find ρ to be significantly different from zero ($p < 0.10$) in column (4) and bordeline (in)significant ($p < 0.125$) in column (1)); correcting for selection bias using the Heckman selection model is thus not strictly necessary for the other three mediator categories. For consistency, Table 3.10 presents the Heckman results for all five models; rerunning columns (2), (3) and (5) using OLS yields results that are very similar to those presented in Table 3.10 (available upon request).

As shown in Table 3.10, we find that having received any of the three treatment significantly strengthened the household's attitude towards waste sorting (see column (1); $p = 0.014$) as well as their belief that waste sorting is important for society (column (3); $p = 0.067$). When we inspect treatment effects on the underlying items, we find that the treatments affect attitudes especially because of their impact on attitude towards in-house organic sorting ($p = 0.001$), while the item probing the respondent's belief in proper waste processing exhibits the largest treatment-induced change ($p = 0.014$) in the sorting benefits mediator category.

Furthermore, we also find some evidence that the treatments improved the treated household's self-efficacy, albeit it that the impact just fails to be statistically significant at conventional levels (column (5); $p = 0.101$). The questions in this index gauged households' appreciation of the difficulty of various issues related to organic waste sorting. The self-efficacy result seems to mainly reflect household improved knowledge of the sorting rules (which waste material should be sorted with what other materials; $p = 0.034$) accompanied with better insight in how well one is sorting waste ($p = 0.083$), with borderline significant impacts on the items probing the difficulty of both bringing the bag to the container and using the container ($p = 0.114$; $p = 0.112$).

Finally, the treatments failed to significantly affect people's moral motivations (column (2); $p = 0.215$), and the same holds for their appreciation of (the lack of) inconvenience of waste sorting (column (4); $p = 0.666$). We therefore find evidence for treatment effects on both household motivations to sort and aspects that might inhibit waste sorting. The survey results suggest that the treatment effects on waste sorting may have been caused by strengthened positive attitudes towards waste sorting, by higher perceived benefits, and to a lesser extent by an increased perceived ability to engage in waste sorting.

	Attitude towards waste sorting	Personal norm	Perceived Societal Benefits	Perceived (lack of in-) Convenience	Perceived Self-efficacy
	(1)	(2)	(3)	(4)	(5)
	mean(standardized items)	mean(standardized items)	mean(standardized items)	mean(standardized items)	mean(standardized items)
Received any of the three treatments	0.335**	0.149	0.217*	0.0446	0.159
	(0.137)	(0.120)	(0.119)	(0.103)	(0.0971)
Used organic waste facilities at least once during baseline	0.314**	0.266**	0.334***	0.233*	0.202**
	(0.129)	(0.116)	(0.1000)	(0.126)	(0.0956)
Regular sorter during baseline	0.183	0.0812	-0.0593	0.164	0.148
	(0.131)	(0.132)	(0.113)	(0.146)	(0.105)
Surface of the dwelling (ln)	0.0136	0.0801	0.125	0.305*	0.262**
	(0.188)	(0.190)	(0.167)	(0.165)	(0.121)
Household with children	0.359***	0.276**	0.0871	0.0821	0.104
	(0.125)	(0.120)	(0.113)	(0.105)	(0.0916)
Participated in baseline survey	0.260**	0.308**	-0.0456	0.117	0.0533
	(0.127)	(0.124)	(0.124)	(0.104)	(0.0896)
Household with at least one member over 65	0.249*	0.224*	0.0841	0.103	0.366***
	(0.134)	(0.132)	(0.138)	(0.125)	(0.0871)
Distance to nearest organic waste container (in m.)	0.000832	0.00373	-0.00249	-0.00145	-0.000457
	(0.00295)	(0.00280)	(0.00295)	(0.00233)	(0.00194)
Constant	-0.879	-1.214	-0.608	-0.942	-1.562***
	(0.751)	(0.776)	(0.710)	(0.744)	(0.527)
Number of households (N)	1093	1093	1093	1093	1093
ρ	0.163	0.122	-0.0952	0.174	0.0631
$\chi_2(1)$ test $\rho = 0$	2.363	1.565	0.574	2.783	0.205
p-value $\rho = 0$	0.124	0.211	0.449	0.0953	0.651

All models estimated using Maximum Likelihood.
Robust standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.10: Treatment effects on indices for five potential treatment mediator categories.

3.6 Conclusions

Households can contribute to reducing natural resource scarcity as well as mitigating climate change by better sorting their waste. While the public benefits of improved waste sorting are substantial, the private benefits are negligible. In this paper, we examine whether providing information about the behavior of neighbors can be leveraged to stimulate households to separately sort their organic waste. We find that providing information about the average weight of organic waste deposited by other households in one's block and in combination with setting group goal, is an effective strategy to persistently increase sorting rates – in terms of the frequency of use of the organic waste collection facilities, as well as in the amount of organic waste deposited. In contrast, we find that providing information about the behavior of a specific fellow resident in the neighborhood, a local 'role model', does not result in similar increases in waste sorting – neither when this information is provided in isolation (as a stand-alone intervention), nor when it is implemented in combination with the information on the actual behavior of the households in their direct vicinity.

Our study contributes to the scant literature on the impact of soft behavioral interventions to improve household waste sorting. We interpret our two key interventions as attempts to harness social learning to improve environmental outcomes in a context where the costs of environmentally friendly behaviors are (perceived to be) considerable, and of which the private monetary benefits are essentially zero. We do acknowledge, however, that our interventions, and then especially the social feedback intervention, also include aspects other than those typically associated with social learning. The fact that in the social feedback treatment information is provided about the actual amount of waste sorted by one's peers may be perceived as a (descriptive) social norm, and the fact that a star evaluation system is used to convey information about actual performance implies that the treatment also has elements associated with collective goal setting. Further research is needed to uncover the exact mechanism by which the provision of information on the behavior of one's peers affects outcomes.

3.7 References

Aadland, D., & Caplan, A. J. (2003). Willingness to pay for curbside recycling with detection and mitigation of hypothetical bias. *American Journal of Agricultural Economics*, *85*(2), 492–502.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (NBER Working Paper No. 24003). National Bureau of Economic Research.

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, *29*(1), 1–11.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9-10), 1082–1095.

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037.

Allison, P. D., & Waterman, R. P. (2002). Fixed-effects negative binomial regression models. *Sociological Methodology*, *32*, 247–265.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Ashraf, N., Bandiera, O., & Jack, B. K. (2014). No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics*, *120*, 1–17.

Baird, S., Bohren, J. A., McIntosh, C., & Özler, B. (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, *100*(5), 844–860.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-hall.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122–147.

Bandura, A. (1988). Organisational applications of social cognitive theory. *Australian Journal of Management*, *13*(2), 275–302.

Bandura, A., & Walters, R. H. (1963). *Social learning and personality development.* New York: Holt, Rinehart and Winston.

Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, *107*(3), 797–817.

Bartl, A. (2014). Moving from recycling to waste prevention: A review of barriers and enables. *Waste Management and Research*, *32*(9), 3–18.

Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy*, *96*(4), 675–700.

Bel, G., & Gradus, R. (2016). Effects of unit-based pricing on household waste collection demand: A meta-regression analysis. *Resource and Energy Economics*, *44*, 169–182.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, *119*(1), 249–275.

Bettinger, E. P., Long, B. T., Ehrenberg, R., Jacob, B., & Murnane, R. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review*, *95*(2), 152–157.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Bonferroni, C. E. (1935). *Il calcolo delle assicurazioni su gruppi di teste* (Tech. Rep.). Rome: Tipografia del Senato.

Bramoullé, Y., & Kranton, R. (2007). Public goods in networks. *Journal of Economic Theory*, *135*(1), 478–494.

Brekke, K. A., Kipperberg, G., & Nyborg, K. (2010). Social interaction in responsibility ascription: The case of household recycling. *Land Economics*, *86*(4), 766–784.

Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics*, *87*(9-10), 1967–1983.

Briguglio, M. (2016). Household cooperation in waste management: Initial conditions and intervention. *Journal of Economic Surveys*, *30*(3), 497–525.

Burlig, F., Preonas, L., & Woerman, M. (2020). Panel data and experimental design. *Journal of Development Economics*, *144*, 102458.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics - methods and applications*. Cambridge: Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge: Cambridge University Press.

Carrus, G., Passafaro, P., & Bonnes, M. (2008). Emotions, habits and rational choices in ecological behaviours: The case of recycling and use of public transportation. *Journal of Environmental Psychology*, *28*(1), 51–62.

Chamizo-Gonzalez, J., Cano-Montero, E. I., & Muñoz-Colomina, C. I. (2016). Municipal solid waste management services and its funding in Spain. *Resources, Conservation and Recycling*, *107*, 65–72.

Czajkowski, M., Zagórska, K., & Hanley, N. (2019). Social norm nudging and preferences for household recycling. *Resource and Energy Economics*, *58*, 101110.

DellaVigna, S., & Linos, E. (2021). *RCTs to scale: Comprehensive evidence from two nudge units* (Working Paper April 2021).

Dijkgraaf, E., & Gradus, R. (2015). Efficiency effects of unit-based pricing systems and institutional choices of waste collection. *Environmental and Resource Economics*, *61*(4), 641–658.

Dijkgraaf, E., & Gradus, R. H. (2004). Cost savings in unit-based pricing of household waste. The case of The Netherlands. *Resource and Energy Economics*, *26*(4), 353–371.

DIME Analytics. (2019). *IETOOLKIT: Stata module providing commands specially developed for impact evaluations.*

Dur, R., & Vollaard, B. (2015). The power of a bad example: A field experiment in household garbage disposal. *Environment and Behavior*, *47*(9), 970–1000.

Dur, R., & Vollaard, B. (2019). Salience of law enforcement: A field experiment. *Journal of Environmental Economics and Management*, *93*, 208–220.

Eble, A., & Hu, F. (2020). Child beliefs, societal beliefs, and teacher-student identity match. *Economics of Education Review*, *77*(19), 101994.

Ferrara, I. D. A., & Missios, P. (2005). Recycling and waste diversion effectiveness: Evidence from Canada. *Environmental & Resource Economics*, *30*, 221–238.

Ferraro, P. J., Miranda, J. J., & Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: Evidence from a randomized environmental policy experiment. *American Economic Review: Papers & Proceedings*, *101*(3), 318–322.

Ferraro, P. J., & Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, *95*(1), 64–73.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Foster, A. D., & Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, *103*(6), 1176–1209.

Fullerton, D., & Kinnaman, T. C. (1995). Garbage, recycling, and illicit burning or dumping. *Journal of Environmental Economics and Management*, *29*, 78–91.

Fullerton, D., & Kinnaman, T. C. (1996). Household responses to pricing garbage by the bag. *American Economic Review*, *86*(4), 971–984.

Fullerton, D., & Wolverton, A. (2000). Two generalizations of a deposit-refund system. *American Economic Review*, *90*(2), 238–242.

Gamba, R. J., & Oskamp, S. (1994). Factors influencing community residents' participation in commingled curbside recycling programs. *Environment and Behavior*, *26*(5), 587–612.

Giaccherini, M., Herberich, D. H., Jimenez-Gomez, D., List, J. A., Ponti, G., & Price, M. K. (2019). *The behavioralist goes door-to-door: Understanding household technological diffusion using a theory-driven natural field experiment* (NBER Working Paper No. 26173). National Bureau of Economic Research.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482.

Gómez-Miñambres, J. (2012). Motivation through goal setting. *Journal of Economic Psychology*, *33*(6), 1223–1239.

Gsottbauer, E., & van den Bergh, J. C. J. M. (2011). Environmental policy theory given bounded rationality and other-regarding preferences. *Environmental and Resource Economics*, *49*(2), 263–304.

Hage, O., Söderholm, P., & Berglund, C. (2009). Norms and economic motivation in household recycling: Empirical evidence from Sweden. *Resources, Conservation and Recycling*, *53*(3), 155–165.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Hogg, D., Favoino, E., Nielsen, N., Thompson, J., Wood, K., Penschke, A., ... Papageorgiou, S. (2002). *Economic analysis of options for managing biodegradable municipal waste* (Final Report to the European Commission). Eunomia Research & Consulting, Scuola Agraria del Parco di Monza, HDRA Consultants, ZREU and LDK ECO on behalf of ECOTEC Research & Consulting.

Holm, S. (1979). Board of the foundation of the scandinavian journal of statistics a simple sequentially rejective multiple test procedure author

(s): Sture holm published by : Wiley on behalf of board of the foundation of the scandinavian journal of statistics stable u. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hornik, J., Cherian, J., Madansky, M., & Narayana, C. (1995). Determinants of recycling behavior: A synthesis of research results. *Journal of Socio-Economics*, *24*(1), 105–127.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference: For statistics, social, and biomedical sciences an introduction.* Cambridge: Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Kinnaman, T. C. (2006). Policy watch: Examining the justification for residential recycling. *Journal of Economic Perspectives*, *20*(4), 219–232.

Kinnaman, T. C., & Yokoo, H. F. (2011). The environmental consequences of global reuse. *American Economic Review: Papers & Proceedings*, *101*(3), 71–76.

Kirakozian, A., & Charlier, C. (2015). *Just tell me what my neighbors do! public policies for households recycling* (GREDEG WP No. 2015-20).

Knussen, C., & Yule, F. (2008). "I'm not in the habit of recycling": The role of habitual behavior in the disposal of household waste. *Environment and Behavior*, *40*(5), 683–702.

Lange, F., Brückner, C., Kröger, B., Beller, J., & Eggert, F. (2014). Wasting ways: Perceived distance to the recycling facilities predicts pro-environmental behavior. *Resources, Conservation and Recycling*, *92*, 246–254.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Linder, N., Lindahl, T., & Borgström, S. (2018). Using behavioural insights to promote food waste recycling in urban households - Evidence from a longitudinal field experiment. *Frontiers in Psychology*, *9*(352), 1–13.

List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*(4), 773–793.

Matheson, T. (2019). *Disposal is not free: Fiscal instruments to internalize the environmental costs of solid waste* (IMF Working Paper No.

WP/19/283). International Monetary Fund.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, *99*(2), 210–221.

Milford, A. B., Øvrum, A., & Helgesen, H. (2015). *Nudges to increase recycling and reduce waste* (Discussion Paper No. 2015–01). NILF Norwegian Agricultural Economics Research Institute.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311–329.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32.

Nomura, H., John, P. C., & Cotterill, S. (2011). The use of feedback to enhance environmental outcomes: a randomised controlled trial of a food waste scheme. *Local Environment*, *16*(7), 637–653.

Nyborg, K., Howarth, R. B., & Brekke, K. A. (2006). Green consumers and public policy: On socially contingent moral motivation. *Resource and Energy Economics*, *28*(4), 351–366.

Oster, E., & Thornton, R. (2012). Determinants of technology adoption: Peer effects in menstrual cup take-up. *Journal of the European Economic Association*, *10*(6), 1263–1293.

Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. *Science*, *325*(5939), 419–422.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, *41*(5), 847–855.

Puhani, P. A. (2000). The heckman correction for sample selection and its critique. *Journal of Economic Surveys*, *14*(1), 53–68.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007, May). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–434.

Sidique, S. F., Lupi, F., & Joshi, S. V. (2010). The effects of behavior and attitudes on drop-off recycling activities. *Resources, Conservation and Recycling*, *54*(3), 163–170.

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

Usui, T. (2008). Estimating the effect of unit-based pricing in the presence of sample selection bias under japanese recycling law. *Ecological*

Economics, *66*(2-3), 282–288.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources*, *33*(1), 127–169.

Viscusi, W. K., Huber, J., & Bell, J. (2011). Promoting recycling: Private values, social norms, and economic incentives. *American Economic Review: Papers & Proceedings*, *101*(3), 65–70.

Vollaard, B., & van Soest, D. (2020). *Breaking habits.*

Vringer, K., van der Heijden, E., van Soest, D., Vollebergh, H., & Dietz, F. (2017). Sustainable consumption dilemmas. *Sustainability*, *9*(6), 1–21.

Wolfolds, S. E., & Siegel, J. (2019). Misaccounting for endogeneity: The peril of relying on the heckman two-step method without a valid instrument. *Strategic Management Journal*, *40*(3), 432–462.

Young, W., Russell, S. V., Robinson, C. A., & Barkemeyer, R. (2017). Can social media be a tool for reducing consumers' food waste? A behaviour change experiment by a UK retailer. *Resources, Conservation and Recycling*, *117*, 195–203.

3.A Appendices

3.A.1 Supplementary materials

Figure 3.A1: Photograph of one of the collective organic waste containers that were introduced in the research area.

Figure 3.A2: Example of one of the four social modelling flyers.

Title Translates as: [Name model] cleans up the dining table. In the final panel the model states 'Sorting waste is fun and good for the environment'.

To the resident of this address
[address]
[PC] SCHIEDAM

Your letter from	Your feature	Treated by	Date
		[CUSTOMER SERVICE]	[DATE]
Regarding		Extention number	Our reference
Food scraps test		[EXT NUMBER]	[REFERENCE NUMBER]

Dear resident of Schiedam West,

In the vegetable, fruit and food (VFF) container closest to you, <container weight> kilos of VFF have been collected separately in the past 4 weeks. The goal for this container is <goal weight> kilos.

With <container weight> kilo your current score is: <star score>
Neighborhood container goal: ★★★★★

In total you will receive 4 letters with feedback on the neighborhood container weight. This is how your neighborhood container performs per month:

<stars1>	<stars2>		
Month 1	Month 2	Month 3	Month 4

<statement>
<statement: how much more for an extra star>

Explanation
In June 2018 you received a letter regarding a national test for vegetables, fruit and food waste (VFF) sorting. Your neighborhood in Schiedam West has been selected to participate. The test aims to find out how we can best support you as a resident sorting VFF.

Progress of test in Schiedam West
All preparations for the test in Schiedam West have now been completed. Surveys were conducted prior to the test. Perhaps you have also been approached for this. This letter is the start of the trial, which will run until about June this year.

Monthly letter
As part of the test, we will regularly inform some of the residents of Schiedam West on the quantity of VFF they deliver separately to the nearest VFF collection container. You belong to the group of randomly chosen residents that we provide this waste information to every month. We hope this helps you to dispose your VFF separately. If you are already sorting, we hope it helps you to keep going.

Why sort VFF waste?
VFF does not belong in residual waste. If you deliver your VFF waste separately, we can use it again as a raw material input for new products. For compost and also for biogas. By sorting your VFF we also have to burn less waste. This means less CO2 emissions and less of a burden on the environment.

More information
We will of course inform you at the end of the test of the results. Do you currently have questions about this research? Then please contact us. You can do this via the website of Irado: [WEBSITE]
Do you have questions about waste sorting? Please contact us via [EMAIL] or [TELEPHONE NUMBER].

Sincerely,

[NAME LOCAL PROJECT LEADER]
Project leader

Tip:
What is allowed in the VFF container?
√ In the VFF container everything that is edible and that grows and flourishes, such as: vegetables, fruit, potatoes, meat and fish, peanut and nut shells, eggshells, food scraps, tea bags, coffee grounds and coffee filters, bread, cheese crusts and small flowers - and plant waste.

X The VFF container is not allowed: cat litter without eco-label, dog and cat hair, wood and thick branches, sand / soil, bird cage sand, (full) vacuum cleaner bags, ash from ashtray and fireplace. And also no traditional plastic bags, garbage bags or pedal bin bags.

gemeente Schiedam

Afval Challenge

pagina 2

Figure 3.A3: Social feedback letter.

	(1)	(2)	(3)	(4)	(5)	(6)
Type of household-container cluster	Pure Control	Mixed Contr./ Mod.	Pure Feedback	Mixed Feedb./Comb.		
Treatment assigned	*Control*	*Control*	*Social feedback*	*Social feedback*	Difference	Difference
	Mean/SE	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(3)-(4)
Average number of household members	2.267	2.186	2.300	2.253	0.081	0.047
	[0.072]	[0.099]	[0.080]	[0.063]		
Share of one person households	0.354	0.382	0.346	0.374	-0.029	-0.027
	[0.022]	[0.038]	[0.030]	[0.030]		
Share of two person households	0.306	0.310	0.306	0.288	-0.004	0.019
	[0.010]	[0.038]	[0.016]	[0.025]		
Share of households with three or more members	0.340	0.308	0.347	0.338	0.033	0.009
	[0.022]	[0.030]	[0.026]	[0.019]		
Share of households with at least one member over 65	0.182	0.239	0.184	0.240	-0.057	-0.056**
	[0.018]	[0.031]	[0.014]	[0.020]		
Share of households with at least one child younger than 3 yrs	0.099	0.066	0.110	0.084	0.033**	0.026
	[0.011]	[0.008]	[0.009]	[0.014]		
Average living surface of the dwelling (in m2)	99.024	97.813	102.361	94.821	1.212	7.540
	[4.957]	[7.835]	[6.164]	[2.600]		
Average tax value of the dwelling (in euro's)	1.35e+05	1.49e+05	1.45e+05	1.33e+05	-1.43e+04	12181.533
	[5271.923]	[17482.566]	[9684.654]	[4949.878]		
Share of owner-occupied dwellings	0.583	0.536	0.561	0.567	0.047	-0.006
	[0.051]	[0.118]	[0.083]	[0.044]		
Share of households living in multi-family dwellings	0.674	0.716	0.724	0.695	-0.041	0.029
	[0.061]	[0.066]	[0.053]	[0.051]		
Average distance to nearest organic waste container (in m)	67.561	75.159	77.162	69.555	-7.598	7.607
	[6.610]	[6.267]	[6.448]	[2.779]		
Average number of unique residual waste disposal days per week	1.799	1.736	1.763	1.846	0.063	-0.083
	[0.047]	[0.057]	[0.071]	[0.060]		
Average number of unique organic waste disposal days per week	0.118	0.099	0.100	0.139	0.019	-0.039
	[0.025]	[0.032]	[0.022]	[0.026]		
Share of households having used the organic waste facilities	0.154	0.160	0.154	0.189	-0.006	-0.035
at least once during the baseline period	[0.014]	[0.040]	[0.014]	[0.021]		
Share of households that regularly used	0.066	0.066	0.053	0.075	-0.000	-0.022
the organic waste facilities during the baseline period	[0.015]	[0.023]	[0.012]	[0.014]		
Number of households	1,043	468	1,103	455		

Table constructed using the iebaltab command of the ietoolkit package (DIME Analytics, 2019).
The t-tests are conducted on the treatment coefficients obtained by regressing each covariate on the treatment indicator. Standard errors are clustered at the household-container level.
***, **, and * respectively indicate significance at the 1, 5, and 10 percent critical level.

Table 3.A1: Balance tests for the household-container clusters to be used for spillover tests IV and V.

3.A.2 Robustness checks main analysis

Negative binomial and probit models

In this Appendix we test whether results of Table 3.3, obtained via OLS, are robust to using econometric models that are able to address the discrete-value nature of our dependent variables: the negative binomial model to address the issue of the frequency-of-use data taking on values between 0 and 7, and the probit model to address the binary nature of the choice to dispose waste at the organic facilities in a specific week.

In our robustness analyses, we either include household or container fixed effects as covariates in the negative binomial models, and always container fixed effects in the probit models. Our choice of the different types of fixed effects is driven by the discussion on how to deal with the 'incidental parameter problem' in nonlinear models with fixed effects (Neyman & Scott, 1948). Specifically, as fixed effects cannot be easily partialled out of the estimation in nonlinear models, they all need to be estimated. This can cause the model parameter estimates to not converge in probability to their true values if the number of cross-sections is large (because the number of fixed effects is increasing in the number of cross-sectional units). This concern is valid for the probit model (Cameron & Trivedi, 2005), and hence we use container-level fixed effects to save on the number of fixed effect parameters to be estimated. In contrast, Allison and Waterman (2002) show that the incidental parameter problem does not bias the estimates of the negative binomial model even if the number of fixed effects is large. We thus probed whether the negative binomial regressions converged when using household fixed effects; if not, we present the results using container fixed effects.

The results of the negative binomial regressions are presented in columns (1), (3) and (5) of Table 3.A2. The coefficients presented are incidence ratios, and hence they capture the treatment effect as measured by the percentage change in the dependent variable. The results obtained using the negative binomial are qualitatively identical to those of the OLS regressions (as presented in Table 3.3). The largest treatment effects are found for the social feedback intervention, those of the social modelling are smaller, and the combined treatment did not improve sorting above and beyond what has been achieved by the social feedback treatment effect. Also in quantitative terms the results are similar. The negative binomial results estimate the social feedback treatment to increase sorting by 69 and 83 percent during and after the intervention period, respectively, as compared to 45 and 71 percent according to the OLS estimates. The difference in estimated treatment effects of the social modelling intervention are even smaller; the negative binomial model documents 29% and 26% increases in the frequency of use in and after the intervention period, respectively, compared to 35% and 30% increases according to the OLS estimates. And regarding the robustness of the treatment estimates on the share of unique households that made use of the facilities per week, we find that the marginal effect estimates of the probit model are virtually identical to the marginal effects obtained using OLS; compare columns (2), (4) and (6) of Tables 3.3 and 3.A2.

	Social feedback Pure Feedback vs. Pure Control clusters		Social modelling within Mixed Control/ Modelling clusters		Social modelling on top of feedback within Mixed Feedback/Combined clusters	
	(1) Disposal Days	(2) Decision to sort	(3) Disposal Days	(4) Decision to sort	(5) Disposal Days	(6) Decision to sort
main						
Social feedback × Intervention period	1.686***	0.0465***				
	(0.174)	(0.00986)				
Social feedback × Post intervention	1.827***	0.0435***				
	(0.302)	(0.0131)				
Social modelling × Intervention period			1.294**	0.0188**		
			(0.158)	(0.00914)		
Social modelling × Post intervention			1.261*	0.0151*		
			(0.173)	(0.00867)		
Combined treatment × Intervention period					1.066	0.0150
					(0.0866)	(0.0142)
Combined treatment × Post intervention					1.031	0.00928
					(0.0976)	(0.0147)
Reference group mean (Intervention period)	0.126	0.0763	0.104	0.0707	0.243	0.149
Reference group mean (Post intervention)	0.125	0.0752	0.116	0.0752	0.244	0.144
Number of observations (N × T)	128760	128760	55560	55560	54900	54900
Number of households (N)	2146	2146	926	926	915	915
Week fixed effects	Y	Y	Y	Y	Y	Y
Household fixed effects	N	N	N	N	Y	N
Container fixed effects	Y	Y	Y	Y	N	Y
$\chi^2(1)$-statistic $[\beta^T = \beta^P]$	0.362	0.0985	0.0775	0.245	0.164	0.456
p-value $[\beta^T = \beta^P]$	0.547	0.754	0.781	0.620	0.686	0.500

Standard errors, clustered at the household-container level in (1) and (2) and at the household level in (3)-(6), in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.A2: Estimates of the main treatment effects (Tests I-III) using non-linear estimation models (the negative binomial model and the probit model for the odd- and even-numbered columns, respectively), aimed at probing the robustness of the treatment estimates obtained using OLS.

Multiple hypothesis testing

In this subsection we analyze whether the results of our main tests, Tests I-III, are robust to correcting for multiple hypothesis testing (MHT). We follow the approach developed by List et al. (2019) using STATA's `mhtexp` command. List et al.'s procedure acknowledges the possible dependence between the p-values of the various tests (think of the test results of an intervention's impact on two alternative measures of essentially the same outcome variable of interest), and hence is better able to detect truly false rejections than, for example, the corrections proposed by Bonferroni (1935) or Holm (1979); see also Romano and Wolf (2005). We perform this procedure for our four key outcome variables: the frequency of organic waste sorting and the share of households using the collective organic waste facilities, in the intervention period as well as in the post intervention period. Table 3.A3 presents the results.

The results are as follows. Having corrected for MHT, the social feedback treatment significantly increases the share of households using the facilities both within the intervention period ($p = 0.032$) as well as thereafter ($p = 0.000$). The treatment's impact on the frequency of of use also remains significantly different after the intervention has been completed ($p = 0.000$), but not so in the period in which the treatment was in the process of being rolled out ($p = 0.155$).

Next, the social modelling impacts turn out to be less robust to MHT – in line with the earlier finding that these impacts are both smaller and less persistent than those of the social feedback treatment. The social modelling treatment's impacts on the frequency of sorting remain borderline significant within and post intervention (with $p = 0.091$ and $p = 0.107$, respectively), but its impact on the share of households using the facilities fail to remain significant both within and after the intervention period ($p = 0.172$ and $p = 0.174$, respectively). And consistent with the earlier results of a lack of an additive effect of the combined treatment compared to the social modelling treatment, none of the MHT-corrected tests are significant ($p = 0.872$, or worse).

Outcome	Test	Treatment 1	Treatment 2	Household-Container Clusters	Difference in means	Multiplicity Adj. p-value
Share of households sorting (Intervention Period)	I	Control	Social Feedback	Pure Control vs. Pure Feedback	0.028	0.032
Share of households sorting (Intervention Period)	II	Control	Modelling	Mixed Modelling/Control	0.032	0.172
Share of households sorting (Intervention Period)	III	Social Feedback	Combined treatment	Mixed Feedback/Combined	0.010	0.872
Share of households sorting (Post Intervention)	I	Control	Social Feedback	Pure Control vs. Pure Feedback	0.045	0.000
Share of households sorting (Post Intervention)	II	Control	Modelling	Mixed Modelling/Control	0.032	0.174
Share of households sorting (Post Intervention)	III	Social Feedback	Combined treatment	Mixed Feedback/Combined	0.001	0.994
Frequency of sorting (Intervention Period)	I	Control	Social Feedback	Pure Control vs. Pure Feedback	0.039	0.155
Frequency of sorting (Intervention Period)	II	Control	Modelling	Mixed Modelling/Control	0.069	0.091
Frequency of sorting (Intervention Period)	III	Social Feedback	Combined treatment	Mixed Feedback/Combined	0.001	0.976
Frequency of sorting (Post Intervention)	I	Control	Social Feedback	Pure Control vs. Pure Feedback	0.075	0.000
Frequency of sorting (Post Intervention)	II	Control	Modelling	Mixed Modelling/Control	0.067	0.107
Frequency of sorting (Post Intervention)	III	Social Feedback	Combined treatment	Mixed Feedback/Combined	0.008	0.981

Table 3.A3: The outcomes of testing for multiple hypothesis biases.

Chapter 4

The impact of real-time consumption feedback on gas and electricity use

4.1 Introduction

In the world-wide effort to reduce global carbon emissions, much of the policy makers' attention is directed towards the residential energy market. Not only is the residential sector responsible for a substantial share of emissions (in the Netherlands, this share is about 15-20 percent; (PBL, 2020)), its level of energy consumption and associated energy costs may be inefficiently high too. Households often do not take measures that are known to increase energy efficiency even if they are cost-effective (e.g. Allcott & Greenstone, 2012; Gillingham & Palmer, 2014), and there is also evidence

This chapter is based on joint work with Kees Vringer, under the working title *'The impact of real-time and disaggregated energy consumption feedback on residential gas and electricity usage'*. We thank foundation !Woon, the Woonbond, Agem Energieloket (former VerduurSaam Energieloket), the municipality of Utrecht, tenant council Albanianae, tenants platform MEVM, housing association UWoon and residents initiative Energierijk Houten for the opportunity to execute the experiment and for their help with the field work. In particular we thank Eef Meijerman, Ingrid Houtepen, Kelly Schwegler, Jaap van Leeuwen, Marion Overberg, Justin Pagden, Jeroen aan het Rot, Belinda Haverkamp, Cees Veerman, Henk Oostland, Pier Schipper, Iris Uittien, Annet Bultman and Katrin Larsen. Furthermore, we thank Ernestine Elkenbracht and Erik van Lidth de Jeude (Quintens) for their excellent project management of the field work as well as Cheyenne Ramada for her excellent research assistance. We thank Netherlands Enterprise Agency (RVO) and The Dutch Ministry for Internal Affairs for their financial project support. Finally, we thank Daan van Soest and Ben Vollaard for their valuable comments.

that they do not respond in a rational manner to energy price changes (Ito, 2014). Some economists have argued that the residential energy market does not just suffer from traditional market failures (such as the carbon emission externality problem associated with energy consumption), but also from behavioral anomalies preventing households to make energy efficient decisions (Tietenberg, 2009; Allcott et al., 2014; Price, 2014). For instance, households have been shown to misperceive the energy cost of different uses (Attari et al., 2010), and to pay limited attention to energy costs when these are not explicitly made salient (Allcott, 2011; Sexton, 2015).

The recognition that households may make mistakes when consuming energy opens up a new set of behavioral strategies to improve energy decision making. Studies analyzing such strategies demonstrate their potential (Allcott & Rogers, 2014; Jessoe & Rapson, 2014; Ito et al., 2018). One strategy in particular, consumption feedback, has received a lot of attention (see the meta-reviews by Abrahamse et al., 2005; Darby, 2006; Fischer, 2008; Ehrhardt-martinez et al., 2010; Delmas et al., 2013; Karlin et al., 2015). These studies generally find that feedback improves the efficiency of energy consumption. Yet, the different impacts documented suggests that feedback mediums vary in their energy saving potential, raising questions about why these effect sizes are different. With the large investments made in smart meter infrastructure, research into the determinants of effective feedback and the mechanism by which it induces savings becomes also increasingly policy relevant.[1]

In this paper we estimate the impact of providing real-time and disaggregated energy consumption feedback on residential gas and electricity consumption. We do so by implementing a field experiment involving over 800 households. We recruit households – living in gas-heated dwellings and interested in receiving technology, an In-Home Display, that is able to provide (close to) immediate insight in the consumption of both heating and cooking (gas) and appliances (electricity) – in seven locations across the Netherlands. Identification of the treatment effect is by means of a random rejection design, via which half of the applicants receive a device, and the

[1]It is European policy to provide at least 80% of consumers with intelligent metering systems (Directive (EU) 2019/944, 2019). For instance, van Gerwen et al. (2010) calculated for the Netherlands that a smart meter roll out would be cost effective if these smart meters would realize energy savings of 3.5%, or more. Other EU countries also expect energy savings of 2-3%, for example in the UK (1.5-5%), Sweden and Belgium (1-2%) (Vringer & Dassen, 2016). However, Vringer and Dassen (2016) concluded that the promotion of smart meters in the Netherlands did not realize the projected savings: actual energy savings amounted to less than 1 percent, presumably due to the low take up of smart meter applications that provide consumption feedback.

other half does not. We evaluate energy savings by comparing daily electricity and gas consumption data that we collected for at least 7 months after the displays were installed. To better understand how the display affected behavior, we provide survey evidence concerning the impact of the device on household knowledge of the largest contributors to their energy bill and their monthly energy charges.

We find the display to induce substantial energy savings. On average, households that receive the display realize gas savings of 6.9 percent and electricity savings of 2.2 percent. The gas savings seem to be highest during colder days, a pattern suggestive of changes in the energy consumption of space heating. Though households save energy on one of the most energy intensive categories within their home, we do not find evidence that households targeted this use due to being better informed of space heating's relative energy intensity compared to other uses. However, we do find evidence that households became better informed of the amount of money spent on gas consumption. As households report to continue actively use the display long after having received it, (sustained) energy cost salience seems to have been an important mediator driving the observed changes in energy consumption.

Our study is thus unique due to its combination of two aspects. First, we complement the feedback literature by providing evidence on an in-home display that distinguishes the consumption by energy-intensive uses (in our case gas, predominantly reflecting heating-related consumption) from the energy use of electric appliances.[2] There is various research suggesting that the impact of electricity feedback on electricity savings can be increased by adding other information as a complement (see e.g. Abrahamse et al., 2005; Tedenvall & Mundaca, 2016). Feedback on the underlying sources of energy consumption has been found to increase the impact of aggregated forms of feedback by 5 percent (Gerster & Andor, 2020), while Schleich et al. (2017) finds that IHD feedback generates an additional 5 percent savings when households are offered the option to also receive web-portal or standard mail feedback. Asensio and Delmas (2015) find feedback to generate 8 percent more electricity savings when provided with reminders of the environmental and health externalities involved with energy consumption. Yet the potential of display feedback on separate energy uses is underexplored. This framing

[2]Electric air-conditioning devices are also very energy-intensive. However, although air-conditioning is becoming increasingly popular in the Netherlands, it remains atypical for a household to have air-conditioning. We therefore assume cooling technologies to be of limited concern in our sample. Based on the seasonal patterns in electricity consumption, we do not find evidence suggesting otherwise, see section 4.2.4.

of feedback information is not only relevant for contexts with dual-fuelled residences, such as the Netherlands, but is also relevant for electricity-only homes, due to the development of new technologies recognizing the footprint of individual uses in metered energy consumption.[3]

Second, we complement the existing line of work by analyzing the effect of display feedback using a field experiment that involves a larger sample size and a longer time period than typical in this literature. Our study involves a sample size of over 800 households monitored for over 7 months post treatment. We so contribute to a number of peer-reviewed studies (e.g. Schultz et al., 2015; Aydin et al., 2018; Westskog et al., 2015; Lynham et al., 2016; Schleich et al., 2017)[4] and a large number of utility pilots and policy reports (see e.g. Faruqui et al., 2010; McKerracher & Torriti, 2013; Ehrhardt-martinez et al., 2010) analyzing the effect of energy in-home displays. The peer-reviewed literature typically involves sample sizes below 150 households (Westskog et al., 2015; Lynham et al., 2016; Aydin et al., 2018) while utility pilots tend to be of larger scale (Faruqui et al., 2010; McKerracher & Torriti, 2013). The study by Schultz et al. (2015) monitors the energy consumption of about 431 households, but over a relatively short period of time (up to three months after the installation of the display).

Both the time since installation and sample size are important features for further study. Treatment exposure generally has been identified as a factor important for feedback impact, although the direction of the effect remains unclear (Darby, 2006; Fischer, 2008; Ehrhardt-martinez et al., 2010). Moreover, the combination of the field experimental method combined with a larger sample size is important. Current research tends to suffer from low statistical power compared to the larger scale utility pilots, but at the same time the peer-reviewed work tends to be more explicit in supporting their claim to causal inference. This study aims to bridge this gap by analyzing the effect of an energy display in a randomized field experiment using a

[3]A notable study analyzing the impact of targeted feedback using in-home displays is the policy report by CER (2011). This report analyzes effective gas demand-side management programs using a randomized controlled trial, with treatment arms of about 300 households. One treatment is involves in-home displays providing gas consumption feedback as part of a combination treatment (also including bimonthly energy billing and a detailed energy usage statement). The authors find this combined treatment to induce 2.9 percent gas savings measured over a period of 11 months. See also Harold et al. (2015) for further analysis.

[4]Note that there are additional studies analyzing high-frequency feedback as a complement to dynamic price incentives (Jessoe & Rapson, 2014; Ito et al., 2018). Since research has shown price incentives to strongly influence the impact of feedback (Faruqui et al., 2010), we exclude these here.

larger sample size and time horizon.

Our results highlight that real-time and disaggregated energy consumption feedback mitigates consumer mistakes. We find evidence that the feedback results in households' becoming better informed of the size of the gas bill. The induced reduction in energy consumption is thus likely to reflect a better alignment of households' actual energy consumption with their privately optimal consumption.

Yet it remains an open question to what degree households depend on continued interaction with the device to remain informed (salience effects), or whether they internalized the information (learning effects) (Jessoe & Rapson, 2014; Lynham et al., 2016). Our experimental design does not allow us to distinguish the (relative) impact of salience and learning as drivers of this result. Instead, we use the survey answers to provide descriptive evidence for the household engagement with the device. Our study so contributes to discussions in the literature regarding the engagement fostered by display devices, and the degree to which they allow the users to learn from the feedback (Buchanan et al., 2015).

The set-up of this paper is as follows. Section 4.2 discusses the experimental design, a random-rejection design with participants recruited at seven different locations. The section also explains how we expect the in-home display to affect household energy behavior. Section 4.3 discusses our energy data set and the empirical strategy we use to analyze this data. Section 4.4 covers the results of this analysis. Section 4.5 analyzes survey answers to gain insight into the determinants of energy savings. Section 4.6 concludes.

4.2 Experimental Design

4.2.1 Background and Experimental setting

When households consume energy, they typically do not receive feedback on their level of energy use or the financial consequences of their choices. This lack of feedback makes it difficult for households to learn the energy cost associated with their consumption decisions, which may lead households to overconsume energy and not properly evaluate the energy costs of new appliances or energy conservation measures. The difficulty to observe this cost is inherent in the way households consume energy: households consume energy indirectly through other consumption decisions, while the ex-post and aggregate nature of energy billing obscures the underlying consumption behaviors that gave rise to the energy bill. In the Netherlands for instance, nearly all households pay their energy costs in advance by a fixed monthly installment, while the actual bill is settled with the advance payment after a year.

With the rise of smart meters, the possibility has opened up to collect high-frequency energy data to be employed in informational technologies, such as in-home displays. These devices, also referred to as energy monitors, do not only allow households to receive better insight into their energy consumption patterns, but also allow households to receive this feedback in at a location within their home of their choosing.

To analyze the effect of an energy in-home display on household energy consumption, we partnered with seven local organisations to recruit a large pool of households interested in receiving a free in-home display to be used for one year. Our partners included municipalities, energy offices and social housing- and tenants associations, operating in seven different areas in the Netherlands. As a result, our experiment took place at seven locations. Recruitment occurred over a period of almost two years (from Fall 2017 until June 2019), and consisted of various approaches, including social media campaigns, leaflets, and advertisements in neighborhood magazines. The display was introduced as a device that provides insight in actual energy usage which could help households cut costs on their energy bill. The materials informed households interested in receiving the device to sign up using a local project website.

The sign-up process included a short pre-experiment questionnaire and the experimental participants were restricted to those that met the requirements for participation: the household should live in one of the partner areas, should not be planning to move within the near future, have a modern (non DSMR 2.2) smart meter that can be accessed by their energy company, and live in a gas-fired dwelling. Last, the household had to agree to share their energy data for research purposes. These recruitment campaigns led to an experimental population of 802 households for which electricity and gas data were available. Figure 4.1 shows the location of the experimental population by postal code. The three largest experimental locations are recognizable on the map as the three largest non-overlapping circles. Identification of the causal impact of the display on energy consumption is by randomizing which households were offered the display. Because our experimental sample consisted of households having applied to receive an display, we essentially use an "random rejection" approach.

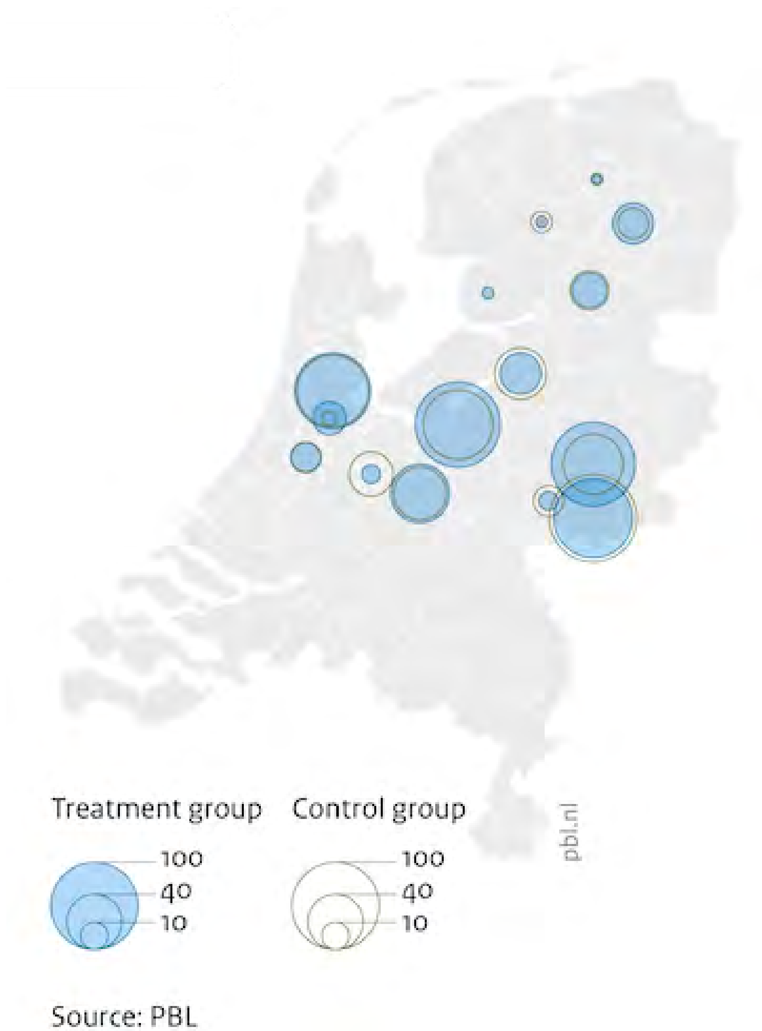


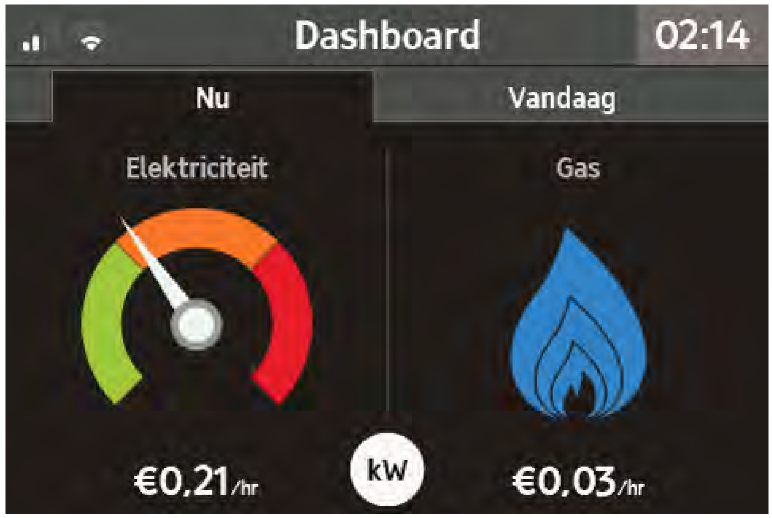
Figure 4.1: Distribution of Experimental population by postal code 2-region.

4.2.2 Treatment

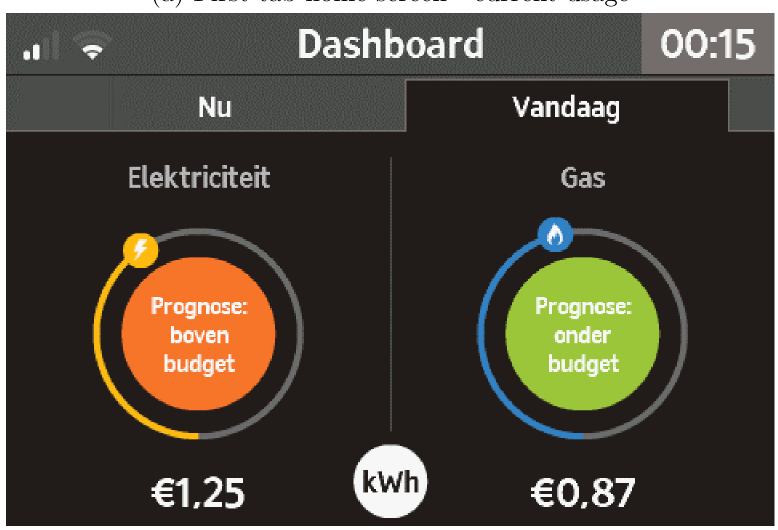
The in-home display analyzed in our study is a small monitor that provides direct feedback on gas and electricity consumption. The direct feature of the device, as defined by Darby (2006), implies that the energy use information has not been processed in some form, and thus can be provided relatively quickly to the user. In our case, the electricity feedback is given (close to) instantaneously, while the feedback on gas consumption is provided after at most 10 minutes. The consumption data is visualized by means of a simple color-coded dashboard. The dashboard allows households to choose whether the feedback is provided in terms of current hourly energy consumption, or in comparison to a daily budgetary goal. In addition, a colored light is fitted to the top of the display that signals the level of current electricity usage as determined by a self-learning algorithm. The display is also able to provide feedback information on past energy consumption (e.g. per week, or per month). Figures 4.2 and 4.3 provide examples of the display and the energy dashboard.

Figure 4.2: The Energy In-Home Display

Arguably, the display feedback increases the salience of a household's energy costs. As in the absence of feedback, the energy use implications of consumption decisions within the home only become apparent at the later moment of billing, it is likely that households pay little attention to this aspect when consuming energy. The display makes energy costs salient at the moment of consumption, by drawing attention to energy costs through both the mounted light and the color-coded dashboard. Salience has been found to affect decision making in various field experimental studies (see for



(a) First tab home screen - current usage



(b) Second tab home screen- today's budget

Figure 4.3: Key functions of the energy in-home display

Translations when reading from left to right, from above to below:
Panel (a):
Now - Today, Electricity - gas.
Panel (b):
Now - Today, Electricity- gas, Forecast: Above budget, Forecast: Below budget.

a review of the literature DellaVigna, 2009).

Besides reminding households of energy costs, the display's informational content may facilitate learning. In particular, as the display provides information on the real-time electricity and gas usage, households are able to update beliefs regarding (the costs of) their aggregate level of electricity and gas consumption. Typically, households are uninformed about the size of their energy bill, providing substantial scope for improvement. Brounen et al. (2013) finds that only about half of Dutch households that participated in a representative survey was able to provide an estimate of their monthly energy charges. A possible reason for this lack of attention are automatic electronic billing plans that allow households to pay for their utilities without needing to review their bill first (Sexton, 2015).

Moreover, with more frequent exposure to consumption feedback households may deduce the importance of energy intensive measures in driving these bills. Households can for instance compare their level of energy use before and after consumption decisions to infer the level of energy use associated with specific actions. As such, the consumption feedback may help households to learn what processes are important contributors to their energy consumption and become better able to identify what behaviors, appliances and technologies are most conducive to conserving energy. This type of learning may be especially facilitated by the gas feedback, as the feedback involves at most three types of processes and behaviors: space-heating, water-heating, and cooking. The electricity feedback in contrast, may in addition reflect many other (electrical) appliances, making it more difficult to judge the energy intensity of the underlying measures. To help households uncover effective energy saving strategies, we provided households with an accompanying booklet, which helps households conduct the necessary calculations to estimate yearly savings. The booklet pays special attention to uncovering the yearly cost of low-energy intensive appliances that always use energy as opposed to high-energy intensive appliances that only create infrequent peaks in usage. Previous studies have found substantial scope for improvement on energy consumption beliefs. Attari et al. (2010) finds that households systematically underestimate the energy costs of processes in the home and the potential for energy savings by a factor of three. In addition, households seem to be overly focused on behaviors as opposed to investments, as they tend to perceive behavioral changes such as switching off lights as the most conducing to conserving energy (Attari et al., 2010).

As false beliefs regarding the profitability of energy saving measures may reflect mistakes in how households process information, the display and the accompanying booklet help households to properly process the energy consumption feedback. Examples of such mistakes, relevant in the domain of energy efficiency, are present bias or bias of concentration (Allcott, 2016). The first bias refers to mistakes in decision making when households overvalue payoffs closer to the present compared to the future when comparing trade-offs between two future dates (O'Donoghue & Rabin, 1999), while the second refers to mistakes resulting from households overweighing the cost of behaviors and devices whose cost accrue in occasional sharp increases as compared to cost that accrue in sustained smaller amounts over time (Koszegi & Szeidl, 2013).

The display aims to prevent households from making such mistakes when processing the energy consumption feedback. The dashboard allows households to see projections of current energy costs relative to an energy budget, with goal setting being an effective strategy in the energy domain to counteract the present bias internality (Andor & Fels, 2018). The provided booklet helps households calculate the (yearly) energy savings of conservation measures. Households that have trouble making these calculations may be helped by this guidance, which could help them more effectively evaluate the savings potential of conservation strategies (see for instance Brent & Ward, 2018). By guiding households through these calculations, the treatment aims to help households form the right beliefs regarding the financial savings potential of energy saving measures.

4.2.3 Treatment assignment

Because of our long period of recruitment, we allocated households to the treatment (being offered a display) or control group by batch. We monitored the number of households that signed up for the project, and regularly grouped newly recruited households into location specific batches within treatment was assigned. We verified whether these households met the participation requirements and checked the households' smart meter compatibility with the display. If the households met the requirements we assigned them to treatment and control.

The original design was to create a treatment and control group of equal size, by randomly assigning half of each batch to treatment. During the project however, we learnt that our treatment display was incompatible with two smart meter types due to a software malfunction.[5] This affected our design, as the number of households with incompatible smart meters

[5]Specifically, at all locations the display was incompatible with DSMR 2.0 and ISKRA smart meters. For one location (number 5) we were able to update the software which solved the incompatibility issues with ISKRA meters.

was sufficiently large to substantially reduce the power of our design when excluded from the project. We therefore asked practitioners out of the field about the meter type assignment procedure and explored the option of using meter type assignment as an instrument. Based on this qualitative evidence, we decided to exclude one type of unsuitable smart meter (DSMR 2.0 meters) from the project. This particular meter was predominantly installed in households that had actively requested to receive their smart meter even before the national roll-out of smart meters had started. We did include a second type of unsuitable meter (ISKRA meters), as we did not find evidence that households were able to self-select into this type of meter. Specifically, both this meter and suitable meters are predominantly installed free-of charge by network operators in the Dutch national smart meter roll-out. In this national program, network operators contact households region by region and offer them a free smart meter replacement. Households cannot choose which smart meter type they receive, the type installed is dependent on the available meter stock at their network supplier. As network operators use multiple meter suppliers to avoid dependence on one supplier, which meter households receive is quasi-random.

The validity of our randomization procedure is therefore predicated on the smart-meter roll-out having been quasi-random. In accordance with the qualitative evidence from the field, we expect a positive but imperfect correlation in meter type among neighbors. Neighboring households are contacted at the same time in the smart meter roll-out and thus receive a meter from the supply-side mix of meters available at the time of meter installation. We find a positive intra-cluster correlation of 0.17 ($p<0.05$) in meter type (suitable, or unsuitable ISKRA meter) installed at the 2-digit postal code level – a neighborhood of about 1750 households.

Taking into account this knowledge about the meter installations, we analyze whether households assigned to either one of the two meter types are comparable, considering that some regions may be over-represented in one of the two meter type groups. Table 4.1 shows the balance test on observables for both households with and without suitable meter in the 6 locations where we relied on meter type assignment. Specifically, we show the two group means, and then test whether these means are significantly different once we control for region fixed effects. Out of the 28 tests, we only find evidence for a slightly different distribution of residences by construction year and a different self-reported baseline energy efficiency score. The group of households with a suitable meter has a 4.5 percentage point larger fraction of homes built between 1975 and 1984, and a slightly smaller fraction (of 6.6 percentage points) of homes that were built after 2014. Though the households with non-suitable meters are more likely to live in older homes compared to households with suitable meters, a characteristic that typically correlates with higher space heating energy requirements, we also find evidence that these households report their energy efficiency score to be higher. This may indicate that these households have more often taken isolation measures effectively equalizing heating demand. Our results therefore do not suggestive of different levels of energy consumption in one of the groups. In addition, we note that the fraction of newly built residences is the only one with a normalized difference larger than 0.25 – the cutoff typically used in this literature above which imbalances are thought to be a concern (Imbens & Rubin, 2015; Imbens & Wooldridge, 2009; Abadie & Imbens, 2011). Based on the results we therefore conclude that we cannot reject the assumption of as-if random meter type assignment on the regional level. Based on observable characteristics and reported energy efficiency, we do not find the energy consumption of households with a display-suitable energy meter to differ from the energy consumption of households without a suitable meter.

We have therefore verified the validity of meter type as instrument for treatment assignment on the regional level. However, we did not use meter type as the sole determinant of treatment assignment. When we only assign households by meter type, we have no control over the size of the treatment arms which could lead the control group size to deviate substantially from the size of the treatment group, negatively impacting power when we make between group comparisons (List et al., 2011). We therefore decided to employ a hybrid design instead. Our randomization process can be formalized as follows. Let N_{jk} denote the number of applicants in randomization batch j in location k. Of those N_{jk}, we first assign all households with an unsuitable meter to the control group. We then assigned households with a suitable meter to treatment with probability $\pi_{jk} \in [0.5, 1]$. Specifically, this probability is decreasing in the number of unsuitable households already assigned to the batch control group (F_{jk}) , and in some cases, decreasing in the overall control group at the respective location. Specifically, $\pi = \frac{0.5}{1 - \frac{C_{jk}}{N_{jk}}}$, where $\frac{C_{jk}}{N_{jk}}$ denotes the relative size of the batch control group (C_{jk}), to the overall batch size, $\frac{C_{jk}}{N_{jk}}$. In most batches C_{jk} equals the number of households with an unsuitable smart meter (F_{jk}). In a few cases, we chose to redefine $C_{jk} = F_{jk} + \alpha_{jk}$ with $\alpha_{jk} > 0$, to enlarge the probability of treatment group assignment within the batch. We set $\alpha_{jk} > 0$ when previous batches contained many unsuitable meters, leading the the cumulative treatment arm to

Variable	N	(1) Total Mean/SE	N	(2) Non-Suitable meter Mean/SE	N	(3) Suitable meter Mean/SE	T-test Difference (2)-(3)	Normalized difference (2)-(3)
Elderly hh	615	0.241 (0.017)	151	0.219 (0.034)	464	0.248 (0.020)	-0.029	-0.068
1 person hh	615	0.275 (0.018)	151	0.291 (0.037)	464	0.269 (0.021)	0.022	0.049
2 person hh	615	0.590 (0.020)	151	0.576 (0.040)	464	0.595 (0.023)	-0.019	-0.038
>2 hh members	615	0.135 (0.014)	151	0.132 (0.028)	464	0.136 (0.016)	-0.003	-0.010
hh with kids <4 years	615	0.107 (0.012)	151	0.119 (0.026)	464	0.103 (0.014)	0.016	0.051
hh with kids 4-18 years	615	0.309 (0.019)	151	0.298 (0.037)	464	0.313 (0.022)	-0.014	-0.031
Primary school	615	0.020 (0.006)	151	0.026 (0.013)	464	0.017 (0.006)	0.009	0.067
Practical education (no tertiary diploma)	615	0.197 (0.016)	151	0.146 (0.029)	464	0.213 (0.019)	-0.068	-0.170
Secondary higher education / Pract. tertiary educ.	615	0.265 (0.018)	151	0.258 (0.036)	464	0.267 (0.021)	-0.009	-0.020
Higher tertiary education (bachelor)	615	0.270 (0.018)	151	0.298 (0.037)	464	0.261 (0.020)	0.037	0.084
Higher tertiary education (master or doctor)	615	0.249 (0.017)	151	0.272 (0.036)	464	0.241 (0.020)	0.030	0.070
An apartment flat	615	0.250 (0.017)	151	0.278 (0.037)	464	0.241 (0.020)	0.037	0.085
Terraced house	615	0.218 (0.017)	151	0.166 (0.030)	464	0.235 (0.020)	-0.069	-0.168
Corner house	615	0.156 (0.015)	151	0.159 (0.030)	464	0.155 (0.017)	0.004	0.010
Semi-detached house	615	0.200 (0.016)	151	0.219 (0.034)	464	0.194 (0.018)	0.025	0.061
Detached house	615	0.154 (0.015)	151	0.166 (0.030)	464	0.151 (0.017)	0.015	0.041
Other (farmhouse,houseboat)	615	0.021 (0.006)	151	0.013 (0.009)	464	0.024 (0.007)	-0.010	-0.073
Before 1945	615	0.180 (0.016)	151	0.205 (0.033)	464	0.172 (0.018)	0.033	0.085
Between 1945-1974	615	0.223 (0.017)	151	0.219 (0.034)	464	0.224 (0.019)	-0.006	-0.013
Between 1975-1984	615	0.211 (0.016)	151	0.245 (0.035)	464	0.200 (0.019)	0.045**	0.109
Between 1985-1994	615	0.158 (0.015)	151	0.192 (0.032)	464	0.147 (0.016)	0.046	0.125
Between 1995-2004	615	0.098 (0.012)	151	0.079 (0.022)	464	0.103 (0.014)	-0.024	-0.081
Between 2005-2014	615	0.067 (0.010)	151	0.046 (0.017)	464	0.073 (0.012)	-0.027	-0.108
After 2014	615	0.063 (0.010)	151	0.013 (0.009)	464	0.080 (0.013)	-0.066*	-0.273
Home own property	615	0.574 (0.020)	151	0.576 (0.040)	464	0.573 (0.023)	0.003	0.006
Energy efficiency score	557	3.885 (0.053)	135	4.022 (0.100)	422	3.841 (0.063)	0.181*	0.144
Energy saving potential score	578	4.336 (0.040)	143	4.266 (0.082)	435	4.359 (0.046)	-0.093	-0.096
Has solar panels before randomisation	616	0.383 (0.020)	151	0.344 (0.039)	465	0.396 (0.023)	-0.051	-0.105

Notes: Balance test comparing the characteristics of the households with suitable and nonsuitable smart meters (ISKRA type meters). The Table excludes outliers and the 186 households at location 5 where we did not rely on meter type treatment assignment. Energy efficiency score denotes the answer to the question: 'How energy-efficient are you?', where the person filling in the survey could give their answer on a scale from 0 ('Not energy efficient') to 6 ('Very energy efficient'). The 'Imagine you receive a energy display. How much energy do you think you will save with this energy display? Households give their answers on a scale from 0 ('Nothing') to 6 ('Very much'). We perform the balance t-tests using a linear regression model where each covariate is regressed on the meter type indicator, with region fixed effects included in all estimations. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Table adapted from table constructed using the iebaltab command of the ietoolkit package (DIME Analytics, 2019).

Table 4.1: Balance test on the two meter types

lag behind the cumulative control arm size at the location.[6] The randomisation dates, batches and accompanying treatment assignments are listed in Table 4.2.

[6] In 2 batches involving 12 households in total (1.5 percent of experimental population), there were supply-side constraints, reducing scope for treatment assignment. We assigned these households to control.

Date of Rand.		Location 1		Location 2		Location 3		Location 4		Location 5		Location 6		Location 7		Total Freq.	Total %
		Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%		
11-01-2018	Control group							9	1.12%							9	1.12%
	Treatment group							10	1.25%							10	1.25%
23-04-2018	Control group											25	3.12%			25	3.12%
	Treatment group											23	2.87%			23	2.87%
11-06-2018	Control group							3	0.37%					35	4.36%	38	4.74%
	Treatment group							5	0.62%					34	4.24%	39	4.86%
23-06-2018	Control group	10	1.25%			10	1.25%					4	0.50%			24	2.99%
	Treatment group	10	1.25%			4	0.50%					3	0.37%			17	2.12%
06-07-2018	Control group							5	0.62%			5	0.62%			10	1.25%
	Treatment group							5	0.62%			4	0.50%			9	1.12%
24-09-2018	Control group							7	0.87%					2	0.25%	9	1.12%
	Treatment group							5	0.62%					5	0.62%	10	1.25%
24-10-2018	Control group	2	0.25%									11	1.37%			13	1.62%
	Treatment group	2	0.25%									14	1.75%			16	2.00%
30-11-2018	Control group											31	3.87%	2	0.25%	33	4.11%
	Treatment group											32	3.99%	7	0.87%	39	4.86%
15-02-2019	Control group							5	0.62%					3	0.37%	8	1.00%
	Treatment group							4	0.50%					9	1.12%	13	1.62%
10-07-2019	Control group			26	3.24%			4	0.50%	88	10.97%	61	7.61%	17	2.12%	196	24.44%
	Treatment group			36	4.49%			4	0.50%	97	12.09%	76	9.48%	24	2.99%	237	29.55%
04-08-2019	Control group			7	0.87%	11	1.37%			1	0.12%					19	2.37%
	Treatment group			5	0.62%											5	0.62%
Grand Total		24	2.99%	74	9.23%	25	3.12%	66	8.23%	186	23.19%	289	36.03%	138	17.21%	802	100.00%

Table 4.2: The number of households randomised by randomisation date and location, expressed in absolute amount (Freq) or percentage of total experimental population (%)

4.2.4 Data

As households that have signed up for the project filled in a baseline questionnaire, we have access to baseline characteristics relating to household energy usage. In addition, we collected energy data during the project by reading the households' smart meters. To evaluate whether our randomization procedure was successful, we discuss these two data sources and evaluate balance.

Baseline characteristics

After excluding households of the unsuitable meter type installed before the national smart meter roll-out, we end up with an experimental population of 802 households in our experiment for whom we have energy data.[7] The descriptives of these households are depicted in Column (1) in Table 4.3. Most households are two-person households (57%), followed by a substantial share of one-person households (30%). One-third of households has one or more children of school age, while only 10 percent has one or more children below the age of 4. About a quarter of households have one or more household members aged over 65. Furthermore, close to half of households have at least one member with tertiary education. Only 2.5% reports primary school as highest education level in the household. About 29% of the households in our sample live in semi- or fully- detached homes, while an about equal share (25%) lives in an apartment. Only few of these residences were built after 2014 (only 5 percent), while close to two-thirds was built before 1984. About half of the homes are owned-occupied. About 40 percent are classified as having no solar panels at baseline, as we did not observe these households to sell electricity before randomisation.

We conduct balance tests on the overall treatment and control group sample to determine whether our randomization procedure resulted overall in two groups that are similar on observables. Columns (2) and (3) of Table 4.3 show the sample means of the control and treatment group, while the final two columns show their (normalized) mean difference. We cluster the standard errors of all tests at the regional level if the respective location employed meter type assignment. We cluster at the household level instead if the location did not assign unsuitable meters to the control group. We

[7] We also exclude 8 outlier households from our analysis as their estimated yearly amount of electricity purchased was above 9000 KWh, a number well above the electricity requirement for a 5-person household living in a large (detached house) in the Netherlands (5*1631,8=8159 KWh based on CBS data). Data based on CBS data for 2018, website consulted on 10-7-2020 https://www.cbs.nl/nl-nl/cijfers/detail/83882ned.

Variable	(1) Total N/[Clusters]	(1) Total Mean/SE	(2) Control group N/[Clusters]	(2) Control group Mean/SE	(3) Treatment group N/[Clusters]	(3) Treatment group Mean/SE	T-test Difference (2)-(3)	Normalized difference (2)-(3)
Elderly hh	801	0.262	384	0.229	417	0.293	-0.063**	-0.144
	[199]	(0.019)	[101]	(0.023)	[110]	(0.024)		
1 person hh	801	0.302	384	0.318	417	0.288	0.030	0.065
	[199]	(0.049)	[101]	(0.051)	[110]	(0.053)		
2 person hh	801	0.567	384	0.552	417	0.580	-0.028	-0.057
	[199]	(0.030)	[101]	(0.034)	[110]	(0.032)		
>2 hh members	801	0.131	384	0.130	417	0.132	-0.002	-0.005
	[199]	(0.022)	[101]	(0.022)	[110]	(0.026)		
hh with kids <4 years	801	0.094	384	0.112	417	0.077	0.035**	0.121
	[199]	(0.007)	[101]	(0.009)	[110]	(0.011)		
hh with kids 4-18 years	801	0.291	384	0.297	417	0.285	0.012	0.025
	[199]	(0.040)	[101]	(0.033)	[110]	(0.048)		
Primary school	801	0.026	384	0.026	417	0.026	-0.000	-0.002
	[199]	(0.008)	[101]	(0.009)	[110]	(0.011)		
Practical education (no tertiary diploma)	801	0.247	384	0.245	417	0.249	-0.005	-0.011
	[199]	(0.034)	[101]	(0.035)	[110]	(0.037)		
Secondary higher education / Pract. tertiary educ.	801	0.282	384	0.289	417	0.276	0.013	0.029
	[199]	(0.026)	[101]	(0.034)	[110]	(0.027)		
Higher tertiary education (bachelor)	801	0.243	384	0.258	417	0.230	0.028	0.064
	[199]	(0.018)	[101]	(0.021)	[110]	(0.019)		
Higher tertiary education (master or doctor)	801	0.201	384	0.182	417	0.218	-0.036*	-0.090
	[199]	(0.045)	[101]	(0.044)	[110]	(0.047)		
An apartment flat	801	0.247	384	0.227	417	0.266	-0.040	-0.092
	[199]	(0.085)	[101]	(0.080)	[110]	(0.094)		
Terraced house	801	0.262	384	0.284	417	0.242	0.042	0.095
	[199]	(0.035)	[101]	(0.037)	[110]	(0.039)		
Corner house	801	0.190	384	0.195	417	0.185	0.011	0.027
	[199]	(0.026)	[101]	(0.028)	[110]	(0.029)		
Semi-detached house	801	0.166	384	0.169	417	0.163	0.006	0.017
	[199]	(0.047)	[101]	(0.050)	[110]	(0.045)		
Detached house	801	0.119	384	0.112	417	0.125	-0.013	-0.039
	[199]	(0.041)	[101]	(0.040)	[110]	(0.044)		
Other (farmhouse,houseboat)	801	0.016	384	0.013	417	0.019	-0.006	-0.049
	[199]	(0.013)	[101]	(0.010)	[110]	(0.017)		
Before 1945	800	0.147	383	0.136	417	0.158	-0.023	-0.063
	[198]	(0.061)	[100]	(0.063)	[110]	(0.061)		
Between 1945-1974	800	0.247	383	0.253	417	0.242	0.011	0.026
	[198]	(0.031)	[100]	(0.032)	[110]	(0.036)		
Between 1975-1984	800	0.230	383	0.245	417	0.216	0.030	0.070
	[198]	(0.031)	[100]	(0.032)	[110]	(0.034)		
Between 1985-1994	800	0.147	383	0.159	417	0.137	0.023	0.064
	[198]	(0.015)	[100]	(0.018)	[110]	(0.019)		
Between 1995-2004	800	0.101	383	0.086	417	0.115	-0.029	-0.096
	[198]	(0.017)	[100]	(0.018)	[110]	(0.021)		
Between 2005-2014	800	0.074	383	0.065	417	0.082	-0.016	-0.062
	[198]	(0.010)	[100]	(0.011)	[110]	(0.014)		
After 2014	800	0.052	383	0.055	417	0.050	0.004	0.020
	[198]	(0.022)	[100]	(0.023)	[110]	(0.023)		
Home own property	801	0.441	384	0.435	417	0.446	-0.011	-0.022
	[199]	(0.112)	[101]	(0.115)	[110]	(0.113)		
Energy efficiency score	733	3.907	349	3.934	384	3.883	0.051	0.041
	[189]	(0.076)	[95]	(0.096)	[106]	(0.085)		
Energy saving potential score	755	4.356	367	4.319	388	4.392	-0.073	-0.077
	[190]	(0.036)	[96]	(0.064)	[106]	(0.033)		
Has solar panels before randomisation	802	0.409	384	0.393	418	0.423	-0.030	-0.061
	[199]	(0.053)	[101]	(0.058)	[110]	(0.054)		

Notes: Balance test using the ISKRA smart meters and the smart meters suitable for the display. Excluding outliers. There are 2 households for which the baseline survey was incomplete. In addition, the energy efficiency score and energy savings potential score have some missing responses as these questions were optional. Energy efficiency score denotes the answer to the question: 'How energy-efficient are you?', where answer was given on a scale from 0 ('Not energy efficient') to 6 ('Very energy efficient'). The 'Imagine you receive a energy display. How much energy do you think you will save with this energy display? Answers were given on a scale from 0 ('Nothing') to 6 ('Very much'). The value displayed for t-tests are the differences in the means across the groups. The t-tests are conducted on the treatment coefficients obtained by regressing each covariate on the treatment indicator. Standard errors are clustered by region if meter type assignment was used at the location the household belongs to, and clustered on the household level otherwise. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Table adapted from table constructed using the iebaltab command of the ietoolkit package (DIME Analytics, 2019).

Table 4.3: Balance test on the two treatment arms

only find 3 out of the 28 tests to be statistically significant, suggesting that the treatment and control group have on average similar characteristics. Importantly, we do not find differences in terms of home size and age, two characteristics that moderate the effect of smart-meter interventions tested as part of Ireland's Smart Metering Gas Consumer Behavioural Trial (Harold et al., 2018).

We have also verified whether the treatment and control groups within locations are comparable, using the same procedure as before but this time comparing the treatment and control groups within each location. The number of significant differences in the 196 tests[8] is not larger than the number of false rejections consistent with the significance criterion applied (test results available on request).

Energy Data

We have data on households' daily (net) amount of gas and electricity purchased, collected over the period from December 7, 2017, until June 14, 2020. The gas data are measured in cubic meters (m^3) while electricity is recorded in Watt hours (Wh). For gas, energy purchased equals energy consumed, but for households with solar panels we rely on smart meter data reflecting energy purchases and energy returned to the grid to calculate the amount of electricity consumed. For those households, we estimate the electricity production capacity of their solar system by measuring the amount of electricity produced as a function of the sun's strength on a number of different days, and use that function to estimate how much electricity they produced on all other days. By subtracting the observed amount of electricity the households puts back on the grid, this analysis provides insight into the household's self-consumption of the electricity it produced. More detailed information about our estimation strategy can be found in Appendix 4.A.1.

Our energy data panel is unbalanced, and consists of more than 346,000 observations. This unbalance is a direct result of our rolling recruitment design, which creates variation in the date at which the smart meter of households become readable. Once the smart meter of a household becomes readable, we track their energy data until the end of the experiment. In addition, some energy observations were set to missing (less than 0.1 percent of observations) as we observed a data entry error to take place in the first data entry after the number of digits of total energy purchased or sold recorded by the smart meter changes.

[8] 7 treatment-control comparisons for 28 variables

To better understand our panel data set, we plot the number of energy observations by calendar date. Figure 4.4 shows the number of observations by calendar day for the gas and electricity data. The number of observations per calendar day increased gradually the during the recruitment period, with a large influx in households signing up for the project just before the end of June 2019, when the final recruitment rounds were concluded. The Figure demonstrates that we had a few technical errors in the data recording infrastructure, causing no data to have been collected at certain calendar dates. The largest energy data readability error occurred in the heating season of 2019/2020, from the start of January to halfway March 2020. As these errors induce missing values for all households, the errors do not affect our ability to causally interpret our treatment effect estimates.

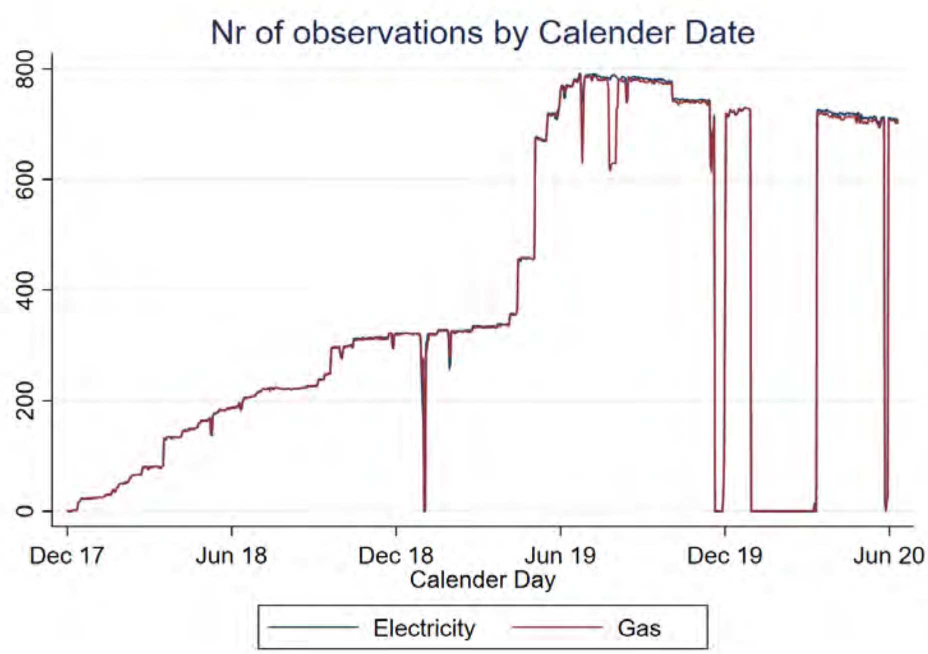


Figure 4.4: Number of observations per calendar

The actual values of daily mean gas and electricity consumption are shown in the two panels of Figure 4.5. We clearly observe seasonality in energy consumption: the larger energy consumption during the cold winter months, and the lower energy requirements in the warmer months. It is impossible to detect treatment effects in the Figures due to the high variability of consumption, and hence we need to resort to regression analysis to formally test for the presence of treatment effects.

4.2.5 Power

We determine the statistical power of our study using the power calculation procedure for two-way difference-in-difference models developed by Burlig et al. (2020). The procedure developed by Burlig et al. (2020) takes into account that the errors of these models typically are correlated over time (Bertrand et al., 2004). The relationship between this correlation and power is complex, as power can be increased by adding pre-treatment obervations that are correlated with post-treatment observations; but extra pre- or post-treatment observations provide less incremental information when they are correlated to those in the existing data-set.

We calculate power analytically using the `pc_analytic` STATA package. To this end, we estimate equation (4.1) using the control group households, and save the error terms (ϵ_{it}). We then model the auto-correlation in the error term as an autoregressive process of the first order, and calculate the Minimum Detectable Effect (MDE) given the observed variance and autoregressive coefficient observed in the error data. To calculate the MDEs for gas and electricity, we assume a conventional level of power (0.8) and the regular significance level (0.05). We perform the calculations assuming a data-set of 802 households, observed for 94 days before the start of treatment and 264 after. We pick these numbers by plotting the number of observations by day until treatment as defined in equation 4.1, and picking the minimum and maximum date at which we observe at least 80 percent of households.[9] We find the MDE to be 0.0732 m^3 for gas, and 142.62 Wh for electricity.

4.2.6 Treatment Compliance

We randomly assigned households to treatment groups by means of a random rejection design. After treatment group assignment, those households assigned to treatment were contacted to schedule display installation. However, not everyone assigned to treatment complied with their treatment status, and thus ended up owning the treatment display. We were able to (successfully) install the display with 83 percent of the treatment group households. The majority of non-installations (75 percent) were a direct result of households having lost interest and not accepting the offer. For a

[9] We therefore ignore that we observe some households for for longer times, but also ignore missing data within this interval.

minority other constraints (such as technical difficulties) prevented installation. The display installations were coordinated by randomisation batch, with the waiting time between randomisation and display installation varying across households. Half of the display installations took place within 64 days, while the some of the other half of households had to wait much longer (see Figure 4.6). Besides the waiting time, the date of installation varies within the sample. As can be seen in Figure 4.7, display installations were regularly scheduled during the recruitment period (until June 2019), with a spike in installations in the months after as then the relative large number of households randomised at the end of this period received the display.

Because of the endogeneity involved in the decision to accept the display, it may be that those who did receive the display differ systematically from those who did not. This could lead the group of households that received a display to be systematically different than the overall treatment group. To analyze whether we find evidence for households with and without display installation to differ in terms of observables, we use a probit model to regress an indicator that equals one if the display was accepted and installed, and zero otherwise, on household and residence characteristics. We control for the display roll-out procedure by including location and randomisation date fixed effects, as proxies for the waiting time until display installation. Column (1) of Table 4.4 demonstrates that only two variables are (borderline) significant. We find that single-member households have a 10 percentage point lower probability of successful display installation when we control for other characteristics (p=0.041), and that home ownership is associated with a similar sized decrease (p=0.125). In the second column we analyze whether these characteristics are collinear with other variables by dropping the non-significant observables from the equation. We now fail to find evidence that home ownership reduces the probability of a successful installation (the coefficient more than halves to -3.7 percentage point, p=0.493) while being a single household is again associated with a decrease of about 10 percent (p=0.024). In conclusion, the treated households seem to only deviate in terms of the share of single households from those who did not receive display installation. As we find no significant evidence highlighting the importance of other household sizes, we find only limited evidence for household composition to correlate with display installation, while we do not find other important predictors of energy usage (such as residence type, year built or highest education level as proxy for income (Brounen et al., 2012)) to correlate with display installation status. Based on observable characteristics, we therefore find treatment and control to be very similar.

We only find a slightly larger share of single households in the treatment group as compared to control, but we do not find reason to believe that energy consumption differs between these two groups.

4.3 Empirical strategy

4.3.1 Intention to Treat Effect

We estimate the impact of the IHD technology using two steps. As a first step, we measure the average impact of offering IHDs (without making the distinction yet of whether a household accepted it, or not) and estimate the intention-to-treat (ITT) effect. In our setting, this effect captures the impact of offering an IHD to a random subset of households that has declared interest in receiving one. Equation (4.1) shows our key specification. Here, y_{irt} denotes the amount of energy consumed by household i on day t in location r. T_{ir} is a variable that equals one if household i in location r has been assigned to the treatment group, and zero otherwise. $POST_{irt}$ is a variable that equals one for household i either from the moment onwards that the IHD had been installed successfully (for the treatment group households) or from the batch's median installation moment onwards (for the households that ended up not accepting the IHD), and zero otherwise. T_{irt} denotes the dynamic treatment indicator, and is defined as $T_{irt} = T_{ir} \times POST_{irt}$.

We include household fixed effects (α_i) as well as location-calendar day fixed effects (γ_{rt}). Finally, X_{irt} is a vector of weather controls that vary over time, and depending on the location, also within locations. The variables included in X_{irt} are daily average temperature, and daily average temperature squared and global solar radiation[10]. The error term is denoted by ϵ_{irt}, which is clustered at the 2-digit postal code (region) level if household i's location r was a location where we assigned households with an unsuitable meter to control, and household level otherwise. The regression equation is as follows:

$$y_{irt} = \beta_{ITT} T_{irt} + \alpha_{ir} + \gamma_{rt} + \delta X_{irt} + \epsilon_{irt}. \tag{4.1}$$

[10]All weather data included in X_{irt} come from the Dutch Meteorological Institute's website (KNMI). Each household is assigned the weather information as collected by the nearest weather station. For four locations all households live closest to the same weather station, creating no within location variation in these variables, while for three locations there are multiple weather stations that account for that location's weather data.

Table 4.4: Correlates of receiving a display.

	(1)	(2)
Has solar panels before randomisation	0.0454	
	(0.0384)	
Elderly hh	-0.0371	
	(0.0416)	
1 person hh	-0.101**	-0.107**
	(0.0492)	(0.0473)
>2 hh members	-0.0112	
	(0.0564)	
Household with children below 18	0.0458	
	(0.0413)	
Bachelor level education or higher	0.0578	
	(0.0405)	
Larger residence type	0.0456	
	(0.0424)	
Home owner	-0.0914+	-0.0369
	(0.0595)	(0.0538)
Residence built after 1984	0.0161	
	(0.0393)	
Observations	408	408

Dependent variable is whether a household has accepted and received the display. Only treatment households are included. Observations by household. Robust standard errors between parentheses. All coefficients are average marginal effects estimated using a probit model. All regressions include display roll-out fixed effects, by including a set of indicators for the locations, and the randomisation dates. The larger residence type refers to corner houses, semi-detached houses, and a small share of unclassified houses (farmhouses,houseboats).

$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.3.2 Treatment Effect on the Compliers

As a second step, we analyze the impact of the display on energy usage on the sample of households that accepted the display, and were able to actually receive the display. To this end, we estimate a Local Average Treatment Effect (LATE), the Average Treatment Effect on the Compliers. To estimate the LATE, we need to assume instrument exogeneity and monotonicity (Angrist et al., 1996).[11] Instrument exogeneity holds because of random treatment assignment. The monotonicity assumption holds as well, since we did not install any displays among control group households. Therefore, having been assigned to treatment makes all households (weakly) more likely to end up owning a display.

We estimate the LATE using 2SLS, and use our dynamic treatment group indicator (T_{irt}) as an instrument for successful display installation. Formally, the 2SLS model is specified as follows:

$$I_{it} = \rho T_{it} + \theta_i + \psi_t + \eta X_{it} + \tau_{irt}, \tag{4.2}$$

$$y_{irt} = \beta_{LATE} \hat{I}_{irt} + \alpha_i + \gamma_{rt} + \delta X_{irt} + \epsilon_{irt}. \tag{4.3}$$

All the variables are defined as before. The new variable, I_{it}, is an indicator that equals one if the display was accepted and installed in the home of household i at some time before time t. We first estimate the probability of the installation, using equation (4.2) and then include this variable ($\hat{I}_{irt}$) as a covariate in the second stage, shown in equation (4.3).

4.4 Empirical Results

4.4.1 Treatment effects on Energy Consumption: ITT estimates

We start with the analysis of the Intention-To-Treat effect as defined in equation (4.1). As a first step, we estimate the equation on the full data-set. Table 4.5 shows the estimation results. We find treatment households to consume 134.7 Wh less electricity (p=0.030) and -0.150 m^3 less gas (p=0.015) per day. These results translate into a 1.8 percent and 5.6 percent reduction in the consumption of, respectively, electricity and gas.

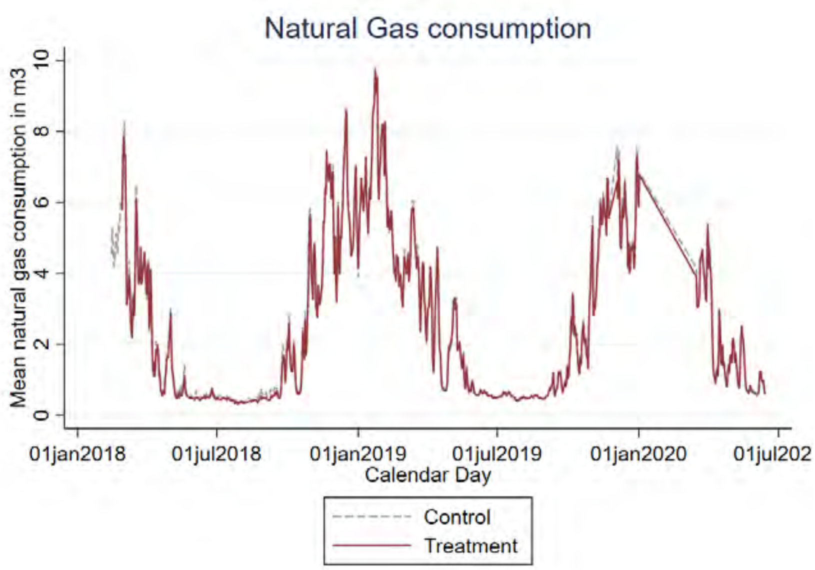
[11] Note that, strictly speaking, also the intention-to-treat effect is a local treatment effect, given that it is estimated on the subsample of households who selected into the experiment.

Table 4.5: Intention-To-Treat Estimates

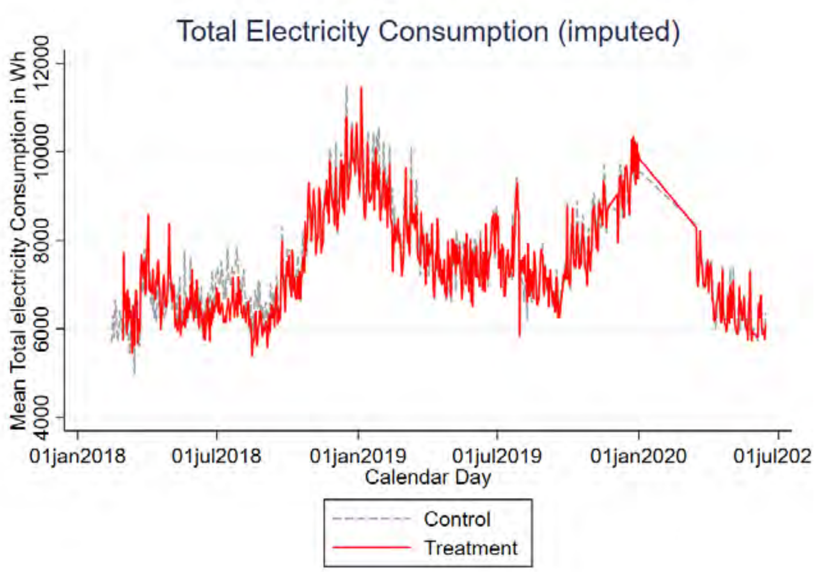
	Electricity	Gas
	(1)	(2)
Treatment	-134.7**	-0.150**
	(61.49)	(0.0613)
Control group mean	7660.794	2.665
R^2	0.652	0.755
Number of hh (N)	802	802
Observations (N $\times$ T)	346321	343375
Day $\times$ location effects	Y	Y
Weather controls	Y	Y
Household Effects	Y	Y

Dependent variables are electricity consumption in Wh and gas consumption in m^3. Observations by household and day. Standard errors between parentheses. Standard errors are clustered by region if meter type assignment was used at the location the household belongs to, and clustered on the household level otherwise. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

As a second step, we analyze heterogeneity of the treatment effect as a function of outdoor temperatures, as energy consumption fluctuates by season (see the Figure 4.5). We split the observations of a household's gas and electricity consumption into four quartiles based on the average outdoor temperature for each day. We then re-estimate equation (4.1) on each of these four subsamples. The estimated treatment effects are plotted in Figure 4.8. For gas, we find treatment effects to be significant in three of the four quartiles. The sample splits depict a clear pattern: the treatment effect is observed to be larger on colder days, and smaller, and measured with increased precision, on warmer days. For electricity, we only find the treatment effects to be significant (p=0.039) in the second quartile consisting of the moderately cold days. We do not find evidence that the electricity treatment effect varies by outdoor temperature.



(a) Gas consumption



(b) Total Electricity Consumption

Figure 4.5: Treatment and Control Energy Consumption by Calendar day

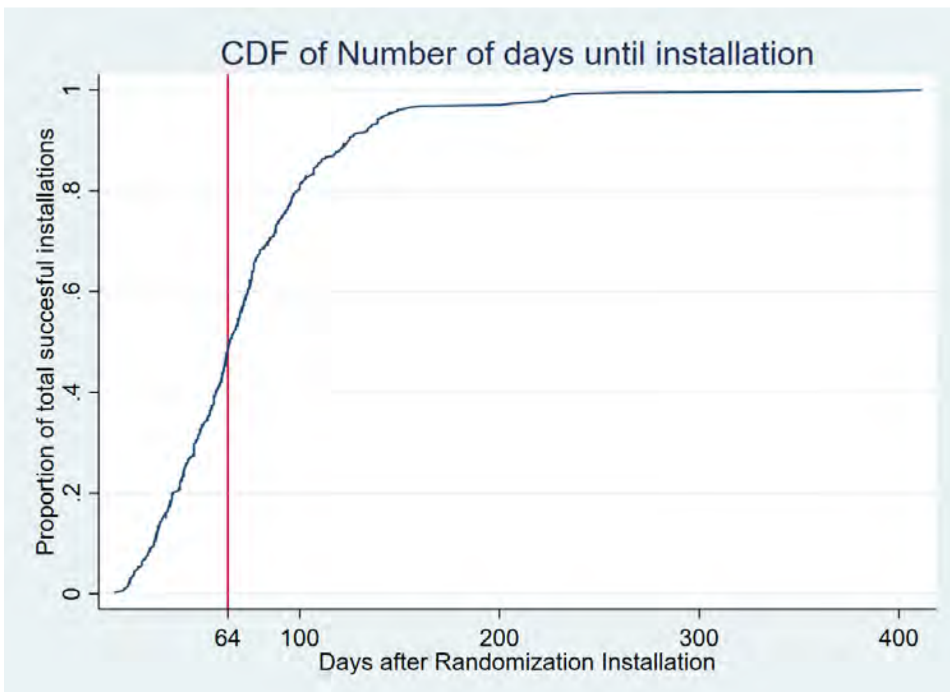


Figure 4.6: The Cumulative Distribution Function (CDF) of the Number of days after randomisation until display installation.

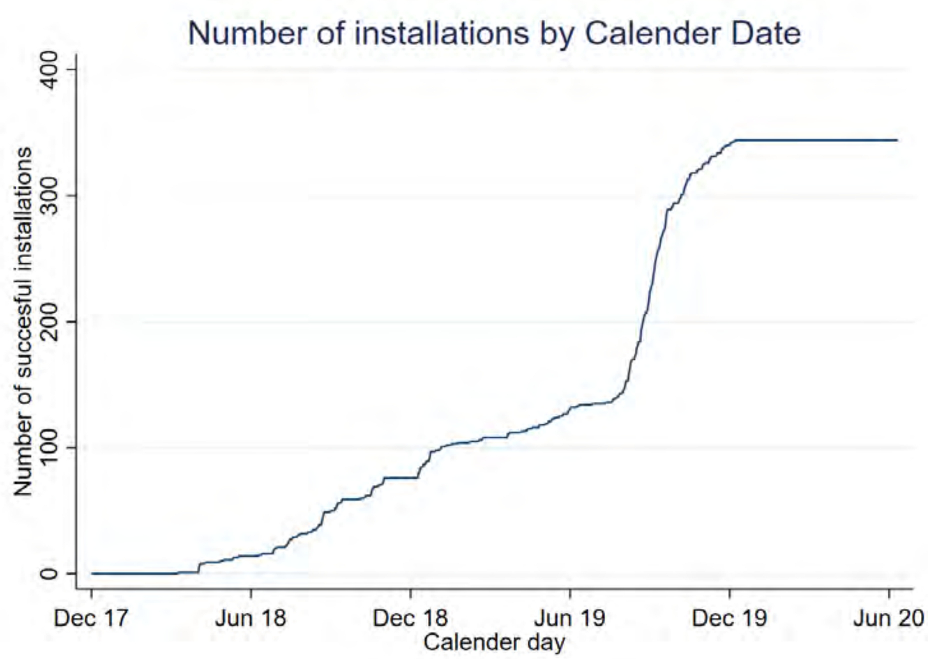
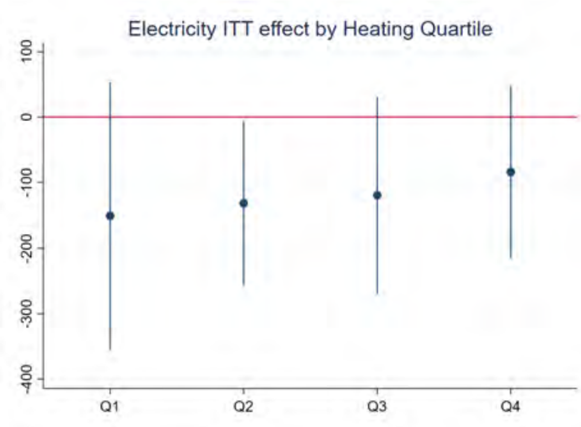
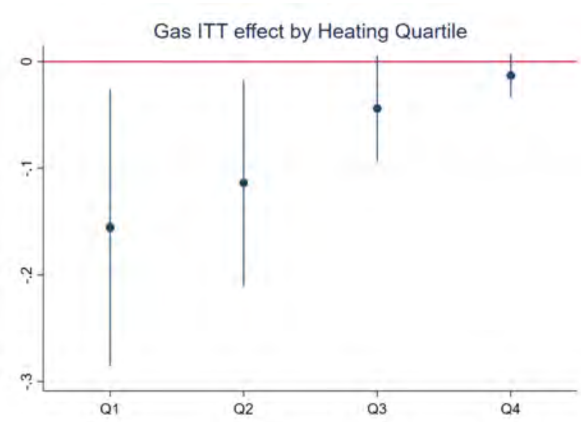


Figure 4.7: Number of installations by calendar date



(a)



(b)

Figure 4.8: Intention-To-Treat Estimates by Heating Quartile

The depicted coefficient is β_{ITT}, obtained by estimating equation (4.1) on heating quartile sample splits. The vertical line above and below the estimate depicts the 95 percent confidence interval. Q1 denotes the first quartile (the 25 percent of observations with the lowest ambient temperature) while Q4 denotes the fourth quartile (the 25 percent of observations with highest ambient temperature). The regressions that are shown in Panel (a) have gas consumption in m^3 as dependent variable, while those in Panel (b) instead use electricity consumption in Wh. All regressions include household and location $\times$ day fixed effects, and control for weather controls. Standard errors are clustered by region if meter type assignment was used at the location the household belongs to, and clustered on the household level otherwise.

4.4.2 Treatment effects on Energy Consumption: IV estimates

We now determine the effects of display installation on energy consumption for the compliers, those households that accepted the display (and for who we were able to install the device) when assigned to treatment. Table 4.6 depicts the LATE estimates obtained using equation (4.3).

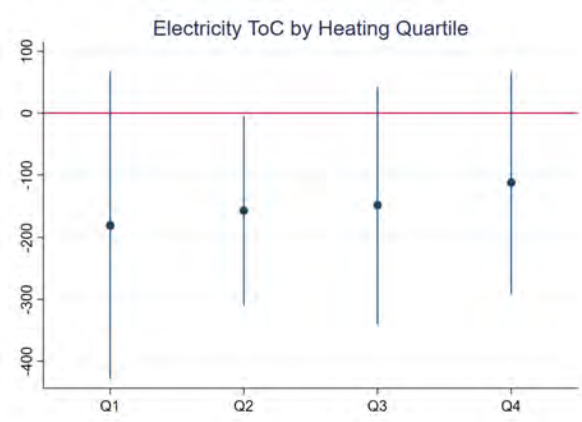
Before moving to the outcome of the equation, we analyze the relevance of our instrument by analyzing the correlation between the instrument (T_{it}) and the instrumented regressor (I_{it}). We rely on Kleibergen-Paap tests to this end, as these tests are valid when regression errors are not independent and identically distributed (i.i.d.) (Kleibergen & Paap, 2006). Both the rk LM-test and the rk Wald F-test reject the null of insufficient relevance, and thus indicate that our model does not suffer from under- nor weak-identification.

Regarding the outcome of the regressions, we find electricity consumption of those who accepted and received the display to go down by 2.2 percent (p=0.035), while their gas consumption is estimated to go down by 6.9 percent (p=0.014) compared to the control group mean. The IV estimates by outdoor temperature quartile are shown in Figure 4.10. The results are qualitatively similar to what we reported previously. As before, the energy savings tend to be higher when the temperature is lower, and this pattern is most prevalent for natural gas savings.

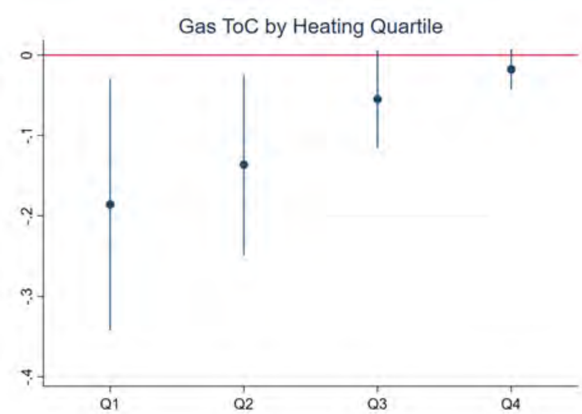
Table 4.6: IV Estimates of Treatment Effect on Compliers

	Electricity	Gas
	(1)	(2)
Instrumented Installation	-166.5**	-0.185**
	(78.23)	(0.0744)
Control group mean	7660.794	2.665
R^2	0.002	0.001
Number of hh (N)	802	802
Observations (N $\times$ T)	346321	343375
Day $\times$ location effects	Y	Y
Weather controls	Y	Y
Household Effects	Y	Y
Kleibergen-Paap rk LM-statistic (1 df)	8.793	8.821
Kleibergen-Paap rk Wald F-statistic	834.979	833.869

Dependent variables are electricity consumption in Wh and gas consumption in m^3. Observations by household and day. Standard errors between parentheses. Standard errors are clustered by region if meter type assignment was used at the location the household belongs to, and clustered on the household level otherwise. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. The Kleibergen-Paap statistics are significantly different from zero, indicating that our specification does not suffer from under- nor weak-identification.



(a)



(b)

Figure 4.10: Treatment effect On Complier (ToC) Estimates by Heating Quartile

The depicted coefficient is β_{LATE}, obtained by estimating equation (4.3) on heating quartile sample splits. Q1 denotes the first quartile (the 25 percent of observations with the lowest ambient temperature) while Q4 denotes the fourth quartile (the 25 percent of observations with highest ambient temperature). The regressions that are shown in Panel (a) have gas consumption in m^3 as dependent variable, while those in Panel (b) instead use electricity consumption in Wh. All regressions include household and location $\times$ day fixed effects, and control for weather controls. Standard errors are clustered by region if meter type assignment was used at the location the household belongs to, and clustered on the household level otherwise.

4.5 How households interact with the in-home display

We find that the display leads to significant energy consumption savings. The percentage energy savings are larger for gas than for electricity consumption, and the variation in treatment effect suggests that the gas savings materialized by a lower level of space-heating consumption. We now turn to exploring the mechanism behind these responses. Did the display help households to make better decisions by informing households of the (relative) energy intensity of different energy uses? Or did the display change behavior by educating households regarding the size of their gas and electricity bills? We conducted a survey to get a better understanding of these motivations.

We sent out the survey in four different rounds. Given the staggered treatment roll-out, this design enabled us to survey households when exposed to the feedback information for an (approximately) similar period of time. Households received the survey invitation at least ten months after they had received notification of their treatment status. At the time of filling in the survey, the vast majority (90%) of households who accepted and received a display had been exposed the the display feedback for more than half a year. Two of the four rounds took place in the non-heating season (one in September 2019, and the other in May 2020); the other two rounds took place within the heating season (in January 2020 and March 2020). Households were invited to participate by means of an e-mail, and survey participation was incentivized with a lottery: twenty gift cards of each 50 euros were raffled off. Our recruitment strategy led to a 55 percent response rate (505 respondents), with the response rate being slightly higher among treated households (61 percent) than among control households (48 percent).[12]

As we do not observe survey answers for all households, we face two challenges when analyzing the survey data. First, the household types that respond to the survey may differ across treatment arms when the decision

[12] We include the full experimental population in the survey analysis, which consists of 926 households. This sample includes the sample analyzed in the data analysis section (N=802) plus a group of households that participated in the experiment but for who we were unable to collect energy data (N=124). A balance test on this full set of households (available on request) shows that all treatment-control differences are below a normalized value of 0.25, and only 1 out of the 27 tested characteristics is significant. We therefore do not find evidence for imbalance on this broader experimental population either. We have coded incomplete survey attempts as nonresponses.

to answer the survey is related to treatment group assignment. Treatment-induced differences in response rates therefore affect the internal validity of our estimates, as it hinders causal interpretation of our estimated treatment effect. Second, when the decision to participate in the survey is non-random, the survey respondents may be unrepresentative of our experimental population. We therefore first address selection into the survey, and then discuss our empirical approach and estimation results.

4.5.1 Selection into the survey

To analyze the decision to participate in the survey, we estimate the following Linear Probability Model:

$$p_i = \mu T_i + \nu X_i + \gamma_i + \varphi_i, \tag{4.4}$$

where p_i is an indicator variable that equals one if household i filled out the survey, and zero otherwise. T_i denotes the treatment indicator, and X_i denotes a vector of household characteristics collected at baseline. We account for variation in survey response due to variation in timing of survey recruitment (and hence also in the time since display installation) by including three sets of fixed effects in vector γ_i. We include indicators for every location, randomisation date, and month at which the survey invitation was sent. As $E[p_i|T_i, X_i, \gamma_i] = P[p_i = 1|T_i, X_i, \gamma_i]$ the regression coefficients indicate the change in survey response probability associated with a certain level of a covariate. The error term is denoted by φ_i. Estimation of equation (4.4) is by using OLS.[13]

Table 4.7 presents the coefficient estimates obtained using equation (4.4). We restrict the sample to those households for which we have energy data at baseline, and include an indicator for solar capacity at baseline. Column (1) shows that households in the treatment group have a 14 percent higher probability to participate in the survey than households in the control group (p=0.000). Moreover, we find that having solar panels at baseline and two household characteristics – coming from a household where one member has completed bachelor-level education or higher, and age (having at least one elderly household member)– to be significant predictors of survey participation. In column (2), we omit the predictors that were not statistically significant to see whether these three variables are collinear with the

[13]Note that the findings are virtually unchanged when using a probit model instead. The added benefit of the Linear Probability Model is that we can interpret the interaction coefficients, for which we are unable to calculate the marginal effects in the probit model.

other characteristics. We do not find evidence for substantial collinearity, as the coefficients are virtually unchanged. Finally, we analyze whether these three main predictors reflect asymmetric non-response among treatment households. To that end, we interact the three predictors with the treatment indicator and add these terms to our estimation equation. We report the estimated coefficients in column (3). We find that the elderly household coefficient substantially loses predictive power when we allow for different selection effects by treatment arm, as the coefficient is no longer significant (p=0.552). The significant elderly in treatment interaction coefficient (p=0.080) suggests that the before-found higher response rates among elderly is primarily driven by the households with elderly members in the treatment group. For the other two characteristics we do not find evidence for a different selection process among treatment households as compared to control.[14]

We thus find evidence for non-random selection into the survey (survey respondents tend to more often have solar panels at baseline, and come from older households, where at least one member enjoyed bachelor education). As age has been found to correlate with outcomes similar to our survey questions (such as conservation practices, knowledge and attitudes (Mills & Schleich, 2012)), and households with elderly members in treatment seem to respond more frequently than their control group counterparts, we control for this variable in our survey analysis to obtain causal treatment effects on the survey respondent population. We explain our identification strategy in more detail in the next section.

4.5.2 Survey evidence

To analyze the effect of the display on survey outcomes, we proceed as follows. We assume that the higher response rate in the treatment arm is not correlated with the potential outcomes of the survey questions once we control for the household having at least one elderly member.[15] Under this assumption, we can estimate the average treatment effect on the survey

[14]We corroborate these results when we analyze selection on the broader population of 925 households for which we have baseline characteristics, irrespective of whether baseline energy data was available. For this sample we do not have information on baseline ownership of PV, and hence we drop this variable (and its treatment interaction) from the analysis. Only the elderly-in-treatment interaction coefficient turns insignificant in this specification ($p = 0.220$).

[15]In the terminology of Gerber and Green (2012), we assume that missingness (not filling in the survey) is independent of the potential survey outcomes.

Table 4.7: Analysis of the decision to respond to the survey invite.

	(1)	(2)	(3)
Treatment group	0.142***	0.141***	0.134**
	(0.0345)	(0.0343)	(0.0599)
Has solar panels before randomisation	0.0884**	0.0896**	0.127**
	(0.0382)	(0.0371)	(0.0535)
Elderly hh	0.110**	0.111***	0.0359
	(0.0427)	(0.0384)	(0.0604)
1 person hh	0.0216		
	(0.0420)		
>2 hh members	-0.0172		
	(0.0537)		
Household with children below 18	0.00540		
	(0.0422)		
Bachelor level education or higher	0.132***	0.127***	0.122**
	(0.0390)	(0.0378)	(0.0531)
Larger residence type	0.0468		
	(0.0410)		
Residence own property	-0.0380		
	(0.0625)		
Residence built after 1984	-0.0101		
	(0.0429)		
Treatment group × Elderly hh			0.136*
			(0.0776)
Treatment group × Has solar panels before randomisation			-0.0707
			(0.0684)
Treatment group × Bachelor level education or higher			0.00142
			(0.0686)
Constant	0.419**	0.405**	0.388**
	(0.195)	(0.188)	(0.187)
N	827	827	827
R^2	0.0896	0.0873	0.0918

Observations by household. Robust standard errors between parentheses. All regressions include indicators for every location, randomisation date, and month at which the survey invitation was sent. The larger residence type refers to corner houses, semi-detached houses, and a small share of unclassified houses (farmhouses,houseboats). This specification includes both households included in the main analysis, and those who were not included due to energy data constraints (leading to N=926). We exclude 99 households for which we do not know whether solar panels were available before randomisation.

$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

respondents as follows:

$$y_i = \alpha + \beta T_i + \gamma Elderly_i + \gamma_i + \epsilon_i. \tag{4.5}$$

where y_i denotes our outcome measure, and $Elderly_i$ is a covariate which equals one if household i includes an elderly member and is zero otherwise. T_i denotes the treatment group, while ϵ_i denotes the error term. In this case, β is the Intention-To-Treat effect on the survey respondent population, with 92% of treatment respondents actually having a display installed. We estimate robust standard errors in all specifications.

How energy was saved

We start with analyzing what behaviors and investments have contributed to the realized energy savings. To this end, we asked household to self-report whether they have undertaken nine types of daily energy saving behaviors and 13 types of energy investments. The survey questions can be found in Appendix 4.A.2, and draw on questions used in Starke (2014); Starke et al. (2020). We count number of behaviors for which a household indicates that they undertake them frequently, and the number of energy efficiency investments households report to have undertaken. We refer to these two sums as the 'habit score' and 'investment score', respectively. Table 4.8 shows the results. Based on the control group answers, we find that households take quite some energy saving measures; they report to engage in, on average, six out of the nine energy conservation behaviors, and to have made eight out of the 13 energy saving investments. We do not find either of the two scores to be statistically significantly higher for the treatment group- not for the overall scores (see Table 4.8), but also not for the underlying items, nor when using an alternative scale that involves more lenient coding of the items (available on request).[16] The positive coefficients on the habit-score suggests, if anything, an intensification of energy habits (such as lowering thermostat settings, taking shorter showers, and removing icing from freezer compartments). Moreover, the coefficient on the investments scale suggests, if anything, a decrease in the number of isolation measures taken.

[16] In the latter scale we broaden our definition of energy conservation behavior to also encompass irregular habits and intentions to undertake a certain behavior and/or investment. The intentions could have explanatory power of the investment behavior in the period after the survey was administered.

Table 4.8: Treatment estimates on reported energy saving behavior

	(1) Habit Score	(2) Investment Score
Treatment group	0.107	-0.084
	(0.160)	(0.189)
Observations	505	505
R^2	0.055	0.086
Control group mean	6.038	7.598

Observations by household. Robust standard errors between parentheses. The habit score and investment score are calculated as the sum of respectively 9 or 13 items. Household receive 1 point per item if they report to always or frequently do the behavior (habit) or have undertaken the investment (investment). The items are listed in Appendix 4.A.2. All specifications control for being an elderly household, and include display roll-out and survey recruitment fixed effects, by including a set of indicators for the locations, the randomisation dates, and the month for which the survey was sent.

$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Information effects on expected energy savings

Next, we turn to the behavioral determinants of the estimated treatment effect. In this section we start by discussing the influence of the display on household judgement of the energy costs. The display may have improved the salience of the households' energy cost and so function like a reminder. But the device may also facilitate learning about the energy intensity of energy uses, or the size of the energy bill. Before we analyze these factors and turn to our results, we first describe the household interaction with the device among those that had an display installed.[17]

Did households regularly interact with the display? And did they monitor their real-time energy usage or did they mostly make use of the daily budget interface (see Figure 4.3)? The survey answers suggest that the display was successful in increasing the salience of energy consumption. Most of the respondents reported that they kept the display in a central place of their home (such as the living room (69%), or the kitchen (16%)), and a large majority of respondents state to have regularly looked at the display information (more than 85%). Yet, the interaction frequency did go down over time. In the first weeks after installation more than 80% of households looked more than five times a week; in the week preceding the survey this had gone down to 52%. Only 15% of households reported to have ceased looking at the display altogether. We therefore find little evidence for disuse, but also found little evidence that households disliked the display or questioned its use.[18] Households mostly use the direct feedback functionality regularly (as reported by three out of five households), while only a minority reports to monitor their daily budget on a regular basis (24 percent). This results suggest that households mostly use the device to make real-time energy consumption salient and are less interested in the functionality framing the cost feedback with regard to a daily budget – at least at the time the survey was administered, at which most had the display for

[17]For this descriptive analysis we ignore 5 households that claimed to not have received an display. We have verified these installations with project partners and have found no evidence that our installation data were inaccurate. We therefore label these responses as respondent mistakes. Also among households that were noted to not have received a display we found conflicting survey answers; overall, however the amount of mistakes is minor (3.6 percent of answers).

[18]Only a minority of 12 percent of households declares to not find the feedback useful, while only 5 percent reported to not find it pleasurable. Therefore, we do not find evidence that the display was an experience good in the terminology of Nelson (1970); a good whose attributes are difficult to assess ex-ante. The households in our experiment were interested in receiving the device, and the vast majority is still actively engaging with the display over half a year post installation.

Table 4.9: Treatment estimates on household estimated energy charges.

	Electricity		Gas	
	(1)	(2)	(3)	(4)
	Estimated Bill	Confidence estimate range 0-4 standardized	Estimated Bill	Confidence estimate range 0-4 standardized
Treatment group	-2.262	0.149^{+}	9.486^{**}	0.103
	(3.969)	(0.094)	(4.530)	(0.094)
Observations	505	505	505	505
R^2	0.064	0.049	0.100	0.030
Control group mean (Non standardized)	46.331	2.108	58.447	2.090

Observations by household. Robust standard errors between parentheses. All specifications control for being an elderly household, and include display roll-out and survey recruitment fixed effects, by including a set of indicators for the locations, the randomisation dates, and the month for which the survey was sent. The confidence question asked households how certain their were of their bill estimate on a five-point scale; with 0 indicating very uncertain, and 4 indicating very certain.
$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

over 10 months.

To investigate whether the display made households better informed energy consumers, we analyze two aspects. First, we analyze whether the display led households to be better informed of their monthly energy charges. We ask households to estimate their (average) monthly gas and electricity payment, and consecutively ask them to indicate how confident they are in their estimates on a scale from 0 (not very certain) to 4 (very certain).[19] Table 4.9 shows the results. For electricity, we do not find convincing evidence that households became more knowledgeable on their monthly charge. We do not find evidence that the estimated amount changed (p=0.569), and only find borderline significant (p=0.112) evidence that households became more confident in their bill estimate. For gas, we do find evidence that households became better informed of their bill. We find significantly higher bill estimates in the treatment group (difference of about 10 Euros, p=0.037), closer to the bill estimate expected by their yearly gas usage.[20] However, we find no evidence for households to be more confident in their gas bill estimate (p=0.273).

Second, we analyze whether the display enabled households to better judge the energy intensity of various technologies and behaviors. Following Attari et al. (2010), we ask households to evaluate the intensity of energy use of a measures with respect to the energy use of other measures.[21] Specifi-

[19] These questions build on the cost awareness measure by Brounen et al. (2013).

[20] The yearly usage among survey respondents is about 987 m^3, this should correspond to a monthly charge of about 86.55 Euros (NIBUD, 2021)

[21] Attari et al. (2010) elicit energy beliefs by asking households to report the (electrical) usage of different measures in terms of light bulbs.

Table 4.10: Treatment estimates on Use by Category Score.

	(1) Use by Category Score range 0-30 standardized
Treatment group	-0.192**
	(0.091)
Observations	505
R^2	0.111
Control group mean (non-standardized)	21.509

The items are listed in Appendix 4.A.2. Observations by household. Robust standard errors between parentheses. All specifications control for being an elderly household, and include display roll-out and survey recruitment fixed effects, by including a set of indicators for the locations, the randomisation dates, and the month for which the survey was sent.
$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

cally, we asked households to rank a number of fairly broad energy consumption categories (heating, lightening, cooking etc.) by their energy intensity. We used this ranking to calculate a Energy Category Score, where every category placed in the correct position provides households with 5 points. For every position the answer deviated from the correct one, a point is subtracted. We sum the scores over the six items, which creates an overall score ranging from 0 to 30. The average score is about 22 points, and we standardize the scores (by substracting the group mean, and dividing by the standard deviation) to facilitate interpretation. The results are shown in Table 4.10. We do not find evidence that households in the treatment group are more knowledgeable on the energy intensity of these categories. In fact, we find evidence of the opposite, as households in the treatment group have a score that is 0.192 standard deviation lower than that of the control group (p=0.035).

To better understand the cause of the lower Use by Category Score, we re-estimate the treatment effect using the ranked position of the items as dependent variable. Table 4.11 demonstrates that treated households on average rank heating 0.321 positions lower than control households (p=0.010), while ranking cooking 0.358 (p=0.006) positions higher in the energy intensity ranking. In other words, treatment households perceive heating to be

Table 4.11: Treatment estimates on rank of items of the Use by Category Score.

	Use by Category Ranking					
	(1)	(2)	(3)	(4)	(5)	(6)
	Bathing and Showering	Heating	Cooking	Lightening	White goods	Brown goods
Treatment group	-0.020	0.321**	-0.358***	0.110	-0.039	-0.014
	(0.134)	(0.125)	(0.129)	(0.108)	(0.113)	(0.133)
Observations	505	505	505	505	505	505
R^2	0.089	0.114	0.065	0.064	0.044	0.057
Control group mean	3.726	1.929	3.571	5.108	2.462	4.203

Observations by household. Robust standard errors between parentheses. A value of 1 means that the household ranked the item as most energy intensive, while a ranking of 6 means the opposite. The questions are listed in Appendix 4.A.2. All specifications control for being an elderly household, and include display roll-out and survey recruitment fixed effects, by including a set of indicators for the locations, the randomisation dates, and the month for which the survey was sent.
$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

less energy intensive compared to control households, and perceive cooking to be relatively more energy intensive. If these beliefs drive energy conservation, we would expect households to mainly conserve on cooking-related energy consumption and not on heating. Yet, as we mostly find gas savings to materialize on cold days, these learning effects do not seem to be driving our natural gas savings results.

So far we find evidence that the experimental households enjoyed the display device. We find disuse to be of minor concern, and document positive evaluations of the device. Moreover, we find that the display made households better informed of the importance of the gas bill in total energy costs. However, we do not find the device to highlight the importance of space heating in particular. In contrast, we found evidence for households to update energy intensity beliefs of heating and cooking in the wrong direction.[22] For electricity, the results are inconclusive. We did not find clear evidence for households to become better informed of the size of the energy bill, nor for learning about the electricity use of appliances. However, due to the broad range of appliances driving electricity usage, we cannot exclude the electricity energy savings to be caused by reductions on finer margins than we studied, such as learning of the energy intensity of specific electronic devices.

[22]This may relate to the location of the display within the household home, as 16 percent of treated households reported to have positioned the device in the kitchen.

Motivational behavioral determinants

So far, we have analyzed whether the display made households better informed, and concluded that the device informed households of the importance of gas costs in total energy costs. Given that households underestimate the gas bill, this better knowledge on gas costs helps households make the trade-off between the benefits and costs associated with energy conservation. We therefore expect welfare to increase, as households make energy decisions, such as investments or the changing of energy use behaviors, of which they expect the financial benefits to cover the (financial) costs. However, an important assumption of this positive welfare effect is that other non-monetary costs are absent. In this subsection we explore whether the display impacted household utility through other channels. Recent behavioral economic research highlights that behavioral interventions may directly impact utility by introducing annoyance costs (Allcott & Kessler, 2019), or a moral cost by appealing to prosocial preferences (Brandon et al., 2017).

As we found disuse to be minor and evaluations of the display to be positive, see section 4.5.2, we do not have reason to believe the annoyance costs to be of importance. The device is easy to unplug, and thus households do not need to be exposed to the feedback if they do not want to. However, the influence on the pro-social motive may be of concern. To analyze the influence of the display on the motive to conserve, we analyze the effects of the display on two types of attitudes. First, we analyze whether the interaction with the display has affected attitudes towards energy conservation. If the households enjoy energy conservation more because of the device, perceive energy conservation as more profitable, or find it easier to conserve energy, they may evaluate energy conservation more positively. Second, we analyze whether the device affects household attitudes towards the Dutch 'energy transition', by asking households about the desirability of greener energy sources and the need for governmental policy in the energy domain. Table 4.12 shows that we fail to find evidence for households attitudes towards energy conservation, or the energy transition to have changed. Digging deeper into the survey responses, we find that households tend to favorably evaluate the energy transition as evidenced by the high control group mean (score 4 out of 5), which reduces the scope of improvement for this variable. The lower control group mean indicates that ceiling effects are less likely to play a role with respect to the household attitude towards energy conservation (control group mean 1.5 out of 5). Yet, we do not find evidence for this attitude to have been moved by the display either.

We therefore conclude that an 'annoyance cost' is absent, and that the

Table 4.12: Treatment estimates on Attitude measures

	Attitudes	
	(1) Energy Conservation mean(std items)	(2) Energy Transition mean(std items)
Treatment group	-0.019	0.057
	(0.060)	(0.067)
Observations	505	505
R^2	0.067	0.079
Control group mean (Non standardized items)	1.458	4.181

Observations by household. Robust standard errors between parentheses. Dependent variables are the mean of the standardized survey items. The items are listed in Appendix 4.A.2. All specifications control for being an elderly household, and include display roll-out and survey recruitment fixed effects, by including a set of indicators for the locations, the randomisation dates, and the month for which the survey was sent.
$^{+}$ $p < 0.125$, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

display did not change the (prosocial) motive to conserve energy. As a result, households need to become better off after receiving the display. Either the purchasing of new energy-saving technologies, or the lowering of energy consumption, are choices made because the household is better informed of the energy costs or energy intensity of uses. We therefore expect consumer welfare to increase. As the reduction in energy consumption decreases carbon-dioxide externalities, we also expect public welfare to increase.

4.6 Conclusion

In this paper, we implemented an RCT to analyze the impact of real-time disaggregated feedback on energy consumption. The feedback was provided by means of an energy in-home display. We recruited households interested in receiving a display that provides insight into metered natural gas (reflecting predominantly heating usage) and electricity consumption (reflecting appliances usage) in seven locations in the Netherlands. We randomly offered such a display to half of the applications; the other half did not receive such offer. We monitored all households' energy consumption for a minimum of 7 months after the display (would have been) installed. After treatment households had been exposed to the device for about half a year, we collected survey evidence on the treatment household interaction with the device and the degree to which treated and control households made mistakes in their judgement of energy costs. Our results show that the energy display leads to gas savings of 6.9 percent, and electricity savings of about 2.2 percent.

We find evidence that households reduced gas consumption due to being better informed of the relatively high gas expenditure. The survey answers indicate that households frequently consulted the device for information on their energy usage, which suggests that sustained cost salience may have played an important role in reducing energy consumption.

Our results demonstrate that this type of feedback mitigates consumer mistakes, which suggests that the display helps households align actual energy consumption with their privately optimal consumption. These findings contribute to the recent literature that stresses the importance of concrete and relevant information to optimize the effect of energy consumption feedback. This may involve providing feedback in a real-time manner (Darby, 2006) but also including concrete energy saving information advice (see e.g. Abrahamse et al., 2005; Gerster & Andor, 2020; Schleich et al., 2017). Our findings suggest that having energy monitors providing more detailed feedback can be effective application of these insights. As a result, these monitors may be an effective complementary measure to policies that stimulate the take-up of smart energy meters.

Though we find evidence that households became better informed (regarding gas costs), we are unable to provide causal evidence regarding the degree to which households depend on continued interaction with the device to remain informed (salience effects), or whether they internalized the information (learning effects). Further research designed to decouple the impact of these channels could provide further insight in these mechanisms that drive the changes in energy decision making. By blocking either one of these two channels, this research can provide further insight into the need for continued feedback exposure. See Jessoe and Rapson (2014); Lynham et al. (2016) for two studies that apply these insights in their experimental design.

4.7 References

Abadie, A., & Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, *29*(1), 1–11.

Abrahamse, W., Steg, L., Vlek, C., & Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, *25*(3), 273–291.

Allcott, H. (2011). Consumers' perceptions and misperceptions of energy costs. *American Economic Review: Papers & Proceedings*, *101*(3), 98–104.

Allcott, H. (2016). Paternalism and energy efficiency: An overview. *Annual Review of Economics*, *8*(1), 145–176.

Allcott, H., & Greenstone, M. (2012). Is there an energy efficiency gap? *Journal of Economic Perspectives*, *26*(1), 3–28.

Allcott, H., & Kessler, J. B. (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics*, *11*(1), 236–276.

Allcott, H., Mullainathan, S., & Taubinsky, D. (2014). Energy policy with externalities and internalities. *Journal of Public Economics*, *112*, 72–88.

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003–3037.

Andor, M. A., & Fels, K. M. (2018). Behavioral economics and energy conservation – a systematic review of non-price interventions and their causal effects. *Ecological Economics*, *148*, 178–210.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Asensio, O. I., & Delmas, M. A. (2015). Nonprice incentives and energy conservation. *Proceedings of the National Academy of Sciences*.

Attari, S. Z., DeKay, M. L., Davidson, C. I., & Bruine de Bruin, W. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences*, *107*(37), 16054–16059.

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

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of*

Economics, *119*(1), 249–275.

Brandon, A., Ferraro, P. J., List, J. A., Metcalfe, R. D., Price, M. K., & Rundhammer, F. (2017). *Do the effects of social nudges persist? Theory and evidence from 38 natural field experiments* (NBER Working Paper Series No. 23277). National Bureau of Economic Research.

Brent, D. A., & Ward, M. B. (2018). Energy efficiency and financial literacy. *Journal of Environmental Economics and Management*, *90*, 181–216.

Brounen, D., Kok, N., & Quigley, J. M. (2012). Residential energy use and conservation: Economics and demographics. *European Economic Review*, *56*(5), 931–945.

Brounen, D., Kok, N., & Quigley, J. M. (2013). Energy literacy, awareness, and conservation behavior of residential households. *Energy Economics*, *38*, 42–50.

Buchanan, K., Russo, R., & Anderson, B. (2015). The question of energy reduction: The problem(s) with feedback. *Energy Policy*, *77*, 89–96.

Burlig, F., Preonas, L., & Woerman, M. (2020). Panel data and experimental design. *Journal of Development Economics*, *144*, 102458.

CER. (2011). *Gas customer behaviour trial findings report* (Information Paper No. CER11180a). Commission for Energy Regulation (CER).

Darby, S. (2006). *The effectiveness of feedback on energy consumption. A review for DEFRA of the literature on metering, billing and direct displays* (Tech. Rep.). Environmental Change Institute.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, *47*(2), 315–372.

Delmas, M. A., Fischlein, M., & Asensio, O. I. (2013). Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012. *Energy Policy*, *61*, 729–739.

DIME Analytics. (2019). *IETOOLKIT: Stata module providing commands specially developed for impact evaluations.*

Directive (EU) 2019/944. (2019). Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU. *Official Journal of the European Union*.

Ehrhardt-martinez, K., Donnelly, K. A., & Laitner, J. A. (2010). *Advanced metering initiatives and residential feedback programs: A meta-review for household electricity-saving opportunities* (Report No. E105). Washington D.C: ACEEE.

Faruqui, A., Sergici, S., & Sharif, A. (2010). The impact of informational feedback on energy consumption - A survey of the experimental evidence. *Energy*, *35*(4), 1598–1608.

Fischer, C. (2008). Feedback on household electricity consumption: A tool for saving energy? *Energy Efficiency*, *1*(1), 79–104.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: design, analysis, and interpretation.* New York: W.W. Norton.

Gerster, A., & Andor, M. A. (2020). *Disaggregate consumption feedback and energy conservation.*

Gillingham, K., & Palmer, K. (2014). Bridging the energy efficiency gap: Policy insights from economic theory and empirical evidence. *Review of Environmental Economics and Policy*, *8*(1), 18–38.

Harold, J., Lyons, S., & Cullinan, J. (2015). The determinants of residential gas demand in ireland. *Energy Economics*, *51*, 475–483.

Harold, J., Lyons, S., & Cullinan, J. (2018). Heterogeneity and persistence in the effect of demand side management stimuli on residential gas consumption. *Energy Economics*, *73*, 135–145.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference: For statistics, social, and biomedical sciences an introduction.* Cambridge: Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Ito, K. (2014). Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing. *American Economic Review*, *104*(2), 537–563.

Ito, K., Ida, T., & Tanaka, M. (2018). Moral suasion and economic incentives: Field experimental evidence from energy demand. *American Economic Journal: Economic Policy*, *10*(1), 240–267.

Jessoe, K., & Rapson, D. (2014). Knowledge is (less) power: Experimental evidence from residential energy use. *American Economic Review*, *104*(4), 1417–1438.

Karlin, B., Zinger, J. F., & Ford, R. (2015). The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin*, *141*(6), 1205–1227.

Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, *133*(1), 97–126.

Koszegi, B., & Szeidl, A. (2013). A model of focusing in economic choice. *Quarterly Journal of Economics*, *128*(1), 53–104.

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, *14*(4), 439–457.

Luthander, R., Widén, J., Nilsson, D., & Palm, J. (2015). Photovoltaic self-consumption in buildings: A review. *Applied Energy*, *142*, 80–94.

Lynham, J., Nitta, K., Saijo, T., & Tarui, N. (2016). Why does real-time information reduce energy consumption? *Energy Economics*, *54*, 173–181.

McKerracher, C., & Torriti, J. (2013). Energy consumption feedback in perspective: Integrating Australian data to meta-analyses on in-home displays. *Energy Efficiency*, *6*(2), 387–405.

Mills, B., & Schleich, J. (2012). Residential energy-efficient technology adoption, energy conservation, knowledge, and attitudes: An analysis of european countries. *Energy Policy*, *49*, 616–628.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311–329.

NIBUD. (2021). *Energie en water.* Retrieved 2021-03-29, from `https://www.nibud.nl/consumenten/energie-en-water/`

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*(1), 223–251.

PBL. (2020). *Klimaat- en energieverkenning 2020* (Tech. Rep.). Den Haag: Planbureau voor de Leefomgeving.

Price, M. K. (2014). Using field experiments to address environmental externalities and resource scarcity: Major lessons learned and new directions for future research. *Oxford Review of Economic Policy*, *30*(4), 621–638.

Schleich, J., Faure, C., & Klobasa, M. (2017). Persistence of the effects of providing feedback alongside smart metering devices on household electricity demand. *Energy Policy*, *107*, 225–233.

Schultz, P. W., Estrada, M., Schmitt, J., Sokoloski, R., & Silva-Send, N. (2015). Using in-home displays to provide smart meter feedback about household electricity consumption: A randomized control trial comparing kilowatts, cost, and social norms. *Energy*, *90*, 351–358.

Sexton, S. (2015). Automatic bill payment and salience effects: Evidence from electricity consumption. *Review of Economics and Statistics*, *97*(2), 229–241.

Starke, A. D. (2014). *With a little help from my friends: Investigating the effectiveness of rasch-based energy recommendations with social endorsements* (Master Thesis). Eindhoven University of Technology.

Starke, A. D., Willemsen, M. C., & Snijders, C. C. (2020). Beyond "one-size-fits-all" platforms: Applying campbell's paradigm to test personalized energy advice in the netherlands. *Energy Research and Social Science*, *59*(October 2019), 101311.

Tedenvall, M., & Mundaca, L. (2016). Behaviour, context and electricity use: Exploring the effects of real- time feedback in the swedish residential sector. *39th IAEE International Conference 'Energy: Expectations and Uncertainty'*(June).

Tietenberg, T. (2009). Reflections - energy efficiency policy: Pipe dream or pipeline to the future? *Review of Environmental Economics and Policy*, *3*(2), 304–320.

van Gerwen, R., Koenis, F., Schrijner, M., & Widdershoven, G. (2010). *Intelligente meters in Nederland: Herziene financiële analyse en adviezen voor beleid* (Tech. Rep.). Arnhem: KEMA.

Vringer, K., & Dassen, T. (2016). *De slimme meter, uitgelezen energie(k)?* (Tech. Rep.). Den Haag: PBL.

Warrink, A., & Vringer, K. (2021). Eigen consumptie pv-installaties van huishoudens - een bepaling op basis van slimme meter gegevens en zon instralingsgegevens.

Westskog, H., Winther, T., & Saele, H. (2015). The effects of in-home displays-revisiting the context. *Sustainability*, *7*, 5431–5451.

4.A Appendix

4.A.1 Electricity Consumption Data

Importantly, our smart meter data does not register (total) energy production, but only energy purchased and sold to the grid. And hence we do not know actual electricity consumption – the sum of electricity purchased and the amount of own solar production that the household consumed. As households do not produce gas, the amount of gas purchased coincides with the amount of energy consumed. However, to analyze effects on electricity consumption, we need to rely on an estimated measure of the own production consumed to infer total electricity consumption.

We estimate the amount of electricity production consumed by analyzing the hourly amount of electricity sold for households with solar capacity. We assume that for every household with solar panels in our data set, there is an hour for which the solar panels produced at maximum capacity and consumption on top of the base load was negligible. This hour may for instance be lunchtime during a particularly sunny day in the summer when none of the household members were at home. We can identify this hour in our data as the time slot at which the amount of production sold was at its maximum level. As we assume there to be no electricity consumption on top of the base load during this hour, we can calculate the underlying amount of electricity production by calculating the base load and adding this amount to the observed amount of electricity sold. We calculate the base load as the average electricity consumption between 3:00 and 5:00 a.m. at night.

Next, we use the maximum solar capacity estimates to uncover the rate at which a household's solar panels transform solar radiation into electricity, and use this so-called 'system production factor' to infer solar production during other time slots. Specifically, we calculate the system production factor by dividing the maximum solar capacity estimates by the solar radiation levels measured during these time slots by the weather station closest to each household home (based on data of the Dutch Meteorological Institute's website). We use this factor to estimate solar production during the other hours in our data by multiplying the term with the corresponding hourly solar radiation levels. Based on these estimates of hourly electricity production, we infer the amount of production consumed by subtracting the amounts of hourly electricity production sold from our production estimates. We find that households on average consume 31 percent of their total electricity production, which is line with the estimates by Luthander et al. (2015) who find the share of solar production consumed to lie between

17 and 38 percent among German residents. More information on the calculation of solar production consumed and the distribution of the base load estimates can be found in Warrink and Vringer (2021).

4.A.2 Survey

Reported Behavior

We asked the respondents whether they engage in the following nine energy saving habits on a 5-point scale (always, often, sometimes, never but having the intention to do so, never and no intention). Together these items were were aggregated into a habit score. The items of the habit score are:

1. Air-drying the laundry outside or on a line
2. Turning lighting off when nobody is in the room
3. Lowering the indoor temperature when nobody is at home
4. Cooking with a lid on the pans
5. Lowering the indoor temperature with one degree
6. Having short showers
7. Letting clothes air instead of washing them
8. Defrosting the fridge regularly
9. Disconnecting appliances instead of putting them on standby

Next, we asked the respondents whether they have invested in the following thirteen energy saving measures in the last 6 months on a 5-point scale (yes, no- measure was already taken, no- but I intend to undertake the investment, no- and have no intention to undertake the investment, I have no idea). As some investments may not be possible for households living in rental residences, the option 'Not applicable' was also included. The measures questioned for the investment score were:

1. Energy-Saving light bulbs
2. Insulation glass
3. Isolated walls

4. Weather strips on doors
5. LED lightning
6. Floor isolation
7. Water saving shower
8. Weather strips on windows
9. Isolation foil in radiators
10. Solar panels
11. Floor heating
12. Heat-pump
13. Replacing energy inefficient appliances

Learning

For the Use by Category Score, we asked the respondents to order six applications in intensity of energy use. The six applications were, in order of importance:[23]

1. Heating the dwelling (4780 kWh)
2. Showering and bathing (1230 kWh)
3. Kitchen en washing appliances - "White goods" (1002 kWh)
4. Household electronics - "Brown goods" (742 kWh),
5. Lightning (704 kWh) and
6. Cooking (400 kWh).

[23] We calculated the order of importance of the six kinds of household energy use as follows: An average household in the Netherlands used in 2015 about 1600 m^2 natural gas, used for heating the dwelling (78%), heating water (18%) and cooking (4%) (Milieucentraal, 2015). The average use for household appliances comes from Mileucentraal (2010). To add up natural gas and electricity we based our calculation of the primary energy requirement; One m^3 natural gas is equivalent to about 3.6 kWh electricity because one m^3 natural gas requires 31 MJ primary energy and 1 kWh electricity requires about 8.5 MJ primary energy.

To calculate how correct the respondents did answer this question, we gave them 5 points for every application in the right position, 4 if they did place the application a position too high or too low, 3 for a deviation of 2 positions, etc.

Attitudes

To compute an index for the household attitude towards energy conservation, we asked the respondents on a 7-point Likert scale (I totally not agree - I totally agree) about their opinion about six statements: Saving energy...

1. ... saves me money
2. ... contributes to a cleaner environment
3. ... is normal
4. ... is a hassle
5. ... is fun
6. ... costs more (money, trouble) than its yield

To determine the household attitude towards the energy transition, we asked the respondents on a 7-point Likert scale (I totally not agree - I totally agree) about their opinion on the energy transition. The index involves six statements:

1. To keep the earth habitable, we can no longer use gas, coal and oil
2. Windmills are needed to combat climate change
3. Climate protection may cost money
4. It is necessary that windmills and solar panels are installed in our landscape
5. It is a good idea that the government plans to decouple the Netherlands from the natural gas network
6. Solar panels on my own roof are needed to combat climate change

Chapter 5

Concluding remarks and policy implications

In this thesis, I analyzed the impact of behavioral interventions on household behaviors aimed at improving environmental outcomes. I found that persuasive appeal and social feedback treatments were effective in improving organic waste sorting in both the short and longer run. A flyer including a social model was also effective during the 9 months after the treatment ended, but its effect seemed less persistent. In addition, real-time disaggregated energy feedback was found to reduce household consumption of both electricity and gas. Evidence for the causal impact of these interventions was provided by means of field experiments.

In this final chapter, I take a step back and evaluate the contributions made by this thesis. First, I discuss how the findings contribute to the literature on the effectiveness of behavioral interventions, and especially on their impact on habitual behaviors. Second, I analyze the relevance of the interventions studied for environmental policy. I examine whether the results are expected to generalize to other settings and whether the treatments are 'scalable' (that is, remain equally efficient from a cost-benefit perspective). I conclude with a brief discussion of how field experiments can be used to resolve uncertainty regarding the net benefits of policies, and how field experiments can support evidence-based policy making.

5.1 Contribution to literature on behavioral interventions

The findings of this thesis highlight the potential of behavioral interventions in fostering improved resource conservation, also in the longer term. Chapter 2 until 4 find persistent effects of behavioral interventions, in domains where households are able to use technology fixes to conserve resources (energy consumption), but also in domains where long-lasting effects necessarily imply sustained behavioral change (residential waste sorting). These results also highlight the potential of behavioral interventions both when markets are present, and when they are absent. There is a direct private financial benefit to conservation in the energy domain but not in the waste sorting domain studied. These insights therefore contribute directly to questions raised in the recent literature about the the longer-term effects of behavioral interventions, especially their ability to change habits (see Brandon et al., 2017; Czibor et al., 2019), and their use in and outside of market settings (Shogren & Taylor, 2008; Kesternich et al., 2017). However, I do acknowledge that the interventions studied are package interventions (DellaVigna & Linos, 2021), involving multiple behavioral instruments, and that more work is necessary to identify which part of the interventions studied was critical for the observed treatment effects.

Although the findings of the three field experiments indicate persistent behavioral change, which suggests the formation of new habits, I am unable to attribute the persistent behavioral change to a reduction in marginal costs of engaging in the behavior due to a change in the habit stock (Becker & Murphy, 1988). In the waste sorting experiments, there is also evidence for other mechanisms that may explain the persistence in the treatment effect. For instance, the persuasive appeal examined in Chapter 2 affected the motivation of households, and so increased the taste for waste sorting. And the social learning interventions of Chapter 3 induced learning that affected beliefs about the benefits of waste sorting and self-efficacy. These effects may reflect a direct influence of the information embedded in the intervention letters. But they may also reflect indirect effects, when behavioral change in response to the information facilitates learning-by-doing or experience good processes (Nelson, 1970; Foster & Rosenzweig, 1995). Future work is needed to pull apart these direct and indirect mechanisms, using clever identification strategies building on a theoretical framework (Card et al., 2011). An interesting paper in this domain, suggesting a theoretical framework for so-called 'experience enhanced goods' is Alpízar et al. (2020).

5.2 Implications for Dutch environmental policy

From a policy perspective, influencing habitual behaviors is attractive as even temporary interventions may have persistent effects. In addition, because the behavior is repeated frequently, even small changes can accumulate into sizeable effects. The findings of this thesis suggest that indeed long-term change can be obtained using simple and cheap interventions. However, to apply the results to other contexts, we need to assess the external validity of the findings. Can we expect the observed changes in environmental behavior to also materialize in other localities, with similar or possibly different populations? What benefits do we expect from a larger-scale policy? And are the average implementation costs subject to, for example, decreasing or rather increasing returns to scale? These questions all relate to the costs and benefits of the interventions, and highlight that there may be uncertainties regarding how these components vary with the size of the intervention.

To analyze the policy implications of this thesis, I guide policy makers through key considerations that may impact the net benefits of a larger scale policy based on the checklist listed in Al-Ubaydli et al. (2021). These recommendations build on insights of the economic model by Al-Ubaydli et al. (2019) and the conceptual framework introduced in Al-Ubaydli, List, LoRe, and Suskind (2017); Al-Ubaydli, List, and Suskind (2017). This framework consists of three categories. The first two categories are the representativeness of the population and situation with respect to the relevant policy environment, the policy target population and situation. The target population refers to the types of individuals whose behaviors a policy aims to change, while the target situation refers to policy features relating to the context (such as the home of the individuals, or the market price of energy) and implemented program (such as a letter containing certain information sent by postal mail). For each chapter, I analyze key aspects that may impact the representativeness of the studied populations and situations with regards to the policy environment. In the discussion, I pay special attention to the possibility that market prices may be influenced when the policy program is of a much larger scale.

The more similar the study environment is to the policy environment, the more likely it is that the results of the policy resemble those estimated in the study. If the policy environment differs, the policy maker needs to be aware of the risk that results may differ. She may decide to lower uncertainty regarding program impact by collecting additional evidence in a setting closer resembling the policy environment. But also when the population and situation are similar, there is a risk that the policy will not have any impact. This risk is especially relevant for surprising results that have not been replicated. To help policy makers decide whether evidence is sufficient to warrant policy implementation, I discuss the final category of the Al-Ubaydli et al. (2021) framework, correct statistical inference. I analyze how practitioners should evaluate statistical evidence to curb the risk that public funds are spent on ineffective policies. In addition, I highlight how experimentation could reduce the risk that a program does not induce the same results in the policy environment. I conclude with a few final remarks regarding the use of behavioral interventions and field experimentation for Dutch environmental policy.

5.2.1 Key factors influencing representativeness of the population and situation for the residential organic waste sorting interventions

The first domain of environmental behavior studied in this thesis is household organic waste sorting. Organic waste sorting is a behavior that could create a substantial reduction in environmental damages. It reduces CO_2 emissions involved in waste incineration, allows for green energy production (in the form of green gas, green electricity and heat reuse), and enables the production of compost. The latter provides valuable nutrients to be used as natural fertilizer, and is often re-used as production input by the farming industry (Bijleveld et al., 2021). As a result, the findings of Chapter 2 and 3 are of direct relevance for Dutch policy relating to the national ambition to transition to a circular economy (Rijksoverheid, 2016). The policy target is to make sure that by 2050, all primary resources consumed are used efficiently and that all “waste materials” are re-used, without generating any pollution (Rijksoverheid, 2016).[1]

With such high ambitions, the consumption and production patterns of virtually all Dutch households and sectors need to change. Urban residents, and then especially those living in multi-family buildings, are of key interest, as sorting rates are generally found to be low among this population segment. Compared to households in single-family dwellings, these households have a higher cost to sorting: they need to walk to drop-off facilities instead of having access to curbside collection programs, and tend to be more space-constrained (indoors as well as outdoors) which hinders the

[1]In addition, there are some carbon dioxide savings associated with the processing of sorted organic waste, which makes this topic also relevant (albeit to a smaller degree) for climate policy, that aims for the Netherlands to be climate neutral by 2050 (Klimaatwet, 2019).

temporary storage of sorted waste streams before they are deposited in the collective waste sorting facilities. The experimental sites of Chapter 2 and 3 were chosen with this policy environment in mind, as they are both urban neighbourhoods where households have access to organic waste by means of drop-off waste facilities.

In these chapters, the persuasive appeal and social learning interventions had the greatest impact. These treatments can yield substantial environmental benefits when provided to a large group of Dutch households. When we assume that the effects generalize to the broad population of all Dutch residents of multi-family dwellings (that likely would make use of drop-off waste facilities), and that the post treatment effect materializes for one year, we expect the social feedback and persuasive appeal treatments to reduce greenhouse gas emissions by about 2300 tons of CO_2 equivalents, and to generate over 5102 ton of compost.[2][3] The social modeling treatment is expected to reduce emissions by about 680 ton CO_2 equivalents per year, and to generate about 1513 ton compost. The CO_2 reductions emerge because of the production of green gas, electricity and heating, which can substitute for fossil-based sources.

There is one important situational feature, however, that must be satisfied. At the studied sites, all the multi-family dwelling households have access to an organic waste sorting program (organic drop-off facilities), and thus are able to sort organic waste, which may not be the case at other multi-family dwelling neighborhoods. If absent, the costs of establishing these facilities need to be incurred before the behavioral interventions are implemented, otherwise households are not able to sort organic waste, even if they wanted to. This investment may be worthwhile, especially when cou-

[2]To calculate the CO_2 reductions, I multiplied the treatment effects in disposal days with the average weight in kilogram/disposal day as estimated in Chapter 3 (1.164) and the CO_2 reduction per kilogram of organic waste (0.15) as estimated by Bijleveld et al. (2021). To calculate the amount of compost produced, I assumed that about a third of the waste weight is processed to compost based on the 2018 waste processing data analyzed in RWS (2020).

[3]In Chapter 2, the post-treatment period was shorter than in Chapter 3. In the former this period was five months (persuasive appeal); versus nine months in the social learning treatments of the latter chapter. If we assume that the persuasive appeal treatment remains at the level measured in the final two months of the monitoring period, the average treatment effect during the 9 months post treatment is (0.0678*3+0.0998*6)/9=0.0891 disposal days. This effect is very close to the average treatment effect in the 9-month post-intervention period of the social feedback treatment (0.0893). I therefore calculate the environmental impact for the social feedback treatment, assuming the one year treatment effect is 0.0893 disposal days, and conclude that the impact for the persuasive appeal is comparable.

pled with one of the behavioral interventions to boost waste sorting rates. The mere introduction of the drop-off facilities led to about 370 gram of sorted organic waste per household per week at the site of Chapter 2, while this amount was less than half this amount (140 gram per household per week) in Chapter 3.[4] The persuasive appeal and social feedback treatments analyzed in these two chapters increased the organic waste weight collected by at least 28 and 74 percent respectively.[5]

The studied interventions in the waste sorting domain could therefore substantially lower environmental and thus public cost. In the next section, I discuss key factors of that may influence the degree to which the results replicate in other settings, and the costs of the program when provided on this larger scale.

Persuasive appeal, reciprocity and the reward

Because the experimental designs allowed for some time between the introduction of organic waste facilities and the start of the interventions, Chapters 2 and 3 were able to analyze treatment heterogeneity across sorting and non-sorting households. Chapter 2 indeed finds treatment impact to differ among those who sorted waste at the time of the intervention, and those who did not sort. This heterogeneity in effect is mainly visible for the reward and reciprocity treatments, where the treatment impact estimates for non-sorter households were four times (or more) smaller than the estimates on the sorter subsample. I therefore expect the average treatment effect to decrease when these treatments are provided to a population of which a smaller fraction of households sorts waste at baseline. The analysis of Section 2.3.2 finds three population characteristics to correlate with baseline organic waste sorting: being a single person household, having young children, and age (being an elderly household). Elderly age was found to positively correlate with organic waste sorting, while the other two characteristics negatively impacted the decision to sort. Only the distance to the waste containers had explanatory power of the situational characteristics. Closer by organic waste facilities and further away residual waste containers are found to increase waste sorting. Therefore in neighborhoods where fewer

[4]I calculated these amounts by assuming that a household disposes 1.164 kilogram organic waste per disposal day, (based on the estimate obtained in Chapter 3), and multiplying this amount by the control group mean number of weekly disposal days in the post-treatment period (0.320 and 0.120 days per week respectively).

[5]Based on the estimation results presented in Section 3.4.5 of Chapter 3, the treatment impact in terms of disposal days presents a lower bound of the treatment impact on the waste weight collected.

elderly households live but single-person households and young children are more prevalent, and where organic waste facilities are further away (while residual containers are closer by), I expect waste sorting rates to be lower, and thus the reward and reciprocity treatments to induce a smaller increase in waste sorting. For the persuasive appeal, I expect the longer run impact to be similar across populations where more or less households are sorting their organic waste. The treatment impact became virtually identical for sorters and non-sorters in the longer run. Another noticeable population characteristic is that average income in the area of Chapter 2 is relatively high at a level of about 1.5 times mean income in the Netherlands (CBS, 2015). Whether and to what extent this affects the treatments' effectiveness (and hence the scalability of the results) is an open question.

An other important feature of the study was that all three interventions involved about the same budget. If there are (dis-) economies of scale, the relative attractiveness of the three interventions may vary with scale. To analyze how the average cost per intervention changes, I analyze the ratio of fixed to variable cost among the treatments. The reciprocity treatment is the one with the highest variable cost (because of the cost of the high-end cutting board that is offered to all households), and the persuasive appeal treatment is the one with the lowest variable cost (as the cost of designing the information materials is quite high, whereas the cost of sending out the information leaflets is relatively low). The persuasive appeal treatment is the winner of the horse race on the basis of its (longer-run) impact in the RCT; taking into account the fact that treating additional households would be both more effective and less costly for the persuasive appeal treatment than for the other two interventions, just reinforces the claim that this treatment is indeed the most (cost-) effective one. I therefore conclude that the persuasive appeal remains the most promising treatment and that its average implementation costs are likely characterized by economies of scale.

Two situational features that receive special attention in the framework by Al-Ubaydli et al. (2021) are spillover and general equilibrium effects. Spillover effects are also referred to as peer effects, or network effects, and refer to the interactions between households that may influence behavior. I do not expect such spillovers to be there. Chapter 3 specifically studied spillovers from households that received a behavioral intervention (social modeling flyer) to their non-treated neighbors, and did not find evidence for these effects to materialize. General equilibrium effects refer to the effects of the interventions on the overall market, or market system, that may be present in a large scale application. For instance, improved waste sorting behavior could lower residential waste collection fees when the policy target population is a high enough share of municipality residents. These fees are levied as yearly lump-sum taxes in the studied area. But general equilibrium effects may also impact costs attributes. For instance, the larger national production of sorted organic waste may decrease the price recycling facilities, including compost producers, are willing to pay for the materials. As this price decrease lowers the financial returns to running the organic waste recycling schemes, this may put upward pressure on the municipality waste collection fees and so may undo (part of) the fee decrease.

Based on these considerations, I conclude that the persuasive appeal treatment remains the most promising one, also for policymakers. Better waste sorting leads to environmental benefits, and may also downward push lump-sum waste collection fees and so create financial benefits. Based on the treatment heterogeneity analysis, I expect the treatment impact of the persuasive appeal to be similar across settings with higher or lower baseline organic sorting rates. In addition, the average cost may decrease with scale when there are economies of scale in the production of the letters. However, there are also some open questions. The downward push on waste collection fees may be undone when the organic waste price changes. An examination of the (national) organic waste market and the plausibility of this scenario may provide further insight in this matter.

Social feedback and social modeling

The same population and situational features discussed for Chapter 2 may influence the impact of the social learning interventions studied in Chapter 3. Social feedback is significantly more effective among households who had already been sorting, while the chapter is unable to formally conclude heterogenous treatment impact to be present for social modeling. The coefficients signs suggest, if anything, the social modeling impact to also be larger among households with sorting experience. This implies that especially for the social feedback treatment, the average treatment effect may be larger in settings where sorting rates are higher at baseline. How these sorter households are distributed across the neighborhood, does not seem to matter much. We did not find evidence that a higher star score (that a household obtains when it has more sorter neighbors) influenced behavior, but only evidence that those who sorted ex-ante responded more strongly. Policy environments where organic sorting facilities are further away, with less elderly but more small children and more single-member households, are expected to have a lower baseline sorting rate (see Section 2.3.2). I thus expect the impact of social feedback to be lower in those environments. On the other hand, when the organic waste containers are closer by, less households have small children, while elderly and multi-member households are more prevalent, the treatment effect may be larger than the effect observed in Chapter 3. Another feature that may influence the representativeness of the experimental population is average income. The average income in the area of Chapter 3 is below the Dutch average (CBS, 2015), but it is unclear whether this impacts treatment effectiveness.

Also the costs of the letter interventions may change in the relevant policy environment. Especially in a larger scale setting, the program costs involve economies of scale. When the social modeling intervention can be sent to larger-sized neighborhoods, the fixed costs of designing and making the flyer can be shared over a larger group of households which decreases the average fixed cost. Also the average variable cost may decrease when municipalities are able to print the flyers more cheaply in bulk. However, the degree to which larger scale applications can exploit economies of scale is likely constrained by the local nature of the treatments' design. For instance, it may be a key program feature that the social models studied were local neighborhood residents. Larger-scale applications that print the same model in flyers provided to different neighborhoods may be able to economize on costs relating to social model recruitment, but may not reach similar results.

For the social feedback treatment, economies of scale considerations may play a larger role. Especially when it is possible to buy the required container-level weighing equipment in a joint purchase order, it may be possible to reduce average program cost by economizing on the fixed cost of the equipment purchase. But the weighing infrastructure is not the only fixed cost. To implement the treatment, it is necessary to know for each household what the nearest organic container is (to create 'container-clusters'), and a data-pipeline needs to be established, where the weight data is processed into container-cluster specific feedback. I therefore expect the fixed costs to be substantially higher for the social feedback treatment compared to the social modeling treatment, especially when the weighing-data infrastructure is not already present. The variable cost is comparable to the social modeling treatment, as it involves the printing and distributing of the letters. For municipalities that do not have container weighing data available, it may be useful to explore whether existing sources of data can provide input for alternative social learning applications. Social feedback could also be given using the number of households disposing organic waste in given month. However, it may be that the impact of the intervention changes when the letters are adapted. Field experiments could be used as a tool to evaluate the impact of these alternative sources of feedback.

Regarding market equilibrium effects, the expected mechanisms are similar as discussed in Section 5.2.1. There may be downward pressure on the residential lump-sum waste collection fees when residential waste sorting sufficiently improves, but the fees may also again increase when the (equilibrium) price of organic waste decreases in response to a larger supply of organic waste. Moreover, also for these treatments I do not expect spillover effects that change the behavior of households that did not receive the letter treatments. There may, however, be peer effects among households that receive social feedback letters. For these households, changes in the sorting behavior of their neighbors changes the content of the letters they receive. We studied behavior in a setting where relatively few households sorted waste at baseline. In policy environments where more households sort their waste, the treatment response is expected to be higher, and this then may feed into a more positive feedback trajectory and further encourage sorting. Specifically, though we did not find evidence that the star content of the first letter mattered for treatment impact, it may be that the content of one of the three subsequent letters actually did matter. These letters provide insight in whether neighbors changed their behavior in response to the group-level feedback, and the size of the change. The letters may (further) encourage waste sorting when the feedback trajectory suggests a stronger group effort in response to feedback. However, in this case the letters may need to contain a different set of star-score rules, when the 5-star goal studied is too easily reachable in this setting and thus does not motivate better group performance.

Based on these features, I conclude that the social feedback treatment is expected to induce a larger increase in organic waste sorting in policy environments with less small children, more elderly and multi-family member households, and closer by organic waste facilities. In those areas, more households are expected to be sorting their organic waste before treatment and thus the treatment effect is expected to be larger. For social modeling, there is more uncertainty about how the results extrapolate, but the coefficient signs suggest also here for treatment impact to be larger among households with sorting experience. The expected benefits of the treatments involve a reduction in environmental costs, but there may also be financial benefits when the waste collection fees decrease in response to better residential sorting. The costs of the interventions may change at scale. I expect the average cost of the social feedback letters to decrease when weighing equipment can be bought in bulk. For social modeling, I expect average cost to decrease when the letters are sent to larger sized neighborhoods. Similar to the interventions of Chapter 2, there may be a general equilibrium effect in the organic waste market. This may influence the price of organic waste and the reduction in the municipality waste collection fee.

5.2.2 Key factors influencing representativeness of the population and situation for the intervention targeting residential energy conservation

The second type of environment-related behavior studied in this thesis is household energy consumption. Also reductions in energy consumption could induce substantial public benefits, as both the production of electricity and natural gas are associated with CO_2 emissions. Given that the residential sector accounts for about a fifth of national emissions (PBL, 2020), interventions that stimulate energy conservation among households may be effective tools for Dutch climate policy (Klimaatwet, 2019; SER, 2013). Especially energy consumption feedback is seen as promising in this regard.

To stimulate the development and uptake of energy consumption feedback, the Dutch government started in 2015 with the replacement of analogue by smart energy meters. This smart meter roll-out was expected to induce about 3.5 percent residential energy savings (van Gerwen et al., 2010). However, Vringer and Dassen (2016) found the energy savings to fall short of this objective, mostly because the uptake of consumption feedback after meter replacement was lower than expected. To improve the savings percentage, a covenant was made between public and private organizations in the Dutch energy market (Ministerie van Economische Zaken, 2017). In this agreement, the involved parties committed to further develop the market for energy feedback devices and to provide improved energy consumption feedback rapports to all Dutch households. However, a recent evaluation by Paradies et al. (2020) found that also these rapports did not induce the savings needed to reach the savings objective. It therefore remains a policy-relevant question how to capitalize on the Dutch smart meter infrastructure and increase residential energy conservation.

Chapter 4 analyzed the effect of a specific type of consumption feedback, namely disaggregated real-time feedback, that is provided by in-home energy displays. The chapter finds the display to induce 2.2 percent electricity and 6.9 percent gas conservation. To provide an indication of the environmental benefits of a larger scale policy, I calculate the amount of CO_2 emissions avoided for the 400,000 households that owned a feedback device making use of real-time energy data in 2019 (RVO, 2020). Assuming that these feedback devices reduced the annual consumption of electricity and gas with respectively 2.2 and 6.9 percent (from a baseline of 2810 KwH electricity and 42 GJ gas (Menkveld et al., 2017)) and using CO_2 intensities of 0.49 kg/KwH and 56.4 kg/GJ respectively (CBS, 2019; Zijlema, 2019)), these households together reduce yearly Dutch CO_2 emissions by about 77 kilotonnes. Assuming a social cost of carbon of 100 Euros per ton, these households reduce social costs by 7.7 million Euros (or about 20 Euros per household).

Real-time disaggregated energy feedback could therefore substantially lower social costs when provided on a larger scale. In the next section, I discuss key factors of that may influence the degree to which the results replicate or differ in the relevant policy environment, and the costs of the program when provided on a larger scale.

Real-time energy consumption feedback

Chapter 4 differs from the first two chapters, in that households self-selected into this experiment. The decision to select into the program depends on household interest in receiving the display, and may be a function of factors as the expected benefits of the display treatment (Roy, 1951) or participation costs (see Al-Ubaydli et al., 2021). As a result, the expected impact of the display treatment may be different for those who decide to sign up for the experiment and received and installed the device, compared to those who did not sign up. The experiment therefore identifies a local average treatment effect (LATE, Imbens & Angrist, 1994) as opposed to the average treatment effect (ATE) that expresses the expected savings on the general population (e.g. the average Dutch household).

When comparing the household and home characteristics of the experimental households to the Dutch average, a few differences stand out. The experimental households live more often in apartments (and hence in smaller homes), and are relatively more likely to be tenants.[6] But the residences are relatively similar to a representative sample in terms of year built. The largest differences are for homes built between 1975-1984 and between 1945-1974. These home types account for 15 and 31 percent of residences (CBS, 2020), and are over- and undersampled by about 8 and 6 percentage points respectively, a pattern suggesting that the participants tend to live in slightly newer homes. This suggests that the scope for energy savings may be larger

[6] 55 percent rent their residence compared to the Dutch average of 43 percent (Stuart-Fox et al., 2019) About 29% of the households in the experiment live in semi- or fully-detached homes, which is 15% below the national average in the Netherlands) (CBS, 2016).

among the general population, since home owners can take more energy-saving measures and the energy (and especially gas) bill tends to be larger among larger and older home types. This interpretation is reinforced by the relatively high (self-reported) frequency of energy conservation behavior and measures taken among the experimental households, a rate that may be lower among the general population.

The most noticeable feature, however, is that households in the experiment are more likely to have solar panels (40 percent) than the average household in the Netherlands (15 percent). Most important for the intervention, this may indicate that households have higher than average energy knowledge, because they are already engaged with their energy consumption (and/or production). The high energy knowledge score measured in the survey may therefore be lower for the average Dutch household. But there may also be differences in other factors that influence the ability of households to become informed by the display ('treatment moderators'). For instance, solar panels may lead to higher household interest in the energy feedback, when households enjoy being notified that their solar panels are producing electricity. The experimental households may also have more positive attitudes towards energy conservation and energy policy than a representative sample. This may lead to both higher household interest in receiving feedback, and an extra motive to conserve energy due to its impact on the environment. On the other hand, the price incentive to conserve energy may be lower among the experimental households than for the average household. Households with solar panels earn income from their solar panels, which reduces their energy bill and when production is large enough, can even result in them not needing to pay anything for their electricity consumption.

If a larger scale policy embeds a similar opt-in subscription feature, I expect the experimental households to be representative of the target population in terms of participation costs. I expect the larger scale program to also only select households with sufficient interest in the display to actively sign-up. The study locations where we recruited households are more or less representative of the Netherlands, and therefore I expect this program to recruit households that are similar in terms of pre-existing energy knowledge, pro-environmental attitudes and prevalence of solar panels (and thus price incentive to conserve energy). If in addition, the displays are installed by a trained volunteer similar to the experiment, I would expect a similar fraction of about one fifth to drop out, and average energy savings to be similar to the savings measured in the experiment. However, the recruited sample may have a larger scope for energy savings when the policy program is able to recruit more home owners, and more residents of larger and older home types. These households are expected to have a larger gas bill and may more severely underestimate their gas bill. When a larger scale policy successfully recruits households that were less prevalent in the experiment, and indeed the display corrects larger mistakes, average energy savings may increase. I expect the program to foster positive net public benefits, given that the yearly energy savings detected in Chapter 4 translate to financial savings of about 85 Euros, while the device costed about 100 Euros per piece. This calculation ignores the extra public benefits that arise from the reduction in carbondioxide emissions. In addition, the average display costs can likely be lowered in a larger bulk order.

When a larger scale policy changes the recruitment feature of the program, the target population may deviate from the experimental population studied in Chapter 4. This creates uncertainty regarding the expected benefits of the program, as the experimental results may not be generalizable. For instance, consumer mistakes may be more prevalent in a representative sample of households, compared to the households that participated in the study who may already have been engaged with their energy consumption before the program started. The size of the mistake, as measured by the gap between the actual and the estimated gas bill, may also be larger. This could foster a stronger energy conservation response when households become better informed of the actual size of their energy bill. The households may also have taken less energy saving measures before participating in the program, and so have more options to conserve energy available.

Although the scope for energy savings may be larger among this population, this does not mean that this potential is realized. To become better informed in response to consumption feedback, households do need to install and use the display. The interest to do so may be lower among the general population. For instance, in the United Kingdom, displays were offered to households at smart meter installation, a method that substantially lowers participation cost. In a rapport evaluating the progress of this roll-out, DECC (2013, 2014) found that 15 percent of households had a display device. A substantial share (about 40 percent) of households who did not receive a device yet indicated to also not be interested in receiving one. Among the households that did already receive a display, one out of five displays was never installed, and another one was installed, but never used.

New experiments can provide insight in how program attributes relating to display distribution impact the expected energy savings of the program. By testing the impact of the way by which households are recruited, or the way by which the displays are distributed, policy makers obtain valuable information about the expected benefits (energy savings) and costs that may guide policy design. For example, it may be optimal to distribute the device to the general population, when the economies of scale in the purchasing of displays are substantial and average energy savings on the representative population are sufficiently large. An alternative strategy could be to only provide the displays to households who actively declare to be interested, or who value the display enough to exert the effort to sign up for the program. The latter strategy may prevent public funds being wasted on devices that may never be used and may be desirable when the average energy savings decrease too much in the first program. But also variations in recruitment methods may merit further study. Different recruitment methods may target a different populations, and some methods may convince more households to participate than others. There may also be practical reasons to study changes in some of the program attributes. For instance, the policy program may want to send the displays to households by mail as opposed to installing them. The supply of trained volunteer labor may be insufficient to cover a large-scale implementation, and educated labor may be too costly.

The type of display used in the program is another important program feature. Chapter 4 analyzed one type of energy display, a simple monitor that made the current consumption of gas and electricity salient. When the policy program distributes a different type of display, results may differ. Unfortunately, there is little randomized evidence on the impact of other energy displays in the Dutch context.[7] I therefore recommend to evaluate whether alternative displays also make the importance of heating salient, to analyze whether these alternative devices can also be expected to induce savings on space heating. If the functionality of other devices differs too much from the display studied, it may be decided to conduct a separate experimental evaluation of their impact.

Finally, a policy maker may want to consider general equilibrium effects to evaluate the impact of the program at scale. For instance, the scaled-up program may influence energy prices if the program is sufficiently large. It is difficult to predict whether this would be the case as many determinants influence these prices. If the reduction in energy usage does influence energy prices, however, we would expect the prices to decrease, which may increase demand for energy in either the short or longer run. These effects would therefore dampen the positive environmental impact calculated before. But there may also be positive general equilibrium effects. Chapter 4 found that

[7] As energy taxes are relatively high in the Netherlands, the price motive to conserve is higher in the Dutch context than, for instance, the U.S. context. But there may also be cultural differences across countries.

the display especially induced energy savings during the coldest days. Since those days are the days of peak energy demand in the Netherlands (Ministry of Economic Affairs of the Netherlands, 2016), these results demonstrate that consumption feedback could be an effective tool to reduce strain on the energy grid during periods of high demand. Moreover, when households become better informed of the size of their gas bill, this may impact the housing market (Brounen et al., 2013). Consumer demand may reduce for homes that carry a larger gas bill, such as poorly insulated homes. When better insulated homes are more in demand, this may provide an extra financial incentive for households to insulate their homes and reduce their energy bill. The better insulated homes reduce carbon dioxide emissions and so further increase the public benefits of the display intervention. However, further research is necessary to analyze whether indeed this spillover on housing demand occurs.

In conclusion, a large scale policy that stimulates the take-up of displays can induce substantial benefits at scale. These benefits involve both private financial benefits from a lower energy bill and public benefits due to the accompanying reduction in carbondioxide emissions. When the displays are provided using an opt-in program such as the one studied, I expect the energy savings to be similar to those estimated in the experiment. As there likely are economies of scale in the purchasing of displays, I expect the private financial benefits to cover the financial cost and thus positive public net benefits. Yet, if the method of recruitment of the program is changed to recruit other types of households, the energy saving results may not be generalizable. Specifically, I expect average energy savings to increase when a larger-scale program is better able to recruit interested households that have more opportunities to conserve energy (e.g. owner occupants, residents of larger and older homes). However, average savings may also decrease when the display is provided to less interested households. Further experimental research may provide insight into the program attributes that maximize the average net public benefit of the display intervention. Moreover, this research may provide insight into whether the average energy savings cover the (likely lower) average costs of these programs. Finally, a larger scale roll-out of energy displays may lead to both positive and negative general equilibrium effects. It may reduce peak-demand of energy during winter time which reduces strain on the energy grid, but may also decrease energy prices which may undo part of the energy consumption decrease. In addition, there may be spillovers on the housing market. Better informed energy consumers may also be better informed consumers in the housing market, increasing housing demand for well-insulated homes.

5.2.3 Properly acknowledging and resolving uncertainty in evidence-based policy making

Correct statistical inference is the final aspect relevant for evaluating the expected impact of the programs studied at scale (Al-Ubaydli et al., 2021). Even if the experimental population and situation are representative of the policy environment, results may not replicate in other settings. Correct inference starts with the awareness that a single study is not conclusive evidence of program impact. Even the positive result of a well-powered study may be a false-positive; and thus may mask a true zero effect. But there are also factors that increase the probability that study findings do not replicate. This probability increases when the p-value in the study was incorrectly reported, for instance when the authors do not correct for multiple hypotheses testing (List et al., 2019). The study findings may also not be representative of true program impact when the study is underpowered, or the literature suffers from publication bias. Studies that are underpowered may not be able to detect program impact, but may also find extreme (too large) effects (Gelman & Carlin, 2014). Publication bias can lead the findings of a literature to not be representative of the true (distribution of) treatment effects as non-significant results do not get published (Young et al., 2008). Al-Ubaydli et al. (2021) advises policy makers to ignore evidence from low-powered studies, to not take p-values at face value and to not rely on a single piece of evidence.

Instead, they recommend policy makers to adopt a critical reading of available statistical evidence, and evaluate a metric such as the post-study probability (PSP) that a result is true (Maniadis et al., 2014). This metric takes the prior, the expected probability that a study result is true given the pre-study evidence, and updates this prior consistent with the strength of the provided result. The evaluation of whether evidence is actionable should thus be taken based on this updated value (the PSP), and not the p-value mentioned in the particular study. Based on this evidence, the policy maker can decide whether the evidence is convincing enough to warrant program implementation in the targeted policy environment, or whether additional statistical evidence should be collected.

The treatment effects estimated in Chapter 2 until 4 are in line with the pre-existing literature in both the waste sorting and energy domain. Behavioral interventions in the domain of residential waste sorting tend to increase sorting by a few percentage points (Nomura et al., 2011; Milford et al., 2015; Linder et al., 2018), while display interventions tend to generate savings in the order of 6-10 percent (e.g. Schultz et al., 2015; Faruqui et al., 2010; Ehrhardt-martinez et al., 2010). I therefore argue that the prior, the pre-study probability that the result of one of these chapters is true, should be in favor of the result being true. However, the exact choice of prior should be decided upon by the policy maker as it relates to the relevant policy environment at stake. If the policy maker is not too concerned that publication bias may be present in the pre-existing literature, she may set the prior equal to the fraction of pre-existing work that finds the same result. Based on the studies listed, this leads to a value close to one. Alternatively, she may set the prior more conservatively at a value closer to 0.5 when she deems the literature to not be well-developed enough. A lower prior may also reflect uncertainty because of differences between the populations and situations studied in the literature and the relevant policy environment. At a prior of 0.5, and probability of type I and II error of 0.05 and 0.2 respectively (given that the evidence of Chapter 2 until 4 had 80% power and the effect was significant at 5 percent significance), the PSP is 0.94, which means that the probability that the found evidence is true is 94 percent.[8] If the policy maker deems this level of confidence in the result sufficient, she may implement the larger scale policy. If she requires a higher level of confidence, she may decide to collect additional evidence. Al-Ubaydli et al. (2021) recommend policy action to be taken when the PSP is at least 0.95 and argue for at least 3 or 4 independent replications of a result before scaling up an intervention.

Being mindful that representativeness of the population, situation and statistical inference, influence program impact allows the policy maker to make an informed evaluation of available evidence. When there is too little proof, she may decide to collect additional statistical evidence by means of additional field experiments before a national roll-out. These replications may also be fitted to uncover specific features of the cost-benefit trade-off. For instance, a new experiment may specifically compare the impact of two programs that vary in the manner by which they distribute displays to Dutch households. In addition, larger-scale experimentation could specifically analyze whether interventions are equally effective, and cost-effective, at scale (Muralidharan & Niehaus, 2017). Larger scale experimentation can also allow for a formal analysis of market equilibrium effects. For instance, by randomly assigning municipalities to treatment (e.g. the social feedback letters) or control (no intervention) it is possible to analyze the impact of the intervention on the municipal waste collection fees.

[8] $PSP = \frac{(1-\beta)\pi}{(1-\beta)\pi+\alpha(1-\pi)}$, where π denotes the prior, $(1-\beta)$ the power of the study, and α the significance level of the result.

5.2.4 Concluding remarks regarding policy lessons

In this thesis, I have analyzed several behavioral interventions that may be attractive for policy to improve environmental outcomes. The interventions had lasting impact in different domains (energy use and waste sorting) and on different types of Dutch households (with incomes below and above the Dutch average).[9] Because of their relatively low costs and relatively easy political implementability, these interventions may be low-hanging fruit and attractive policy tools to complement existing environmental policy.

The evidence is promising, but it is clear that the interventions studied are by no means sufficient to reach the circular economy and green house gas emissions targets for the year 2050 in the Netherlands (Rijksoverheid, 2016; Klimaatwet, 2019). Therefore, questions on how to further improve the environmental impact of these interventions remain relevant. More households need to receive and use the displays in order to facilitate larger national energy savings, but it is uncertain how to optimally distribute the display devices to maximize program impact. Also in the waste sorting experiments, substantial scope for improvement was left untapped. A large number of households never sorted organic waste, also not after the interventions were provided. More research is needed to understand whether a lack of interest in either waste sorting or consumption feedback is rational, or whether there are mistakes in judgement that prevent households to participate. Research into these hidden barriers, and how to overcome them, can provide valuable insights that may feed into more effective versions of the interventions studied and the programs by which they are implemented. The research avenue discussed in Section 5.1 may be promising in this respect, as these barriers may relate to the exposure-enhanced mechanism at play. For instance, households may underestimate the influence of habit formation, or neglect to incorporate learning-by-doing or experience good processes in their decision making.

Field experiments can be used to analyze these questions. Coupled with insights from the literature on scaling, these experiments can be designed to uncover relevant information for the particular policy question at stake. The evidence they provide lowers uncertainty regarding the expected net benefits of policies, which reduces the risk that ineffective policies are adopted. Field experiments can therefore play an important role in the movement

[9] Average income in the area of Chapter 2 is about 1.5 times mean income in the Netherlands, while the average income of the area of Chapter 3 is below the Dutch average (CBS, 2015). In both areas we found behavioral treatments to induce persistent behavioral change.

to evidence-based policy making (Al-Ubaydli et al., 2021).[10] In addition, field experimental evidence can speed up political processes by being more convincing compared to other (especially non-causal) empirical methods. But there are also costs involved. Generating field experimental evidence takes time, as data needs to be collected, and project partners may need to be found to implement a desired experiment. The question of whether to implement a field experiment therefore ultimately revolves around the public benefits of the stronger empirical evidence, and whether these benefits exceed the public costs of collecting this evidence.

[10] In this paradigm, policies are based on evidence (Davies, 2012), and scarce public funds are only spent on what actually works, in a way that ideally maximizes the public benefits of each Euro spent.

5.3 References

Alpízar, F., Bernedo Del Carpio, M., Ferraro, P. J., & Meiselman, S. (2020). *Exposure enhanced goods and technology disadoption: Evidence from a randomized controlled trial with resource-conserving technologies* (Working Paper, January 14).

Al-Ubaydli, O., Lee, M. S., List, J. A., Mackevicius, C. L., & Suskind, D. (2021). How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy*, *5*(1), 2–49.

Al-Ubaydli, O., List, J. A., LoRe, D., & Suskind, D. L. (2017). Scaling for economists: Lessons from the non-adherence problem in the medical literature. *Journal of Economic Perspectives*, *31*(4), 125–144.

Al-Ubaydli, O., List, J. A., & Suskind, D. (2019). *The science of using science: Towards an understanding of the threats to scaling experiments* (Working Paper No. 25848). NBER.

Al-Ubaydli, O., List, J. A., & Suskind, D. L. (2017). What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review*, *107*(5), 282–286.

Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy*, *96*(4), 675–700.

Bijleveld, M., Beeftink, M., Bruinsma, M., & Uijttewaal, M. (2021). *Klimaatimpact van afvalverwerkroutes in Nederland: CO_2-kentallen voor recyclen en verbranden voor 13 afvalstromen* (Tech. Rep. No. 20.190400.163). Delft: CE Delft.

Brandon, A., Ferraro, P. J., List, J. A., Metcalfe, R. D., Price, M. K., & Rundhammer, F. (2017). *Do the effects of social nudges persist? Theory and evidence from 38 natural field experiments* (NBER Working Paper Series No. 23277). National Bureau of Economic Research.

Brounen, D., Kok, N., & Quigley, J. M. (2013). Energy literacy, awareness, and conservation behavior of residential households. *Energy Economics*, *38*, 42–50.

Card, D., DellaVigna, S., & Malmendier, U. (2011). The role of theory in field experiments. *Journal of Economic Perspectives*, *25*(3), 39–62.

CBS. (2015). *Kerncijfers wijken en buurten 2015.* Retrieved 2021-03-22, from `https://www.cbs.nl/nl-nl/cijfers/detail/83220NED`

CBS. (2016). *Vier op de tien huishoudens wonen in een rijtjeshuis.* Retrieved from `https://www.cbs.nl/nl-nl/nieuws/2016/14/vier-op-de-tien-huishoudens-wonen-in-een-rijtjeshuis`

CBS. (2019). *Rendementen en CO_2-emissie van elektriciteitsproductie* *in Nederland, update 2019.* Retrieved from `https://www.cbs.nl/nl-nl/achtergrond/2021/08/rendementen-en-co2-emissie-van-elektriciteitsproductie-in-nederland-update-2019`

CBS. (2020). *Voorraad woningen; gemiddeld oppervlak; woningtype, bouwjaarklasse, regio.* Retrieved 2021-03-23, from `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/82550NED/table?fromstatweb`

Czibor, E., Jimenez-Gomez, D., & List, J. A. (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal*, *86*(2), 371–432.

Davies, P. (2012). The state of evidence-based policy evaluation and its role in policy formation. *National Institute Economic Review*, *219*(1), 41–52.

DECC. (2013). *Quantitative research into public awareness, attitudes, and experience of smart meters (wave 3)* (Tech. Rep.).

DECC. (2014). *Quantitative research into public awareness, attitudes, and experience of smart meters (wave 4)* (Tech. Rep.).

DellaVigna, S., & Linos, E. (2021). *RCTs to scale: Comprehensive evidence from two nudge units* (Working Paper April 2021).

Ehrhardt-martinez, K., Donnelly, K. A., & Laitner, J. A. (2010). *Advanced metering initiatives and residential feedback programs: A meta-review for household electricity-saving opportunities* (Report No. E105). Washington D.C: ACEEE.

Faruqui, A., Sergici, S., & Sharif, A. (2010). The impact of informational feedback on energy consumption - A survey of the experimental evidence. *Energy*, *35*(4), 1598–1608.

Foster, A. D., & Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, *103*(6), 1176–1209.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (Sign) and type M (Magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467.

Kesternich, M., Reif, C., & Rübbelke, D. (2017, July). Recent trends in behavioral environmental economics. *Environmental and Resource Economics*, *67*(3), 403–411.

Klimaatwet. (2019). *Klimaatwet.* Retrieved 2021-03-23, from `https://wetten.overheid.nl/BWBR0042394/2020-01-01`

Linder, N., Lindahl, T., & Borgström, S. (2018). Using behavioural insights to promote food waste recycling in urban households - Evidence from a longitudinal field experiment. *Frontiers in Psychology*, *9*(352), 1–13.

List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, *22*(4), 773–793.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, *104*(1), 277–290.

Menkveld, M., Rietkerk, M., Mastop, J., Tigchelaar, C., & Straver, K. (2017). *Besparingseffecten van slimme meters met feedbacksystemen en slimme thermostaten* (ECN Notitie No. 17-017).

Milford, A. B., Øvrum, A., & Helgesen, H. (2015). *Nudges to increase recycling and reduce waste* (Discussion Paper No. 2015–01). NILF Norwegian Agricultural Economics Research Institute.

Ministerie van Economische Zaken. (2017). *Convenant 10 PJ energiebesparing gebouwde omgeving* (Tech. Rep.).

Ministry of Economic Affairs of the Netherlands. (2016). *Energy report: Transition to sustainable energy* (Tech. Rep.). Ministry of Economic Affairs.

Muralidharan, K., & Niehaus, P. (2017). Experimentation at scale. *Journal of Economic Perspectives*, *31*(4), 103–124.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311–329.

Nomura, H., John, P. C., & Cotterill, S. (2011). The use of feedback to enhance environmental outcomes: a randomised controlled trial of a food waste scheme. *Local Environment*, *16*(7), 637–653.

Paradies, G., Dreijerink, L., & Menkveld, M. (2020). *Effectmeting verbeterd verbruiks- en kosten overzicht* (TNO-rapport No. 2020 P10380). TNO.

PBL. (2020). *Klimaat- en energieverkenning 2020* (Tech. Rep.). Den Haag: Planbureau voor de Leefomgeving.

Rijksoverheid. (2016). *Nederland circulair in 2050* (Tech. Rep.).

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146.

RVO. (2020). *Monitoringrapportage 2019 convenant Gebouwde Omgeving* (Tech. Rep.). Rijksdienst voor ondernemend Nederland.

RWS. (2020). *Afvalverwerking in Nederland: gegevens 2018* (Tech. Rep.). Utrecht: Rijkswaterstaat (RWS).

Schultz, P. W., Estrada, M., Schmitt, J., Sokoloski, R., & Silva-Send, N. (2015). Using in-home displays to provide smart meter feedback about household electricity consumption: A randomized control trial comparing kilowatts, cost, and social norms. *Energy*, *90*, 351–358.

SER. (2013). *Energieakkoord voor duurzame groei* (Tech. Rep.).

Shogren, J. F., & Taylor, L. O. (2008). On behavioral-environmental economics. *Review of Environmental Economics and Policy*, *2*(1), 26–44.

Stuart-Fox, M., Blijie, B., Ligthart, D., Faessen, W., & Kleinepier, T. (2019). *De woningmarkt en leefbaarheid in krimpgebieden. Uitkomsten van het WoonOnderzoek Nederland (WoON) 2018 en CBS data* (Tech. Rep.). Delft: ABF Research.

van Gerwen, R., Koenis, F., Schrijner, M., & Widdershoven, G. (2010). *Intelligente meters in Nederland: Herziene financiële analyse en adviezen voor beleid* (Tech. Rep.). Arnhem: KEMA.

Vringer, K., & Dassen, T. (2016). *De slimme meter, uitgelezen energie(k)?* (Tech. Rep.). Den Haag: PBL.

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*(10), 1418–1422.

Zijlema, P. (2019). *Berekening van de standaard CO_2 -emissiefactor aardgas t.b.v. nationale monitoring 2020 en emissiehandel 2020* (Tech. Rep.). Utrecht: RVO.

This dissertation studies the impact of behavioral interventions on waste sorting and energy conservation, two domains where sustained environmental conservation has the potential to substantially reduce social costs. The interventions are evaluated by means of field experiments. The first essay investigates the relative impact of behavioral interventions versus neoclassical interventions. It finds that interventions that draw on extrinsic motivations have an immediate and sizable effect on waste sorting behavior, but also that the average treatment effects attenuate steeply over time. In contrast, the essay finds equally sizeable yet long-lasting effects of a treatment designed to increase households' intrinsic motivation to sort waste. The second essay analyzes the effect of social learning interventions. It considers two interventions, one aimed at leveraging social learning via role models and a second one via feedback on the prevalence of organic waste sorting in the household's direct vicinity. The essay finds that both interventions increase waste sorting in the short run, but only the social feedback's impact is long-lasting. The third essay analyzes residential energy consumption, and how real-time disaggregated consumption feedback corrects consumer mistakes in this domain. The essay finds this feedback, provided by way of in-home displays, to reduce household energy consumption. The savings are largest on gas consumption, and the evidence suggests the effect to reflect reductions in space heating. The three essays are preceded by an introductory chapter that introduces the topic of study and the field experimental methodology. The dissertation closes with a concluding chapter that reviews its contribution to the literature on behavioral interventions and its implications for Dutch environmental policy.

Mirron Adriana Boomsma ('s-Gravenhage, the Netherlands, 1993) obtained a Bachelor of Science degree in Economics & Business Economics from Maastricht University in 2014, with a one semester visit to the Hong Kong University of Science and Technology (2013). She continued her studies as a Research Master in Economics student at the CentER Graduate School of Tilburg University. After graduating in 2016, she joined the Economics Department as a PhD candidate, on a joint project between the Tilburg Sustainability Center (TSC) and The Netherlands Environmental Assessment Agency (PBL). During her doctoral studies, she spent a semester at the University of Chicago through the Aart de Zeeuw scholarship program. Her academic supervisors were prof. dr. Daan van Soest and dr. Ben Vollaard.

ISBN: 978-90-5668-655-0
DOI: 10.26116/center-lis-2113

TILBURG SCHOOL OF ECONOMICS AND MANAGEMENT

CentER Dissertation Series