WORKING PAPER

THE PROOF OF THE PUDDING IS IN THE HEATING: A FIELD EXPERIMENT ON HOUSEHOLD ENGAGEMENT WITH HEAT PUMP FLEXIBILITY

Baptiste Rigaux
Sam Hamels
Marten Ovaere

May 2025
2024/1101

GHENT UNIVERSITY

Department of Economics

Faculty of Economics and Business Administration
Campus Tweekerken, St.-Pietersplein 5, 9000 Ghent - BELGIUM

The Proof of the Pudding is in the Heating: A Field Experiment on Household Engagement with Heat Pump Flexibility

— Accepted version, May 2025 —
— The final version is published in *Energy Economics*,
available at https://doi.org/10.1016/j.eneco.2025.108565 —
© 2025. This manuscript version is made available under the CC BY-NC-ND 4.0 license.

Baptiste Rigaux[1,*], Sam Hamels[1], Marten Ovaere[1]

[a] *Department of Economics, Ghent University, Tweekerkenstraat 2, Gent, 9000, Belgium*

Abstract

As renewable energy grows, flexible electricity demand becomes essential. We conducted a field experiment with nine heat pumps in well-insulated homes near Ghent, Belgium. During 287 flexibility interventions, we remotely deactivated heating until indoor temperatures reached predefined thresholds or households manually overruled the intervention. After initiating a flexibility event, the heat pump power is initially lowered by 250 W on average per unit in the fleet. As some heat pumps in the fleet reactivate, they consume more power to restore their threshold temperatures, triggering a rebound effect that gradually reduces net power savings achieved. On average, net power savings become zero after 18 hours, followed by a rebound period. Overall heat pump consumption was reduced by around 1 kWh per event, stabilizing 36 hours after the event start. If flexibility activation is timed strategically, up to €1.1 can be saved through price arbitrage, assuming wholesale prices at energy-crisis-level, while the capacity benefits value can be up to $175. Smart heating algorithms could further increase savings generated by all value streams. Colder weather significantly influences savings, by increasing the power available for flexibility but also amplifying rebound effects. This flexibility came with moderate comfort impacts: on average, indoor temperatures were 0.38 °C lower during interventions. However, 19% of interventions were manually overruled when larger temperature drops occurred, with households citing discomfort, illness, or occupancy as factors on an online dashboard. These findings suggest that flexible residential heating can support renewable energy integration with moderate comfort impacts.

Keywords:
Electricity Demand, Flexibility, Direct Load Control, Field Experiment, Household, Heat Pump, Thermal Comfort
JEL: Q40, Q41, D12

*Corresponding author.
Email addresses: baptiste.rigaux@ugent.be (Baptiste Rigaux), sam.hamels@ugent.be (Sam Hamels), marten.ovaere@ugent.be (Marten Ovaere)

1. Introduction

Electricity demand flexibility is crucial for the energy transition, especially with the growing variability of renewable sources like wind and solar. Achieving targets such as renewable energy accounting for 32% of gross final energy consumption in the EU by 2030 (Council of the European Union, 2018) requires changes not only in electricity supply but also in demand, with households and industries shifting their electricity consumption to periods when the sun shines or the wind blows. As the transition to renewables accelerates, projections show that the EU's flexibility resources for managing daily electricity consumption peaks will need to double between 2021 and 2030, reaching up to 362 TWh per year (European Environment Agency, 2023). Various economic incentives exist to support households in this transition, typically falling into two categories: implicit and explicit flexibility. Implicit flexibility occurs through time-varying prices and tariffs that encourage households to consume more when renewable supply is high and prices are low (due to the low marginal cost of renewable generation) and less when renewable supply is low and prices rise. Explicit flexibility programs, on the other hand, offer direct compensation from third parties, such as aggregators or Distribution System Operators (DSOs), for adjusting consumption to meet system or grid needs (Nouicer et al., 2020). Implicit and explicit flexibility benefit households, but they also produce wider societal benefits. Indeed, reduced system consumption peaks lower the need for costly, carbon-intensive peak power plants and fuel imports, while also decreasing the demand for reserve power plants and transmission capacity (Fischer and Madani, 2017). This, in turn, contributes to grid stability and minimizes the risk of power outages in a system increasingly based on renewable supply.

The adoption of low-carbon technologies, like electric vehicles (EVs) and heat pumps (HPs), is viewed as a critical step toward increasing residential demand flexibility. Heat pumps, in particular, hold significant potential for flexibility, as space and water heating—representing 78.3% of European households' final energy consumption (Eurostat, 2022)—can be adjusted by a few hours to better match electricity supply. This potential was recently demonstrated in practice by a study in the UK, where Bernard et al. (2024) found that HPs reduce gas use by 90% and CO_2 emissions by 36% in households where they are installed.

However, despite economic incentives like time-varying prices, many households may fail to fully exploit this flexibility due to several well-documented barriers. The behavioral economics literature on flexibility highlights for instance the status quo bias—where people tend to keep current habits even when incentives are offered—and risk aversion and loss aversion, where perceived risks and potential losses are too high to trigger changes (see, e.g., (Darby and McKenna, 2012; Frederiks et al., 2015; Good et al., 2017; Hobman et al., 2016; Good, 2019)). Additionally, bounded rationality is a recognized barrier when the complexity or volume of available information about incentives overwhelms households, often leading to satisficing behavior, meaning that simple heuristics for adjusting electricity consumption patterns result in sub-optimal gains (see (Good, 2019; Good et al., 2017; Gyamfi et al., 2013; Hobman et al., 2016; Kim and Shcherbakova, 2011; Frederiks et al., 2015)). This tendency is amplified when the incentives to shift electricity consumption trigger frequent and repeated adjustments, leading to what is referred to as response fatigue (Kim and Shcherbakova, 2011). Moreover, the delayed financial gains from adjusting consumption contribute to time-discounting behavior (see (Frederiks et al., 2015; Good, 2019; Hobman et al., 2016)). In addition to their timing, the magnitude of these gains may also play a role (Fabra et al., 2021): as noted by Bailey et al. (2024), they may sometimes be too small to offset the households' opportunity cost of the time they invest in adjusting their consumption.

Nowadays, manufacturers are addressing these challenges by developing tools such as smartphone apps and automation systems, which simplify control and reduce the effort needed to adjust electricity consumption. This raises two critical questions, central to our study: how willing are households to give up control over their assets, and how do they respond to potential discomfort, such as thermal discomfort? Accurately estimating the flexibility potential of households requires a better understanding of these behavioral barriers.

This paper studies the flexibility potential of residential HPs through a field experiment. Our treatment interventions involve temporarily switching off HPs until one of three predefined stopping scenarios is triggered: the indoor temperature drops below a certain threshold, the domestic hot water (DHW) tank temperature falls below 40 °C, or the household overrides the intervention via an online platform, where they must provide a reason for doing so. In cases where interventions were notified a day in advance, households could also override preemptively. As such, this setup mimics what is commonly referred to as 'direct load control' (DLC), where "a utility has the right to control the customers' appliances, usually based on a contract which is useful for peak demand reduction or emergency situation handling" (Kostková et al., 2013). By implementing a default

and decentralized automation of the HPs, our experiment simplifies the typically complex decision-making process required for households to respond to flexibility incentives, and helps alleviate some of the behavioral barriers discussed above.

We conducted 287 flexibility interventions over two winter seasons (2022-2023 and 2023-2024). Before the first season, a survey revealed that the participants were generally richer, more environmentally conscious, and living in better-insulated homes compared to the Belgian average. This is expected, as less than 5% of Belgian households are equipped with a HP (Rosenow et al., 2022), and these are typically installed in homes with above-average insulation. The survey also explored how households formed expectations about flexibility and their comfort temperature preferences during interventions.

We find that the flexibility interventions unfold in two distinct phases. In the first phase, power consumption decreases to almost zero as the HP is temporarily switched off. In the second phase, power consumption increases beyond the usual levels as the HP compensates to return to the setpoint temperature (i.e., the predefined temperature chosen by the household on their thermostat). This phenomenon, known as a rebound peak, has been shown to occur for instance when many automated flexible assets resume normal operation at the same time after periods with high day-ahead prices or high time-of-use tariffs (Muratori et al., 2014; Ludwig and Winzer, 2022; Dewangan et al., 2022; Tomat et al., 2022).

To quantify HP flexibility during interventions, we measure five key variables: the duration of how long the HP remains turned off (hours), the power reduction over the course of the intervention (kW), the total decrease in electricity consumption during the intervention (kWh), the increase in electricity consumption after the intervention (kWh), and the financial savings resulting from the intervention (€). We conduct our analysis at two levels: the individual HP level, focusing on periods when HPs are blocked (referred to as 'interventions'), and the fleet level, capturing the aggregated response of both blocked HPs and those that have already got unblocked after an intervention is initiated in a fleet of HPs (referred to as 'flexibility events').

First, we analyze the interventions at the individual HP level. We find that interventions lasted an average of 12.8 hours before meeting the criteria for one of the stopping scenarios. A regression analysis of intervention duration shows that the strongest predictor of duration is the indoor temperature at the start of the intervention, with a one-degree increase extending the duration by 2.1 hours on average over the whole sample. Similarly, a higher initial DHW tank temperature and higher outdoor temperature positively influence intervention length. We did not find any evidence that households altered their behavior by raising the indoor temperature prior to pre-notified interventions. This is consistent with a post-experiment survey, where households indicated that they did not place high importance on being notified of upcoming interventions. Interestingly, the time of day when the intervention started had no measurable effect on its duration. During the interventions, HP power consumption dropped to around 50 W, as some electricity is still required to operate the HP's electrical boards and circulatory pumps. Using the average HP power profile outside intervention and rebound periods as a counterfactual, we estimate that HP power consumption is reduced by 84% on average during an intervention, translating to an average reduction in electricity consumption of 3.2 kWh over the duration of an intervention. However, this flexibility comes with significant post-intervention rebound effects, requiring an average of 607 W of additional power consumption in the first post-intervention hour, or 2.4 kWh of electricity consumption over 16 hours to restore the indoor and DHW temperatures back to the user setpoint. A regression analysis with household fixed effects on rebound energy consumption within 16 hours after intervention stop shows that it is primarily driven by the difference between the household's setpoint temperature at the end of the intervention and the actual indoor temperature. Each one-degree increase in this difference is associated with an average increase of 0.80 kWh in rebound electricity consumption (borderline significant) in rebound energy consumption. Additionally, for each degree increase in the average outdoor temperature during the 16-hour period, rebound consumption decreases by 0.28 kWh. Similar to intervention duration, the time of day when the intervention stops has no significant effect on rebound consumption.

Second, we broaden the scope of our analysis to study fleet-level behavior, aggregating all HPs into a single profile to quantify flexibility over time relative to the start of the intervention. We refer to this as flexibility events. Unlike the analysis of individual interventions, which focuses only on HPs during their blocked periods, flexibility events capture the aggregated response of the entire fleet after heating is temporarily deactivated. During a flexibility event, some HPs remain off, while others gradually get unblocked at different times—depending on building and household characteristics—and experience rebound effects. On average, the power reduction at the start of a flexibility event is around 250 W per HP, gradually decreasing to 0 W after around 18 hours. Beyond this point, the fleet-level rebound peak occurs, during which the fleet consumes

more electricity than it would have without the intervention. Flexibility events last longer than individual interventions because individual HPs reactivate at different points in time, resulting in a lower average fleet-level rebound than at the individual HP level, thereby allowing still-blocked HPs to offset the unblocked HPs' rebound impact and extend the period of the fleet-level net power reduction. We find a clear relationship with weather conditions. When average temperatures within the first 18 hours of an event are below 3 °C, the initial power reduction reaches 600 W per HP, with energy consumption reduced by 4.5 kWh in the first phase. However, the rebound phase causes total net savings per event to drop to only around 1 kWh after 36 hours. In contrast, when average temperatures are above 9 °C, consumption decreases by 1.5 kWh, with no rebound observed, resulting in sustained savings. Note that all houses in our sample have an energy performance label A, indicating a high degree of thermal insulation and impacting both the observed power reductions and intervention durations.

To monetize flexibility events, we simulate financial gains using real day-ahead prices in Belgium. We calculate gains by assuming a flexibility event is initiated at each hour of the winter seasons 2022-2023 and 2023-2024. Using outdoor temperature data, each event is assigned to one of four average power reduction profiles—specific to defined temperature ranges—and matched with day-ahead electricity prices to estimate net savings. On average, net savings at 36 hours after the event start amount to €0.13 per event per HP. However, targeting periods of high and volatile electricity prices can increase net savings to €1.09 per event per HP. In practice, flexibility aggregators are likely to adopt targeted strategies that maximize financial gains. Our back-of-the-envelope calculation shows that a substantial number of events under high-price conditions are required to reduce the payback period for a smart thermostat (enabling HP flexibility) to around 10 years—specifically 40 to 50 events per year under 2022–2023 winter price levels. Moreover, aggregators are expected to target additional value streams. For example, capacity benefits can reach up to $175 per HP.

Finally, we assess the discomfort experienced by households during the interventions. Our findings indicate that most interventions resulted in a modest temperature decrease of 0.38 °C within a household, representing the average change across the entire intervention period. However, in cases where interventions were manually overruled, the temperature drop was significantly larger, reaching 1.06 °C by the end of the intervention (i.e., at the moment of overrule). This suggests that households' observed overruling behavior aligns with their subjective perceptions of discomfort, though this is not consistent across all interventions, as the correlation between temperature drop and manual overrule is weak. This is further supported by the reasons provided by households upon overruling, which can be categorized into three main categories: "Too low indoor temperature" (accounting for 65% of manual overrides); followed by "Sickness/health concerns", when illness triggers a desire for thermal comfort (20.5%), and "Presence at home", when an individual, usually working or studying at home, requires higher temperatures (16.5%). In a post-experiment survey, households reported experiencing slight to moderate discomfort, with an average score of 2.4 on a 1–5 Likert scale (*No discomfort - Extreme discomfort*), and confirmed a strong preference for being offered the option to overrule interventions, averaging 3.9 on a 1–5 scale (*Not important at all - Extremely important*).

Related literature and contributions.

Our study contributes to the growing literature that studies the responses of households to incentives to make their electricity consumption more flexible. Research on time-varying prices and other forms of remuneration as strategies to shift consumption away from peak periods often shows weak or limited reactions. For instance, Herter and Wayland (2010) find a 5% reduction in residential electricity consumption under a critical peak pricing scheme. More recently, Fabra et al. (2021) show that Spanish households do not significantly react to a real-time pricing scheme, while Enrich et al. (2024) found that consumption during peak periods only reduced by an average of 9% when the aforementioned scheme was later replaced by a time-of-use tariff. These relatively low reactions may be partly explained by a lack of awareness about incentives among the households involved (Fabra et al., 2021; Enrich et al., 2024). This suggests a need for technologies that make price incentives more effective and salient, such as in-home displays or luminous orbs, which can lead to stronger household responses compared to the absence of such technologies, as shown by (Jessoe and Rapson, 2014; Allcott, 2011). Additionally, numerous studies identify automation as a key enabling technology, reducing or even eliminating the need for households to manually adjust their electricity consumption to incentives (Herter et al., 2007; Faruqui and Sergici, 2010; Bollinger

and Hartmann, 2015, 2019; Harding and Lamarche, 2016; Harding and Sexton, 2017).[1]

While the aforementioned literature primarily focuses on overall household electricity consumption, evidence suggests that these conclusions extend to programs directly targeting heating flexibility. In the absence of automation, recent studies indicate that HP usage remains highly price-inelastic. For example, a 2024 Swiss randomized control trial by Boogen and Winzer (2024) implemented steep electricity price increases—from a typical 0.3 CHF/kWh to an average of 4 CHF/kWh (up to 7 CHF/kWh)—but found only a 13.8% drop in consumption. A follow-up survey revealed that households primarily achieved this reduction through adjustments to heating settings. This suggests that automation might further enhance responsiveness. This is supported by Brewer (2023), whose survey on hypothetical heating costs shows that households would not substantially lower their thermostats in response to higher prices (with an elasticity between -0.005 and -0.014), preferring bearing the extra costs instead.

In the behavioral economics literature on flexibility too, a consensus is emerging in favor of automation, which is seen as a way to reduce the cognitive load imposed on households and to help eliminate well-documented barriers outlined above such as status quo bias, bounded rationality, and response fatigue (Frederiks et al., 2015; Darby and McKenna, 2012; Good et al., 2017). Acknowledging advancements in broader energy economics research, including studies on the behavioral aspects of residential flexibility, our paper positions itself within an emerging literature focusing on field trials of automated heating flexibility. In their study, Bailey et al. (2024) compare two automation treatments under time-varying price conditions. They find that households with access to an app to manage appliances (thermostats, hot water heaters, EV chargers) reduced peak consumption by about 5%, similar to those adjusting manually. However, the reduction jumps to 26% for households whose appliances were automatically controlled by the utility as part of the treatment. Similarly, Blonz et al. (2025) show that automating thermostats with time-of-use pricing reduces the electricity consumption of air conditioning (A/C) units by 63% during peak periods in summer, resulting in savings of CA$0.21 per household per day of activation, with most participants experiencing no discomfort. Extending this, Fu et al. (2024) study the implementation of a time-of-use tariff in California, where electricity prices doubled during summer peak hours. Analyzing the response of households with smart thermostats, they show that this led to an average increase in air conditioner thermostat setpoints of 1.04 °F (0.6 °C), reducing runtime and lowering the utility's electricity load by 5% on the hottest summer days. Finally, Kane et al. (2024) conducted a recent field experiment using Wi-Fi-controlled switches to automatically manage A/C units during flexibility interventions, resulting in an 8.5% reduction in household electricity demand and a 2.3% reduction in CO_2 emissions during events.

The benefits of automation have also been demonstrated specifically in the context of HP flexibility. For example, Jensen et al. (2018) test the automatic and continuous adjustment of HPs of eight households based on electricity prices, weather forecasts and user-defined temperature preferences. They found that such automation reduced household energy costs by 6.8% to 16.9%, while avoiding significant discomfort. Similarly, Bernard et al. (2024) study a large-scale implementation of a specific time-of-use tariff for HPs on 6,631 households. Although automation was not a specific feature of the experiment, most participants reported using smart thermostats to automate their HPs, leading to a 50% reduction in electricity consumption during peak (expensive) periods and a significant shift to off-peak periods. This resulted in savings of up to 18% on households' yearly bill. In another pilot study by the same research center, 30 four-hour flexibility interventions were tested on 43 HPs (Centre for Net Zero and Nesta, 2024). Participants were allowed to set their own minimum and maximum temperature thresholds using smart thermostats, enabling homes to be preheated to the maximum during the first two hours of an intervention and then cooled down to the minimum during the blocking period. The study found a 74% reduction in HP consumption during the blocking period, with no significant rebound consumption afterward. Indoor temperatures were on average 0.85 °C higher during preheating. Only 9% of events were overruled by participants in advance, mainly to avoid unnecessary preheating when they were away, and just

[1] In addition to studies focusing on incentivizing long-lasting consumption flexibility, another literature stream investigates short-term responses to critical electricity shortages under extreme weather or supply disruptions. In this context, evidence regarding emergency conservation appeals is mixed. On one hand, Holladay et al. (2015) report that press-released appeals during peak periods can paradoxically increase energy use. On the other hand, Brewer and Crozier (2023) found that a targeted text alert in Michigan prompting households to reduce air-conditioners setpoints—amid low temperatures and gas supply disruptions—resulted in a significant drop in indoor temperatures, with a 45% increase in households at or below the setpoint of 65 °F suggested in the alert. Similarly, Leighty and Meier (2011) observed substantial consumption reductions in Alaska following an avalanche, and Kimura and Nishio (2016) documented a 15% decrease in electricity use after the 2011 earthquake and tsunami in Japan.

2% were overruled during events due to discomfort.

One effective way to ensure participation and acceptance needed for automated flexibility is to allow participants to override interventions (Karjalainen, 2013; Xu et al., 2018). However, these overrides reduce intervention efficiency and can lead to missed power reductions (Tomat et al., 2022): for example, a summertime flexibility program in which utilities adjusted A/C setpoint temperatures saw its energy savings cut in half by overrules (Wildstein et al., 2023). Override rates across households or interventions in A/C flexibility programs vary between trials, ranging from 12.9% to 38.5% of total events (Wildstein et al., 2023; Tomat et al., 2022; Sarran et al., 2021), while Ouf et al. (2024) report that over 75% of participants in a Quebec DLC trial overrode interventions more than 10 times per year. Ouf et al. (2024) attribute these overrules equally to thermal discomfort, event length, and general inconvenience, with Wildstein et al. (2023) reporting that override rates increase with event duration.

Despite this growing body of evidence on automated flexibility, there are three well-identified gaps which make up our contributions to the literature. First, to the best of our knowledge, our paper, along with Bernard et al. (2024), is one of the first to provide quantitative estimates of the potential for HPs to deliver flexibility in real-world settings. This includes both load shedding and monetary savings, while integrating actual human behavior and preferences. Existing studies on residential flexibility are often constrained by the theoretical nature of these programs and rely heavily on methods like choice experiments for investigating stated preferences (Richter and Pollitt, 2018; Ruokamo et al., 2019; Harold et al., 2021; Yilmaz et al., 2021, 2022) or on modeling approaches. For example, modeling in Georges et al. (2017) showed that a fleet of 100 HPs can provide an average flexibility potential of 510 W per unit during "downward modulation" events. However, Good (2019) states that existing works, including models, often assume "well-informed, rational actors". As a result, they may overlook real-world human behavior, especially when households make trade-offs involving thermal comfort. As Bernard et al. (2024) note, "there is currently no causal evidence on how heat pump interventions, particularly time-of-use tariffs, affect actual energy demand in practice.". Our study expands on this by providing further empirical evidence of HP flexibility.

Second, unlike most other studies, our treatment design does not explicitly rely on user-defined temperature setpoints or standards for comfort temperatures to guide interventions[2]. Instead, we set lower temperature thresholds exogenously, allowing us to explore household reactions in a manner agnostic to specific user-defined preferences, aiming to cover a broader range of the comfort spectrum. While future flexibility schemes may indeed enable households to specify comfort boundaries ex-ante or may include algorithms to minimize comfort impact, Zhang et al. (2016) highlight that standards for comfort temperatures can be overly conservative, while Aghniaey and Lawrence (2018) suggest that households can develop habits and tolerance for wider temperature ranges over time.

Third, our setup is specifically designed to test whether participants behave consistently with how their comfort is affected by the interventions. As such, we contribute to studies on households' thermal comfort. Only recently has research on residential flexibility begun to focus on the role of thermal comfort, including in determining households' economic gains (Da Fonseca et al., 2021). Although some of the empirical works mentioned above acknowledge the role of comfort—often indicating that flexibility has no significant impact on this front—this is not the case for all studies (e.g., (Bernard et al., 2024)). Furthermore, even when thermal comfort is discussed, it is rarely used to provide insights into the behavior observed during the experiment. To our knowledge, no existing research has directly compared households' subjective perceptions of discomfort with objective discomfort metrics, as we do in our study through the use of a proxy metric for discomfort. Moreover, we contribute to the literature on manual overrules in heating flexibility programs by analyzing household-provided text data on their motivations to overrule, rather than inferring their behavior from HP measurements alone.

The remainder of this paper is structured as follows: Section 2 details the experimental design and data collection. Section 3 presents the methods used to construct and graphically represent counterfactuals of key HP consumption variables during intervention periods, as well as the regression specifications. Section 4 presents the empirical results from both perspectives—flexibility interventions and flexibility events—quantifying changes in power and energy consumption over time, estimating financial savings, and discussing the impact on household comfort. Finally, Sec-

[2]See (Da Fonseca et al., 2021) for a review of various standards and metrics used to evaluate thermal discomfort in flexibility studies, and for instance, the standard established by the American Society of Heating, Refrigerating and Air-conditioning Engineers (ASHRAE), encountered in studies like (Kaspar et al., 2024).

tion 5 summarizes key findings, concludes and provides policy recommendations for shaping the future of residential flexibility.

2. Experimental setting and data

2.1. Experimental setting

The experiment was carried out in collaboration with Energent, a Ghent-based local energy cooperative with around 2,000 members, developing renewable energy projects[3]. The cooperative selected nine participating households based on previous related and successful projects and coordinated the installation of hardware devices making it possible to remotely monitor and steer their HPs. All HPs in the sample are of the air-to-water type. Our research's target period is in the winter months, when HPs are used to heat homes. Therefore, the experiment spans two heating seasons (HS): HS1 spans November 21, 2022 to April 15, 2023, while HS2 spans October 30, 2023 to March 24, 2024. The outdoor temperatures during both heating seasons were slightly milder than average, with a mean of 7.5 °C in HS1 and 8.3 °C in HS2, compared to the historic average of 6.9 °C for the same period in Belgium (see Appendix A.1).

Temporarily deactivating HPs can help balance supply and demand in an electricity system increasingly based on renewable generation. To explore how this affects electricity consumption and household behavior, our experiment examined the impact of flexibility interventions designed to replicate future flexibility management schemes of HPs. Specifically, the flexibility interventions involved momentarily turning off HPs in our sample, until they are reactivated when one of three predefined scenarios occurs. These scenarios are:

- The participants manually overrule the intervention via an online dashboard, where they are prompted to briefly state their reasons. This overrule interrupts the ongoing intervention and any intervention scheduled within the next 24 hours.
- The indoor temperature reaches a scheduled threshold value (see below).
- The temperature of the domestic hot water tank (DHW) inside the HP falls below 40 °C. This was done to avoid any impact on the ability of households to take a hot shower or bath whenever they wanted. In our sample, three out of the nine HPs, referred to as 'decoupled' HPs, were technically capable of turning off space heating separately from sanitary hot water provision. Therefore, this scenario was only applied to the 'non-decoupled' HPs, where space heating and DHW could only be controlled together.

When one of these scenarios is met, the HP resumes normal operation, i.e., restoring the space heating setpoint to the value selected by the household (e.g., 21 °C) or going into standby if specified in a clock setting by the user.[4]

For each of the two heating seasons, a common schedule of 32 interventions was designed in advance, with interventions sent simultaneously to all HPs. Across both heating seasons, this resulted in up to 64 scheduled interventions per HP, depending on when it was added to the sample. The interventions vary by:

- The intervention's start time (2 a.m., 8 a.m., 2 p.m. or 8 p.m.)[5].
- The intervention's day of the week.
- The intervention's indoor temperature threshold value (16 °C, 17 °C, 18 °C or 19 °C).

The schedule was randomized across these dimensions and the interventions were distributed throughout the heating seasons, across both week- and weekend days. In HS1, households were informed of all 32 interventions one day in advance via text message. In HS2, notifications were sent for only a randomly selected half (16) of the 32 interventions, delivered via e-mail (presented

[3]See `https://energent.be/`.

[4]Note that our experiment used Smart Grid Ready (SGR) to communicate with the HPs, which is a standardized communication interface to either request or force HPs to turn on or off. It includes four possible commands, of which we only used "force off", instructing the HP to cease its operations. Due to limitations in the SGR standard—notably, the lack of control over the setpoint temperature to which rooms are heated—, it was impossible in our experimental setup to perform interventions where buildings are pre-heated to a higher setpoint before a HP is turned off, as done in (Centre for Net Zero and Nesta, 2023, 2024)

[5]Variation in the intervention start time ensures that we test households' responses to interventions at different times of the day, i.e., different levels of home occupation, including when the entire household is likely to be present.

in Appendix B). This randomization avoided correlations with other intervention characteristics, allowing us to test the hypothesis that households strategically reacted to receiving notification e-mails.

Appendix A.2 presents an overview of the sample composition for each heating season. In terms of sample attrition, Household 5 withdrew from the experiment near the end of January 2024, citing diminished motivation and reporting that, as energy prices normalized following the intense phase of the energy crisis during HS1, they no longer saw a compelling reason to continue participating. Additionally, Household 4 did not participate in HS2 due to technical issues, Household 7 was equipped with the proper hardware only shortly before the start of HS2, and Household 8 joined late in January 2023 (during HS2). Notably, none of the households that enrolled and had their HP integrated with the steering platform refused to participate at the last minute.

Overall, there were eight HPs in each heating season, resulting in a final sample of 287 interventions across both seasons (168 in HS1 and 119 in HS2)—somewhat lower than initially planned. This shortfall is due to several factors. First, connection losses between the HPs and the online platform prevented some interventions from taking place correctly. Also, for some HP brands, data collection relied on the manufacturer's API, which was occasionally interrupted by updates or limitations on the number of API calls allowed within a short period. Interventions with substantial missing data on core variables, such as indoor temperature or HP power consumption, were excluded from the final sample.[6] Furthermore, households turning off their HP for extended periods, such as during holidays, also contributed to the reduction in the number of completed interventions. However, the final dataset of 287 interventions is still random across the starting hour of the day, day of the week, and indoor temperature threshold, as shown in Appendix C.

A few factors may explain why our household sample size was modest. First, HP adoption is still uncommon in Belgium (Rosenow et al., 2022). Second, although many studies on residential flexibility have used smart thermostats to control heating, ventilation, and air conditioning (HVAC) units remotely[7], these devices do not directly measure HP power consumption, which is essential to fully assess flexibility potential. Moreover, Peffer et al. (2024) show that the HP's built-in proprietary thermostat system offers the precise control needed for optimal, continuous variable-speed adjustments. Therefore, instead of relying on (smart) thermostats, we sent an experiment coordinator to each household to install the necessary hardware to connect the HPs to the experiment online steering platform. Third, the novelty of flexibility programs makes it still a challenge to recruit participants, especially among those who are not typically environmentally conscious. For instance, a survey conducted in Quebec by Ouf et al. (2024) found that over 60% of respondents were "not at all familiar" with these programs. As a result, securing participation from many households with clean data proved logistically challenging and limited our sample size.

Although both the sample composition and weather conditions differed somewhat between HS1 and HS2 (see Appendix A.1), a cross-season comparison of key experimental results shows that most of our findings are robust across the two heating seasons (see Appendix D). This also suggests that any unobserved within-household changes (e.g., variations in household size or renovation works) did not significantly impact the results.

2.2. Data

2.2.1. Survey data

Before the start of the first heating season, all nine participating households were invited to complete a detailed survey. The survey explained the upcoming flexibility interventions to their HPs and gathered data on their socio-demographic and housing characteristics. It also investigated attitudinal and behavioral factors, including stated preferences for comfort temperatures during and outside the flexibility interventions. One household did not complete the entire survey.

Table 1 presents the key characteristics of the participating households.[8] The findings indicate that our sample is not representative of the wider population in terms of socio-demographic, housing and likely pro-environmental characteristics. The participants can be viewed as generally larger, more educated and better-off households than the Belgium average. They live in bigger, newer

[6] Long interventions, especially on decoupled HPs, are likely underrepresented in the final sample as they are more prone to connection losses or missing data, resulting in their exclusion from the final dataset.

[7] See: (Sarran et al., 2021; Tomat et al., 2022; Wildstein et al., 2023; Blonz et al., 2025; Fu et al., 2024) (all based on the Ecobee "Donate Your Data" program), or (Kane et al., 2024).

[8] In addition, the survey also revealed that all participants are equipped with solar panels, and the yearly electricity consumption reported by seven respondents averages about 5286 kWh (with a standard deviation $\sigma = 1629$ kWh, i.e., a moderate respondent-variability of $CV = 0.31$).

and more energy-efficient[9] homes, which they also predominantly own rather than rent. They demonstrate a good understanding and awareness of the concepts related to the experiment, and can be considered as environmentally conscious. This selection bias is anticipated, as the limited HP ownership in Belgium (Rosenow et al., 2022), likely attracts a demographic that differs from the general population.

The small sample limits how much we can generalize about participant heterogeneity, but some variability is observed across key characteristics. Household size shows moderate variation (standard deviation $\sigma = 0.92$, and the coefficient of variation[10]: CV = 0.27), while the number of children under six is more variable ($\sigma = 0.74$, CV = 1.19), likely reflecting differences in household age composition. In contrast, all behavioral metrics exhibit low variability: CV = 0.09 for understanding flexibility-related concepts, CV = 0.15 for pro-environmental behavior, and CV = 0.09 for engagement in electricity-saving practices. Despite some sociodemographic differences, participants show homogeneous attitudes toward energy use and sustainability, and share a strong understanding of flexibility concepts and consistent pro-environmental behavior and energy-saving practices.

Table 1: Participants' characteristics

	Total respondents	Sample statistics	National average	
Household characteristics				
Mean household size (persons)	8	3.38	2.25	[a]
Mean number of children < 6 years old	8	.63	—	
Share of respondents and/or partner employed full time	8	100%	76.5%	[b]
Share of respondents and/or partner holding a university degree	8	100%	50.0%	[c]
Share of households with total monthly income > €5,000	8	62.5%	—	[d]
Participants' housing characteristics				
Share residing in urban or suburban environment	8	87.5%	85.5%	[e]
Share residing in a semi-detached house	8	37.5%	42.1%	[f]
Share residing in an apartment	8	0%	22.9%	[g]
Share residing in a home surface > 150 m^2	8	50%	—	[h]
Share residing in a home built after 2006	8	37.5%	12.5%	[i]
Share home has been energy-retrofitted	8	75%	—	
Share energy performance certificate rated A	6	100%	—	[j]
Behavioral metrics				
Understanding of flexibility-related concepts [*]	9	3.39	—	
Pro-environmental behavior [**]	8	4.59	—	
Frequency of engagement in electricity-savings practices [***]	8	4.06	—	

[a] (Statbel, 2024b); [b] Of the population aged over 18 (Eurostat, 2024b).; [c] Of the 25-34 years old (Statbel, 2024a).; [d] The average household disposable income in Belgium is $47,446 a year, i.e., about €3,760/month in 2022 (OECD, 2024). [e,f,g] (Eurostat, 2024c); [h] The average home size is 145.5 m^2 (Eurostat, 2024a).; [i] (Statbel, 2024c); [j] The average energy performance in non-retrofitted homes built after 2006 is between 160 (apartments) and 190 (single-family houses) kWh/m^2.year in primary energy use. Both correspond to a rating of B (See: `https://www.epcwaarde.be/epc-score/`, 21 August 2024).
[*] Scale 1-4 (*Never heard of it - I know a lot about it*) based on (Li et al., 2017); [**] Scale 1-5 (*Strongly disagree - Strongly agree*) and items based on (Bauwens and Devine-Wright, 2018); [***] Scale 1-5 (*Never - Always*) and items based on (Herabadi et al., 2021). Appendix E presents the full list of statements.

2.2.2. Heat pump data

The HP sample includes two brands: Daikin and Viessmann. Each HP has been equipped for the experiment with a piece of hardware that provides us with a rich panel dataset at the 5-minute level on the power consumption of the HP, the indoor temperature inside the house (along with the setpoint temperature selected by the household), the outdoor temperature (measured by a sensor placed outside, away from direct exposure to the Sun), the temperature of the water in the DHW tank, and an indicator showing whether the HP is currently blocked by an ongoing intervention.

[9] In Belgium's energy performance certificate rating system, a rating of A means a primary energy use of less than 100 kWh/m^2 per year. The rating ranges from A (best) to F (worst).
[10] The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean.

Participants were offered the opportunity to monitor these variables (except for the intervention indicator) on an online dashboard developed by the project partner EnergieID, a Belgian company specialized in developing tools to monitor energy consumption[11].

3. Methods

3.1. Intervention treatment effect and counterfactual outcome

While our sample lacks a separate control group, the random scheduling of interventions (see Appendix C) supports using non-intervention periods to construct a counterfactual. Therefore, we estimate intervention effects on outcomes (e.g., HP power consumption or indoor temperature) by comparing HP operation during interventions to its operation during non-intervention periods. To ensure the comparison is sound, we compare observations within the same household, time-of-day bin, and outdoor temperature category. We adopt two counterfactual construction approaches, depending on whether outcomes are analyzed with respect to absolute time (e.g., time of day) or relative time (e.g., time elapsed since intervention start or stop).

Further validation of the counterfactuals is provided in Appendix F, where we show that our chosen counterfactual outperforms simpler models and that residual variation is largely attributable to random noise rather than systematic bias.

Direct comparison of daily profiles.

We begin by defining the average treatment effect on the treated[12] (ATT) for the intervention, which quantifies the causal effect of the intervention on the outcome y. For each household h, time-of-day bin t, and outdoor temperature category T, the ATT is defined as:

$$\begin{aligned} ATT_{h,t,T} &= E[y^1_{h,t,T} - y^0_{h,t,T} \mid d_{h,t,T} = 1] \\ &= E[y^1_{h,t,T} \mid d_{h,t,T} = 1] - E[y^0_{h,t,T} \mid d_{h,t,T} = 1]. \end{aligned} \tag{1}$$

where $d_{h,t,T}$ denotes the treatment status of household h at time t (equal to 1 if an intervention is ongoing and 0 otherwise). A straightforward estimate for $E[y^1_{h,t,T} \mid d_{h,t,T} = 1]$ is given by the average outcome during interventions:

$$\bar{y}^1_{h,t,T} = \frac{1}{n(S^1_{y,h,t,T})} \sum_{\tilde{y} \in S^1_{y,h,t,T}} \tilde{y}, \tag{2}$$

where $S^1_{y,h,t,T}$ is the set of all observations of y for household h, time-of-day bin t, and temperature category T during interventions, and $n(S^1_{y,h,t,T})$ is the number of such observations.

In contrast, the counterfactual mean $E[y^0_{h,t,T} \mid d_{h,t,T} = 1]$ is not directly observed. However, using the randomization of intervention schedules, we assume that the intervention is independent of the potential outcomes. This assumption allows us to pair treated observations with comparable non-intervention data from the same household, time-of-day bin, and outdoor temperature category. Specifically, we first classify each day into four temperature categories based on the average outdoor temperature: (i) ≤ 3°C, (ii) > 3°C to ≤ 6°C, (iii) > 6°C to ≤ 9°C, and (iv) > 9°C. Second, we estimate the counterfactual outcome as:

$$\bar{y}^0_{h,t,T} = \frac{1}{n(S^0_{y,h,t,T})} \sum_{\tilde{y} \in S^0_{y,h,t,T}} \tilde{y}, \tag{3}$$

where $S^0_{y,h,t,T}$ is the set of all non-intervention observations for household h, time-of-day bin t, and temperature category T. To avoid bias in the counterfactual, observations too close to an intervention (within 20 minutes before[13] or 16 hours after[14]) are first removed from the set $S^0_{y,h,t,T}$.

As a result, the ATT for household h, time-of-day bin t, and temperature category T is given by eq. (4).

$$ATT_{h,t,T} = \bar{y}^1_{h,t,T} - \bar{y}^0_{h,t,T}. \tag{4}$$

This estimate can then be averaged over t and T to obtain the household-level ATT.

[11] See `https://www.energyid.eu/en`.

[12] Because our sample consists only of treated households (whose counterfactual outcomes are constructed from their own non-intervention data since there is no control group), and given that they are not representative of the general population, we identify the ATT.

[13] Heat pumps typically respond to the blocking signal within a few seconds, but this margin ensures minimum bias.

[14] This margin ensures that any potential rebound in power consumption following an intervention is not captured by the counterfactual.

Alignment of counterfactual outcomes on relative time.

When outcomes are analyzed relative to the intervention (e.g., time elapsed since start or end rather than by clock time), each observation is assigned its corresponding relative time, denoted t_{rel}. This allows consistent comparison of treatment effect dynamics across interventions, irrespective of when they occur. It is especially useful for assessing the overall effect of a unified intervention signal on all HPs, as t_{rel} captures the common response of the HP fleet, even if some units remain blocked while others resume normal operation.

Specifically, for any given t_{rel}, we gather all observations of outcome y, occurring at that relative time and across households and interventions, and denote this set by $S_{y,t_{\text{rel}}}$. The observed outcome at t_{rel}, averaged over all households, time-of-day bins, and outdoor temperature categories, is then given by:

$$\bar{y}(t_{\text{rel}}) = \frac{1}{n(S_{y,t_{\text{rel}}})} \sum_{\tilde{y} \in S_{y,t_{\text{rel}}}} \tilde{y}, \tag{5}$$

where $n(S_{y,t_{\text{rel}}})$ denotes the number of observations in $S_{y,t_{\text{rel}}}$. Note that we do not restrict $S_{y,t_{\text{rel}}}$ to a particular treatment status because, at a given t_{rel}, some HPs may still be blocked while others have resumed operation.

Similarly, to construct the aligned counterfactual outcome at t_{rel}, we proceed in two steps. First, each observation (regardless of its treatment status) is paired with the previously computed counterfactual outcome $\bar{y}^0_{h,t,T}$, defined as the sample average from non-intervention observations for the same household h, time-of-day bin t, and outdoor temperature category T (given by eq. (3)). This results, for each value of t_{rel}, in a collection of these previously computed counterfactual outcomes, denoted by the set $S_{\bar{y}^0_{h,t,T},t_{rel}}$. Second, the aligned counterfactual outcome at relative time t_{rel} is defined as the average over this set:

$$\bar{y}^0(t_{\text{rel}}) = \frac{1}{n(S_{\bar{y}^0_{h,t,T},t_{\text{rel}}})} \sum_{\tilde{y} \in S_{\bar{y}^0_{h,t,T},t_{\text{rel}}}} \tilde{y}, \tag{6}$$

where $n(S_{\bar{y}^0_{h,t,T},t_{\text{rel}}})$ denotes the number of observations in this set.

Given the limited number of clusters in the data (nine households), bootstrapped standard errors that fully account for both intra-household and inter-household variability would lack practical interpretability. Therefore, we assume that observations in $S_{\bar{y}^0_{h,t,T},t_{\text{rel}}}$ are independent and estimate the standard errors using within-set variability:

$$SE\Big(\bar{y}^0(t_{\text{rel}})\Big) = \frac{SD(S_{\bar{y}^0_{h,t,T},t_{\text{rel}}})}{\sqrt{n(S_{\bar{y}^0_{h,t,T},t_{\text{rel}}})}}, \tag{7}$$

where $SD(S_{\bar{y}^0_{h,t,T},t_{\text{rel}}})$ denotes the standard deviation in $S_{\bar{y}^0_{h,t,T},t_{\text{rel}}}$. While this may underestimate the true variability, it provides a tractable approximation of the uncertainty in the aligned counterfactual outcome.

3.2. Regression specifications

Robust intervention treatment effect from fixed-effects regression.

While the ATT estimate in eq. (4) is theoretically robust, it can be affected by data imbalance. In particular, households with more intervention observations (i.e., longer or more frequent interventions) may disproportionately influence the average treated outcome, $\bar{y}^1$ and, similarly, the estimate of the counterfactual outcome, $\bar{y}^0$, may be largely driven by households with fewer or shorter interventions. To address this potential bias, we obtain a more robust ATT estimate of interventions on outcome y (e.g., indoor temperature or heat pump power consumption), averaged over all households, time-of-day bins, and temperature categories via the following regression model that controls for household-specific characteristics invariant across interventions:

$$y_{h,t} = \alpha_h + \beta \cdot I_{h,t} + \varepsilon_{h,t}, \tag{8}$$

where y is observed at the 5-min-of-day-level, α_h captures the household fixed effects (FE) and $\varepsilon_{h,t}$ is the error term. The indicator $I_{h,t}$ is equal to 1 if the observation at time t for household h corresponds to an intervention (and 0 otherwise). Given the randomization of the intervention schedule, we identify β as the ATT on y and interpret it as the average within-household change in y caused by interventions over their entire duration, controlling for time-invariant household characteristics.

We estimate the model in eq. (8) using linear regression. To address the small number of clusters (nine households), we estimate wild cluster bootstrapped standard errors (with 100,000 repetitions) at the household level to obtain p-values from the empirical distribution of the bootstrapped estimates (MacKinnon et al., 2023; Cameron et al., 2008).

Regression of intervention duration.

A key aspect of flexibility interventions is their duration, before they are automatically or manually terminated. Both endogenous factors (e.g., household comfort preferences) and exogenous factors (e.g., weather conditions, time of day when the intervention is initiated) may influence duration. To estimate the effect of these factors on duration $d_{i,h}$ (in hours) of intervention i on household h, we estimate the following regression specification:

$$\begin{aligned} d_{i,h} =& \beta_1 \cdot D_{\text{notif},i,h} + \beta_2 \cdot T^0_{\text{in},i,h} + \beta_3 \cdot \min(T^{\leq 5\text{h}}_{\text{out},i,h}) + \beta_4 \cdot \Delta T^{\text{daily}}_{\text{out},i,h} \\ &+ \beta_5 \cdot \text{TOD}_{\text{2AM},i,h} + \beta_6 \cdot \text{TOD}_{\text{8AM},i,h} + \beta_7 \cdot \text{TOD}_{\text{8PM},i,h} \\ &+ \mathbb{1}\{\text{FE}=0\} \cdot \Big(\beta_8 \cdot T^0_{\text{DHW},i,h} \cdot \delta_{\text{decoupled}} + \beta_9 \cdot T^0_{\text{DHW},i,h} \cdot (1-\delta_{\text{decoupled}}) \cdot \delta_{\{T^0_{\text{DHW},i,h} \geq 40^\circ\text{C}\}} \Big) \\ &+ \mathbb{1}\{\text{FE}=1\} \cdot \beta_{10} \cdot T^0_{\text{DHW},i,h} + \mathbb{1}\{\text{FE}=1\} \cdot \alpha_h + \mathbb{1}\{\text{FE}=0\} \cdot \beta_0 + \varepsilon_{i,h}, \end{aligned} \tag{9}$$

where $D_{\text{notif},i,h}$ is a dummy variable indicating whether intervention i on household h was pre-notified; $T^0_{\text{in},i,h}$ and $T^0_{\text{DHW},i,h}$ are the indoor and DHW tank temperatures at the start of the intervention[15]; $\Delta T^{\text{daily}}_{\text{out},i,h} = T^{\max}_{\text{out},i,h} - T^{\min}_{\text{out},i,h}$ reflects the variability in outdoor temperature during the intervention day; and $\text{TOD}_{\text{2AM},i,h}$, $\text{TOD}_{\text{8AM},i,h}$, and $\text{TOD}_{\text{8PM},i,h}$ are dummies for the intervention start time (with 2 p.m. as the base category). The dynamic component is captured by $\min\left(T^{\leq 5\text{h}}_{\text{out},i,h}\right)$, the minimum outdoor temperature observed within five hours after the intervention starts[16].

We distinguish between models with and without household FE by using the indicator functions $\mathbb{1}\{\text{FE}=0\}$ (for models without FE) and $\mathbb{1}\{\text{FE}=1\}$. In models that include household FE (α_h in eq. (9)), the estimates capture only within-household variation, by controlling for household characteristics that are invariant across interventions, such as insulation levels and comfort temperature preferences. Because the decoupled dummy $\delta_{\text{decoupled}}$ is invariant within a household, the DHW temperature at the start of the intervention enters the FE specification as a single parameter: $T^0_{\text{DHW},i,h}$. Note that models with household FE are estimated without an overall constant term.

The model given by eq. (9) tests the following hypotheses:

(i) Day-ahead intervention notification, $D_{\text{notif},i,h}$, shortens interventions either by prompting preemptive overruling or by increasing overall intervention salience.

(ii) Higher initial indoor temperature, $T^0_{\text{in},i,h}$, and DHW tank temperature, $T^0_{\text{DHW},i,h}$, lengthen interventions by reflecting greater stored energy in the home.

(iii) A higher minimum outdoor temperature, $\min\left(T^{\leq 5\text{h}}_{\text{out},i,h}\right)$, lengthens interventions by slowing the depletion of stored energy via heat exchanges.

(iv) Greater outdoor temperature variability, $\Delta T^{\text{daily}}_{\text{out},i,h}$, influences intervention duration by potentially increasing indoor temperature fluctuations and prompting households to overrule interventions to maintain comfort.

(v) The time of day at which an intervention starts (captured by the TOD dummies) affects its duration.

This model is estimated using linear regression using wild cluster bootstrapped standard errors (100,000 repetitions) at the household level to account for the small number of clusters and potential error correlation within clusters. We report results for both specifications, with and without household FE.

[15] In the non-FE specification ($\mathbb{1}\{\text{FE}=0\}$), the parameter $T^0_{\text{DHW},i,h}$ is split into two terms to distinguish between decoupled HPs (for which interventions turn off only the ambient heating) and non-decoupled HPs (for which interventions disable both ambient and DHW heating).

[16] This specification has been found to yield slightly higher adjusted R^2 values than alternative parametrizations (e.g., using the average outdoor temperature during the intervention).

3.3. Graphical representation

To facilitate interpretation of the figures, the results of the empirical analysis, including the mean estimates across the sample and the confidence intervals constructed using the approach described above, are smoothed in the figures using local polynomial regressions. With the polynomial degree set to 0 (identified as the optimal value), this approach corresponds to a Nadayara-Watson regression (see (Nadaraya, 1964; Watson, 1964)). The bandwidths are set to their optimal cross-validated values, and the weights are determined by the Epanechnikov kernel. This produces a locally weighted smoothing of the sample averages, where the kernel assigns weights within the bandwidth.

4. Results

4.1. Effects of flexibility interventions on individual heat pumps

In this section, we analyze the key factors summarizing the impact of flexibility interventions at the individual HP level. We examine why and when interventions ended, quantify their impact on both ambient temperature and HP power usage, and characterize the post-intervention increase in electricity usage.

4.1.1. Duration of interventions

Table 2 shows a breakdown of the duration d (in hours) of the 287 studied interventions for each of the three reasons for stopping an intervention. The most common scenario was the automatic stop triggered by the DHW temperature threshold amongst non-decoupled HPs, accounting for 201 of all stops. Fifty-four interventions were manually stopped by households, including eight that were preemptively overruled. Thirty-two interventions ended when the indoor temperature threshold was reached. Specifically, twenty-four stopped at the 19 °C threshold, six at the 18 °C threshold, and only two at the 17 °C threshold. Notably, the 16 °C threshold was never reached. This indicates that households avoided letting their homes cool down too much, as shown by the histogram of indoor temperatures during non-intervention periods presented in Appendix A.3.

Table 2: Intervention durations d (in hours) for each reason for stopping an intervention

Reason for intervention stop	N	Percent	Average d	95% CI	Min	Median	Max
DHW temperature threshold	201	70.0%	12.68	11.21, 14.15	0	11.51	82.08
Manual overrule	54	18.8%	15.51	9.82, 21.20	0	11.49	135.80
Indoor temperature threshold	32	11.2%	9.30	5.35, 13.26	0	4.71	34.67
Full sample	287	100%	12.84	11.30, 14.37	0	11.17	135.80

Table 2 shows that interventions in the full sample lasted 12.84 hours on average. Interventions ending by a manual overrule last the longest, while interventions ending because of the indoor temperature threshold last the shortest. However, three t-tests with unequal variances show that the differences in average intervention duration across the stopping scenarios are not significant at the 5% level.

These averages conceal significant variation among individual interventions, as shown by the histogram in Figure 1. Approximately 15.3% of interventions lasted less than one hour and 12.5% of interventions over 24 hours.

To study the source of this variation, we estimate the model in eq. (9). Table 3 presents the results for four models, ranging from a parsimonious specification without household FE to the full model with household FE.

The results reveal consistent general trends across models. Although the significance and magnitude of the estimates vary across specifications, they show expected signs. In particular, intervention duration increases with the indoor temperature at the start, the DHW temperature at the start (both of which reflect the home's thermal energy available for the intervention—supporting Hypothesis (ii)) and the minimum temperature within five hours after the start (which affects the rate at which it depletes—supporting Hypothesis (iii)). Another key general trend is that the dummy variable for day-ahead notification of the intervention is insignificant in all models where it is included. This result fails to support Hypothesis (i) and suggests that preemptive overrules by households in response to notifications were minimal, or that households did not consistently increase their thermostat temperature in preparation for an intervention. As a result, this variable is not relevant and excluded from further analyses. Similarly, daily outdoor temperature variability does not significantly affect intervention duration at the 5% level, failing to support Hypothesis (iv). However, evidence at the 10% effect suggests that each 1 °C increase in variability (i.e.,

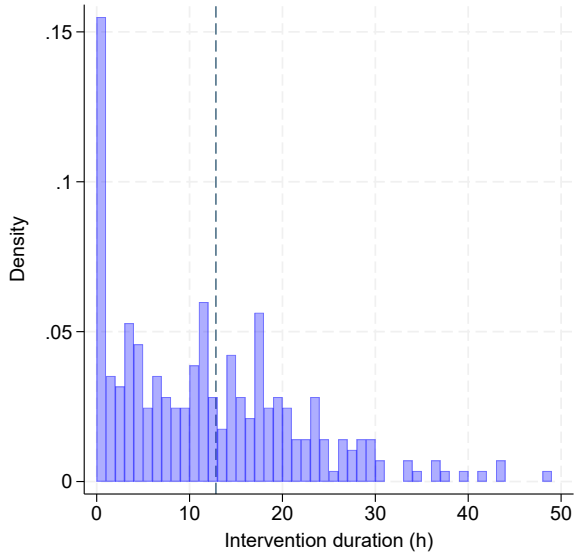


Figure 1: Histogram of the intervention durations d (in hours) of flexibility interventions ($N = 287$). The dotted line indicates the mean duration. The histogram is truncated at $d = 50$ hours to exclude outliers.

the difference between daily maximum and minimum outdoor temperatures) extends interventions by approximately 9–11 minutes[17].

Estimates remain robust within the same type of models. Among the non-FE models, Model (2) achieves the highest adjusted R^2. It shows that the indoor temperature at the start has the largest marginal effect: a one-degree rise in T^0_{in} increases the intervention by 2.1 hours on average. Outdoor temperature also has a positive effect, with each additional degree in $\min(T^{\leq 5\text{h}}_{\text{out}})$ increasing the duration by about 35 minutes. The effect of a one-degree increase in DHW temperature at the start is not significant for non-decoupled HPs[18] but contributes an average 21-minute increase in intervention duration on decoupled HPs.

Among the FE models, our preferred specification is Model (4) which adds the TOD dummies, with a drop of less than 2% in adjusted R^2 compared to Model (3). The marginal effect of the minimum outdoor temperature is similar to the non-FE models: a one-degree increase in $\min(T^{\leq 5\text{h}}_{\text{out}})$ extends the intervention by approximately 41 minutes on average. However, adding household FE further changes the estimates for parameters typically affected by household HP usage patterns. For example, a one-degree increase in the initial DHW temperature extends the intervention by 34 minutes on average within a household. The estimate for the indoor temperature at the start is borderline significant ($p = 0.09$), likely because there is little variation in indoor temperature within households, as most avoid letting their homes get too cold, as noted earlier. Still, within a household, a one-degree increase in T^0_{in} extends the intervention by nearly 2.7 hours at the 10% level. Interestingly, the time of day when the intervention starts does not significantly affect its duration at the 5% level. However, Model (4) shows evidence at the 10% level that interventions starting at 8 p.m. are shorter than those starting at 2 p.m. by an average 2.8 hours. This may be because interventions starting in the evening do not benefit from cooking activities or solar irradiance (even after controlling for outdoor temperature), which contribute to reheating the home during interventions initiated over the afternoon, thereby extending their duration. As a result, Hypothesis (v) is not supported at the 5% level, possibly due to data constraints, while evidence at the 10% level is mixed across different start times.

[17] This finding at the 10% level is consistent with the observation that variability is driven primarily by the daily maximum outdoor temperature, which extends interventions, rather than by the daily minimum. The correlation coefficients are found as follows: $\rho\left(\Delta T^{\text{daily}}_{\text{out},i,h}, T^{\max}_{\text{out},i,h}\right) = 0.66$ and $\rho\left(\Delta T^{\text{daily}}_{\text{out},i,h}, T^{\min}_{\text{out},i,h}\right) = -0.26$, both significant at the 1% level.

[18] It should be noted that the variation in non-zero observations for $T^0_{\text{DHW}} \cdot (1 - \delta_{\text{decoupled}}) \cdot \delta_{\{T^0_{\text{DHW}} \geq 40°\text{C}\}}$ is limited, where the 95% CI ranges narrowly from 46.97 to 47.83 °C (213 observations). The 95% CI for $T^0_{\text{DHW}} \cdot \delta_{\text{decoupled}}$ spans 39.94 to 42.97 °C (50 observations).

Table 3: Linear regression results for intervention duration (in hours)

	(1)	(2)	(3)	(4)
Pre-notification indicator		-2.969	-1.830	-2.080
		(0.660)	(0.748)	(0.726)
Initial indoor temperature (°C)	1.945**	2.107**	2.578	2.738+
	(0.008)	(< 0.01)	(0.124)	(0.089)
Min. outdoor temp. (within 5h, °C)	0.560*	0.585*	0.619*	0.687*
	(0.031)	(0.035)	(0.018)	(0.016)
Daily out. temp. variability (°C)	0.148+	0.170+	0.187+	0.195
	(0.054)	(0.090)	(0.081)	(0.228)
Initial DHW temp. (decoupled, °C)	0.354*	0.345**		
	(0.016)	(0.008)		
Initial DHW temp. (non-dec., ≥ 40°C)	0.197+	0.184		
	(0.070)	(0.133)		
Initial DHW temp. (overall, °C)			0.579+	0.573*
			(0.063)	(0.035)
2 AM start dummy				-1.203
				(0.414)
8 AM start dummy				-0.545
				(0.786)
8 PM start dummy				-2.781+
				(0.096)
Constant	-42.101**	-42.837**		
	(< 0.01)	(< 0.01)		
Household-FE	No	No	Yes	Yes
Adj. R-Square	0.137	0.142	0.193	0.190
N observations	286	286	286	286

Linear regression estimates. The unit of observation is the intervention. The full model is specified in eq. (9). Models (1) and (2) are estimated without household fixed effects (FE), whereas Models (3) and (4) include them. The start time dummies use 2 p.m. as the reference category. P-values (in parentheses) are derived from wild cluster bootstrapped standard errors (100,000 repetitions) clustered at the household level (nine clusters in total). $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$. 'DHW tank' refers to the domestic hot water tank within the heat pump (HP). For decoupled HPs, interventions turn off only the heating (allowing the HP to reheat its DHW tank), whereas for non-decoupled HPs, interventions disable all heating services and automatically stop when the DHW temperature falls below 40°C.

4.1.2. Reduction in indoor temperature during interventions

Figure 2 (left panel) shows the average indoor temperature daily profiles during and outside intervention periods, across all HPs. During non-intervention periods, indoor temperature remains stable throughout the day, rising from about 20.4 °C in the morning to 20.8 °C by late afternoon, averaging 20.59 °C over the day. Interventions reduce this daily average to 20.43 °C, i.e., a small but statistically significant difference of around 0.15 °C at the 1% level (95% CI: 0.14, 0.16 °C), as shown by a t-test for equality of means with unequal variances.

However, this result does not account for household-specific characteristics or data imbalance. To address this, we estimate the FE regression model in eq. (8) with $y_{h,t} = T_{in,h,t}$ (the indoor temperature in household h at time t), resulting in a larger ATT of -0.38°C (95% CI: -0.66, -0.10°C; $p = 0.02$). This represents the average within-household reduction in indoor temperature due to interventions over their entire duration, after controlling for household characteristics invariant across interventions.

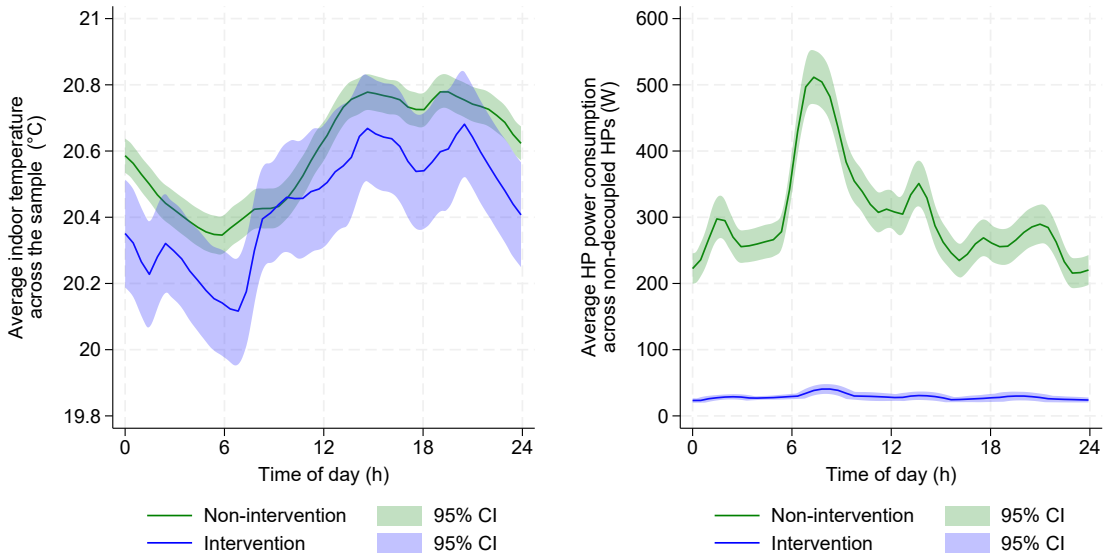


Figure 2: Daily profiles of average indoor temperature (left panel, in °C) and heat pump (HP) power (right panel, in W; non-decoupled units only) during and outside intervention periods, averaged across HPs and heating seasons. The profiles are smoothed using local polynomial regression of degree 0 for the mean and confidence intervals. Standard errors reflect the variability of the mean in 5-min-of-day bins, assuming independence among observations.

4.1.3. Reduction in power consumption during interventions

Figure 2 (right panel) shows the average reduction in HP power consumption during intervention periods, a key outcome of flexibility programs, for non-decoupled units only[19]. On non-intervention days, households typically begin heating their homes around 6 a.m., resulting in a peak power of around 500 W at 7 a.m., due to the demand for both space and domestic water heating. Throughout the day, HP power consumption averages 325 W.

During interventions, HP power does not become zero, but is reduced by an average 271 W (statistically significant at the 1% level; 95% CI: 268, 274 W), based on a two-sample t-test for equality of means with unequal variances), resulting in an average consumption of 53 W over the entire HP sample, i.e., a reduction of 84%. This residual power consumption supports the HP's essential functions, including maintaining electrical circuits, circulatory pumps, and its connection to the online control platform.

Similarly to Section 4.1.2, a robust estimate of the ATT on power consumption is obtained from the FE regression model in eq. (8) with $y_{h,t} = P_{h,t}$ (the HP power consumption in household h at time t). This yields an ATT of -292 W (95% CI: $-390, -200$ W; $p < 0.01$), representing the average within-household reduction in HP power consumption due to interventions, averaged over their duration.

Appendix A.5 shows that HP power consumption is negatively correlated with outdoor temperature, as one expects, with a Pearson correlation of -0.25 ($p < 0.01$) estimated across the entire sample. The profiles for different outdoor temperature ranges indicate that, as outdoor temperatures increase, power peaks shorten, and overall power consumption decreases, while the overall shape of the daily profile (in terms of overall peak timing and pattern) remains consistent.

4.1.4. Reduction in electricity consumption during interventions

The amount of electricity consumption reduction during an intervention is calculated as:

$$\Delta E_{i,h}^{during} = \int_{t_0}^{t_f} \left(P_{cf,i,h}(t) - P_{obs,i,h}(t)\right) \mathrm{d}t \tag{10}$$

Where $\Delta E_{i,h}^{during}$ represents the decrease in electricity consumption (in kWh) observed in household h during intervention i, from its start at t_0 to its end at t_f; $P_{cf,i,h}(t)$ is the household-specific counterfactual power consumption (as defined in eq. (3)) during intervention i if no intervention had occurred, and $P_{obs,i,h}(t)$ is the observed (residual) consumption. The histogram of $\Delta E_{i,h}^{during}$

[19] Appendix A.4 shows power profiles during interventions for the entire sample, including both decoupled and non-decoupled HPs. Decoupled units occasionally exhibit multi-kW power spikes to heat water for DHW usage or to above 60 °C to prevent the spread of Legionella bacteria, but these do not contribute to heating the ambient space.

for all interventions is shown in Figure 3. On average, 3.22 kWh (95% CI[20]: 2.81, 3.63 kWh) of electricity was saved during the interventions. As expected, it is correlated with the intervention duration d, with a Pearson coefficient of 0.54 (significant at the 1% level).

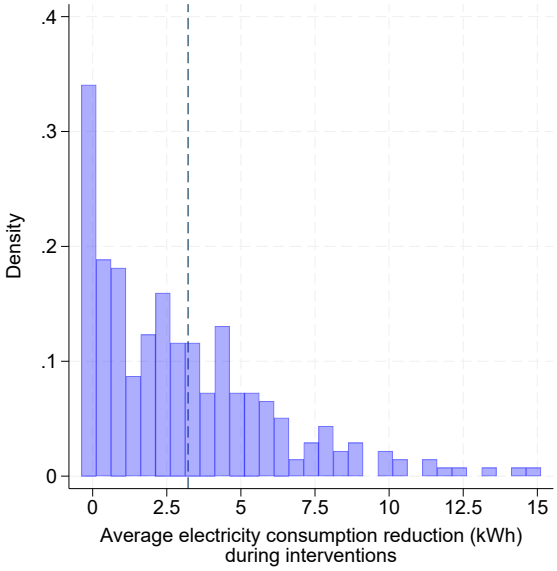


Figure 3: Histogram of the average decrease in electricity consumption during each intervention ($\Delta E_{i,h}^{during}$, in kWh), defined in eq. (10) and computed using the average daily heat pump consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual (defined in eq. (3)). The dotted line indicates the mean energy consumption reduction. The histogram is truncated at 15 kWh to exclude outliers.

4.1.5. Increase in electricity consumption after the interventions

The estimate $\Delta E_{i,h}^{during}$ reflects the reduction in energy consumption during an intervention only. However, when the intervention ends, the HP resumes operation from a lower indoor and/or DHW temperature (depending on whether the unit is decoupled) than the household's typical thermostat setpoint for that time of day. This leads to increased power consumption during the post-intervention period, as the HP works to restore the home and/or the DHW tank temperatures to the setpoint. This phenomenon, often referred to as the rebound peak in the literature (Muratori et al., 2014; Ludwig and Winzer, 2022; Dewangan et al., 2022; Tomat et al., 2022), is calculated for each intervention as a function of t, the time relative to intervention stop:

$$\Delta P_{i,h}(t) = P_{obs,i,h}(t) - P_{cf,i,h}(t) \tag{11}$$

A positive value for $\Delta P_{i,h}(t)$ indicates that the HP consumption (in Watts) is higher than it would have been, had no intervention occurred. Additionally, this extra power consumption can be integrated over time to quantify the additional electricity consumption at a period Δt after the intervention stop :

$$\Delta E_{i,h}^{after}(\Delta t) = \int_{t_f}^{t_f+\Delta t} \left(P_{obs,i,h}(t) - P_{cf,i,h}(t)\right) \mathrm{d}t \tag{12}$$

Figure 4 illustrates the rebound peak in terms of additional power consumption (left panel, in Watts) estimated via eq. (11) and energy consumption (right panel, in kWh) estimated via eq. (12) with $\Delta t = 16h$ after intervention stop. While in principle the optimal Δt for capturing rebound dynamics should vary based on the specific temperature conditions of each intervention, we use a fixed 16-hour window for all interventions in our analysis. As shown in Appendix A.6, where we plot the equivalent of Figure 4 with the window extended to 40 hours, most of the rebound dynamics occur within the first hours, with only minor changes after 16 hours, justifying our approach.

We observe that most of the rebound effect occurs within the first eight hours after the intervention, with approximately one hour required for the rebound to settle in (as shown in the left

[20]These confidence intervals reflect variability across interventions but likely underestimate the true variability, as they do not account for intra-household correlation or variability around the mean counterfactual power level at each time point.

panel). On average, the excess power consumption jumps to around 700 W just after the intervention stops, with an average of 607 W (95% CI[21]: 571, 642 W) in the first post-intervention hour, but quickly drops, with an average of 270 W (95% CI: 260, 279 W) within the first eight hours. The average electricity consumption rebound within the first 16 hours is 2.40 kWh (95% CI: 1.96, 2.85 kWh).

We conclude that an intervention consists of two phases. During the first phase, consumption decreases by an average of about 3.2 kWh. In the second phase, the rebound phase following the intervention stop, the consumption increases by approximately 2.4 kWh on average. Overall, the net effect of an intervention, accounting for the rebound over the 16 hours post-stop, is an average reduction of approximately 0.8 kWh in HP electricity consumption, as measured 16 hours after the return to normal operation.

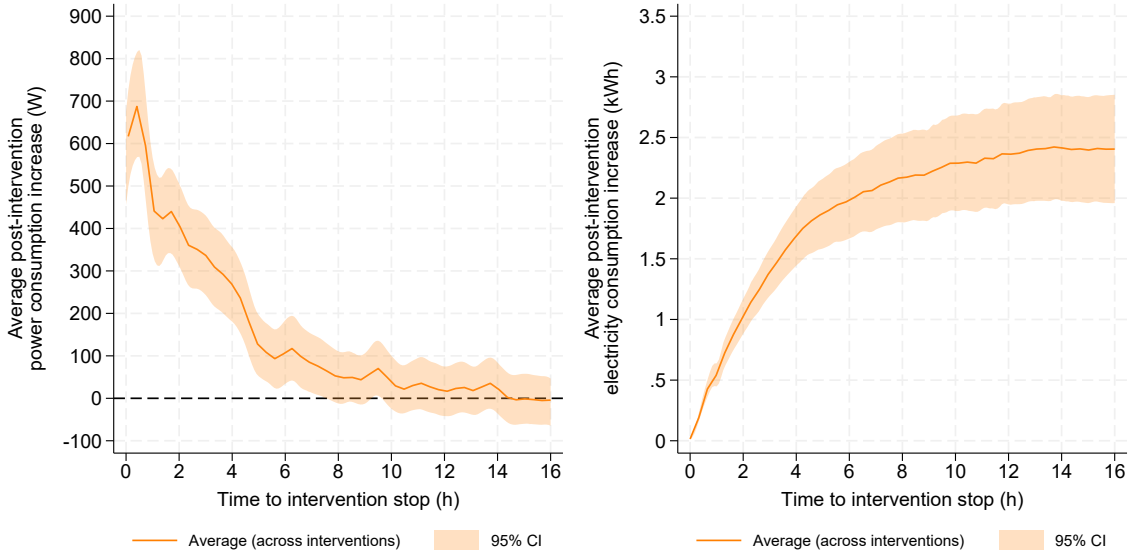


Figure 4: Average increase in heat pump (HP) power consumption during the post-intervention rebound period ($\Delta P_{i,h}(t)$ in W, left panel) and the corresponding energy consumption ($\Delta E_{i,h}^{after}(\Delta t)$ in kWh, right panel) across all interventions. The averages are defined respectively in eq. (11) (left) and eq. (12) (right) and computed using the average daily HP consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual, aligned by time to intervention stop (see eq. (6)). Standard errors reflect the variability of the means in 5-min-to-intervention-stop bins, assuming independence among observations (eq. (7)). The means and confidence intervals are smoothed using local polynomial regression of degree 0.

To study the factors influencing rebound energy consumption, Appendix G presents a regression analysis of $\Delta E_{i,h}^{after}$ at 16 hours post-intervention, with outdoor conditions during the rebound period and indoor temperature at intervention stop as regressors. The model is estimated with wild cluster bootstrap to address the small number of clusters. Results show that rebound consumption depends not on the indoor temperature at intervention stop itself, but on the difference between the indoor temperature and the setpoint temperature[22]. In Model (3) with household FE, each 1 °C increase in this difference is associated with an average increase in rebound consumption of 0.80 kWh within a household at 16 hours post-intervention (borderline significant, $p = 0.06$). Additionally, a 1 °C rise in the average outdoor temperature over the 16 hours post-stop period reduces rebound by 0.28 kWh (significant at the 1% level). Notably, the time of day when the intervention ends has no significant effect once these factors are controlled for.

4.2. Effects of flexibility events on a fleet of heat pumps

For a flexibility program operator, understanding individual HP responses to interventions is secondary to assessing the behavior of an entire fleet of flexible HPs. Once an intervention is initiated fleet-wide, the operator needs reliable estimates of aggregate power reductions, their evolution over time, and the resulting financial savings from reduced consumption during peak hours. This shift in focus acknowledges that, at any point after an intervention begins, some

[21]These confidence intervals reflect variability across interventions but likely underestimate the true variability, as they do not account for intra-household correlation, autocorrelation structures in errors over time, or variability around the mean counterfactual power level at each time point.

[22]The temperature set on the thermostat by the user, which would have been reached without the intervention.

HPs may have already resumed operation, thereby consuming more power than the counterfactual during their rebound period each, while the fleet as a whole may still achieve a net power reduction. This section analyzes these fleet-level dynamics during what we refer to as 'flexibility events', combining interventions and HPs into a single profile relative to time to intervention start.

4.2.1. Fleet-level power consumption profiles during flexibility events

The left panel of Figure 5 shows the aggregated power consumption per HP in the fleet, averaged over the entire intervention sample and across both heating seasons, relative to the time to intervention start. The figure compares the average fleet power consumption per HP to the control HP power level. The fleet-level control consumption, as defined in eq. (6), is derived by aligning all HP-specific counterfactual consumption values during interventions by time to intervention start. As a result, the fleet-level control shows periodic large and smaller peaks, corresponding to the two main—morning and midday—consumption peaks observed in the right panel of Figure 2. Additionally, the right panel of Figure 5 shows the resulting average net power reduction, calculated as the difference between the intervention and control power levels at each value of time to intervention start.

In the pre-flexibility event period, the power consumption of soon-to-be-blocked HPs aligns closely with the control power (green curve, left panel). Once the event begins and interventions are initiated on all HPs in the fleet, the average power consumption per HP quickly drops to just under 100 W (left), resulting in a power reduction of around 250 W at the start of the intervention (right). As the event progresses, some HPs resume normal operation, while others remain blocked by the still-ongoing intervention, leading to a gradual increase in the average fleet power consumption (blue curve, left). This gradually reduces the fleet-level power reduction over time (right) until it eventually turns negative. Around 18 hours after the event starts, the fleet's average power consumption (left) exceeds the control consumption. At this point, the flexibility event no longer reduces electricity consumption, and the fleet begins consuming more electricity than usual. This may be viewed as the fleet-level equivalent of the rebound period in the period following the end of flexibility interventions on individual HPs. Similarly, the average power reduction drops to 0 W after about 18 hours (right). Over the unsmoothed intervention data, the net power reduction averages 251 W during the first hour of a flexibility event (95% CI: 234, 270 W), decreasing to 103 W on average over the first 18 hours (95% CI: 95, 112 W). Between 18 and 36 hours, the fleet-level rebound is limited to just 44 W on average (95% CI: 39, 49 W), and up to only 134 W. These values align closely with the trends depicted in the locally smoothed plots.

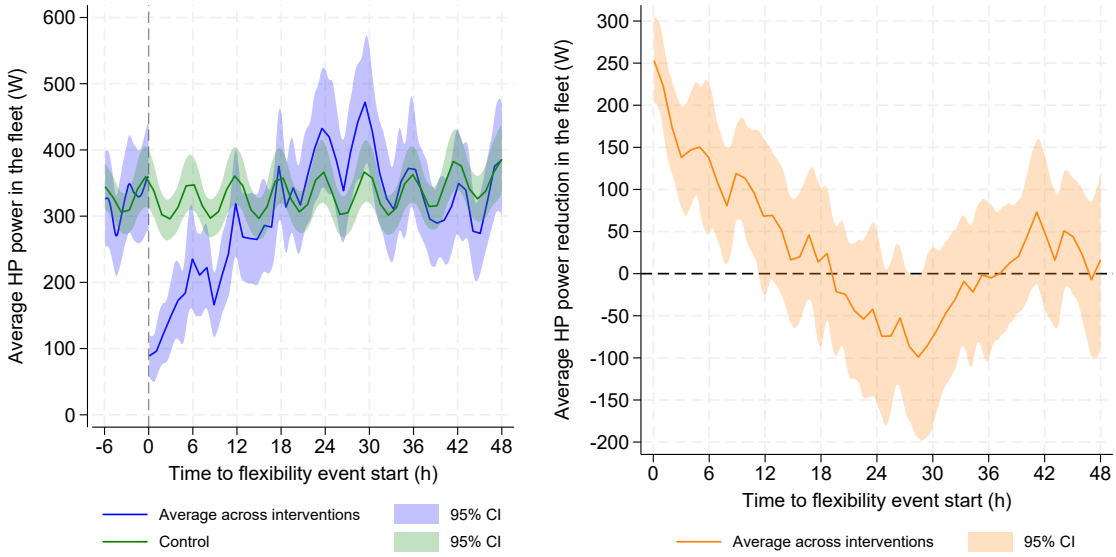


Figure 5: Average heat pump (HP) power consumption per unit in the fleet relative to the time of intervention start (left panel) and the resulting net power reduction (right panel), averaged across all interventions. The control curve is computed using the average daily HP consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual, aligned by time to intervention start (see eq. (6)). Standard errors reflect the variability of the means in 5-min-to-intervention-start bins, assuming independence among observations (eq. (7)). The means and confidence intervals are smoothed using a local polynomial of degree 0, with the bandwidth of the intervention curve (left panel) set to the optimal value over the entire plotted period.

These quantitative results differ from those in Section 4.1.3. Specifically, in the left panel, the minimum power consumption per HP immediately after the intervention starts is about 100 W,

which is higher than the average residual consumption of 53 W observed at the individual HP level. Similarly, in the right panel, the maximum power reduction of around 250 W is lower than the ATT estimate of a reduction of 292 W (see Section 4.1.3), which reflects the effect of interventions on HP power within a household, controlling for household-invariant characteristics. These differences arise because the fleet-level analysis of flexibility events includes data for interventions that stop immediately[23] or shortly after being initiated. Additionally, some HPs take longer to adjust to reduced operation during interventions, particularly when starting from a high-power state (e.g., heating the DHW tank just before an intervention begins). As a result, the fleet's power reduction at the start of the intervention is lower than at the level of flexibility interventions on individual HPs in Section 4.1.3.

A further difference between flexibility interventions and events is their duration: flexibility events reduce power consumption for 18 hours on average, significantly longer than interventions (see Section 4.1.1). This is because the fleet-level reflects the balance between HPs that remain blocked and those that become unblocked during the event. Blocked HPs reduce their consumption by a constant amount (292 W each, on average), while unblocked ones experience individual rebound effects. However, as HPs become unblocked at different times (e.g., due to differences in insulation levels), a staggering occurs at the fleet level. This leads to an average excess consumption from unblocked HPs in the fleet (shown in Appendix A.7) of only 110 W over the first 18 hours, enabling still-blocked HPs to compensate for the unblocked ones over a longer period. The natural staggering in the return to normal operation observed in our setup aligns with the approach studied in (Müller and Jansen, 2019) to mitigate rebound effects in flexible HPs.

The two distinct phases of flexibility interventions at the individual HP level (see Section 4.1.5) are also reflected in the fleet-level analysis of flexibility events, as shown in Appendix A.8, which further shows that flexibility events reduce HP power consumption on average by around 1 kWh, as measured 36 hours after the start of the flexibility event across the fleet.

As HPs operate at higher power levels when outdoor temperatures are lower, there is more electricity consumption available to reduce during the first phase of flexibility events. To account for weather variability, we proceed in two steps. First, we categorize flexibility events into four outdoor temperature groups (< 3 °C, 3 - 6 °C, 6 - 9 °C, > 9 °C), now using the average outdoor temperature within the first 18 hours of the event. Each individual intervention is then assigned to one of these event categories by matching the average temperature within 18 hours after the intervention starts to the corresponding temperature bin for the event.

Second, we construct the fleet-level power profile of flexibility events for each temperature category by aligning both the observed power profiles and the household- and temperature-specific counterfactuals relative to the time to flexibility event start, as in eq. (6). Appendix A.9 presents the figures analogous to the left and right panels of Figure 5. In particular, Figure A.17 shows that, as outdoor temperatures get colder, the initial power reduction in the fleet increases, reaching up to 600 W under 3 °C.

Following this approach, Figure 6 shows the average[24] electricity consumption reduction per HP in a fleet relative to the time to flexibility event start, across four outdoor temperature ranges. Flexibility events occurring in very cold conditions (average outdoor temperature within 18 hours after the start below 3 °C) achieve the highest reductions in the first phase of the event, reaching 4.5 kWh reduction on average at around 23 hours after the event start. However, the post-event rebound considerably reduces these gains in the second phase.

The post-event rebound appears to diminish consumption reductions less as outdoor temperatures increase, aligning with the regression results at the individual HP level discussed in Section 4.1.5. Notably, for events between 6 and 9 °C, the rebound effect appears to offset much of the consumption reductions, leaving virtually no net reductions 36 hours after the event start. This may be because outdoor temperatures between 6 and 9 °C are warm enough that HPs already operate at moderate power levels, thereby limiting kWh reductions achieved in the first phase. At the same time, they are still cool enough to trigger significant rebound effect afterwards, which reduces these small gains during the second phase. Above 9 °C, no rebound effect is observed, suggesting that the rebound's effect at the flexibility event level is not continuous across temperature ranges. At these mild temperatures, the fleet does not require significant additional electricity to restore temperature setpoints, as heat loss to the outside environment is limited.

At 36 hours after the start, all events except those between 6 and 9 °C reduce net consumption

[23] Because the intervention's initial conditions already meet the conditions for one of the termination scenarios.

[24] Due to the scarcity of very cold days in the sample, the confidence intervals for these estimates are not shown as they are wider and harder to interpret. We restrict our discussion to the averages.

by 1 to 1.5 kWh on average, with slightly higher net average reductions for events between 3 and 6 °C compared to those below 3 °C. Finally, while net reductions tend to level off at 36 hours after the event start on average across the sample (see Appendix A.8), the exact point at which they stabilize for low-temperature events remains unclear.

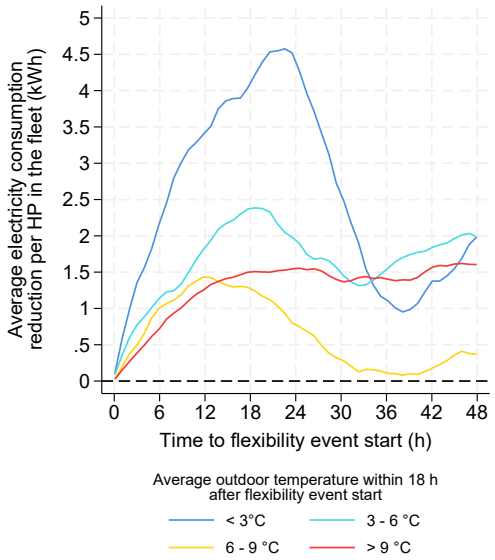


Figure 6: Average electricity consumption reduction (in kWh) per heat pump (HP) in the fleet, categorized by four outdoor temperature ranges. Reductions are calculated using the average HP consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual (aligned by time to intervention start, see eq. (6)), based on outdoor temperatures within the first 18 hours after event start: below 3 °C, 3-6 °C, 6-9 °C, and above 9 °C. The means are smoothed using a local polynomial of degree 0.

4.2.2. Monetary valuation of a flexibility event in a fleet of heat pumps

We estimate the monetary impact of flexibility events across a fleet of HPs using day-ahead electricity prices. We calculate the monetary value of simulated events initiated at each hour throughout the 2022-2023 and 2023-2024 heating seasons (a total of 7,622 hours, i.e., almost as many events, as last hours of HS2 lack sufficient data for 36-hour events), using actual data on day-ahead electricity prices and outdoor temperatures to capture the relationship between outdoor temperature and its effect on HP power reduction at a fleet-level. As day-ahead electricity prices are negatively correlated with outdoor temperature, with a correlation coefficient of -0.46 ($p < 0.01$) throughout the 2022-2023 and 2023-2024 heating seasons, lower temperatures not only influence energy consumption reductions during events but also frequently coincide with higher electricity prices.

The calculation involves two steps. First, using historical outdoor temperature data for Belgium (Royal Meteorological Institute of Belgium, 2024), we compute the average temperature over the 18 hours following each event start to assign it to an appropriate electricity consumption reduction profile from Figure 6. Second, we estimate the cost savings by matching the dataset of simulated events with actual day-ahead electricity prices for Belgium (European Network of Transmission System Operators for Electricity, 2024). To explicitly account for post-intervention increases in electricity consumption, we calculate the average cost savings at two time points: 18 hours and 36 hours after each event start.

On average, we find that flexibility events result in savings of €0.282 (95% CI[25]: €0.274,0.289) after 18 hours per HP and event, which decrease to €0.125 (95% CI: €0.122, 129) at 36 hours. However, aggregators can achieve higher savings by targeting specific high-price periods. To study how often large savings are attainable, we first rank all savings for each event in ascending order to reveal the proportion of savings that are at or below a given level. This is reflected in the left panel of Figure 7, which plots the empirical quantile function of savings across all flexibility events simulated hourly throughout HS1 and HS2. The x-axis shows the percentile of simulated events (which also corresponds to the percentile of time, since each hour in the simulation serves as the

[25]The confidence intervals in this section reflect the variability of mean savings across events but likely underestimate the true variability, as they do not account for variability around the mean counterfactual power level at each time point during the simulated events.

start of a new event), and the curve displays the highest savings observed up to that percentile. In the top 5% of events ($n = 380$), net savings at 36 hours increase significantly, averaging €0.58 per HP and event, and averaging €0.76 in the top 1 % of hours ($n = 76$). The maximum observed savings are €1.09 per HP and event. Remarkably, cited savings above the 95th percentile occurred almost entirely in November and December 2022, at the peak of the electricity crisis.

Around 10% of events ($n = 730$) resulted in negative savings at 36 hours, averaging −€0.03 and reaching as low as −€0.41. The majority of such events occurred when the average temperature over the first 18 hours of the event was between 6 and 9 °C, where consumption reductions are entirely offset by rebound consumption (see Figure 6). Besides, during these events, the average day-ahead price over the 18-hour reduction period was slightly lower than the average price over the full 36-hour event, making rebound consumption more costly than the reductions. Such price dynamics may happen more frequently in this temperature category as it is the one that shows the weakest correlation between outdoor temperatures and day-ahead prices, estimated at just −0.08 (although significant at the 1% level). Overall, this reinforces the importance of targeting the events in a smart way, so that aggregators avoid scenarios where flexibility events result in extra costs rather than generate savings.

The right panel of Figure 7 shows that, as average day-ahead prices increase during the first phase, the gap between savings at 18 hours and 36 hours after the intervention widens. This aligns with the findings from Figure 6, where very cold temperatures (associated to high prices) amplify this gap due to a larger rebound consumption. However, 36-hour savings still increase with higher prices, as expected. The slight deviation at very high prices is due to these bins containing more events occurring below 3 °C, where net savings at 36 hours on average are slightly lower than those achieved by events in the 3 to 6 °C range (see Figure 6).

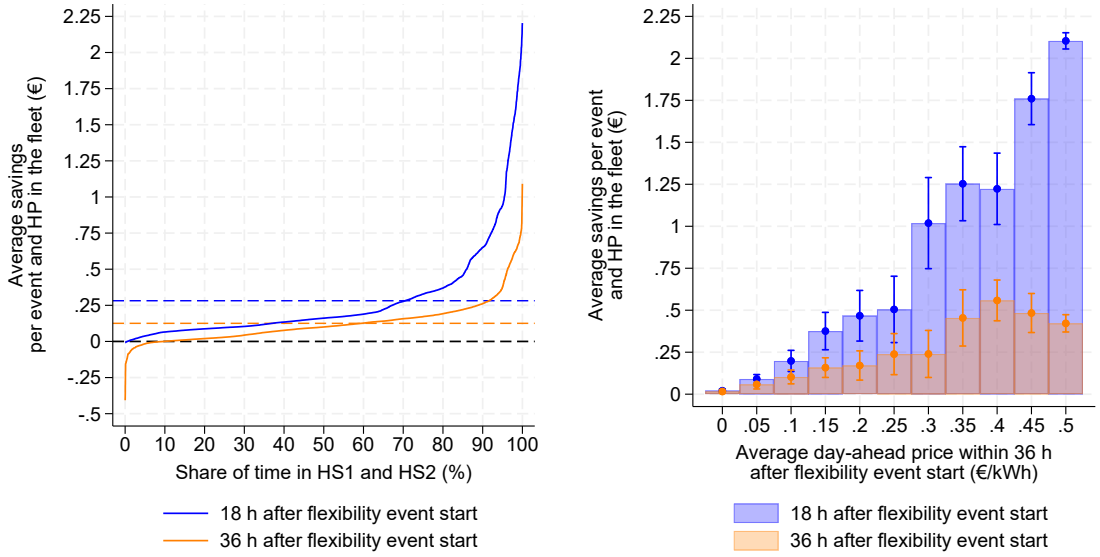


Figure 7: Average savings per event and per heat pump (HP) in the fleet at 18 and 36 hours after the flexibility event start. Simulated flexibility events are initiated at each hour throughout heating seasons (HS) 1 and 2, using the temperature-specific average consumption reduction profiles from Figure 6. Left: quantile function of savings by share of time across HS1 and HS2, smoothed using a local polynomial of degree 0. Horizontal dotted lines indicate the sample averages. Right: histogram of savings binned by the average day-ahead electricity price within the first 18 and 36 hours after the event starts. Confidence intervals at the 95% level, represented by error bars, capture the variability in mean savings within each bin, excluding variability around the HP counterfactual power level.

To compare these savings with the costs of making a HP flexible, we perform a back-of-the-envelope calculation. For simplicity, we assume that an aggregator (or DSO) has ex-ante perfect information on day-ahead wholesale hourly prices and schedules a fixed number of flexibility events each winter heating season, ensuring at least a 48-hour gap between the start times of consecutive events. The aggregator selects the hour with the highest average net savings for the first event, the second highest for the next, and so on (i.e., the events are not evenly distributed within the heating season). We then compare the cumulative savings generated from this optimal allocation of events to the cost of making a HP flexible: via, e.g., the purchase and installation costs of a smart thermostat with demand-response capabilities to receive steering signals from utilities.

Let C_0 denote the initial investment costs in a smart thermostat and $C_{y,n}$, the cumulative annual savings per HP from participating in n flexibility events in year (heating season) y. The net present value (NPV) per HP over Y years is given by eq. (13).

$$NPV_n(Y) = \sum_{y=0}^{Y} \frac{C_{y,n}}{(1+r)^y} - C_0, \tag{13}$$

where r is the discount rate. We assume that the annual savings are constant over time, i.e.: $C_{1,n} = C_{2,n} = ... = C_{Y,n} \equiv C_n$ for a given n, with n constant over years[26].

The payback period (PP_n) is obtained by setting $NPV_n(Y) = 0$. Without discounting ($r = 0$):

$$PP_n(r = 0) = \frac{C_0}{C_n}, \tag{14}$$

while, for $r \neq 0$, it is given by[27]:

$$PP_n(r \neq 0) = \frac{\log\left(\frac{C_0}{C_n}(\rho - 1) + 1\right)}{\log(\rho)} - 1, \tag{15}$$

with $\rho = (1+r)^{-1}$. This expression is defined only if $C_n > C_0 \frac{r}{1+r}$, otherwise the savings are too low to offset the investment under discounting.

Figure 8 illustrates the average payback period (in years) for making a HP flexible via a smart thermostat under both $r = 0$ (left panel) and $r = 5\%$ (right panel) discount rates, across different numbers of flexibility events per heating season. Flexibility events are optimally timed, with at least 48 hours between them, and results are presented separately for HS1- and HS2-electricity price levels (with n remaining constant over the years). The figure shows results for a thermostat cost of €160 (with sensitivity analyses for €120 and €200)[28]. Under HS1-level prices, around 40 events per heating season are required to achieve a payback period under 10 years without discounting, whereas about 50 events are required yearly when a 5% discount rate is applied. This means that, under the savings calculated in this section, significantly high prices and numerous events are required to make the payback period reasonable. Time discounting, and higher investment costs in a thermostat, significantly extend the payback period and, under certain conditions (e.g., few events or high thermostat prices in HS2—see right panel), may even render it undefined (i.e., savings are never high enough).[29]

It is important to note that the savings in this section (and up to €1.09 per flexibility event) are based solely on the volatility of day-ahead prices. Furthermore, in systems with capacity markets[30], an additional value stream arises from the capacity value during peak events. At a capacity value of $700 per kW (Bollinger and Hartmann, 2015, 2019), an average flexible HP in our sample could save up to around $175 per year in investment in additional peak generation capacity, provided that post-intervention increases in consumption are appropriately managed.

In practice, households with a HP will have access to a wide range of different value streams. Specifically, these include: (i) value streams related to the financial incentives to which the household is directly exposed (e.g., optimizing consumption under dynamic tariffs, maximizing solar self-consumption or reducing costs associated with using the electricity distribution grid); and (ii) value streams related to financial incentives that can only be captured through an aggregator acting as an intermediary (e.g., monetizing flexibility in intraday, ancillary services, or balancing markets, or participating in capacity remuneration mechanisms). The highest financial benefit would be achieved through "value stacking"—which involves using advanced algorithms to continuously monitor trade-offs among different financial opportunities, identify potential synergies and ultimately steer the HP in a globally optimal way 24 hours per day and 365 days per year. Such

[26] As shown in Appendix A.10, the level of cumulative savings from this allocation of events differ between the two heating seasons. For instance, while $n = 50$ events per season would generate average yearly savings of about €16.9 under HS1-level electricity prices—due to the energy crisis and colder weather overall—, they amount to only €7.4 per year on average under HS2-level prices.

[27] Let $\sigma_Y \equiv \sum_{y=0}^{Y}(1+r)^{-y} = \sum_{y=0}^{Y}\rho^y$ with $\rho \equiv (1+r)^{-1}$. One can show that $\rho\sigma_Y = \sigma_Y - 1 + \rho^{Y+1}$, so that (if $\rho - 1 \neq 0$, i.e., $r \neq 0$): $\sigma_Y = (\rho^{Y+1} - 1)/(\rho - 1)$.

[28] These values are motivated by Energy Star-certified smart thermostats featuring a "Time of Day Usage" option. Out of 66 models reported on the Energy Star website, nine have their price listed, averaging about $160 with a standard deviation of approximately $40. All nine thermostats listed offer some level of demand-response capability. See: `https://www.energystar.gov/productfinder/product/certified-connected-thermostats/results`, 20 March 2025.

[29] It should however be noted that smart thermostats might generate savings through other channels, and for instance through a finer and more optimal control over heating settings overall, thereby lowering the total payback period of investing in a smart thermostat.

[30] Which is not the case in Belgium.

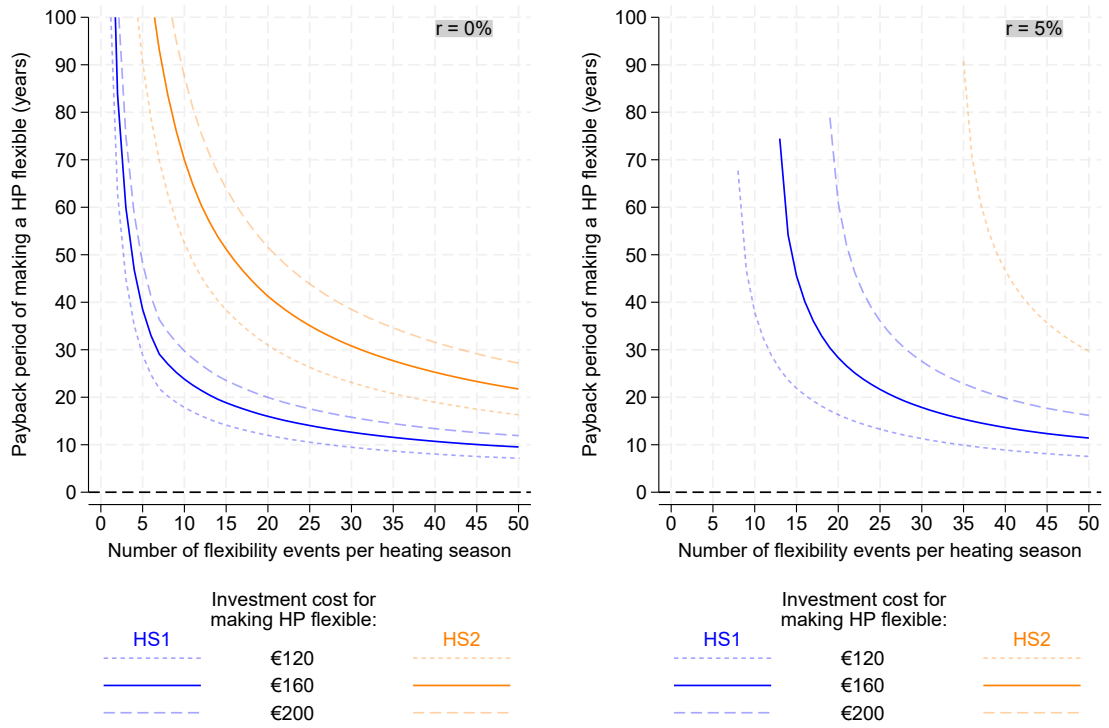


Figure 8: Average payback period (years) for making a heat pump (HP) flexible via a smart thermostat as a function of the number of flexibility events per heating season. Left: without discounting, $r = 0\%$ (cf., eq. (14)). Right: with a constant discount rate $r = 5\%$ (cf., eq. (15)). Curves correspond to thermostat prices of €120, €160, and €200, using day-ahead electricity prices from heating season 1 (HS1, 2022–2023) and heating season 2 (HS2, 2023–2024) for calculating the savings per HP. Flexibility events are optimally allocated (at least 48 hours apart) and savings are computed at 36 hours after event start to include the rebound period effect. Savings through flexibility events are assumed constant over the years.

practices are already implemented[31] in large-scale battery parks and even home batteries, but will be more challenging to implement with HPs due to the additional complexities related to building physics, human behavior and comfort.

4.3. Comfort impact and household responses

In this section, we analyze households' overriding pattern during ongoing interventions, reflecting their subjective perception of discomfort. A thematic analysis of comments left at the time of overruling provides qualitative insights. We also examine whether subjective discomfort aligns with a quantitative proxy for discomfort. Finally, we present post-experiment survey results to understand participants' responses to flexibility interventions and the features they considered important when scaling such programs.

4.3.1. Thematic analysis of reasons for manual overrides

We analyze the reasons households provided on the online dashboard when overriding flexibility interventions. To classify them, a thematic analysis was conducted with the assistance of a large language model[32] (OpenAI, 2024). The analysis identified eight (possibly overlapping) categories of manual overrules. The three main categories are: "Too low indoor temperature" ($N = 35$; 65%), "Sickness/health concerns" when users request the interruption of the intervention or the cancellation of any upcoming intervention due to individuals being temporarily vulnerable to low temperatures ($N = 11$; 20.5%)—this category contains half of the preemptive overrules—, "Presence at home" with individuals present at home requiring comfort to work or study ($N = 9$; 16.5%). Five further categories are identified, although with fewer occurrences each: "Need for

[31] Illustrative examples from the commercial sector include—but are not limited to—Lifepowr/Flexio (`https://www.lifepowr.io/en`, `https://flexio.lifepowr.io/en`), Yuso (`https://yuso.com/be/en/batteries/local-battery-storage`), or Tesla's Autobidder software (`https://www.tesla.com/support/energy/tesla-software/autobidder`) (Accessed: 8 April 2025).

[32] The input prompts are available with the replication codes.

comfort during the weekend" ($N = 3$; 5.5%), "Low water temperature" ($N = 3$; 5.5%), "Returning home" when participants return to a colder home after being away ($N = 2$; 3.5%), "Visitors" when households expect guests ($N = 2$; 3.5%), and "Technical issue" to report errors with the installation ($N = 1$; 2.0%). Households' responses to drops in DHW temperature (in non-decoupled HPs) are further evidenced by morning overrides observed in one household, likely driven by increased hot water demand during that time of day. This supports the automatic stop implemented when the DHW temperature falls below 40 °C. The morning period from 7 to 12 a.m. accounts for 41% of all manual overrules (excluding preemptive ones), making it the most predominant time of day for overruling.

Interestingly, ongoing interventions were overruled at relatively mild indoor temperatures, averaging 19.6 °C (95% CI: 19.3, 19.9 °C) over all overrules. This can be compared to the pre-survey results, where households were asked to report their minimum comfortable indoor temperature during interventions[33]. The average reported temperature was 18.5 °C across the nine participating households but varied significantly: from 14 °C (Household 9, who overruled at 19.0 °C on average over non-preemptive overrules) to 20 °C (Households 2, 3, 5 and 6, who overruled at 20.2 °C on average over non-preemptive overrules).

While this highlights discrepancies between stated and revealed preferences, the indoor temperature at the moment of overruling does not fully capture household discomfort at that stage of the intervention. The temperature drop, i.e., the difference between the initial indoor temperature at the intervention start and the final temperature must also be considered. Moreover, the placement of thermal sensors—typically in a central room—may not always reflect conditions if households activities occur in colder rooms. The location of the thermostat is indeed a key factor affecting the potential of flexibility events, as noted by (Centre for Net Zero and Nesta, 2023).

4.3.2. Quantitative proxy for household discomfort and manual overruling patterns

Beyond the average temperature reduction during an intervention or the temperature at the time of overrule, a more representative proxy for discomfort is the temperature drop attributable to each intervention. This is calculated as the difference between the initial indoor temperature at the start of the intervention, $T_{in}(t^0)$, and at its end, $T_{in}(t^f)$, for each intervention i on household h:

$$\Delta T_{i,h}^{drop} = T_{in,i,h}(t^0) - T_{in,i,h}(t^f) \tag{16}$$

We find that the average temperature drop across the entire intervention sample equals 0.69 °C (95% CI: 0.59, 0.78 °C). As expected, this value, which reflects the total impact of the intervention, is higher than the average temperature reduction of 0.16 °C observed throughout the duration of interventions (see Section 4.1.2). Additionally, in the subsample of interventions that were automatically stopped (by either the T_{in} or T_{DHW} thresholds), the temperature drop shows a moderate correlation with the indoor temperature at the start of the intervention ($r = 0.40$, $p < 0.01$) and with the intervention duration ($r = 0.56$, $p < 0.01$); the negative correlation with the average outdoor temperature during the intervention is close to significance ($r = -0.12$, $p = 0.06$).

To analyze whether household overruling patterns align with the proxy for discomfort in eq. (16), Figure 9 presents a histogram of the temperature drop for each intervention, distinguishing between those that were automatically stopped and those that were manually overruled (excluding eight preemptive overrules, as they do not result from discomfort induced by an intervention).

Most interventions resulted in a positive temperature drop from the initial value, indicating that households experienced actual discomfort. A few outliers, including one automatic stop with a drop of −4.3 °C, likely occurred when interventions were followed by sunny weather or high home occupancy with activities like cooking, both of which contribute to reheating the home, resulting in a negative $\Delta T_{i,h}^{drop}$. The peak observed at a zero temperature drop reflects the interventions that were too short in duration to cause significant deviations from the initial temperature (see Section 4.1.1).

The average temperature drop across automatically stopped interventions is 0.62 °C (95% CI: 0.51, 0.72 °C), but was considerably higher across manually stopped ones, at 1.06 °C on average (95% CI: 0.83, 1.29 °C). The difference between the two means is statistically significant at the 1% level (around 0.45 °C on average, with 95% CI: 0.19, 0.70 °C). This indicates that the discomfort households experience at the time of manual overrule is substantially higher than the discomfort

[33]This temperature was not intended to be used as a feature in the experiment, and households were informed of this.

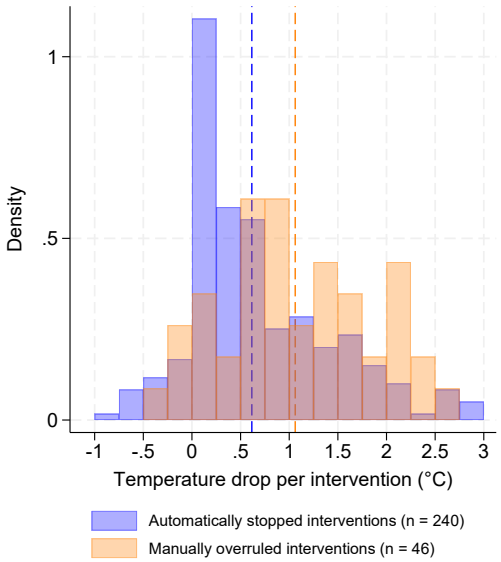


Figure 9: Histogram of the temperature drop induced by interventions ($\Delta T_{i,h}^{drop}$), by automatically stopped and manually overruled interventions. The drop is calculated using eq. (16). Preemptive overrules are excluded from the subsample of manually overruled interventions. The vertical dotted lines indicate the mean temperature drop for each subsample. The histogram is truncated at -1 °C to exclude one outlier.

built up during automatically stopped interventions[34]. This aligns with the thematic analysis of comments left at overrule, which showed that most manual overrules are motivated by thermal discomfort.

Although this might suggest rational behavior from households—overruling interventions when discomfort is greater—it is important to note that the correlation between a dummy variable for manually overruled interventions and the temperature drop is weak, at only 0.20 ($p < 0.01$). Households did not consistently overrule interventions with larger temperature drops, which could be partly explained by their absence from home during some of these interventions.

Finally, a correlation analysis (see Appendix H) shows no strong evidence of habituation (defined as a decline in manual overrule frequency with an increasing number of interventions) at the household level.

4.3.3. Post-experiment participants' feedback

In May 2024, a few weeks after the last HS2 intervention, we sent a short survey to participants; to gather feedback on their experience. Eight of the nine households that participated in HS1 or HS2 responded.

Households rated how interventions affected their indoor comfort on a 1-5 scale (*No discomfort* - *Extreme discomfort*), with an average score of 2.4 indicating only slight to moderate discomfort. This aligns with our finding that interventions did not cause large temperature drops on average. Discomfort from reduced DHW availability was rated minimal, averaging 1.1, suggesting that the automatic 40 °C trigger was effective.

Regarding flexibility program features, households rated advance notification as only slightly to moderately important, averaging 2.5 on a 1-5 scale (*Not important at all* - *Extremely important*), whereas the ability to overrule interventions was rated very important (averaging 3.9). They also valued real-time communication via the dashboard, as shown by messages left on the platform to ask questions or report technical issues, rather than to override interventions. In some cases, households even requested the overruling of potential upcoming interventions within 24 hours, even when none were scheduled.

Finally, when asked about strategies to reduce discomfort during interventions, two households reported opting for warmer clothing. This aligns with the regression results in Section 4.1.1, which fail to support Hypothesis (i), showing no evidence that households increased the thermostat setting

[34]This finding is consistent with results from a robust regression analysis that includes household FE and is restricted to households that overruled interventions at least once. The coefficient for the manual overrule dummy (excluding preemptive overrules) is 0.24 °C ($p = 0.03$; 95% CI: 0.03, 0.53 °C, based on 100,000 wild cluster bootstrap replications), indicating that the average within-household temperature drop was 0.24 °C greater for manually overruled interventions, controlling for household characteristics invariant across interventions.

prior to a notified event.

5. Conclusions

5.1. Summary of key results

Table 4 provides an overview of the main experimental results presented in this paper. A cross-season comparison is available in Appendix D.

Table 4: Overview of key experimental results (full sample)

Variable	Refer to Section	Full sample
Total number of interventions		287
Intervention stop reasons:		
DHW temperature threshold	4.1.1	70.0%
Manual overrule		11.1%
Indoor temperature threshold		18.8%
Mean intervention duration (hours)		12.8 (11.3, 14.4)
ATT on indoor temperature during interventions (°C) [a]	4.1.2	-0.38 (-0.66, -0.10)
ATT on heat pump power during interventions (W) [a]	4.1.3	-292 (-390, -200)
Mean electricity consumption reduction during interventions (kWh)	4.1.4	3.22 (2.81, 3.63)
Mean rebound electricity consumption within 16 hours post-intervention (kWh)	4.1.5	2.40 (1.96, 2.85)
Mean rebound power consumption within 1 hour post-intervention (W)		607 (571, 642)
Mean maximum per-unit power reduction within 1 hour of a flexibility event on a fleet (W)	4.2.1	252 (234, 270)
Mean time until fleet-level rebound (hours) [b]		≈ 18
Mean indoor temperature drop from intervention start to end (°C)	4.3.2	0.69 (0.59, 0.78)
Mean additional temp. drop: manual vs. automatic overrules (°C) [c]		0.45 (0.19, 0.70)

The 95% confidence intervals are in parentheses.
[a] Estimated via linear regression with household fixed effects using 100,000 wild bootstrap replications to cluster standard errors at the household level. [b] Based on visual inspection of local polynomial regression plots, identifying the approximate rebound time. [c] Excludes preemptive overrules.

5.2. Discussion and implications

This field experiment with nine participating households is one of the first practical implementations of a residential electricity flexibility scheme, with a specific focus on flexible heating. The experimental design is suited for studying the interaction between the technical capabilities of heat pumps and the comfort boundaries of the households involved, which together determine the real-world flexibility potential. Our analysis builds on two complementary perspectives: individual heat pump flexibility and aggregated fleet-level flexibility.

First, our experiment was designed so that interventions could either stop automatically or be manually overruled by the user. This mixed approach provides participants with some degree of control, which is likely to become a standard feature of future residential flexibility schemes. We observed that on average, individual interventions lasted around 12.8 hours, with the primary constraint being the demand for domestic hot water. Further, our findings indicate that intervention duration is significantly influenced by the initial heat pump and outdoor temperature conditions, but is not affected by the specific time of day the intervention is initiated. Although interventions reduce electricity consumption, they also lead to rebound effects of up to 700 W as the intervention ends and the heat pump resumes operation.

Second, we present the fleet-level dynamics, focusing on the average contribution of each heat pump during "flexibility events", i.e., coordinated actions across a fleet of assets. These results are particularly valuable, as they offer more realistic insights into the flexibility potential of heat

pumps, considering that some units return to normal operation earlier than others or have their interventions manually overruled, which reduces the maximum achievable power reduction. This perspective reveals that the power consumption of a fleet of heat pumps can be reduced by up to 250 W per heat pump, gradually decreasing to 0 after 18 hours on average, before the fleet consumes more electricity than usual to restore temperature setpoints. On average, each flexibility event yields a net electricity reduction of approximately 1 kWh, measured 36 hours after the event start. The 18-hour phase of reduced consumption during flexibility events is considerably longer than for individual interventions, due to the natural staggering of heat pumps returning to normal operation. This staggering smooths rebound effects, which peak at just 130 W and average 50 W over the 18-hour rebound period.

While it could be argued that large impacts on household comfort would result in a low acceptability of heat pump flexibility, creating a barrier to its large-scale adoption in the future, our findings suggest otherwise. In fact, heat pumps can be turned off for several hours without a noticeable impact on indoor temperatures, as demonstrated by the vast majority of the hundreds of interventions performed in our experiment. On average, the temperature decreased by just 0.69 °C by the end of an intervention, although larger drops were observed in cases of manual overrides. These overrides, along with feedback gathered after the experiment, indicate that households value and actively use their ability to intervene and restore comfort levels. They suggest that such a feature is an important prerequisite for encouraging greater participation of households in heat pump flexibility schemes. It reflects that households are not homogeneous actors who predictably allow or reject flexibility but that their decisions are instead driven by multiple factors, including sick children and dinner parties, which all influence the time-dependent and context-specific acceptability of flexibility.

The key driver expected to motivate heat pump owners to operate them flexibly is the potential for financial savings. These savings depend on the different stakeholders and their abilities to access value streams like price volatility or ancillary services; our study focuses on savings from day-ahead price volatility only. We show that colder outdoor temperatures, frequently coupled with higher day-ahead prices, provide more power reduction potential as heat pumps operate at higher levels, but also result in greater heat loss during rebound periods. Accounting for this trade-off, net savings amount to €0.13 on average per heat pump and flexibility event, occasionally reaching much higher savings up to €1.09 during periods of extreme price volatility. Importantly, because these interventions have minimal impact on household comfort, it becomes feasible to implement them daily throughout the heating season. Automating such interventions allows many small individual savings to add up to substantial annual reductions in heat pump running costs, encouraging more households to participate in flexibility. This is particularly true as automation minimizes the cognitive burden on users by reducing the effort required, as evidenced by the many interventions we conducted with only a few manual overrules.

Our findings support the following key policy recommendations. First, policymakers—possibly in collaboration with electricity retailers and academia—should continue and expand support for larger pilot programs on heat pump flexibility. Although our sample is limited to only nine households, our findings suggest that HP flexibility can address different challenges in a decarbonized electricity system. However, larger-scale studies that include a more diverse range of household types are essential to validate these benefits before business models mature and large-scale implementation rolls out.

Second, beyond supporting pilot programs, policymakers should consider regulatory adjustments to foster heat pump flexibility and maximize its benefits. Flexibilization of domestic demand is critical for the energy transition, as it can help managing national peaks in electricity consumption and address periods of low wind and/or solar production, both of which typically only last a few hours. However, although limited by the sample size, our findings suggest that heat pump flexibility alone may not suffice during extended periods of low renewable generation, such as a two-week 'Dunkelflaute'. Interestingly, households in our study did not strategically adapt to day-ahead notifications of flexibility interventions and reported minimal importance for receiving such notifications in the post-experiment survey. This supports the potential for unannounced, shorter-term flexibility to respond to rapid changes in electricity system conditions. Overall, heat pump flexibility is a relatively inexpensive way to increase the energy system efficiency, as it builds on assets that households are already adopting, without requiring costly new infrastructures for electricity production and storage. In this regard, regulatory measures—such as improving interoperability, implementing standardized communication protocols, and ensuring that heat pumps are equipped out of the box with separate controls for space and domestic hot water heating—can further ensure that the efficiency gains from heat pump flexibility are fully realized.

Third, practitioners, such as energy suppliers, aggregators, heat pump manufacturers, and sys-

tem and grid operators benefit from progress in research as well. Heat pump flexibility is shaped not only by technical capabilities of assets, but also on in-situ factors like building characteristics, which affect event duration, temperature drop rates, and heat loss driving rebound consumption. User behavior, such as setpoint values choice, influences how much of the reduction in consumption is sustained after the rebound period. Additionally, indoor and outdoor temperatures also play a key role in determining flexibility potential and event duration. Our results indicate that practitioners should account for these factors when assessing the net impact of fleet-wide heat pump flexibility, especially regarding rebound effects, as explored in our study. Fleet-level results show that a staggered return to normal operation after flexibility interventions can significantly enhance the benefits of heat pump flexibility. While this staggering occurred naturally in our setup, it may become beneficial for practitioners to develop dedicated algorithms to achieve this at scale. Furthermore, developments in software are also necessary to enable the separation of control between space and domestic hot water heating.

There is ample room for further work on this topic. For instance, one straightforward way to address the question of acceptability of flexible heating involves studying interventions that pre-heat homes (e.g., by 1-2 °C above user setpoint) before interrupting heating during high-price periods. Such interventions could not be tested within the constraints of heat pumps in our sample but hold potential for better understanding comfort thresholds. Similarly, future research could focus on interventions targeting heat pumps with decoupled space and water heating, once these become more common. Besides, as heat pump adoption scales up, it will be essential to study how our findings generalize across different building and household characteristics. In particular, less-well-insulated homes may represent a significant portion of future flexibility potential if high-temperature heat pumps gain popularity as replacements for fossil-fueled systems. In such homes, heat pump flexibility is likely to remain viable without substantially impacting comfort, though the dynamics will differ, with higher power reductions but presumably shorter intervention durations. Exploring these in future research could provide additional insight to policymakers and practitioners on the design of flexibility schemes that make the electricity demand of heat pump users more price-responsive, with minimal impact on comfort.

Data availability

All data and code necessary to replicate the results are available on GitHub, except for the free-text entries from participants during manual overrules, which have been omitted for privacy reasons. The repository also includes the input prompts used for the AI-assisted thematic analysis of these comments. The data and code can be accessed at: `https://github.com/RigauxBaptiste/The_Proof_of_the_Pudding_is_in_the_Heating.git`.

Acknowledgments

This research was supported by the "FlexSys" (A Flexible electricity System contributing to security of supply) project funded by the Energy Transition Fund of the Belgian federal government, managed by the FPS Economy, SMEs, Self-employed and Energy. M. Ovaere was funded by Research Foundation - Flanders (FWO) (mandate no. 12B7822N). B. Rigaux was funded by Research Foundation - Flanders (FWO) (mandate no. 11Q4224N).

The authors are grateful to Joannes Laveyne and Nicolas Van Damme for their helpful discussions during this research, and for the technical support in setting up and operating the field experiment. The authors also thank two anonymous reviewers for their insightful suggestions and valuable feedback, which have strengthened the manuscript.

Author contributions

B.R.: Conceptualization, data collection, software, formal analysis, writing – original draft, review and editing, visualization, data curation.
S.H.: Funding acquisition, conceptualization, data collection, coordination, supervision, writing – original draft, review and editing.
M.O.: Funding acquisition, conceptualization, coordination, supervision, writing – original draft, review and editing.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used OpenAI's ChatGPT to assist with: the thematic analysis of comments left at overrule by participating households overall, the optimization of the analysis code, and the improvement of the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Aghniaey, S., Lawrence, T.M., 2018. The impact of increased cooling setpoint temperature during demand response events on occupant thermal comfort in commercial buildings: A review. Energy and Buildings 173, 19–27. doi:10.1016/j.enbuild.2018.04.068.

Allcott, H., 2011. Rethinking real-time electricity pricing. Resource and Energy Economics 33, 820–842. doi:10.1016/j.reseneeco.2011.06.003.

Bailey, M., Brown, D.P., Wolak, F.A., Shaffer, B., 2024. Centralized vs Decentralized Demand Response: Evidence from a Field Experiment. Working Paper .

Bauwens, T., Devine-Wright, P., 2018. Positive energies? An empirical study of community energy participation and attitudes to renewable energy. Energy Policy 118, 612–625. doi:10.1016/j.enpol.2018.03.062.

Bernard, L., Hackett, A., Metcalfe, R., Schein, A., 2024. Decarbonizing Heat: The Impact of Heat Pumps and a Time-of-Use Heat Pump Tariff on Energy Demand. Working Paper w33036. National Bureau of Economic Research. Cambridge, MA. doi:10.3386/w33036.

Blonz, J., Palmer, K., Wichman, C.J., Wietelman, D.C., 2025. Smart thermostats, automation, and time-varying prices. American Economic Journal: Applied Economics 17, 90–125. doi:10.1257/app.20210618.

Bollinger, B.K., Hartmann, W.R., 2015. Welfare Effects of Home Automation Technology with Dynamic Pricing. Working Paper 3274. Stanford Graduate School of Business. URL: https://www.gsb.stanford.edu/faculty-research/working-papers.

Bollinger, B.K., Hartmann, W.R., 2019. Information vs. Automation and Implications for Dynamic Pricing. Management Science 66, 290–314. doi:10.1287/mnsc.2018.3225.

Boogen, N., Winzer, C., 2024. Households' Willingness to Curtail Electricity Usage During Winter Shortages – A Field Experiment. Working Paper. SSRN. doi:doi.org/10.2139/ssrn.5037079.

Brewer, D., 2023. Household responses to winter heating costs: Implications for energy pricing policies and demand-side alternatives. Energy Policy 177, 113550. doi:https://doi.org/10.1016/j.enpol.2023.113550.

Brewer, D., Crozier, J., 2023. Who Heeds the Call to Conserve in An Energy Emergency? Evidence from Smart Thermostat Data. Working Paper. SSRN. doi:10.2139/ssrn.4360751.

Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. The Review of Economics and Statistics 90, 414–427. doi:10.1162/rest.90.3.414.

Centre for Net Zero and Nesta, 2023. Automating Heat Pump Flexibility: Results From a Pilot. URL: https://www.centrefornetzero.org/papers/automating-heat-pump-flexibility-results-from-a-pilot.

Centre for Net Zero and Nesta, 2024. HeatFlex: the untapped potential of heat pump flexibility. URL: https://www.centrefornetzero.org/papers/heatflex-the-untapped-potential-of-automated-heat-pump-flexibility.

Council of the European Union, 2018. Directive (EU) 2018/2001 of the European Parliament and of the Council. URL: http://data.europa.eu/eli/dir/2018/2001/oj.

Da Fonseca, A.L., Chvatal, K.M., Fernandes, R.A., 2021. Thermal comfort maintenance in demand response programs: A critical review. Renewable and Sustainable Energy Reviews 141, 110847. doi:10.1016/j.rser.2021.110847.

Darby, S.J., McKenna, E., 2012. Social implications of residential demand response in cool temperate climates. Energy Policy 49, 759–769. doi:10.1016/J.ENPOL.2012.07.026.

Dewangan, C.L., Singh, S., Chakrabarti, S., Singh, K., 2022. Peak-to-average ratio incentive scheme to tackle the peak-rebound challenge in TOU pricing. Electric Power Systems Research 210, 108048. doi:10.1016/j.epsr.2022.108048.

Enrich, J., Li, R., Mizrahi, A., Reguant, M., 2024. Measuring the impact of time-of-use pricing on electricity consumption: Evidence from Spain. Journal of Environmental Economics and Management 123, 102901. doi:10.1016/j.jeem.2023.102901.

European Environment Agency, 2023. Flexibility Solutions to Support a Decarbonised and Secure EU Electricity System. Technical Report. European Environment Agency. doi:10.2800/104041.

European Network of Transmission System Operators for Electricity, 2024. Day-Ahead Prices Transparency. URL: https://transparency.entsoe.eu/transmission-domain/r2/dayAheadPrices/show. Accessed: 29 July 2024.

Eurostat, 2022. Final energy consumption in households by type of end-use - quantities. URL: https://ec.europa.eu/eurostat/databrowser/view/nrg_d_hhq/default/table?lang=en. Data extract for year 2022. Accessed: 16 October 2024.

Eurostat, 2024a. Average size of dwelling by household composition and degree of urbanisation. URL: https://ec.europa.eu/eurostat/databrowser/view/ilc_lvho31__custom_12632371/default/table?lang=en. Data coverage: 2023 to 2023. Accessed: 21 August 2024.

Eurostat, 2024b. Distribution of population aged 18 and over by part-time or full-time employment, income group and sex - EU-SILC survey. URL: https://ec.europa.eu/eurostat/databrowser/view/ilc_lvhl04__custom_12638500/default/table?lang=en. Data coverage: 2003 to 2023. Accessed: 21 August 2024.

Eurostat, 2024c. Distribution of population by degree of urbanisation, dwelling type and income group - EU-SILC survey. URL: https://ec.europa.eu/eurostat/databrowser/view/ilc_lvho01__custom_12637832/default/table?lang=en. Data coverage: 2003 to 2023. Accessed: 21 August 2024.

Fabra, N., Rapson, D., Reguant, M., Wang, J., 2021. Estimating the Elasticity to Real-Time Pricing: Evidence from the Spanish Electricity Market. AEA Papers and Proceedings 111, 425–429. doi:10.1257/pandp.20211007.

Faruqui, A., Sergici, S., 2010. Household response to dynamic pricing of electricity: a survey of 15 experiments. Journal of Regulatory Economics 38, 193–225. doi:10.1007/s11149-010-9127-y.

Fischer, D., Madani, H., 2017. On heat pumps in smart grids: A review. Renewable and Sustainable Energy Reviews 70, 342–357. doi:10.1016/j.rser.2016.11.182.

Frederiks, E.R., Stenner, K., Hobman, E.V., 2015. Household energy use: Applying behavioural economics to understand consumer decision-making and behaviour. Renewable and Sustainable Energy Reviews 41, 1385–1394. doi:10.1016/j.rser.2014.09.026.

Fu, Z., Novan, K., Smith, A., 2024. Do time-of-use prices deliver energy savings at the right time? Journal of Environmental Economics and Management 128, 103054. doi:10.1016/j.jeem.2024.103054.

Georges, E., Cornélusse, B., Ernst, D., Lemort, V., Mathieu, S., 2017. Residential heat pump as flexible load for direct control service with parametrized duration and rebound effect. Applied Energy 187, 140–153. doi:10.1016/j.apenergy.2016.11.012.

Good, N., 2019. Using behavioural economic theory in modelling of demand response. Applied Energy 239, 107–116. doi:10.1016/j.apenergy.2019.01.158.

Good, N., Ellis, K.A., Mancarella, P., 2017. Review and classification of barriers and enablers of demand response in the smart grid. Renewable and Sustainable Energy Reviews 72, 57–72. doi:10.1016/j.rser.2017.01.043.

Gyamfi, S., Krumdieck, S., Urmee, T., 2013. Residential peak electricity demand response—Highlights of some behavioural issues. Renewable and Sustainable Energy Reviews 25, 71–77. doi:10.1016/j.rser.2013.04.006.

Harding, M., Lamarche, C., 2016. Empowering consumers through data and smart technology: Experimental evidence on the consequences of time-of-use electricity pricing policies. Journal of Policy Analysis and Management 35, 906–931. doi:10.1002/pam.21928.

Harding, M., Sexton, S., 2017. Household response to time-varying electricity prices. Annual Review of Resource Economics 9, 337–359. doi:10.1146/annurev-resource-100516-053437.

Harold, J., Bertsch, V., Fell, H., 2021. Preferences for curtailable electricity contracts: Can curtailment benefit consumers and the electricity system? Energy Economics 102, 105454. doi:10.1016/j.eneco.2021.105454.

Herabadi, A.G., Kadarusman, Y.B., Yachinta, C., 2021. Effect of environmental optimism on responsible electricity consumption with price concern as a moderator. Psychological Research on Urban Society 4. doi:10.7454/proust.v4i2.128.

Herter, K., McAuliffe, P., Rosenfeld, A., 2007. An exploratory analysis of California residential customer response to critical peak pricing of electricity. Energy 32, 25–34. doi:10.1016/j.energy.2006.01.014.

Herter, K., Wayland, S., 2010. Residential response to critical-peak pricing of electricity: California evidence. Energy 35, 1561–1567. doi:10.1016/j.energy.2009.07.022. demand Response Resources: the US and International Experience.

Hobman, E.V., Frederiks, E.R., Stenner, K., Meikle, S., 2016. Uptake and usage of cost-reflective electricity pricing: Insights from psychology and behavioural economics. Renewable and Sustainable Energy Reviews 57, 455–467. doi:10.1016/j.rser.2015.12.144.

Holladay, J.S., Price, M.K., Wanamaker, M., 2015. The perverse impact of calling for energy conservation. Journal of Economic Behavior & Organization 110, 1–18. doi:10.1016/j.jebo.2014.11.008.

Jensen, R.H., Kjeldskov, J., Skov, M.B., 2018. Assisted Shifting of Electricity Use: A Long-Term Study of Managing Residential Heating. ACM Transactions on Computer-Human Interaction 25, 25:1–25:33. doi:10.1145/3210310.

Jessoe, K., Rapson, D., 2014. Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use. American Economic Review 104, 1417–1438. doi:10.1257/aer.104.4.1417.

Kane, N., Khanna, S., Martin, R., Muûls, M., Sinha, P., Saha, S.K., 2024. Leveraging Automation and Incentives to Enhance Power Demand Flexibility. Technical Report. Imperial College London. URL: https://www.imperial.ac.uk/media/imperial-college/research-centres-and-groups/hitachi-decarbonisation/Tata_Powbal_Imperial-Report.pdf.

Karjalainen, S., 2013. Should it be automatic or manual—The occupant's perspective on the design of domestic control systems. Energy and Buildings 65, 119–126. doi:10.1016/j.enbuild.2013.05.043.

Kaspar, K., Nweye, K., Buscemi, G., Capozzoli, A., Nagy, Z., Pinto, G., Eicker, U., Ouf, M.M., 2024. Effects of occupant thermostat preferences and override behavior on residential demand response in CityLearn. Energy and Buildings doi:10.1016/j.enbuild.2024.114830.

Kim, J.H., Shcherbakova, A., 2011. Common failures of demand response. Energy 36, 873–880. doi:10.1016/j.energy.2010.12.027.

Kimura, O., Nishio, K.I., 2016. Responding to electricity shortfalls: Electricity-saving activities of households and firms in Japan after Fukushima. Economics of Energy & Environmental Policy 5, 51–72. URL: https://www.jstor.org/stable/26189398.

Kostková, K., Omelina, L., Kyčina, P., Jamrich, P., 2013. An introduction to load management. Electric Power Systems Research 95, 184–191. doi:10.1016/j.epsr.2012.09.006.

Leighty, W., Meier, A., 2011. Accelerated electricity conservation in Juneau, Alaska: A study of household activities that reduced demand 25%. Energy Policy 39, 2299–2309. doi:10.1016/j.enpol.2011.01.041.

Li, R., Dane, G., Finck, C., Zeiler, W., 2017. Are building users prepared for energy flexible buildings?—A large-scale survey in the Netherlands. Applied Energy 203, 623–634. doi:10.1016/j.apenergy.2017.06.067.

Ludwig, P., Winzer, C., 2022. Tariff Menus to Avoid Rebound Peaks: Results from a Discrete Choice Experiment with Swiss Customers. Energies 15, 6354. doi:10.3390/en15176354.

MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2023. Cluster-robust inference: A guide to empirical practice. Journal of Econometrics 232, 272–299. doi:10.1016/j.jeconom.2022.04.001.

Muratori, M., Schuelke-Leech, B.A., Rizzoni, G., 2014. Role of residential demand response in modern electricity markets. Renewable and Sustainable Energy Reviews 33, 546–553. doi:10.1016/j.rser.2014.02.027.

Müller, F., Jansen, B., 2019. Large-scale demonstration of precise demand response provided by residential heat pumps. Applied Energy 239, 836–845. doi:10.1016/j.apenergy.2019.01.202.

Nadaraya, E.A., 1964. On estimating regression. Theory of Probability & Its Applications 9, 141–142. doi:10.1137/1109020.

Nouicer, A., Meeus, L., Delarue, E., 2020. The Economics of Explicit Demand-Side Flexibility in Distribution Grids: The Case of Mandatory Curtailment for a Fixed Level of Compensation. RSCAS Working Paper 2020/45. European University Institute, Florence School of Regulation. URL: https://cadmus.eui.eu/bitstream/handle/1814/67762/RSCAS%202020_45.pdf?sequence=1&isAllowed=y.

OECD, 2024. Household Disposable Income. URL: https://data-explorer.oecd.org/vis?lc=en&df[ds]=DisseminateArchiveDMZ&df[id]=DF_DP_LIVE&df[ag]=OECD&df[vs]=&av=true&pd=2022%2C2022&dq=BEL%2BOECD%2BOAVG.HHDI...A&to[TIME_PERIOD]=false&vw=tb. OECD Data Archive. Indicator: Household Disposable Income (2022). Accessed: 21 August 2024.

OpenAI, 2024. ChatGPT (version o1 preview) [AI language model]. URL: https://chat.openai.com/. Accessed: 16 October 2024. The input prompt is available with the replication code and data.

Ouf, M.M., Osman, M., Bitzilos, M., Gunay, B., 2024. Can you lower the thermostat? Perceptions of demand response programs in a sample from Quebec. Energy and Buildings 306, 113933. doi:10.1016/j.enbuild.2024.113933.

Peffer, T., Aczel, M., Heinemeier, K., Pingatore, C., Chung, E., 2024. Smart Thermostats plus Heat Pumps: Incompatible? Or just need counseling?, in: Proceedings of the 2024 ACEEE Summer Study on Energy Efficiency in Buildings, American Council for an Energy-Efficient Economy (ACEEE), Pacific Grove, CA. URL: https://escholarship.org/uc/item/02b1g4p8.

Richter, L.L., Pollitt, M.G., 2018. Which smart electricity service contracts will consumers accept? The demand for compensation in a platform market. Energy Economics 72, 436–450. doi:10.1016/j.eneco.2018.04.004.

Rosenow, J., Gibb, D., Nowak, T., Lowes, R., 2022. Heating up the global heat pump market. Nature Energy 7, 901–904. doi:10.1038/s41560-022-01104-8.

Royal Meteorological Institute of Belgium, 2024. Open Data - Royal Meteorological Institute of Belgium. URL: https://opendata.meteo.be/. Accessed: 29 July 2024.

Ruokamo, E., Kopsakangas-Savolainen, M., Meriläinen, T., Svento, R., 2019. Towards flexible energy demand – Preferences for dynamic contracts, services and emissions reductions. Energy Economics 84, 104522. doi:10.1016/j.eneco.2019.104522.

Sarran, L., Gunay, H.B., O'Brien, W., Hviid, C.A., Rode, C., 2021. A data-driven study of thermostat overrides during demand response events. Energy Policy 153, 112290. doi:10.1016/j.enpol.2021.112290.

Statbel, 2024a. Level of Education. URL: https://statbel.fgov.be/en/themes/work-training/training-and-education/level-education#news. Published: 28 March 2024. Accessed: 21 August 2024.

Statbel, 2024b. Structure of the population - Households. URL: https://statbel.fgov.be/en/themes/population/structure-population/households. Published: 5 June 2024. Accessed: 21 August 2024.

Statbel, 2024c. T04_21_BE_POC_CL - Nombre de logements classiques selon l'époque de construction, Nombre de logements classiques au 01-01-2021 - CENSUS - 2021. URL: https://statbel.fgov.be/sites/default/files/files/documents/Census2021/T04_POC_BE_FR.XLSX. Last updated: 19 April 2024. Accessed 21 August 2024 via: https://statbel.fgov.be/fr/themes/census/logement/epoque-de-construction.

Tomat, V., Vellei, M., Ramallo-González, A.P., González-Vidal, A., Le Dréau, J., Skarmeta-Gómez, A., 2022. Understanding patterns of thermostat overrides after demand response events. Energy and Buildings 271, 112312. doi:10.1016/j.enbuild.2022.112312.

Watson, G.S., 1964. Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 26, 359–372. URL: http://www.jstor.org/stable/25049340.

Wildstein, P.J., Craig, M.T., Vaishnav, P., 2023. Participant overrides can halve the reliability value of direct load control programs. Energy and Buildings 299, 113606. doi:10.1016/j.enbuild.2023.113606.

Xu, X., fei Chen, C., Zhu, X., Hu, Q., 2018. Promoting acceptance of direct load control programs in the United States: Financial incentive versus control option. Energy 147, 1278–1287. doi:10.1016/j.energy.2018.01.028.

Yilmaz, S., Chanez, C., Cuony, P., Patel, M.K., 2022. Analysing utility-based direct load control programmes for heat pumps and electric vehicles considering customer segmentation. Energy Policy 164, 112900. doi:10.1016/j.enpol.2022.112900.

Yilmaz, S., Cuony, P., Chanez, C., 2021. Prioritize your heat pump or electric vehicle? Analysing design preferences for Direct Load Control programmes in Swiss households. Energy Research & Social Science 82, 102319. doi:10.1016/j.erss.2021.102319.

Zhang, F., De Dear, R., Candido, C., 2016. Thermal comfort during temperature cycles induced by direct load control strategies of peak electricity demand management. Building and Environment 103, 9–20. doi:10.1016/j.buildenv.2016.03.020.

Appendix A. Additional figures and tables

Appendix A.1. Average monthly temperatures during the experimental period

Table A.5: Average monthly temperatures (in °C)

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	Average
Heating season 1 [a]	N/A	8.0	4.8	6.9	7.2	8.7	10.8	7.7
Heating season 2 [a]	11.8	8.4	7.5	4.7	9.3	10.2	N/A	8.7
Historic average [b]	11.3	7.2	4.3	3.7	4.2	7.1	10.4	6.9

[a] Source: own data.
[b] Average monthly temperature for the period 1991-2020. Source: Royal Meteorological Institute, `https://www.meteo.be/fr/climat/climat-de-la-belgique/normales-climatiques-a-uccle/temperature/temperature-moyenne`, 26 September 2024.

Appendix A.2. Sample composition

Table A.6: Household participation in the first and second heating seasons

ID	Decoupled dummy	HS1 participation	HS2 participation	Total number of interventions
1	0	●	●	52
2	0	●	●	28
3	0	●	●	56
4	1	●	○	19
5	0	●	◐	34
6	0	●	●	55
7	1	○	●	10
8	0	◑	●	12
9	1	●	●	21
Total	3	8	8	287

○ indicates that the household did not actively participate in the heating season (HS). ◑ (◐) indicates that the household actively participated in the first (second) part of the HS. The black area size does not represent the number of interventions.

Appendix A.3. Histograms of indoor temperature and heat pump power in non-intervention periods

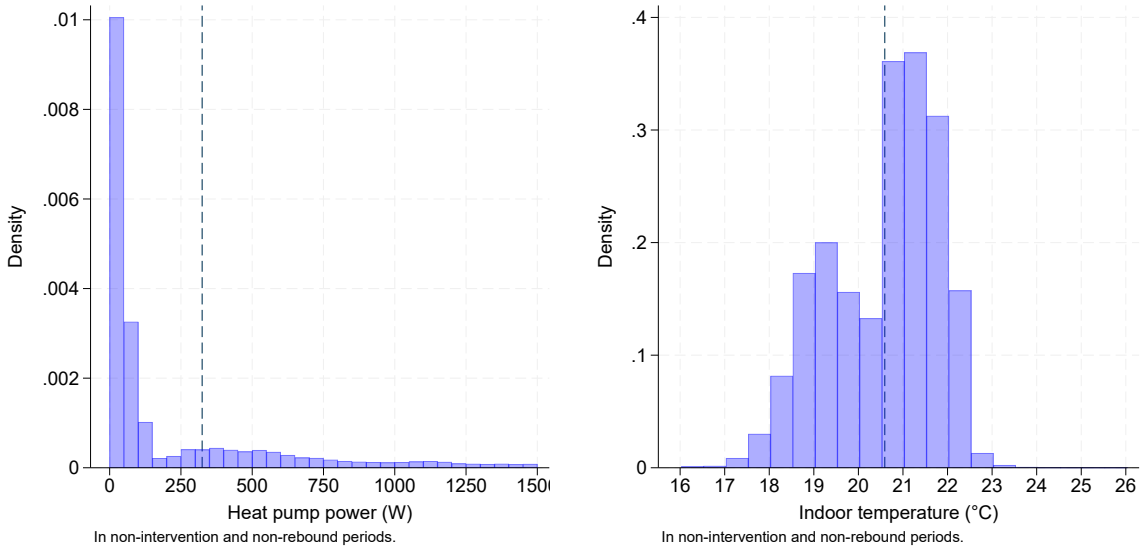


Figure A.10: Heat pump measurements during non-intervention periods. Left: power (in W), truncated at 1500 W. Right: indoor temperature (in °C), shown between 16 and 26 °C to exclude outliers. Vertical dotted lines indicate the sample averages.

Appendix A.4. Average heat pump daily profile (entire heat pump sample)

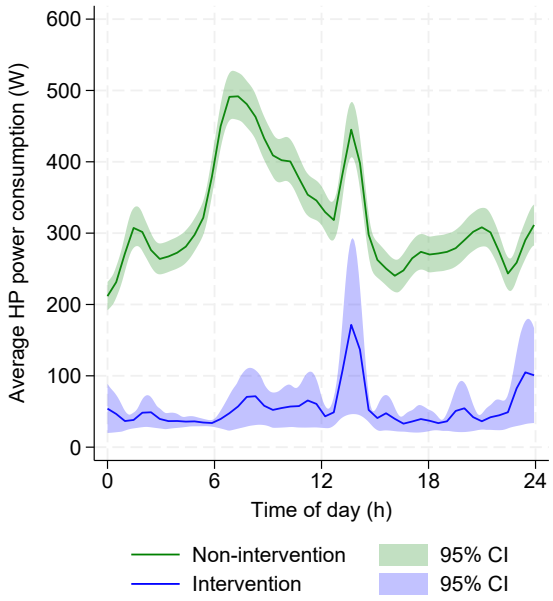


Figure A.11: Daily profile of average heat pump (HP) power during and outside intervention periods, averaged across all HPs and heating seasons. The profiles are smoothed using local polynomial regression of degree 0 for the mean and confidence intervals. Standard errors reflect the variability of the mean in 5-min-of-day bins, assuming independence among observations.

Appendix A.5. Comparison of average heat pump daily profiles across different outdoor temperature ranges

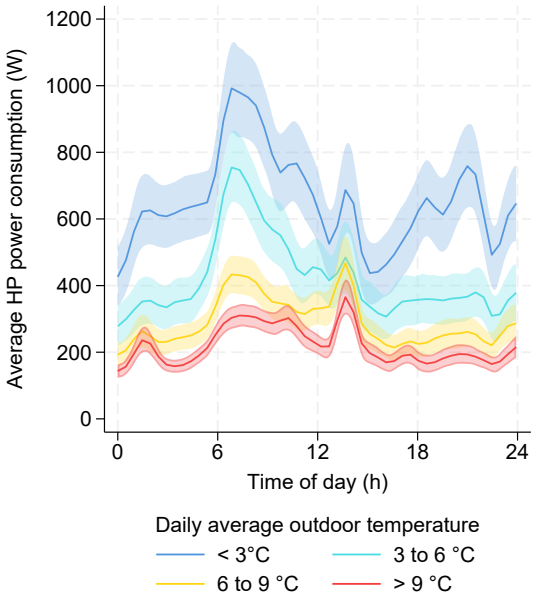


Figure A.12: Daily profile of average heat pump (HP) power during non-intervention periods (in W), averaged across all HPs and heating seasons, and categorized by different daily average outdoor temperatures. The profiles are smoothed using local polynomial regression of degree 0 for the mean and confidence intervals. Standard errors reflect the variability of the mean in 5-min-of-day bins, assuming independence among observations.

Appendix A.6. Intervention rebound period: extended duration plots

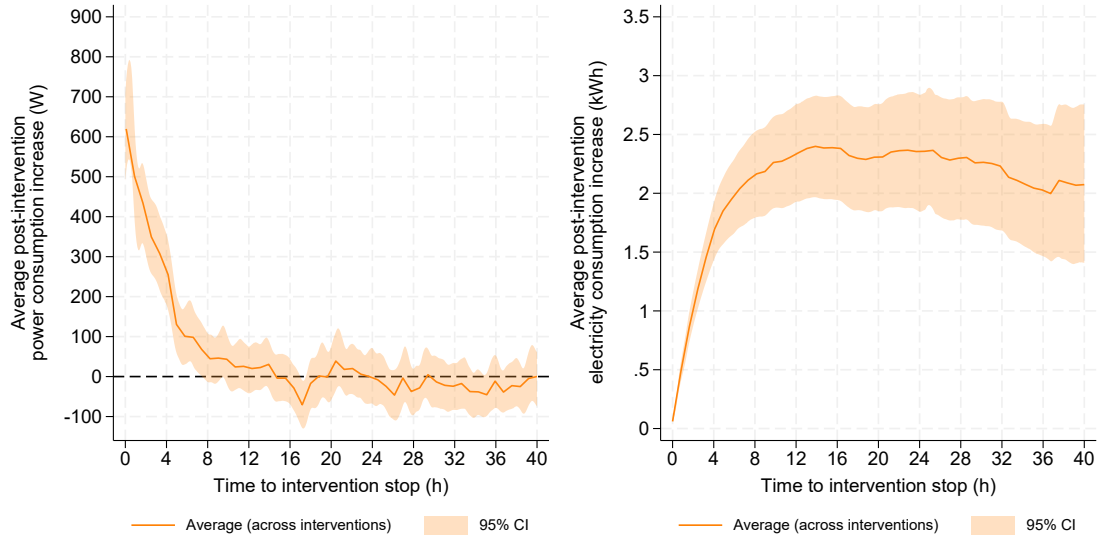


Figure A.13: Average increase in heat pump (HP) power consumption during the post-intervention rebound period (left panel, in W) and the corresponding energy consumption (right panel, in kWh) across all interventions. This figure is similar to Figure 4 but extends the plots until 40 hours in the post-intervention period. The averages are computed using the average daily HP consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual, aligned by time to intervention stop (see eq. (6)). Standard errors reflect the variability of the mean in 5-min-to-intervention-stop bins, assuming independence among observations. The means and confidence intervals are smoothed using local polynomial regression of degree 0.

Appendix A.7. Rebound consumption of heat pumps resuming normal operation at the fleet level

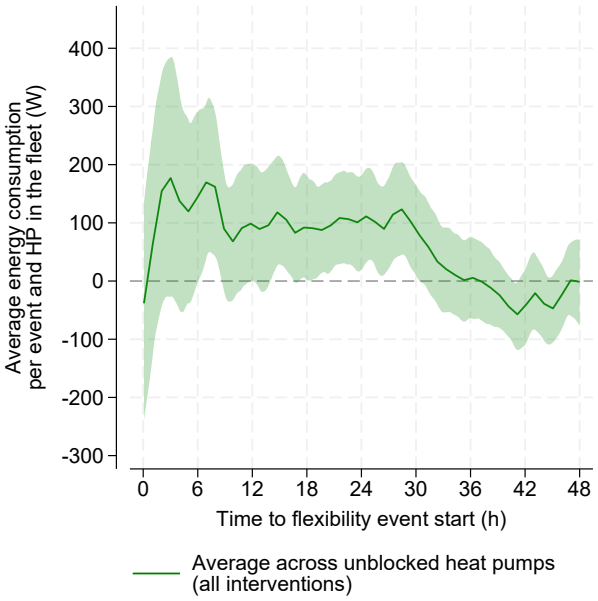


Figure A.14: Average heat pump (HP) power consumption (in W) per unit in the fleet relative to the time of intervention start for unblocked HPs that have completed their intervention and undergo rebound consumption. Standard errors reflect the variability of the means in 5-minute-to-intervention-start bins, assuming independence among observations (eq. (7)). The means and confidence intervals are smoothed using a local polynomial of degree 0.

Appendix A.8. Illustration of the two phases of flexibility events

Figure A.15 shows the average per-HP electricity reduction (in kWh) relative to the event start. The two phases observed in individual HP-level flexibility interventions are also identified at the fleet level. First, consumption decreases by up to 2 kWh during the first 18 hours. Then it rebounds, tending to level off at a net savings of about 1 kWh around 36 hours after the event start—marginally higher, but within the same order of magnitude as the net reduction estimated at 16 hours post-intervention at the individual HP level.

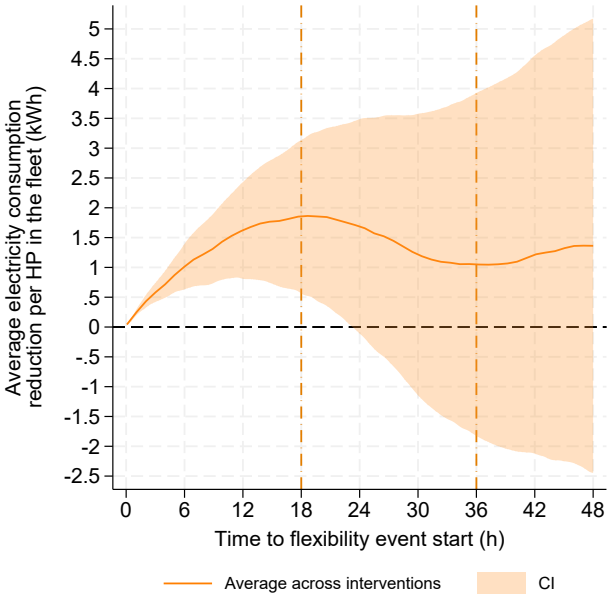
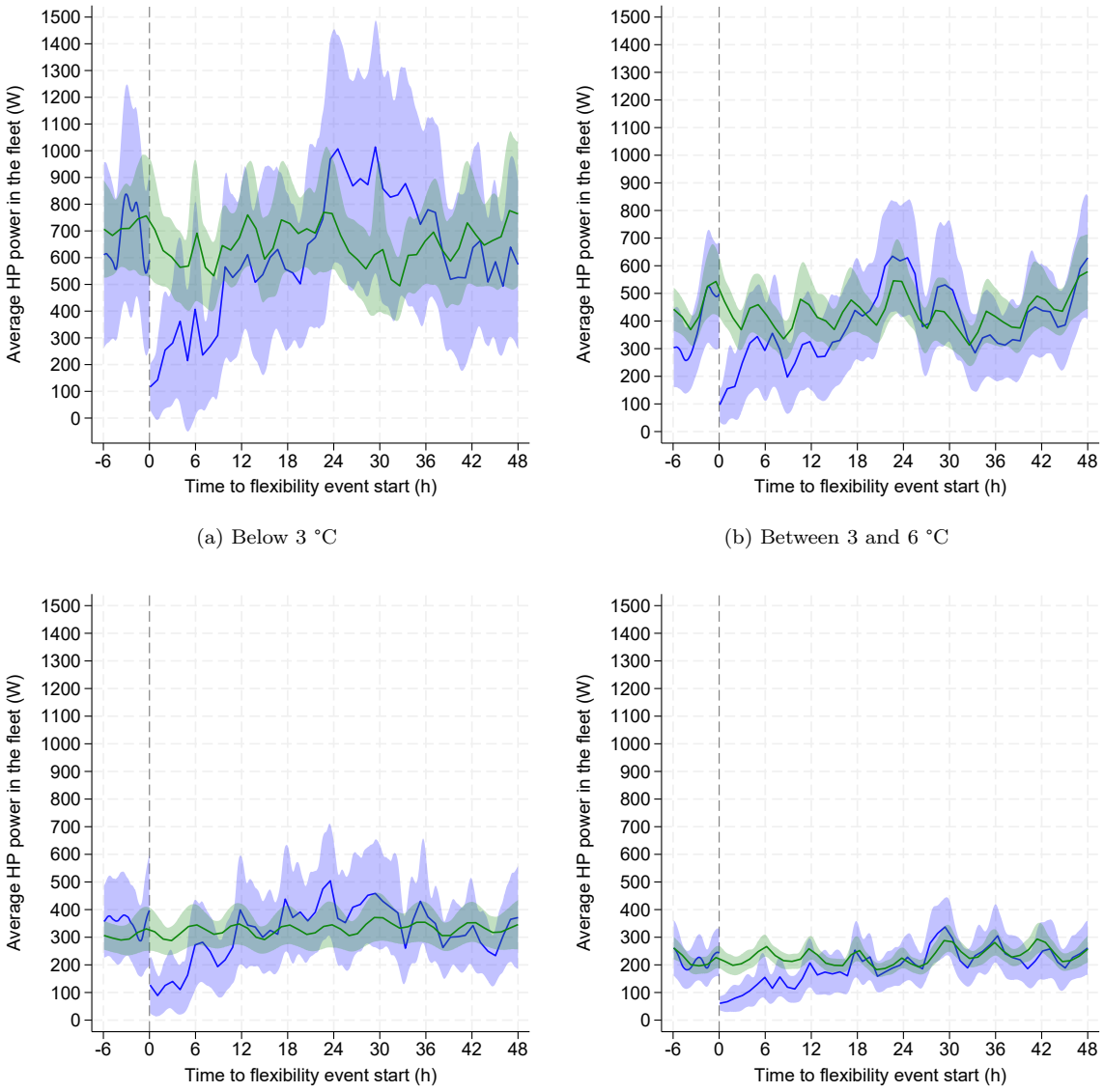


Figure A.15: Average electricity consumption reduction (in kWh) per heat pump (HP) in the fleet across all interventions. Calculated using the average HP daily consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual, aligned by time to intervention start (see eq. (6)). Vertical dotted lines mark the transition from the first phase (reduced consumption) to the rebound period and the start of the stabilizing phase. Confidence intervals are derived from the upper and lower limits of the power reduction estimates. The means and confidence intervals are smoothed using a local polynomial of degree 0.

Appendix A.9. Fleet-level power consumption profiles during flexibility events: heterogeneity across average outdoor temperature



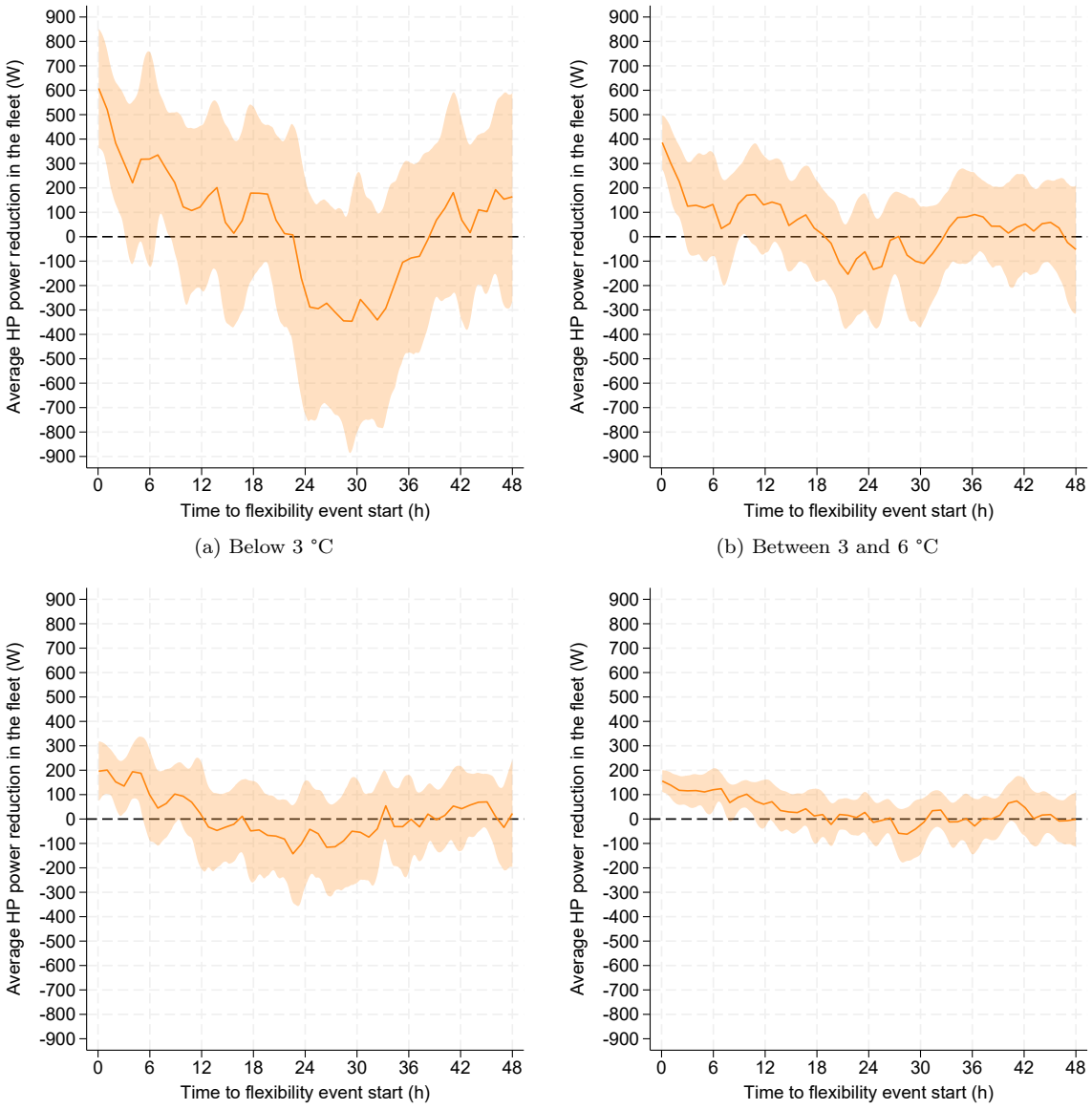
(a) Below 3 °C

(b) Between 3 and 6 °C

(c) Between 6 and 9 °C

(d) Over 9 °C

Figure A.16: Average heat pump power (HP) consumption (in W) per unit in the fleet relative to the time of flexibility event start, categorized by four outdoor temperature ranges. The temperature categories are based on the average outdoor temperature within the first 18 hours after the event starts: below 3 °C, 3 to 6 °C, 6 to 9 °C, and above 9 °C. The control curve is computed using the average daily HP consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual, aligned by time to intervention start (see eq. (5)). Standard errors reflect the variability of the means in 5-minute-to-intervention-start bins, assuming independence among observations (eq. (7)). The means and confidence intervals are smoothed using a local polynomial of degree 0, with the optimal bandwidth of the intervention curve calculated over the entire plotted period.



(a) Below 3 °C

(b) Between 3 and 6 °C

(c) Between 6 and 9 °C

(d) Over 9 °C

Figure A.17: Net average heat pump (HP) power reduction in the fleet relative to the time of flexibility event start, categorized by four outdoor temperature ranges. The temperature categories are based on the average outdoor temperature within the first 18 hours after the intervention starts: below 3 °C, 3 to 6 °C, 6 to 9 °C, and above 9 °C. The net reduction is calculated as the difference between the observed power consumption and the counterfactual, where the counterfactual is the average daily HP consumption profile for the same household, time-of-day bin and outdoor temperature category as the counterfactual, aligned by time to intervention start (see eq. (5)). Standard errors reflect the variability of the means in 5-minute-to-intervention-start bins, assuming independence among observations (eq. (7)). The means and confidence intervals are smoothed using a local polynomial of degree.

Appendix A.10. Back-of-the-envelope calculation of yearly savings through a smart allocation of flexibility events

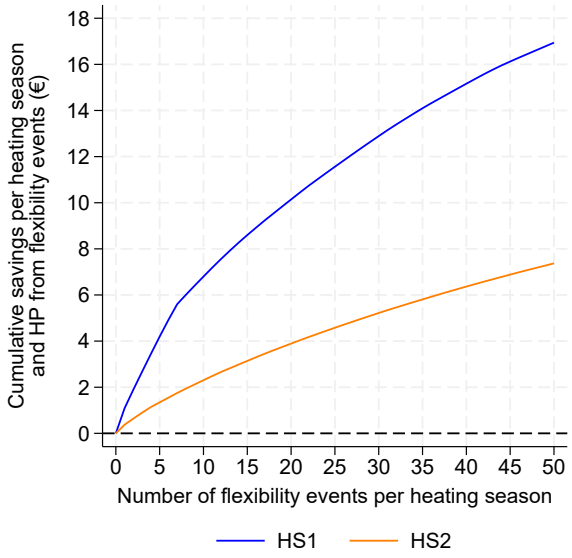


Figure A.18: Average cumulative annual savings (C_n in €, as defined in eq. (15)) per heat pump (HP) as a function of the number of flexibility events n per heating season. Curves correspond to day-ahead electricity prices observed during heating season 1 (HS1, 2022–2023) and heating season 2 (HS2, 2023–2024). Flexibility events are optimally allocated (with a minimum 48-hour interval), and annual savings are assumed constant for a given n and constant across years.

Appendix B. Notification e-mail in heating season 2

During heating season 2, half of the 32 scheduled interventions were notified a day-ahead to all participating households. This notification consisted in the following e-mail (translated from Dutch), sent by the research team:

Object: Notification FlexSys test

Content:

Dear FlexSys participant

Thank you once again for participating in the tests we are conducting in the FlexSys project with your heat pump. We would like to remind you that half of the the tests conducted this heating season will be preceded by a notification, one day in advance. With this message, we are sending you such a notification: a FlexSys test will be conducted tomorrow.

The operation of your heat pump will be temporarily blocked, while, of course, the internal safety mechanisms of the device remain unaffected.

The blockage will stop automatically when the temperature of the house or the buffer tank for sanitary hot water reaches a predetermined lower limit, or will stop immediately when you use the override button via the COFY-box platform.

We refer to the e-mail from [*experiment coordinator at the cooperative*] dated 26/10/2023 for more information about the tests.

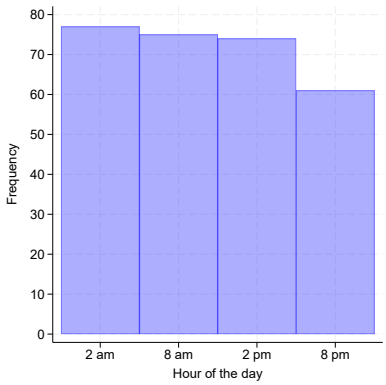
Do not reply to this automatically sent e-mail. If you have any questions or comments, you can e-mail them to:

[*research team's coordinator name and contact details*]

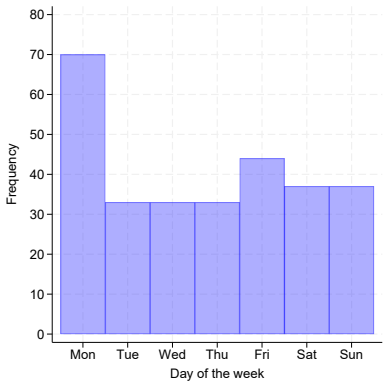
By participating in this test, you are contributing to our research on the potential of smart demand response, for which we thank you sincerely.

The FlexSys team

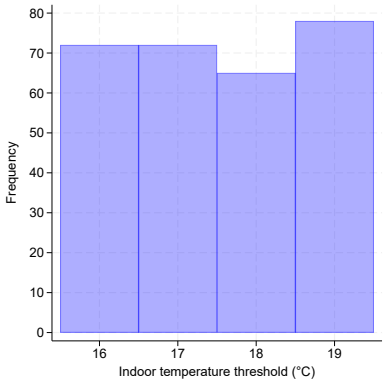
Appendix C. Distribution of the interventions per start time, indoor temperature threshold value and day of the week



(a) Distribution of intervention's starting hour of the day, TOD



(b) Distribution of intervention's starting day of the week, day



(c) Distribution of intervention's indoor temperature threshold, $T_{in}^{\text{thres.}}$

Figure C.19: Distribution of interventions by their characteristics in the intervention schedule

The software SPSS[35] was used to construct an orthogonal design to ensure no correlation between the main effects of the following intervention characteristics: the hour of the day (four levels) and the day of the week (seven levels) at which the intervention is initiated, as well as the indoor temperature threshold used for the automatic trigger back to normal operation (four levels). Additionally, all interventions in HS1 were notified a day in advance, and the schedule in HS2 included a notification status (two levels, with half of the HS2 interventions being notified a day-ahead). While 28 is the least common multiple of the levels of each characteristic, it results in a large imbalance in the combinations, as the 16 combinations of time of day and temperature threshold cannot be evenly distributed among 28 interventions. Instead, SPSS generated a 32-size array, repeating the first value of the day of the week characteristic several times to minimize the imbalance while keeping correlations in the main effects statistically insignificant. As a result, twice as many interventions were scheduled to start on Mondays compared to any other day.

In the dataset of interventions successfully achieved in practice, the Pearson correlation coefficients between the main effects remain statistically insignificant, consistent with the theoretical schedule design. Across the entire intervention sample, the correlation coefficients are: $r(T_{in}^{\text{thres.}}, TOD) = 0.006$, $r(T_{in}^{\text{thres.}}, day) = 0.018$, $r(TOD, day) = 0.051$, $r(T_{in}^{\text{thres.}}, D_{\text{notif}}) = -0.095$, $r(day, D_{\text{notif}}) = 0.023$, $r(TOD, D_{\text{notif}}) = -0.053$. All six correlation coefficients are insignificant at the 5% level, meaning that the 287 studied interventions are random across the starting hour of the day, day of the week, and indoor temperature threshold and notification status.

[35]IBM Corp. Released 2023. IBM SPSS Statistics for Windows, Version 29.0.2.0. Armonk, NY: IBM Corp.

Appendix D. Cross-season comparison of key results

Table D.7 compares key experimental results for the full sample, heating season 1 (HS1), and heating season 2 (HS2). The last column reports the p-value for the difference between seasons, computed under the assumption of independent samples.[36] For relevant outcomes, counterfactuals were constructed separately for each heating season. P-values marked [i,ii] are derived from linear regressions using wild cluster bootstrapped standard errors clustered at the household level. Most outcomes do not exhibit statistically significant differences between the two seasons, with only two exceptions.

First, the mean rebound power consumption within one hour post-intervention is significantly lower in HS2 (at the 1% level). This finding aligns with earlier results showing that rebound power is negatively associated with outdoor temperature (see Appendix G). Since outdoor temperatures in HS2 were milder than in HS1 (see Appendix A.1), this likely explains the reduction.

Second, the mean maximum power reduction within one hour post-intervention is also significantly lower in HS2 (at the 3% level), although the absolute values are very close. This may reflect slightly lower HP operating levels just before interventions during the milder HS2 period.

Despite these differences, both results are similar in magnitude and direction across the two seasons, suggesting that the overall findings are robust.

[36] Although this assumption is not entirely accurate, as some households participated in both seasons.

Table D.7: Cross-season comparison comparison of key experimental results between heating season 1 (HS1) and heating season 2 (HS2)

Variable	Refer to Section	Full sample	HS1	HS2	p-value of the difference HS1-HS2
Total number of interventions		287	168	119	—
Intervention stop reasons:					
DHW temperature threshold	4.1.1	70.0%	78.0%	58.8%	—
Manual overrule		11.1%	6.5%	17.6%	
Indoor temperature threshold		18.8%	15.5%	23.5%	
Mean intervention duration (hours)		12.8 (11.3, 14.4)	12.4 (10.8, 14.1)	13.4 (10.5, 16.3)	0.58
ATT on indoor temp. during interventions (°C) [a]	4.1.2	-0.38 (-0.66, -0.10)	-0.41 (-0.69, -0.06)	-0.34 (-0.75, -0.08)	0.51 [i]
ATT on heat pump power during interventions (W) [a]	4.1.3	-292 (-390, -200)	-316 (-414, -168)	-259 (-382, -158)	0.82 [i]
Mean electricity consumption reduction during interventions (kWh)	4.1.4	3.22 (2.81, 3.63)	3.30 (2.77, 3.82)	2.87 (2.23, 3.51)	0.31
Mean rebound electricity consumption within 16 hours post-intervention (kWh)	4.1.5	2.40 (1.96, 2.85)	2.45 (1.90, 2.99)	2.56 (1.79, 3.33)	0.81
Mean rebound power consumption within 1 hour post-intervention (W)		607 (571, 642)	692 (647, 738)	486 (428, 543)	$<$ **0.01**
Mean maximum per-unit power reduction within 1 hour of a flexibility event on a fleet (W)	4.2.1	252 (234, 270)	255 (235, 275)	226 (209, 243)	**0.03**
Mean time until fleet-level rebound (hours) [b]		$\approx$ 18	$\approx$ 17	$\approx$ 21	—
Mean indoor temperature drop from intervention start to end (°C)	4.3.2	0.69 (0.59, 0.78)	0.75 (0.61, 0.90)	0.59 (0.48, 0.70)	0.08
Mean additional temperature drop: manual vs. automatic overrules (°C) [c]		0.45 (0.19, 0.70)	0.72 (0.38, 1.06)	0.19 (0.14, 0.52)	0.10 [ii]

The 95% confidence intervals are in parentheses. All p-values for the comparison of means between HS1 and HS2, except [i,ii], are derived from unpaired t-tests with unequal variances (Welch's test).

[a] Estimated via linear regression with household fixed effects using 100,000 wild bootstrap replications to cluster standard errors at the household level. ; [b] Based on visual inspection of local polynomial regression plots, identifying the approximate rebound time. ; [c] Excludes preemptive overrules.

[i] P-value for the interaction term between the intervention indicator and heating season, which captures the additional effect of interventions in HS2 relative to HS1. This is obtained from a linear regression of the outcome on the intervention indicator, heating season, and their interaction, using 100,000 wild bootstrap replications to cluster standard errors at the household level. ; [ii] Similarly, p-value for the interaction term between manually overruled interventions and heating season, capturing the additional effect in HS2 relative to HS1, obtained from a linear regression with a wild bootstrap procedure (100,000 replications, clustering standard errors at the household level).

Appendix E. Likert-scale questions in the pre-experiment survey

Appendix E.1. Participants' comprehension of flexibility-related concepts

Participants' comprehension of concepts related to residential electricity flexibility was probed similarly to (Li et al., 2017). Specifically, we used a similar 1–4 scale (1: "Never heard of it", 2: "I have heard of it, but I do not understand the concept", 3: "I know a bit about the concept", 4: "I know a lot about the concept") and asked about the following items, which were randomly ordered and presented in Dutch in the survey:

- "Energy transition",
- "Smart home",
- "Electricity flexibility",
- "Demand-response programs".

Appendix E.2. Participants' attitudes towards the environment

To probe respondents' attitudes towards the environment, the following four relevant items were extracted from a previous related study on Belgian energy cooperative members (Bauwens and Devine-Wright, 2018) and probed via a 1–5 scale (1: "Strongly disagree", 2: "Disagree", 3: "Neither agree, nor disagree", 4: "Agree", 5: "Strongly agree") (randomly ordered and in Dutch in the survey) on the following items:

- "I want to feel that I am personally contributing to the protection of the environment.",
- "I am concerned about climate change.",
- "I am the type of person who cares about the environment.",
- "I see myself as an environmentally conscious consumer.".

Appendix E.3. Participants' propensity to engage with electricity-saving habits

The frequency at which respondents engage with typical electricity-conserving practices were probed, following what was done in (Herabadi et al., 2021) from which we extracted the following nine items (probed on a 1–5 scale: 1: "Never", 2: "Rarely", 3: "Sometimes"; 4: "Often", 5: "Always" ; randomly ordered and presented in Dutch in the survey):

- "I make sure the lights are off before I leave a room.",
- "I use natural light as a light source.",
- "I use energy-saving lamps (e.g., LED lamps).",
- "I unplug the power plug when not in use.",
- "I turn off PCs/laptops when they are not in use (turned off, not in sleep mode).",
- "I choose electronic devices (not lighting) that use the least energy even if they are a bit more expensive to purchase.",
- "I make sure that the refrigerator door is not open too long.",
- "I set a moderate temperature for my heating system.",
- "I reduce the use of warm water for bathing (e.g. use cold water in warm/hot weather)."

Appendix F. Counterfactual power consumption accuracy and bias diagnosis

In this Appendix, we assess the validity of the counterfactual power consumption generated by eq. (3). We test whether the counterfactual reproduces the patterns observed in the non-intervention data[37]. First, we compare the counterfactual's root mean square error (RMSE) across different counterfactual models. Second, we analyze prediction errors by regressing them on observable variables to detect potential bias. Because our main analysis focuses on hourly trends during interventions—which typically last about 13 hours—we aggregate the 5-minute data into multi-hour bins. This reduces high-frequency noise and highlights the sustained trends relevant to the analysis of interventions.

Appendix F.1. Counterfactual power consumption errors

Following eq. (3), the counterfactual heat pump (HP) power consumption model used in the main analysis computes the average consumption within each 5-minute-of-day bin for each household and for each daily average outdoor temperature category (with four categories in total). We then match these counterfactual values to each observation in the intervention dataset that shares the same household, time-of-day bin, and temperature category. To assess the goodness of fit, we define the prediction error for each observation k as: $\epsilon_k = P_k - \hat{P}_k$ and compute the RMSE over the non-intervention dataset as:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(P_k - \hat{P}_k)^2}{N}} \tag{F.1}$$

Here, N is the total number of observations in the training dataset, P_k is the observed power consumption for observation k and $\hat{P}_k$ is the corresponding counterfactual power consumption predicted by eq. (3). RMSE is expressed in Watts, is always non-negative and increasingly penalizes larger deviations between observed and predicted values.

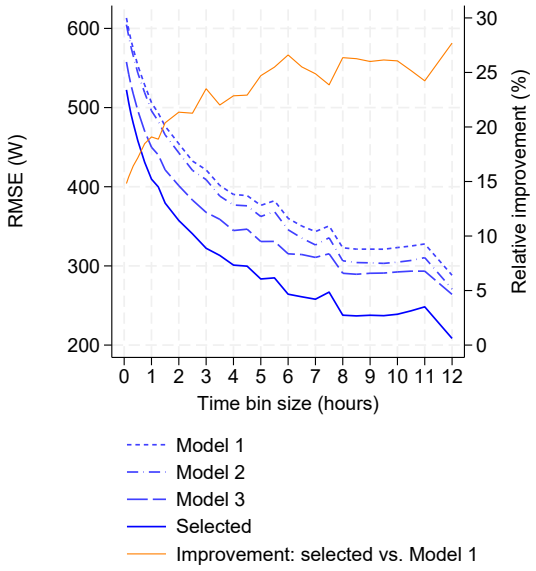


Figure F.20: Root mean square error (RMSE, in W) of predicted counterfactual heat pump (HP) power consumption, for different models and increasing levels of time aggregation. Model 1 uses the global sample mean as counterfactual; Model 2 uses the within-household mean; Model 3 uses the within-household and time-of-day mean; and the 'Selected' model (used throughout this paper) further includes daily average outdoor temperature categories (cf., eq. (3)). RMSE is calculated on the non-intervention dataset, excluding observations too close to the start or end of interventions. Relative improvement of the Selected model is expressed as the RMSE percentage reduction relative to Model 1 (%). Time aggregation is applied by grouping 5-minute observations into larger bins (reestimating the models each time), aligning the analysis with the study's focus on hourly intervention patterns.

Figure F.20 shows how RMSE evolves as observations are aggregated into longer intervals. We compare four models: Model 1 is the most naive model and uses the overall sample mean across

[37]To reduce bias, observations too close to the start (within 20 minutes) or to the end (within 16 hours) of an intervention are removed from this non-intervention dataset.

all households and time periods as the counterfactual; Model 2 uses the within-household average; Model 3 uses the within-household and time-of-day bin mean. The last model, referred to as 'Selected', uses the within-household, time-of-day bin, and daily average outdoor temperature bin as the counterfactual (see eq. (3)). It corresponds to the counterfactual defined in Section 3.1 and used throughout Section 4. Each model[38] is re-estimated for each time bin.

We observe that the root mean square prediction error of the counterfactuals predicted by all models decreases as data are aggregated in larger bins. From 5-minute observations to 12-hour aggregation, RMSE is approximately halved for all models. While the first three models (which omit matching of the counterfactual on the daily average outdoor temperature) perform similarly, the 'Selected' model—which includes the temperature pairing—performs considerably better than Model 3 at all multi-hour bin sizes. Overall, our Selected model reduces Model 1's RMSE by around 20–30% under multi-hour aggregation. In conclusion, these results validate our choice of counterfactual model: although it may not capture 5-minute fluctuations, its performance improves substantially at larger time bins compared to models that do not match on outdoor temperature categories.

Appendix F.2. Regression analysis of prediction errors

Our panel dataset indexes each observation k by household h and datetime t. For each observation (h, t) at the 5-min-of-day level, we now define the prediction error as $\epsilon_{h,t} = P_{h,t} - \hat{P}_{h,t}$. To diagnose potential bias in the counterfactual HP power consumption, we estimate the linear regression model specified in eq. (F.2).

$$\begin{aligned}\epsilon_{h,t} =& \beta_1 \cdot T_{\text{in},h,t} + \beta_2 \cdot T_{\text{DHW},h,t} + \beta_3 \cdot T_{\text{out},h,t} + \beta_4 \cdot \min\left(T^{\text{daily}}_{\text{out},h,t}\right) + \beta_5 \cdot D_{\text{HS2},h,t} \\ &+ \sum_{\tau=1}^{23} \gamma_\tau \mathbb{1}\{HOD(t) = \tau\} + \sum_{d=1}^{6} \delta_d \mathbb{1}\{DOW(t) = d\} + \alpha_h + \varepsilon_{h,t}\end{aligned} \tag{F.2}$$

In eq. (F.2), $T_{\text{in},h,t}$ is the indoor temperature, $T_{\text{DHW},h,t}$ is the domestic hot water tank temperature, $T_{\text{out},h,t}$ is the outdoor temperature, $\bar{T}^{\text{daily}}_{\text{out},h,t}$ is the average daily outdoor temperature, all measured at household h and time t. The variable $D_{\text{HS2},h,t}$ is a dummy indicator equal to 1 during heating season 2 (with heating season 1 as the reference). The indicator functions $HOD(t)$ and $DOW(t)$ capture fixed effects for the hour of day[39] and day of week respectively, and α_h represents household fixed effects.

In this diagnostic approach, our primary focus is on the overall explanatory power of the model and on the statistical significance of the coefficients, rather than the magnitude or sign of individual estimates. Table F.8 presents regression results for four models based on eq. (F.2) at different binning levels, starting with the unbinned data (5-minute-of-day level) and up to 12-hour aggregation. We find that binning the data slightly increases the overall adjusted R^2, suggesting that aggregation helps clarifying the underlying patterns, although explanatory power remains low in all models. Across the models, the absolute value of statistically significant estimates decreases with larger bin sizes. Outdoor temperature (or the daily minimum temperature, significant in Model (1)) consistently explains some of the variation in prediction errors and is significant across models (although borderline significant in Model (4)). Similarly, the difference between heating seasons is statistically significant.

In summary, these results suggest that the counterfactual is somewhat biased with respect to heating season and outdoor conditions. However, given that all models exhibit very low explanatory power (with the highest adjusted R^2 being about only 3% in Model (3)), we conclude that the bias is negligible and that most of the variation in the errors is due to random noise, likely because of the limited sample size.

To conclude this Appendix, the RMSE and bias analyses confirm that the counterfactual HP power consumption generated by eq. (3) is sufficiently accurate for our intervention analysis. Although more advanced methods (e.g., machine learning or matching) may further improve accuracy, their benefits are likely negligible given our sample size constraint. It is worth noting that our approach relies on two underlying assumptions: that prediction errors are scale-invariant (given the RMSE definition) and that any bias is linear.

[38] For Models 3 and 4, the duration of the time-of-day bin matches the time bin size on the x-axis.

[39] When using daily bins of p hours, the hour-of-day fixed effects terms in eq. (F.2) are redefined as $\sum_{\tau=0}^{\frac{24}{p}-1} \gamma_\tau \mathbb{1}\{HOD(t) = \tau\}$.

Table F.8: Linear regression estimates for prediction errors at different temporal aggregation levels

	Unbinned	Binned		
	(1)	(2)	(3)	(4)
	5-min-of-day	4 hours	8 hours	12 hours
T_{in}	-7.743	-16.071	-10.392	-6.568
	(0.689)	(0.339)	(0.491)	(0.624)
T_{DHW}	-5.798^{+}	-0.027	0.914	2.312
	(0.066)	(0.948)	(0.297)	(0.192)
T_{out}	-13.826**	-9.499**	-6.888**	-6.656^{+}
	(< 0.01)	(< 0.01)	(0.008)	(0.075)
$\min\left(T_{\text{out}}^{\text{daily}}\right)$	6.497**	2.846	0.447	0.643
	(0.004)	(0.159)	(0.811)	(0.815)
D_{HS2}	60.790**	58.559**	57.377**	53.924*
	(< 0.01)	(< 0.01)	(< 0.01)	(0.014)
Fixed effects:				
Hour-of-day [a]	Yes	Yes	Yes	Yes
Day-of-week	Yes	Yes	Yes	Yes
Household	Yes	Yes	Yes	Yes
Adj. R-Square	0.013	0.020	0.026	0.031
N observations	509,304	10,960	5,662	3,951

Linear regression estimates for counterfactual prediction errors $\epsilon_{h,t}$ at different temporal aggregation levels (specification in eq. (F.2)). Data are aggregated over time within each household-day. Model (1) uses unbinned (5-minute-of-day) data, while Models (2)–(4) use data aggregated into 4-, 8-, and 12-hour bins, respectively. All models include fixed effects for hour-of-day (or bins), day-of-week, and household; heating season 1 serves as the reference category for D_{HS2}. P-values (in parentheses) are based on wild cluster bootstrapped standard errors (100,000 repetitions), clustered at the household level (nine clusters in all models). $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$.

[a] In Models (2)-(4), multi-hour binning redefines the hour-of-day fixed effects into binned categories (with one omitted to avoid collinearity).

Appendix G. Regression analysis of the rebound energy consumption in the post-intervention period

In this Appendix, we analyze the factors influencing rebound energy consumption in the post-intervention period ($E^{16\text{h}}_{\text{rebound},i}$, in kWh), defined as the additional electricity required within 16 hours after an intervention i for the heat pump (HP) of household h to return to user setpoints. We specify the regression model as:

$$\begin{aligned} E^{16\text{h}}_{\text{rebound},i,h} =& \beta_1 \cdot \overline{T}^{\leq 16\text{h}}_{\text{out},i,h} + \beta_2 \cdot \Delta T^{f}_{\text{setpoint},i,h} + \beta_3 \cdot T^{f}_{\text{in},i,h} \\ &+ \beta_4 \cdot \text{TOD}^{f}_{\text{AM},i,h} + \beta_5 \cdot \text{TOD}^{f}_{\text{evening},i,h} + \beta_6 \cdot \text{TOD}^{f}_{\text{night},i,h} \\ &+ \mathbb{1}\{\text{FE} = 0\} \cdot \beta_0 + \mathbb{1}\{\text{FE} = 1\} \cdot \alpha_h + \varepsilon_{i,h} \end{aligned} \tag{G.1}$$

Where $E^{16\text{h}}_{\text{rebound},i,h}$ denotes the additional energy consumption (in kWh) observed within 16 hours after the intervention ends, relative to the counterfactual derived from the average daily HP consumption profile for the same household, time-of-day bin, and outdoor temperature category during non-intervention periods, aligned by time to intervention stop (see eq. (5)). $T^{f}_{\text{in},i,h}$ is the indoor temperature at the end of the intervention, while $\Delta T^{f}_{\text{setpoint},i,h} = T_{\text{setpoint,i,h}} - T^{f}_{\text{in},i,h}$ measures its deviation from the thermostat setpoint specified by the household. $\overline{T}^{\leq 16\text{h}}_{\text{out},i,h}$ represents the average outdoor temperature within the 16-hour rebound period[40]. The three TOD dummies indicate whether the intervention ended in the morning (6 a.m. - 12 p.m.), evening (6 p.m. - 12 a.m.) or night (12 a.m. - 6 a.m.), with the afternoon as the baseline category. Finally, α_h represents household FE, capturing within-household variation in the parameters by controlling for household characteristics that remain invariant across interventions.[41]

We estimate the model specified in eq. (G.1) via a linear regression and account for the small number of clusters (nine heat pumps) by using wild cluster bootstrap (100,000 repetitions) at the HP-level to compute cluster-robust standard errors. We report the p-values derived from the empirical distribution of the bootstrapped estimates. Table G.9 presents the results for four models, from a parsimonious specification to the full model with household FE in eq. (G.1). All estimates show the expected signs and remain robust across the four specifications, with the exception of the intercept in Model (1). Rebound energy consumption is found to decrease as outdoor temperatures increase (as heat loss during the rebound period is then less pronounced) and to increase when the indoor temperature at the end of an intervention is below the setpoint (although evidence for the latter is mixed, with estimates being borderline significant in models that include household FE).

Interestingly, when comparing Models (1) and (2)—both estimated without household FE—, the coefficient for T^{f}_{in} in Model (1) is insignificant, whereas replacing it with $\Delta T^{f}_{\text{setpoint},i,h}$ in Model (2) results in a statistically significant parameter. This suggests that the rebound is actually driven by the difference between the indoor temperature at the end of an intervention and the user setpoint, rather than by the value of the indoor temperature. Hence, $\Delta T^{f}_{\text{setpoint},i,h}$ better captures the true data generation process.

Including household FE results in higher adjusted R^2 in Models (3) and (4). In Model (3) (achieving highest adjusted R^2), $\Delta T^{f}_{\text{setpoint},i,h}$ has the highest marginal effect on $E^{16\text{h}}_{\text{rebound},i,h}$, although the estimate is only borderline significant. A one-degree increase in the difference between the user's setpoint temperature and the indoor temperature at the end of an intervention is associated with an average rebound increase of 0.80 kWh within a household ($p = 0.06$), as HPs compensate for larger temperature deviations. Additionally, a one-degree increase in the average outdoor temperature within the rebound window reduces rebound by 0.28 kWh on average within a household (strongly significant at the 1% level). Once these parameters are controlled for, the TOD dummies included in Model (4) are insignificant, indicating no evidence that the period of the day when the intervention ends affects rebound consumption.

[40] The average outdoor temperature marginally improved adjusted R^2 compared to other parametrizations, such as the minimum temperature.

[41] Additional specifications were tested, including models with explicit parameters for DHW temperature at the end of the intervention. However, the corresponding estimates were found to be insignificant for both decoupled and non-decoupled HPs. This aligns with expectations: for decoupled units, the rebound is unrelated to DHW reheating, as the hot water buffer is already allowed by the system to be reheated during the intervention if needed. For non-decoupled units, as most interventions stopped due to the DHW automatic threshold, there is only limited variability in T^{f}_{DHW} around 40 °C at the end of an intervention, resulting in an insignificant parameter.

Table G.9: Linear regression results for energy consumption rebound 16 hours after intervention stop (in kWh)

	(1)	(2)	(3)	(4)
$\overline{T}_{\text{out}}^{\leq 16\text{h}}$	-0.345**	-0.315**	-0.276**	-0.283**
	(< 0.01)	(< 0.01)	(< 0.01)	(< 0.01)
T_{in}^{f}	-0.256			
	(0.572)			
$\Delta T_{setpoint}^{f}$		1.046**	0.795+	0.774+
		(0.007)	(0.061)	(0.073)
$\text{TOD}_{\text{AM}}^{f}$				0.042
				(0.923)
$\text{TOD}_{\text{evening}}^{f}$				-0.619
				(0.286)
$\text{TOD}_{\text{night}}^{f}$				0.032
				(0.961)
Constant	10.203	4.643**		
	(0.377)	(< 0.01)		
Household-FE	No	No	Yes	Yes
Adj. R-Square	0.148	0.262	0.308	0.303
N observations	261	261	261	261

Linear regression estimates (specification in eq. (G.1)). Models (3) and (4) include household fixed effects. The reference category for TOD^{f} is afternoon (12 a.m. - 6 p.m.). P-values (in parentheses) are obtained from wild cluster bootstrapped standard errors (100,000 repetitions) clustered at the household level (nine clusters for all models). $^{+}p < 0.1$,$^{*}p < 0.05$, $^{**}p < 0.01$.

Appendix H. Correlation analysis of intervention habituation over time

In this Appendix, we study whether households overrule interventions less frequently over time, which we refer to as habituation. Due to sample limitations, we cannot control for exogenous factors (e.g., weather conditions) or endogenous factors (e.g., temperature drops as a proxy for discomfort). As a result, this analysis is illustrative and does not determine the specific channels through which habituation may occur, such as learning effects (where households adjust their comfort expectations based on experience) or response fatigue (where logging into the experiment platform becomes perceived as increasingly burdensome). We make the following assumptions:

(i) Habituation follows a linear pattern within a given period: either a single heating season (HS) or the entire experiment timeframe.

(ii) Habituation is driven by the cumulative number of interventions experienced, as reflected by the intervention sequence number.

(iii) Habituation is an individual behavior and should be examined at the household level.

Based on these assumptions, we quantify habituation by computing the correlation between a binary indicator for a manual overrule and the intervention sequence number (i.e., the order in which each household experienced a given intervention)[42]. Table H.10 presents these correlation coefficients, with rows corresponding to individual households and columns representing different timeframe definitions. The first two columns show within-season habituation, ranking interventions in the order they were experienced within HS1 and HS2 separately. The third column pools interventions across both heating seasons, ranking them in the order they were experienced overall, to assess overall habituation[43].

Three households never used the overrule button. Since all participants were clearly informed and reminded about this option, we interpret this as an indication that these households were not significantly discomforted by the interventions (or at least not beyond their expectations)[44].

Among households that overruled, only one correlation coefficient is statistically significant at the 5% level: household 3 in HS1, which suggests a moderate tendency to overrule interventions experienced earlier. However, this result is based on only three overrules and should be interpreted with caution.

Overall, the small number of overrules per household limits statistical power. Therefore, we relax assumption (iii) and pool all households, revealing a weak within-season habituation pattern that is statistically significant for HS1 (not for HS2 or the pooled data). This suggests that, households overall were initially more engaged with the overrule button when interventions were new, likely because these interventions represented a more noticeable change in their heating[45].

To conclude this Appendix, we find no strong evidence of habituation at the household level, likely due to the small sample size. A weak habituation emerges for heating season 1 when pooling all households, but this result is purely illustrative. A more robust analysis with a larger sample could use a logit regression of the manual overrule dummy on explanatory variables (e.g., temperature drop, outdoor temperatures, household fixed effects, and day-of-week fixed effects) to better capture the factors influencing overrule behavior.

[42] Because some households joined the experiment later or because certain interventions failed to initiate for some households, the same calendar date for an intervention may correspond to different sequence numbers. For example, the intervention on January 30, 2023, was the eighth intervention for household 4 but the first for household 8.

[43] This distinction does not apply to households 4 and 7, which participated in only one heating season each.

[44] A Welch's t-test shows that households 1, 2, and 8 experienced a lower average temperature drop per intervention compared to the other households: 0.32 °C ($N = 91$) compared to 0.86 °C ($N = 195$). The difference is statistically significant at the 1% level. No clear patterns in household characteristics were found to explain this discrepancy.

[45] This is unlikely to result from systematically higher discomfort at the start of the season. Indeed, we find no significant correlation between the temperature drop per intervention and the intervention sequence number within a heating season, despite later interventions typically occurring in warmer months (e.g., February or March; cf., Appendix A.1).

Table H.10: Correlation between manual overrule and intervention sequence number per household and heating season

ID	Correlation between manual overrule and intervention sequence number		
	Within HS1	Within HS2	Across both HS
1	/	/	/
2	/	/	/
3	**-0.46*** **(0.01)** $n = 3$	-0.16 (0.43) $n = 9$	0.09 (0.49) $n = 12$
4	0.16 (0.50) $n = 15$	—	0.16 (0.50) $n = 15$
5	-0.02 (0.94) $n = 5$	0.40 (0.22) $n = 6$	0.30^{+} (0.08) $n = 11$
6	/	-0.15 (0.45) $n = 3$	0.16 (0.24) $n = 3$
7	—	0.04 (0.92) $n = 3$	0.04 (0.92) $n = 3$
8	/	/	/
9	0.29 (0.57) $n = 3$	-0.15 (0.58) $n = 7$	-0.08 (0.73) $n = 10$
Full sample	**-0.18*** **(0.02)** $n = 26$	-0.15 (0.11) $n = 28$	-0.08 (0.19) $n = 54$

Correlation coefficients between a binary indicator for manual overrules and the intervention sequence number, reflecting the order in which each household experienced the interventions. Column 1 shows results for experiment heating season 1 (HS1), column 2 for heating season 2 (HS2), and column 3 pools interventions across both seasons, ranking them by the order in which they were experienced throughout the experiment. P-values are in parentheses, and n indicates the number of manual overrules for each household ID and season. Households that did not participate in a given HS are marked "—", while those that never used the overrule button in a given HS are marked "/". $^{+}p < 0.1$,$^{*}p < 0.05$