NBER WORKING PAPER SERIES

CRITICAL PEAK REWARDS: EVIDENCE FROM A NATIONWIDE DEMAND RESPONSE PROGRAM

Louise Bernard
Robert D. Metcalfe
Andrew Schein

Working Paper 35485
http://www.nber.org/papers/w35485

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2026

We thank Joe Grainger, from Centre for Net Zero (CNZ), for his critical project management. We thank Simona Burchill, Aaron Cawte, Jessica Colleen, Francesca Corby, Alex Schoch, Kieron Stopforth, and many others at Octopus Energy Limited (OEL), who carried out OEL's implementation of Demand Flexibility Service and our natural field experiment. Their commitment to these twin implementations was remarkable. We thank Lucy Bourke, from CNZ, for her logistics support and background research. We are very grateful for comments and review from Andy Hackett, Gareth Jones, Izzy Woolgar, and Lucy Yu, from CNZ; Iona Penman and Jessica Colleen, from OEL; and seminar participants at Toulouse School of Economics and the World Congress of Environmental and Resource Economists. Conflicts of Interest: AS and LB are paid employees of Octopus Energy Group Limited (OEG), working as researchers in CNZ, and RM is a consultant for CNZ. CNZ is an autonomous research institute within OEG, which also owns OEL. The three authors made final decisions regarding experimental design, data analysis, and writing; all mistakes herein are ours alone. AEARCTR-0012451, AEARCTR-0013031, AEARCTR-0013069, AEARCTR-0013068, AEARCTR-0013034. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Louise Bernard, Robert D. Metcalfe, and Andrew Schein. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Critical Peak Rewards: Evidence from a Nationwide Demand Response Program
Louise Bernard, Robert D. Metcalfe, and Andrew Schein
NBER Working Paper No. 35485
July 2026
JEL No. Q40

ABSTRACT

Demand flexibility at scale is likely to be an important part of the future energy system. This study evaluated the world's first nationwide domestic and commercial demand flexibility program. We analyzed the policy using a natural field experiment, with a sample of 2.7 million, assigning 119,999 customers to a control group and the rest to receive randomized encouragement to sign up to the demand flexibility program. The program paid consumers to reduce their demand below a predicted baseline at peak demand across 16 events (i.e., when the system was stressed) during the Winter of 2023–24. Customers who opted into events reduced demand during events by 23.1% among compliers, at an average cost for the system operator of £2,325 per MWh. The adoption of solar panels and batteries, heat pumps, and EVs all increased demand response.

Louise Bernard
Centre for Net Zero
louise.bernard@centrefornetzero.org

Andrew Schein
Centre for Net Zero
andrew.schein@centrefornetzero.org

Robert D. Metcalfe
Columbia University
School of International
and Public Affairs
and Centre for Net Zero
and also NBER
rdm2176@columbia.edu

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/12451,
https://www.socialscienceregistry.org/trials/13031,
https://www.socialscienceregistry.org/trials/13069,
https://www.socialscienceregistry.org/trials/13068,
https://www.socialscienceregistry.org/trials/13034

1 Introduction

The effective management of energy supply and demand is a major responsibility for grid operators. Failure to do so leads to extreme price spikes and, in the worst cases, blackouts. Despite the risk such disruption poses to economic well-being (Allcott et al., 2016, Burgess et al., 2020, Burlando, 2014, Cole et al., 2018, Fisher-Vanden et al., 2015, Fried and Lagakos, 2023, Gertler et al., 2016), blackouts are frequent occurrences across the world, both in developing and developed countries. While their causes vary, contributing factors include uncertainties in energy imports (Fotis et al., 2023, Złotecka and Sroka, 2018), the variable nature of renewable energy generation (Carreras et al., 2021, Gowrisankaran et al., 2016, Karaduman, 2021, Masood et al., 2018, Wolak, 2022, Yan et al., 2018), the growing incidence of extreme temperature events (Feng et al., 2022, Jahn et al., 2022, Panteli and Mancarella, 2015), and voltage regulation issues (Morão, 2026). Even small shortfalls in supply can cause cascading effects to all consumers in a given market (Borenstein et al., 2023).

We investigated a nationwide policy that created a market to pay consumers to reduce electricity use during periods of system stress, when blackouts could occur or carbon-intensive generation would otherwise be required. This approach — which we refer to as "critical peak rewards"[1] — contrasts with traditional demand response programs that rely on higher prices, often limiting participation and raising equity concerns. To our knowledge, this was the first nationwide program to offer direct financial rewards for voluntary demand reduction. We were able to randomize encouragement to sign up to this market across 2.7 million customers, offering the largest ever natural field experiment on energy demand.

The policy, known as the UK's Demand Flexibility Service (DFS), was created in response to the 2022–23 energy crisis to test consumer willingness to reduce demand during peak periods. Designed and implemented by the system operator in the UK, the National Energy System Operator (NESO; at the time known as National Grid ESO), DFS began as an emergency measure in Winter 2022–23, continued in 2023–24, and transitioned into a regular market service in 2024–25. Our nationwide natural field experiment ran during Winter 2023–24, across 16 events called during periods of peak electricity demand.[2] Events varied in duration, ranging from one to one and a half hours.

[1] We thank Lucas Davis for this suggestion.

[2] By peak demand, we mean electricity usage during times when the British power grid was most at risk of supply being insufficient to meet total demand. These periods generally occur from 16:00–19:00, Monday through Friday.

Events were organized by the system operator but provided by energy retailers and other aggregators of customers. The system operator gave providers a financial incentive ranging from £150 to £6,000 for every megawatt-hour (MWh) of reduced energy demand – with an average accepted bid of £3,100/MWh. Providers were then free to determine how to use their payment from the system operator, allocating some or all of it to incentivizing consumers to reduce demand. Octopus Energy Limited – the energy utility with whom we partnered – offered customers financial incentives ranging from £1.75 to £4 per kWh of demand reduction.

We focused on peak-demand reduction events carried out by this partner energy utility. As of June 2023, this utility was the third-largest domestic electricity provider in Great Britain, serving 16.9% of the domestic market (Ofgem, 2023), and was the most active demand flexibility provider in terms of participant numbers and delivered demand reduction. Of the 16 events conducted during Winter 2023–24 – comprising two live events formally called by NESO in response to system stress (29 November and 1 December 2023) and 14 coordinated test events – we focused on the 12 in which our partner utility bid, had its bid accepted, and successfully delivered opt-in notifications, as well as one additional event in which its bid was rejected but it delivered the event nonetheless.[3]

The natural field experiment used a randomized encouragement design for household and business customers, in which some were invited to sign up for the demand response program while others were not. This design allowed us to identify the causal impact of the intervention on electricity demand and was implemented without upfront selection into the sample. The experimental sample was every household or business that was eligible to take part (e.g., with a smart meter and permission to receive communications from the retailer). In total, we invited approximately 2.6 out of the 2.7 million households in our sample to participate, while 119,999 households formed a control group that received no encouragement. Within this control group, 40,000 households were randomly selected for late encouragement, enabling assessment of potential fatigue or diminishing responsiveness over the winter.

We recovered causal effects in the context of the system operator's critical peak rewards

[3]Live DFS events were formally called by NESO when forecast system margins fell below operational thresholds, triggering demand reduction procurement through the market mechanism. The 29 November and 1 December 2023 events were live NESO-called events. The remaining 14 events were conducted as coordinated test events designed for the research trial. Two events were excluded from our main analysis due to communication issues: one on 12 December 2023, for which notifications were sent only via WhatsApp, and one on 1 March 2024, when technical problems prevented broad opt-in access. The 14 March 2024 event, delivered despite the bid being rejected by NESO, is included in our demand analysis but excluded from the welfare analysis as it falls outside the standard DFS market mechanism.

program by exploiting a two-stage structure of consumer engagement – and, specifically, randomizing encouragement in the first stage of this engagement. Households were first required to explicitly agree to enroll in the utility's flexibility program (hereafter, a one-time "sign-up"). Conditional on sign-up, customers were then required to explicitly agree to participate in individual events (hereafter, event-specific "opt-in") in response to opt-in notices sent by the utility. Opt-in notices typically (a) communicated the price incentive associated with participation in a given event, and (b) provided customers with a hyperlink through which they could opt in to that specific event. In practice, opt-in notices were delivered via email, notifications through the utility's mobile application, or messaging-based updates such as WhatsApp.[4]

1.1 Primary findings

First, we found a substantial reduction in electricity usage during peak events. Among our compliers, we found that opt-in caused electricity consumption to fall by 23.1% during event windows. Again among our compliers, signing up to the program overall caused electricity consumption to fall by 6.8% during events (signed up customers did not opt in to every event). We found that this demand reduction was sharply concentrated during event times, with no evidence of change in consumption outside events. We found no short-run rebound effects nor long-run changes in electricity use.

Second, we found no detectable participant fatigue across the 13 events of Winter 2023–24. Households randomized to receive the invitation later in the trial responded similarly to those who received it earlier, suggesting that treatment effects remained stable over time. This result supports the notion that short-notice, event-specific interventions can be repeatedly deployed without eroding effectiveness, at least for a somewhat limited number of events over the year or season. Similarly, households randomized to receive a "chaser" sign-up encouragement email did not have significantly different demand response versus households who signed up without the chaser.

Third, we investigated the impact of low-carbon technology adoption on demand reduction. Using a staggered difference-in-differences design exploiting within-household changes around the timing of low-carbon technology adoption, we were able to isolate the causal effect of adoption itself. We examined the impact of 1) EVs, 2) heat pumps, and 3) solar and battery. EV adoption modestly increased demand response, while heat pump

[4]Customers could receive multiple opt-in notices for the same event or across different channels (e.g., email and mobile push notifications).

and solar and battery adoption both approximately doubled response in comparison to no technology.[5] The increase in demand response arose due to installation, rather than reflecting pre-existing differences across households, indicating that the adoption of these low-carbon technologies operated as a mediator rather than a moderator of treatment effects.

Fourth, demand response was strikingly homogeneous across household characteristics. Households who signed up during the first (non-experimental) iteration of the program (Winter 2022–23) were more likely to participate, but the LATE of *opting in* was not different between "veterans" and "newcomers". Demand response did not vary significantly by home energy efficiency rating, homeownership status, whether a customer resided in a house versus a flat (i.e., an apartment), or whether the customer was on a prepayment versus credit contract.

Fifth, higher incentives increased participation. We randomized the incentive level in one of the demand response events. In this event, the higher payment (£2.25/kWh vs £1.75/kWh) increased opt-in rates by approximately nine percentage points (36.7% vs 27.9%), implying an opt-in elasticity with respect to the incentive of approximately 1.10. We lacked statistical power to detect differences in demand reduction.

Finally, our welfare analysis showed that the societal value of DFS depended heavily on system conditions and how the program was used operationally. Under normal operating conditions in Winter 2023–24, and when treated primarily as a substitute for routine balancing actions, the program yielded a marginal value of public funds (MVPF) of about 0.63. This reflects the fact that remuneration for demand reduction substantially exceeded the cost of alternative balancing actions in most events. However, the program's value increases sharply under conditions of system stress. When demand reductions are instead valued as avoiding involuntary disconnection, using the UK's statutory value of lost load (£6,000/MWh), the MVPF rose to around 3.08. These results suggest that DFS was not welfare-enhancing as a routine tool at observed prices but could deliver substantial welfare gains as a form of reliability insurance in periods of genuine supply scarcity. Expressed in capacity terms, the program appeared cost-competitive with conventional firm capacity: the reductions it delivered cost around £28/kW per year, in line with the most recent T-4 clearing price.

[5] Heat pump adoption raises baseline event consumption substantially, so the net effect on grid demand during events is still positive even after the opt-in response.

1.2 Contribution to the literature

This paper contributes to an extensive literature on how best to manage energy supply and demand. Prior research has examined a range of pricing mechanisms and contracts, focusing, for instance, on peak-time pricing (Houthakker, 1951, Joskow, 1976) and interruptible contracts, whereby households receive payments or pay lower prices for energy in exchange for acceptance of service interruptions during periods of grid constraints (Allcott et al., 2016, Baldick et al., 2006, Tan and Varaiya, 1993). Research on peak-time and critical-peak pricing has been especially plentiful. However, despite numerous smaller-scale evaluations of the impact of peak-time pricing on energy demand (e.g., Andersen et al. (2019), Bollinger and Hartmann (2020), Caves and Christensen (1980), Caves et al. (1984), Jessoe and Rapson (2014), Wolak (2007)), there exists no causal evidence on the efficacy of nationwide critical peak demand reduction programs.

The closest study to ours is Fowlie et al. (2021), who randomized opt-in versus opt-out enrollment into time-of-use and critical peak pricing tariffs among 174,000 households in Sacramento.[6] Using a similar instrumental variables strategy based on randomized encouragement to identify local average treatment effects, our estimated opt-in demand reduction of 23.1% was close to the 26.4% reduction observed for "active joiners" in their critical peak pricing program. In absolute terms, however, Sacramento households achieved larger reductions (0.658 kW compared to our 0.1773 kW), reflecting higher baseline consumption in US households likely due to larger homes and greater use of air conditioning during summer peak periods.[7] Fowlie et al. (2021) also documented persistent "spillover" reductions on non-event peak days. In contrast, we found that reductions were concentrated within events and did not extend to neighboring hours or days. Participation rates in our setting exceeded 50%, compared to around 20% in Sacramento, suggesting that critical peak rewards may be more attractive to consumers than critical peak pricing. In terms of

[6]Research has also explored interventions that are not based on pricing. For example, this includes conservation appeals and social comparisons during peak consumption periods (Bergquist et al., 2023, Brandon et al., 2019, Brewer and Crozier, 2025, Ito et al., 2018), in addition to responses to national energy crises such as that of Germany following the Russia-Ukraine war (Moll et al., 2023). There is also a literature on real-time and peak-time pricing (Wolak, 2011), including evaluations of critical peak pricing with rebates, where customers were paid based on how far their hourly consumption fell below a prespecified reference level during events. This approach is similar in spirit to our intervention, though implemented at a much smaller scale (1,245 customers) and limited to Washington, D.C. Related work by Ito et al. (2023) used a randomized take-up incentive for dynamic electricity pricing to estimate marginal treatment effects and marginal treatment responses, showing that price-elastic consumers who generate larger welfare gains were more likely to self-select.

[7]In separate work, Burkhardt et al. (2023) used appliance-level data from 280 Texas households to show that critical peak pricing reductions are driven substantially by air conditioning, offering mechanism evidence that our settlement-metering data cannot provide.

heterogeneity, we found robust responses among prepayment customers, whereas Fowlie et al. (2021) found lower responsiveness among low-income households. Finally, we identified a causal role for solar and battery adoption, whereas households with those technologies were excluded from the Sacramento sample.

A complementary recent field experiment is Bailey et al. (2025), who randomized peak-event timing across approximately 1,000 households enrolled at a Canadian utility. Lacking an opt-in margin, their active-response arms reduced consumption by around 5% during events, close to our 7% sign-up estimate. Among customers estimated to have "always responded", they calculated a response of around 26%, in the range of our local average treatment effect on opting into events. The two designs sit at different points on the field-experiment spectrum, and we see them as mutually reinforcing. Theirs is a framed field experiment: the sample was drawn from households that had already adopted the utility's app and were recruited into an explicit trial, with a few hundred households per arm. Ours is a natural field experiment, run across a population of roughly 2.7 million households who experienced the intervention as an ordinary feature of their energy service. That difference moves us closer to several of List (2025)'s conditions for external validity, as we discuss below. And, because the intervention was a live commercial service rather than a bespoke experimental program, it also lets us ground the welfare analysis in the real costs of delivering flexibility.

A second contribution is to the external validity and policy relevance of demand response estimates. Our design satisfies four conditions for external validity identified by List (2025). First, external unconfoundedness holds because treatment – receipt of an encouragement email – was randomly assigned, ensuring that selection into the study did not confound our estimates of program effects. Second, external overlap is satisfied by the scale and composition of our sample: with 2.7 million customers spanning all regions of Great Britain and diverse housing, income, and tariff arrangements, our sample broadly mirrors the household population that any nationwide demand flexibility program would target. The caveat is that the partner supplier — at the time the third-largest in the UK — has a customer base traditionally skewed toward early adopters with green-tech engagement, so our sample is likely somewhat more engaged than a fully random draw of the UK population. Reassuringly, our heterogeneity checks show that more recently acquired households (less selected on engagement) deliver similar demand reductions to long-tenured participants, suggesting this selection does not materially shape the headline estimates. Third, parallelism is supported by the fact that our experiment was embedded within the actual, operational DFS program rather than a standalone research exercise.

The incentives, communications, and opt-in mechanisms faced by our experimental sample were identical to those faced by all other program participants, minimizing the gap between study conditions and real-world deployment. Fourth, investigator neutrality is satisfied, as customers were unaware of the experiment or their assignment, and observation was limited to the retailer's typical settlement metering. Together, these features mean that our estimates are likely to correspond closely to the policy-relevant effects of a scaled, real-world demand flexibility program. This is particularly valuable because the program we study was the first nationwide domestic demand flexibility program of its kind.

The remainder of this paper is structured as follows. Section 2 describes the experimental design, supplementary analyses, and data sources. Section 3 presents the results from our natural field experiment and associated quasi-experimental analyses. Section 4 uses these causal impacts to evaluate the welfare impacts of the program. Finally, Section 5 concludes with a summary of findings and policy implications.

2 Experimental design and analysis framework

2.1 The Treatment: The Demand Flexibility Service

The Demand Flexibility Service (DFS) was a market created and administered by the National Energy System Operator (NESO) to procure voluntary reductions in electricity demand during periods of system stress. When the forecast system margin – the buffer between available generation capacity and expected peak demand – fell below an operational threshold, NESO called a DFS event for a specified window, typically one to one and a half hours in duration, in order to procure demand-side reductions that restored adequate headroom between supply and load. Energy retailers and other customer aggregators could submit bids to NESO offering to deliver a specified quantity of demand reduction at a given price. If NESO accepted a bid, the provider received a payment per megawatt-hour (MWh) of verified demand reduction delivered, with accepted bids during Winter 2023–24 ranging from £150 to £6,000 per MWh. Providers were then free to determine how to deploy this payment, including how much to pass through to the consumers whose behavior generated the reduction.

From the consumer's perspective, participation involved two distinct stages. First, customers needed to enroll in their provider's demand flexibility program, a one-time sign-up

that registered them as eligible to participate in future events. Second, ahead of each individual event, enrolled customers received an opt-in notification, delivered via email, mobile app push notification, or messaging services, that specified the financial incentive on offer and provided a mechanism through which the customer could confirm participation in that specific event. Customers who opted in were expected to reduce their electricity consumption during the event window; those who did not opt in faced no obligation or penalty. Our partner utility, Octopus Energy, offered customers between £1.75 and £4 per kilowatt-hour of demand reduction during Winter 2023–24.

Payments to consumers were based on measured demand reduction relative to a customer-specific counterfactual baseline, constructed from recent historical smart meter readings for the same half-hour periods as the event. The baseline followed the methodology of the regulated BSC P376 modification, calculated mechanically from a fixed historical window of comparable non-event days matched by day type and time of day, with no in-day or weather adjustment.[8] Demand reduction was then calculated as the difference between this baseline and actual metered consumption during the event window: consumers whose consumption fell below their baseline received a payment proportional to the estimated reduction, while those who consumed at or above it received nothing.

In principle, a customer could try to raise their baseline by consuming more on the qualifying pre-event days, inflating their measured reduction. In practice this was largely self-defeating: the additional energy consumed during the baseline window was itself paid for at retail rates, and any offsetting gain on the event day depended on the size of the inflated baseline relative to the event incentive, which for most customers did not work out in their favor.

A further concern arises from the structure of the payment itself. Because consumption below baseline was rewarded while consumption above it carried no penalty, a customer faced no downside to opting in, which created scope for two forms of selection. A customer might opt in opportunistically, anticipating low consumption – for instance, when away from home – without altering their behavior. Or, they might opt in on every event, knowing they would retain only the favorable outcomes. Both would tend to inflate measured reductions relative to any genuine behavioral response.

These concerns are a central motivation for our paper. Our estimate of flexibility comes from the randomized encouragement rather than the baseline, so the counterfactual is the randomized control group's realized consumption, not a customer-specific baseline that

[8]The modification specifies a trimmed mean, using ten weekday baselines for weekday events and four weekend baselines for weekend events.

selection can distort. Opportunistic or indiscriminate opt-in can still affect program costs, but our findings show little evidence of a wedge between remunerated and real flexibility. In Appendix A11 we compare mean event-window consumption between households not yet signed up and signed-up households that did not opt in; the two groups had very similar baselines, with reductions that averaged to zero.

The specific characteristics of DFS events, including the notice period, remuneration level, and timing, varied considerably across the winter, as summarized in Figure 1. This variation is central to several of our analyses and is described below. The number of households signed up ranged from 1 million to 1.4 million. In other words, of the 2.6 million households signed up to the demand flexibility program from all providers[9], approximately half were in our sample. On average, 37.4% of signed-up households opted in to each event. The notice period varied from 1 day to as little as 3 hours, and the remuneration offered ranged from £1.75 to £4 per kWh. Events were typically scheduled during UK peak demand hours, with most events centered between 4:30pm and 6:00pm. Earlier events tended to be longer, have longer notice periods, and have higher remuneration.

2.2 Field experimental design

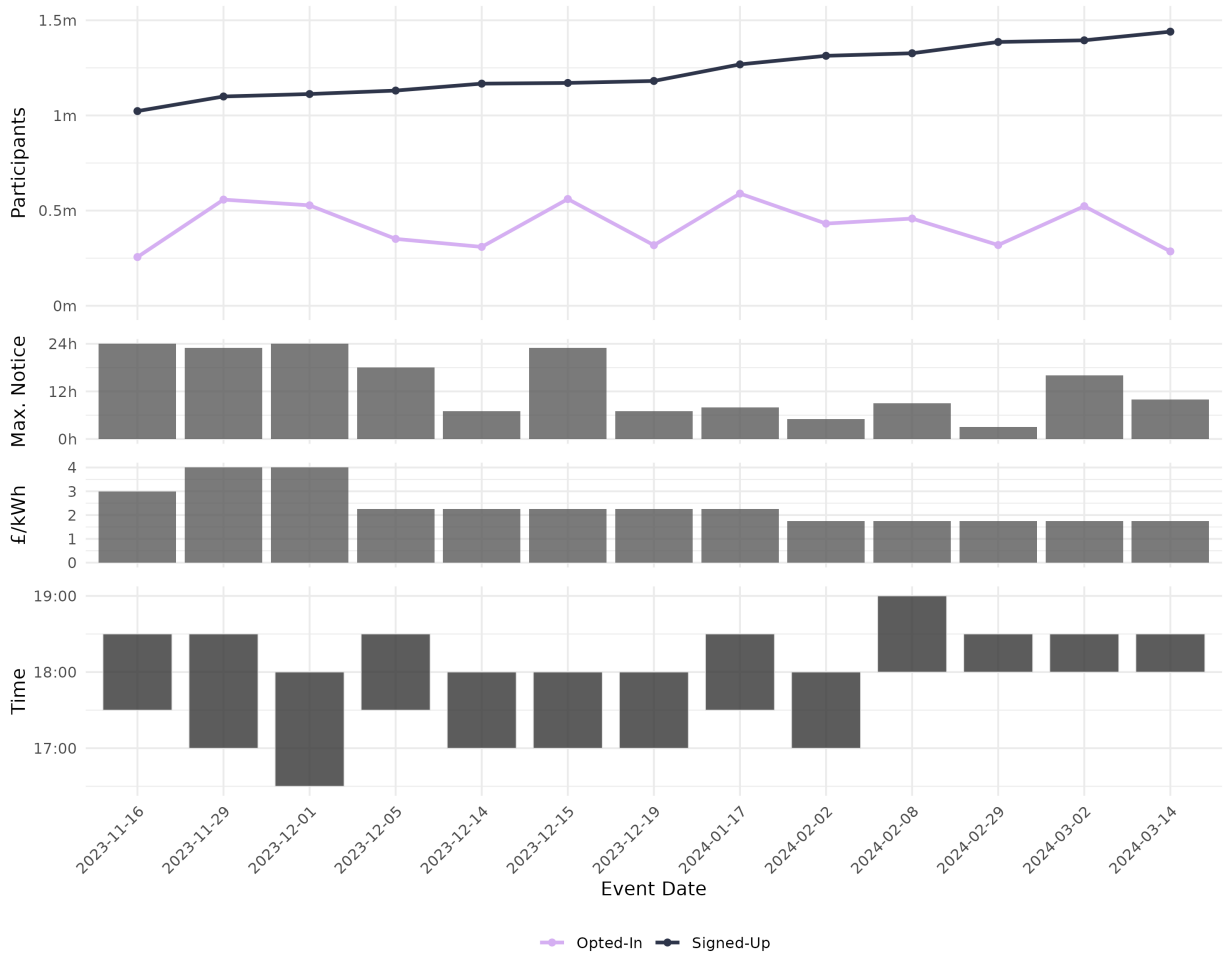
We used a randomized encouragement design to estimate the causal impact of the program on energy demand. We partnered with the implementing energy utility, Octopus Energy Limited, to randomize invitations to sign up for the program, withholding them from a randomly selected control group.[10] Households in the control group were not prevented from enrolling in the program through other channels, including prior participation, information obtained via social media, in-app banners, or word of mouth.

We estimated three treatment effects: (i) the *intent-to-treat* (ITT) effect, comparing outcomes between households that were encouraged and those that were not; (ii) the *Local Average Treatment Effect* (LATE) on sign-up, capturing the causal effect for compliers who enrolled because of the encouragement; and (iii) the LATE on opt-in, focusing on compliers who enrolled and opted into events because of the encouragement.

[9]It is a coincidence that 2.6m customers were in our treatment (encouragement) group, and also that across all DFS providers 2.6m customers ultimately signed up to the demand response program.

[10]Program eligibility was governed by the utility's operational requirements. Customers needed a smart meter settled on a half-hourly basis and at least one full day of half-hourly meter data to sign up. To remain eligible for a given event, the utility required receipt of at least 80% of smart-meter readings over the preceding 15 days. Eligibility extended to both direct-debit and smart prepayment customers, including customers on smart time-varying tariffs and tariffs bundled with managed electric-vehicle charging or export arrangements, but excluded customers already participating in another Demand Flexibility Service scheme.

Figure 1: Description of Events

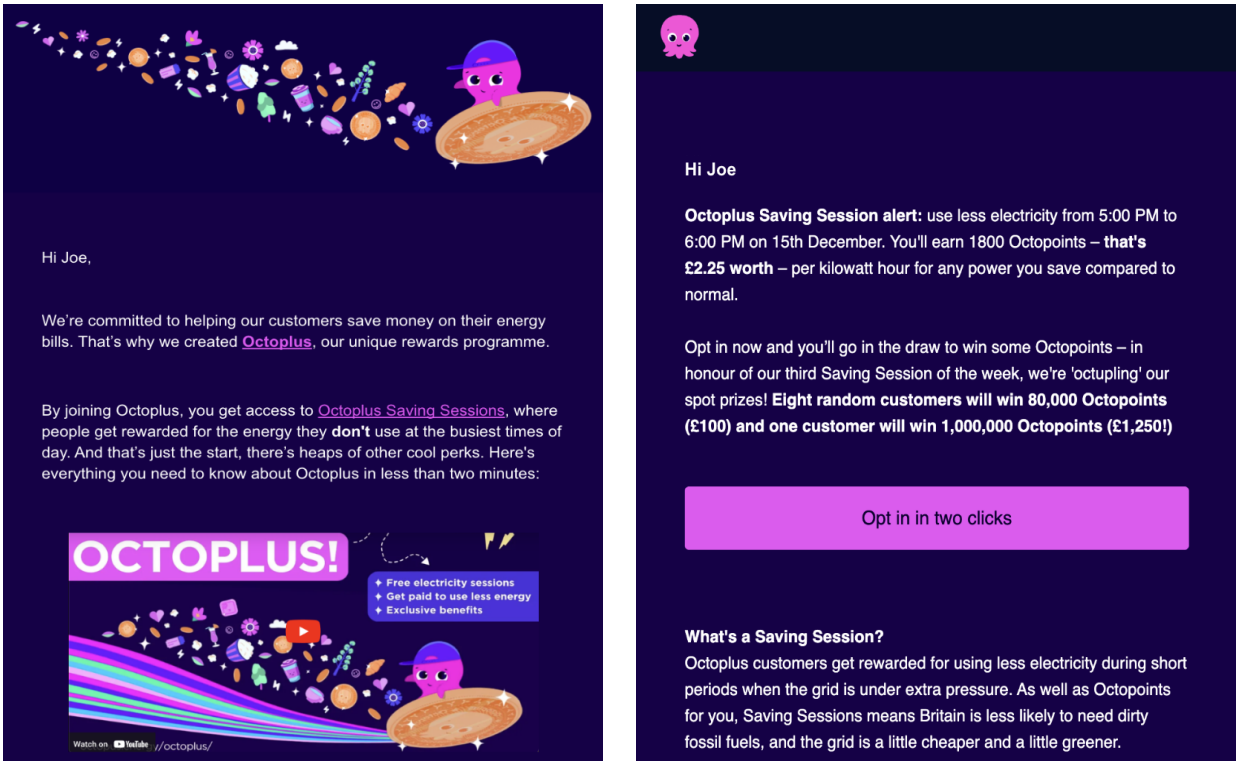


Notes: The top panel shows the total number of households signed up to each event (dark grey line) alongside (and always above) the number of households opting in (light purple line). The second panel reports the notice period, measured in hours between first notification sent and the start of the event. The third panel shows the remuneration offered per event, expressed in £ per kWh of reduced consumption. Finally, the bottom panel displays the timing of each event during the day. Together, the figure highlights substantial variation across events in participation, notice, incentives, and timing, which we exploit in the subsequent analysis.

We assigned households to treatment and control by stratified random sampling, withholding encouragement from the control group. We began the roll-out of encouragement on 24 October 2023.[11] Customers were contacted between 24 October 2023 and 13 November 2023 by batches. Communication was withheld from a control group of 119,999 households (approximately 4.45% of the total sample). To enhance balance and external valid-

[11]Just prior to the first event, a small number of customers who had previously participated in the previous year's version of the program were inadvertently contacted. We show that our results are robust to previous participation in the program in Table A14.

Figure 2: Example communication materials used in our natural field experiment



(**a**) Sign-up encouragement (**b**) Event opt-in notification email

Notes: Panel (a) shows the initial sign-up encouragement sent to customers in our treatment group (control group customers received no such email). Panel (b) shows the subsequent opt-in notification sent prior to each individual demand flexibility event. Customers received "Octopoints" for participating and reducing their consumption; 800 Octopoints equaled £1.

ity, treatment assignment in the main trial was preceded by stratified random sampling. Stratification was based on households' Grid Supply Point (GSP) group[12] and customer payment type (credit versus prepayment).

We detected a high share of always-takers; despite this pattern, we observed a persistent 20 percentage point gap in sign-up rates between the treatment and control groups (Figure 3). One distinctive aspect of our two-stage least squares analysis, compared to most randomized encouragement designs, is that we observed all the households repeatedly over the 13 events. The strength of our instrument (encouragement) varied over time as the compliance rate difference between the control and treatment groups changed. In other words, the strength of our first stage was driven by the wedge between treatment and control group sign-up rates, and this difference changed as the event season progressed.

[12]There are 14 GSP groups in Great Britain; see https://data.nationalgrideso.com/system/gis-boundaries-for-gb-dno-license-areas for a visual representation.

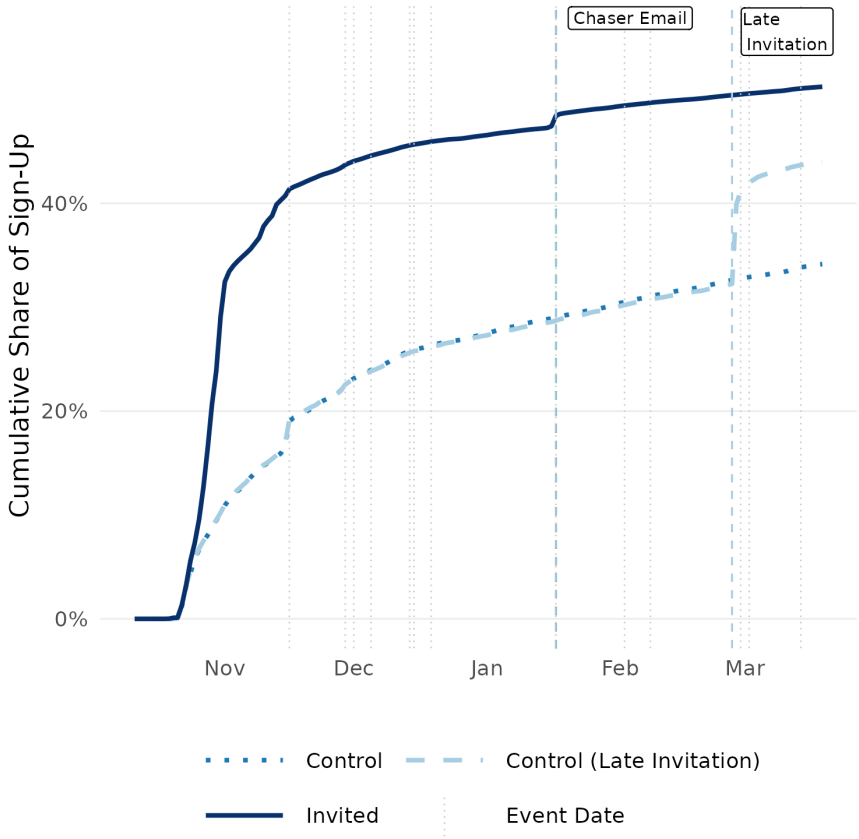
The relationship between encouragement and opt-in was even more complex, as it reflects the relationship between encouragement and sign-up, further diluted by the variation in opt-in (among signed up customers) from event to event.

The two-stage least squares estimates are consistent under the standard exclusion restriction. For the sign-up LATE, the required condition is that the invitation affects event-period consumption only through sign-up. For the opt-in LATE, the condition is narrower: invitation must affect event-period consumption only through event-specific opt-in. Because signed-up customers receive event-specific notices containing the timing, incentive, and opt-in link for each event, as described in Section 2, three alternative channels are worth taking seriously. Sign-up could shift event-window consumption through (i) general salience of the program, (ii) information transmitted by the event notice itself, or (iii) unremunerated reductions by signed-up customers who do not click the opt-in link. Each of these would affect consumption without passing through the act of opt-in, and would therefore violate the opt-in exclusion restriction.

The evidence in Appendix A11 speaks directly to these channels. Comparing event-window consumption across three groups — not-yet-signed-up households, signed-up households that did not opt in to the event, and opted-in households — we found that the not-yet-signed-up and signed-up-but-not-opted-in groups behaved essentially identically: baselined "reductions" average 0.005 kW in both groups, against 0.2 kW for opted-in households, and the small event-by-event deviations from zero have opposite signs early and late in the season, consistent with weather-driven baselining bias rather than a behavioral response (Figure A5, rows 1 and 2). If salience or notice receipt were generating meaningful unremunerated reductions among signed-up customers, we would expect row 2 to lie systematically below row 1; it does not. We treat this as a direct exclusion-restriction diagnostic rather than only as a baselining check. Finally, we also show that encouraged customers are not more likely to change their peak-time electricity usage (Section 3.4), or switch to another tariff or provider (Appendix A6).

The impact of two sub-trials is also visible in Figure 3. A randomized chaser email had only a modest effect (Appendix A5.1), while a late invitation sent to a randomly selected one-third of the control group halved the gap between invited and non-invited households (Appendix A5.2) within that late-invited one-third. All of the randomizations were pre-registered with their own pre-analysis plans.

Figure 3: Sign-Up Rates by Trial Arm



Notes: This figure illustrates the cumulative share of households that signed up for the demand flexibility program across the three primary experimental arms: the invited (treatment) group in solid dark blue, the original control group in dotted medium blue, and the late invitation subset of the control group in dashed light blue. The dashed vertical lines indicate the timing of specific sub-trials – the chaser email campaign in mid-January and the late invitations issued in late February – and the dotted light grey vertical lines indicate individual demand flexibility events. A persistent gap of approximately 20 percentage points in sign-up rates is observed between the treatment and control groups throughout the trial period, providing a strong first stage for our two-stage least squares analysis.

2.3 Quasi-experimental analysis

We complemented our experimental analysis with a quasi-experimental strategy to identify a mediator of demand-side flexibility. Using a staggered difference-in-differences design that exploited within-household variation in the timing of low-carbon technology adoption, we established that solar panel and battery installation causally increased event-time flexibility, ruling out pre-existing selection and identifying low-carbon technology ownership as a mediator, rather than a moderator, of demand response.

2.4 Data

We collected several sources of data on households' electricity consumption and household characteristics. Consumption data was recorded at the half-hourly level before and during the trial. We obtained customer characteristics, including Energy Performance Certificate (EPC) ratings, house type (e.g., detached, semi-detached, terrace, flat), geographic location, and tariff type at the start of the trial. We also recorded whether the property had an export meter, indicating solar panels or home batteries. Other key variables included estimated annual consumption (EAC), a variable energy retailers in Great Britain hold for all households to provide annual cost estimates and estimate consumption in the absence of meter readings, and whether households had a prepayment meter versus a credit meter. To capture socioeconomic context, we matched households' postcodes to attributes about their area, in particular the Index of Multiple Deprivation (IMD), a composite measure produced by the UK Government that ranks small geographic areas in England according to relative deprivation across multiple domains, including income, employment, health, education, and housing. Finally, we conducted surveys to collect additional details, such as the presence of low-carbon technologies, household structure, home-ownership status, number of rooms, occupation, presence of dependents, and behavioral changes related to energy use.

2.5 Descriptive statistics

We assessed the balance of our treatment and control with respect to pre-trial customer characteristics. The treatment and control groups were well-balanced across all variables (see Column 1, Table A3). However, when we focused on participants who actively signed up (Column 2, Table A3), we observed that past participants were significantly more likely to sign up again. Households with characteristics indicating higher affluence also exhibited higher sign-up rates — homeowners, house dwellers (vs. flat residents), and those with high energy efficiency scores were also more likely to sign-up. Households on Time-of-Use (ToU) tariffs or with electric heating, who may be familiar with demand shifting, also participated at higher rates. Interestingly, households using prepayment meters[13] were much more likely to participate, potentially driven by the appeal of cost-saving programs.

Finally, in Column 3, Table A3, we examined the correlation between customer charac-

[13]Prepayment meters are often used by individuals who face challenges paying their energy bills or prefer to manage their energy costs closely.

teristics and the quantity of opt-ins across all events (as discussed previously, households had to actively opt in before each event to be remunerated for their demand reduction). The results show similar patterns to the correlations we saw with sign-up. Past participants opted in more, suggesting the potential for periodic implementation of such programs with sustained participation (but also suggesting a potential "sorting" of the *most* engaged households into the first iteration of the program). Prepayment households and households on ToU tariffs and/or with electric heating also opted in more frequently, as did homeowners and residents of houses (rather than flats).

These findings highlight the role of sorting on characteristics — whether driven by existing familiarity with demand response programs, cost-saving motives, or environmental consciousness — in shaping participation. This underscores the importance of our randomized encouragement design in isolating the impact of flexibility programs on electricity usage. In the next section, we will discuss how the experimental design enabled us to capture the causal effects of program invitations on participation and flexibility.

3 Results

This section presents our main experimental results (Section 3.1.1). We then examine mediators of demand-side response like low-carbon technology adoption (Section 3.2), and moderators such as household and event characteristics (Section 3.3). We then explore how response changes over time (Section 3.4) and through the sub-trials (Section 3.5).

3.1 What magnitude of demand response did households deliver?

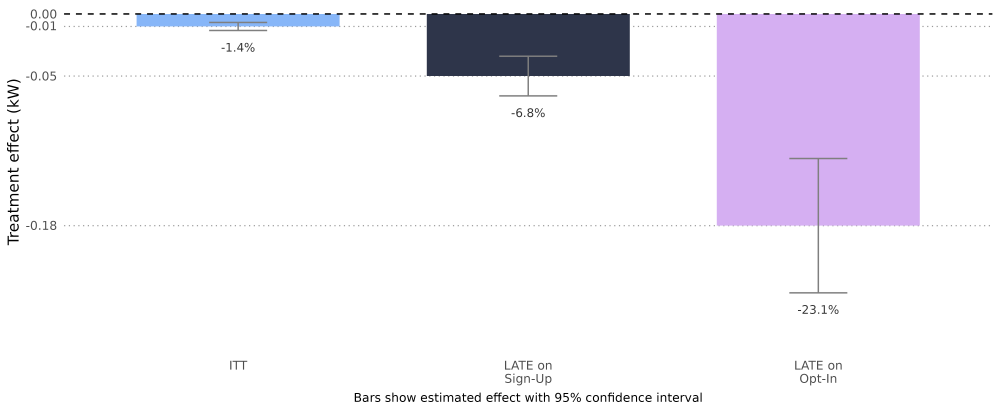
3.1.1 Demand response during events

First, we estimated the reduction in electricity consumption during events for households in the main treatment group, defined as those invited to join the program, relative to households in the control group, who were never invited. We focused on the intent-to-treat (ITT) effect of the invitation itself. This estimand captures the average impact of being offered the program and avoids selection concerns that arise from conditioning on opt-in status, as discussed in Section 2.5. We then estimated the effect of participation among compliers using a Local Average Treatment Effect (LATE) framework (Equation (5)).

To verify that the email encouragement indeed boosted program participation as in-

tended, we examined the first-stage outcomes. As shown in Figure 3 and quantified in Table A4, receiving an invitation email caused a large jump in enrollment: we see 20 and 5 percentage point differences in sign-up and opt-in rate respectively between the control and the treated group over the full duration of the trial. These differences confirm that the encouragement was effective in driving program sign-ups and event participation, providing a strong first stage for the two-stage least squares analysis.

Figure 4: Impact of the Main Trial Invitation on Consumption (kW)



Notes: This figure reports estimated treatment effects on electricity consumption during event periods as pre-specified primary analysis. Bars show point estimates for the intention-to-treat (ITT), the LATE on sign-up, and the LATE on opt-in, expressed in kWh per hour (kW). Error bars represent 95% confidence intervals. Percentage labels indicate the proportional change in consumption relative to the control group's average consumption (0.7660 kWh per hour) during the event window; note that this average includes always-takers in the control arm who signed up through channels other than the invitation.

We show the results of the intent-to-treat analysis, and Local Average Treatment Effects on sign-up and opt-in, in Figure 4 and Table A5. Being invited to participate in events reduced electricity consumption by 0.01 kW during events. Using the non-invited group as a baseline[14], this reduction represented 1.4% of average hourly consumption, indicating that simply encouraging households to join the flexibility program led to energy savings during peak events, even though many did not actually sign up, and even though many in the control group *did* enroll, through the persistent wedge in enrollment rate between the two groups.

Households who *signed up* to the program in response to the encouragement reduced their demand by 0.0520 kW during events (6.8% of control group demand during events). Opting in led to a reduction of 0.1773 kW (23.1%). Both are local average treatment effects

[14] As we implemented a randomized encouragement design, we do not have a pure control group; i.e., some always-takers in the non-invited group signed up and opted in.

identified directly by the randomized encouragement design: they describe the behavior of *compliers*, the households whose participation was induced by the invitation. In Appendix A10 we compare these complier LATEs to the response of always-takers (households that signed up or opted in without any encouragement) and find that always-takers reduced demand somewhat more, consistent with the pattern documented by Fowlie et al. (2021) that more motivated participants respond more strongly. Reweighting compliers and always-takers by their population shares yields an average reduction of 0.2146 kW (28.1%) per opted-in household. This last quantity is not a complier LATE and was not part of our pre-analysis plan: it is an exploratory exercise that relies on the additional assumption that compliers and always-takers in a universal campaign would respond as estimated in Appendix A10, and the always-taker response is itself a descriptive comparison rather than a causally identified effect. We report it because it is the natural quantity for estimating aggregate demand reduction. Table 1 summarizes how each estimand is defined.

Table 1: Main estimands

Estimand	Estimate (kW)	% of control	Subpopulation
ITT of invitation	−0.0105	−1.4%	Encouragement impact among all invited households
LATE on sign-up	−0.0520	−6.8%	Sign-up compliers (household-event)
LATE on opt-in	−0.1773	−23.1%	Opt-in compliers (household-event)
Participant average	−0.2146	−28.1%	Compliers + always-takers, reweighted

Notes: Estimates from Table A5 (kW specification) and Table A17. Control mean was 0.766 kWh per hour during the event window. The first three rows are causal estimands identified by the randomized encouragement design. The participant average reweights the complier opt-in LATE and the always-taker response by their respective population shares (33% and 67%, respectively); it is the relevant quantity for estimating aggregate demand reduction from the program (Section 4).

Our main analysis uses net consumption (electricity imported from the grid minus electricity exported by households with solar generation). For completeness, we re-ran the analysis using import and export as separate outcomes. Results were similar (Table A6), confirming that our findings are robust to the definition of consumption.

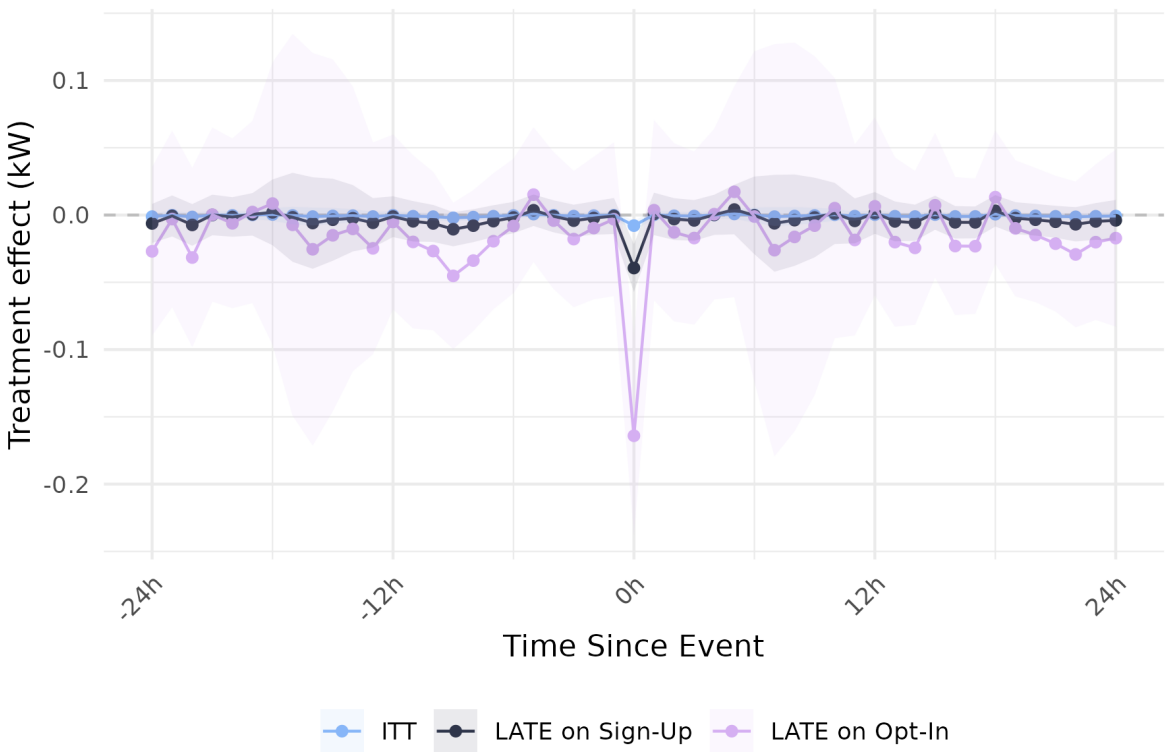
3.1.2 Displacement versus destruction

We next investigated how households reduced demand during events: did they shift consumption to nearby hours (i.e., load displacement), or did they forego usage entirely (i.e., load destruction)? To examine this, we estimated our main models over a symmetric 48-hour window around each event, breaking the day into hourly bins – 24 hours before, the

event hour(s), and 24 hours after. We excluded six of the 13 trial events because their windows overlapped with another event in the ± 24 hour estimation window, leaving seven events in the event-study sample.[15]

In Figure 5, we plot the estimated treatment effects (ITT and LATEs) for each hour in the 48-hour window. Reductions were sharply concentrated during the event periods. Outside of the event window – both before and after – we observed no systematic increase in consumption. In other words, we detected limited evidence of pre-event anticipatory shifting, nor of deferred usage.

Figure 5: Event Study: Impact of Treatment Before, During, and After Events



Notes: Each point reports the estimated treatment effect for one hourly bin relative to the event, based on our pre-specified event-study analysis. Negative values on the horizontal axis denote pre-event hours and positive values denote post-event hours. The point at 0 reports the impact during the event window itself and is of similar magnitude to the main-sample estimate, even though this event-study analysis excludes events with overlapping windows. Shaded areas show 95% confidence intervals.

This pattern conflicts with survey evidence. We conducted a survey among program participants (sent to 4,100 households, of whom 1,012 (24.7%) responded). The great majority – 87.1% – said they "adjusted electrical appliance usage at home" – but this answer

[15]The included events took place on 16 November 2023, 5 December 2023, 19 December 2023, 17 January 2024, 2 February 2024, 8 February 2024, and 14 March 2024.

explicitly included delaying, changing, *or* avoiding use of the appliance(s). Most importantly, a substantial minority – 23.4% – of respondents said they explicitly scheduled their appliances to run before or after the event.

We acknowledge that this conflict between our empirical results and the survey evidence may partly be due to insufficient power to detect very diffuse substitution. Using the rule-of-2.8[16], the minimum detectable effect in the best-powered pre- and post-event hours is on the order of roughly 30–40% of the estimated event effect. This implies that, even restricting attention to non-overlapping events, the design is powered to detect only fairly large displacement effects concentrated in a single hour. In other words, the event-study estimates rule out large rebound or pre-shifting responses, but cannot detect smaller patterns of intertemporal substitution spread across surrounding hours.

For policy design, the key result is that the intervention created no economically meaningful secondary peaks in the hours around events. These could otherwise strain the grid in periods surrounding events. This is especially notable because these periods can also be sensitive: while the event coincided with the system's absolute peak, nearby hours are also often relatively constrained, so avoiding new peaks in these other hours may be important for overall grid stability.

3.1.3 Habit creation analysis

We next examined whether the intervention had any lasting effects on energy use outside of events, i.e., did households begin to reduce consumption in general? A related but opposing possibility is that households may have shifted their consumption from peak times on event days to peak times on non-event days. To test these hypotheses, we ran our main models on non-event days to examine whether there was any systematic difference in peak-time electricity consumption between the treatment and control groups outside of scheduled events.

There was no statistically significant difference in peak-time electricity usage on non-event days (Figure A1). Estimated coefficients were statistically indistinguishable from zero, except for the first days of the trial.

From a policy perspective, the absence of habit formation is notable. It suggests that the treatment effect is event-specific and tightly concentrated, rather than causing general energy conservation or behavioral spillovers. While habit formation might be desirable in

[16]The "rule of 2.8" is a standard approximation for the minimum detectable effect under a two-sided 5% test with 80% power: MDE $\approx (1.96 + 0.84) \times$ SE $\approx 2.8 \times$ SE.

some contexts, its absence here strengthens the case that our estimated effects are driven by direct responses to opt-in and financial incentives, not broader lifestyle changes.

3.2 A key mediator: low-carbon technology adoption

Low-carbon technology adoption can plausibly affect program participation and demand response over the course of the trial, placing it on the causal pathway between treatment and demand response. Low-carbon technology adoption is therefore a candidate *mediator* – a channel through which demand response is delivered. Isolating its causal role requires a design that exploits the within-household timing of installation rather than a simple interaction.

We examined how post-enrollment low-carbon technology adoption affected households' demand response flexibility. We focused on the 17,085 signed-up households that installed a low-carbon technology after the first event (16 November 2023): 10,658 installed an EV charger, 2,209 installed solar photovoltaic panels with battery storage, and 4,628 installed an air-source heat pump. Some households adopted multiple technologies; the sum of category counts therefore exceeds the total.

Adopters differed from non-adopters on observable characteristics — higher property values, higher estimated annual consumption, higher EPC ratings — and adoption timing also varies within technology type (Table A15). A simple comparison would therefore confound the causal effect of adoption with selection. We address this using a TWFE staggered difference-in-differences estimator that interacts installation status with opt-in:

$$Y_{it} = \beta_0 \,\mathrm{Installed}_{it} + \beta_1 \,\mathrm{OptedIn}_{it} + \beta_2 \,\mathrm{OptedIn}_{it} \times \mathrm{Installed}_{it} + \beta_3 \,\mathrm{HDD}_{it} + \delta_t + \alpha_i + \varepsilon_{it} \quad (1)$$

where δ_t are event fixed effects, α_i are household fixed effects, and HDD_{it} is heating degree days in the household's GSP region. β_0 captures the direct technology effect (consumption change holding opt-in fixed); β_1 captures the opt-in effect before installation and β_2 captures how adoption modifies the opt-in–consumption relationship — the additional flexibility unlocked by the technology.

The first key concern in our setup is endogenous adoption timing. Installation timing could be planned in response to DFS event conditions or expected payouts. In practice this is implausible on both magnitude and timing grounds. An opted-in household would save 0.1773 times an average of £2.9 per kWh for an hour event – against installed costs of approximately £8,000–£14,000 for solar-plus-battery, £7,000–£13,000 for an air-source

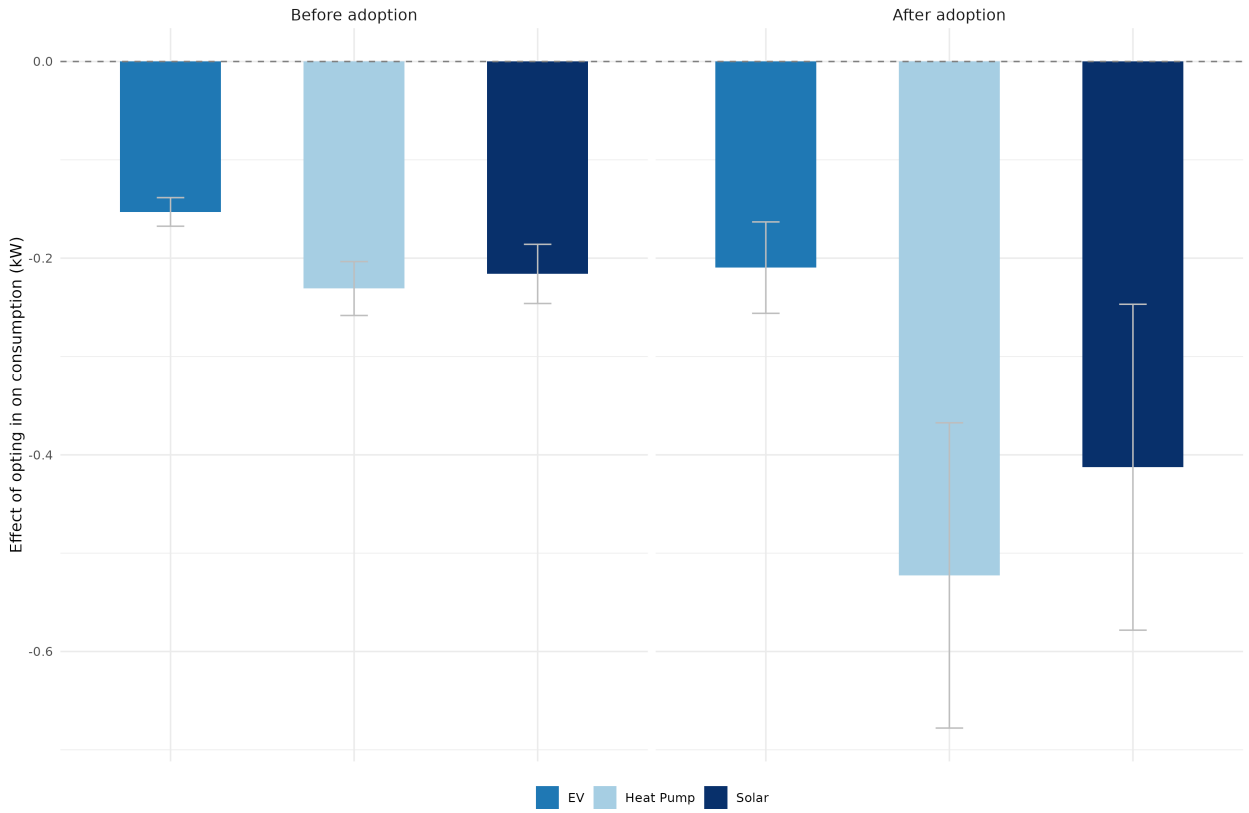
heat pump, and £800–£1,500 for an EV charger (excluding the EV itself). Installation lead times are also long: typical home surveys, supply, and installer scheduling produce delays of weeks to several months between the decision to adopt and the installation date. Households therefore cannot meaningfully shift the timing of installation in response to in-season DFS conditions, and the financial scale of DFS rewards is too small relative to LCT capital costs to drive the adoption decision itself. We view this as direct support for treating within-household installation timing as exogenous to event conditions conditional on household and event fixed effects. We also show the lack of pre-trends in Figure A3.

An additional concern is endogenous increase in opt-in. Because installation raises opt-in rates, $\mathrm{OptedIn}_{it}$ is a post-treatment variable and conditioning on it could in principle reintroduce selection: the pool of opted-in events post-installation is compositionally different, augmented by households whose participation was induced by the technology. Identification of β_2 therefore requires an additional assumption beyond parallel trends. We assume that the consumption reduction from opting in is *invariant to opt-in propensity*. Under this assumption: it does not matter that installation-induced participants have lower or higher opt-in propensity, because propensity is uncorrelated with the size of the demand response. This assumption is plausible in our setting. Once a household opts in, the demand reduction is largely mechanical: a smart EV charger pauses automatically, a heat pump is switched off, a battery discharges to offset net imports. The technology executes the response, so the magnitude of demand response is determined primarily by the size of the flexible load rather than by the household's underlying engagement level. We cannot rule out violations to this assumption entirely, however. If installation-induced opt-in compliers are systematically more responsive, this bias would inflate $\hat{\beta}_2$.

A related concern is that households opt in strategically on events where their baseline consumption is unusually high, which would inflate β_1 and $\beta_1 + \beta_2$ independently of any technology effect. As discussed in Section 2, we found no evidence of strategic event selection. Table A16 shows the TWFE model outputs, visualized in Figure 6.

EV chargers: EV adoption had no effect on event consumption outside of opt-in periods ($\hat{\beta}_0 = 0.006$ kW, insignificant). Before installation, EV adopters reduced consumption by 0.15 kW when opting in. After installation, the interaction term adds a further 0.06 kW reduction, suggesting a modest increase in demand response consistent with the ability to pause smart charging during events.

Figure 6: TWFE Estimates: Impact of Low-Carbon Technology Adoption on Consumption During Events



Notes: Each panel shows the opt-in coefficient from Equation (1) estimated separately for EV, heat pump, and solar adopters. This analysis is exploratory; LCT adoption was not part of any pre-analysis plan. *Opted-In* ($\hat{\beta}_1$): effect of opting in before installation. *Installed* $\times$ *Opted-In* ($\hat{\beta}_2$): opted-in plus the additional consumption reduction when opting in after adoption. Error bars are 95% confidence intervals with household-clustered standard errors. See Table A16 for full regression outputs.

Heat pumps: Heat pump adoption substantially raised event consumption ($\hat{\beta}_0 = 0.68$ kW), reflecting the electric heating load added by the technology. However, opted-in households after installation reduced consumption by an additional 0.29 kW relative to the pre-adoption opt-in effect (a more than doubling of reduction) — consistent with households temporarily switching off their heat pump to shed load during events.

Solar panels and batteries: Solar adoption reduced event consumption by 0.38 kW ($\hat{\beta}_0$), reflecting lower net imports due to on-site generation. After installation, opted-in solar adopters reduced consumption by a further 0.20 kW, approximately doubling their reduction, consistent with battery discharge during events.

3.3 Moderators and heterogeneity analysis

In addition to our analysis of low-carbon technology adoption as a key mediator, we examined possible treatment moderators. In particular, we examined a range of potentially important pre-existing household characteristics – consumption level, tariff, dwelling type – that may be correlated with the size of the response but are not changed by the program (Section 3.3.1). We also examined how treatment effects vary with temperature and remuneration during the event (Section 3.3.2).

3.3.1 Treatment effect heterogeneity

We examined a range of pre-existing household characteristics that may be correlated with the size of response but are not changed by the program. Because none were randomly assigned, we read the patterns below as indicative rather than causal – they may reflect genuine differences in responsiveness, but could also be driven by unobserved confounders. Of the dimensions in Table 2, estimated annual consumption, prior participation, energy efficiency (EPC), Index of Multiple Deprivation, time-of-use tariff, prepayment, prior supplier, and region were pre-specified in our pre-analysis plan; property value, floor area, dwelling type, occupancy tenure, urban/rural location, and income are exploratory additions. Table 2 summarizes the full set, with complete regression estimates in Table A14; here we discuss only the dimensions where response actually differed.

The clearest pattern was that response scaled with the size of a household's electricity demand. Participants with above-average annual consumption delivered significantly larger absolute reductions – roughly double the full-sample average treatment effect – but in proportional terms (the log specification) their savings were similar to everyone else's (Table A14). Property value and floor area showed the same pattern. For context, the mean EAC for standard domestic profile (single-rate domestic customers, rather than dual-rate customers) in 2023 was 3,365 kWh, according to national EAC estimates provided by the energy regulator to all retailers. This is lower than the mean EAC in our sample (approximately 4,120 kWh).

One exception among tariff characteristics was dual rate customers: these customers, who are disproportionately likely to heat their homes electrically, reduced demand roughly twice as much as others in proportional terms (Table A14). This is consistent with direct electric top-up heating being curtailable during events, though we cannot rule out that higher baseline consumption also contributes.

Table 2: Summary of treatment-effect heterogeneity by household characteristic

Characteristic	Treatment-effect heterogeneity
Characteristics related to size of home and typical consumption	
Estimated annual consumption	Larger absolute reduction; proportional reduction similar
Property value	Like consumption – absolute up, proportional flat
Floor area	Like consumption – absolute up, proportional flat
Prior program experience	
Participated in 2022–23 season	No difference in event response; affects sign-up only
Housing and geographical characteristics	
Energy efficiency (EPC)	No significant difference
Dwelling type	No significant difference
Occupancy tenure (own vs rent)	No significant difference
Urban vs rural	No significant difference
Index of Multiple Deprivation (LSOA level)[a]	No significant difference
Income (MSOA level)[a]	No significant difference
Region (Figure A2)	Sign-up rates vary; per-event reduction similar (ITTs differ, LATEs alike)
Supplier and tariff characteristics	
Time-of-use tariff	No significant difference
Dual-rate tariff (Economy 7)[b]	Larger absolute reduction; larger proportional reduction
Prepayment meter	No significant difference
Prior supplier (Bulb transfer)[c]	No significant difference
Tenure with energy provider	No significant difference

Notes: Each row reports the interaction between the invitation to participate and the household characteristic. These characteristics were not randomly assigned, so the estimates indicate associated variation in response rather than causal moderators. "Absolute" effects are measured in levels (kW per opt-in); "proportional" effects come from the log consumption specification and express reductions relative to baseline usage. Full regression estimates (percent change, by subgroup level, with significance stars) for every characteristic above are reported in Table A14; regional sign-up and opt-in patterns are shown separately in Figure A2.

[a] LSOAs (Lower Layer Super Output Areas) average 1,000 to 3,000 residents. MSOAs (Middle Layer Super Output Areas) average about 7,000 residents.

[b] Economy 7 is a legacy UK dual-rate electricity tariff that charges lower rates during 7 off-peak night-time hours (approximately midnight–7am) and higher rates during the day. It was designed for storage heaters and immersion water tanks that charge up overnight and release heat during the day. Economy 7 customers are disproportionately likely to have electric storage heaters as their primary heating.

[c] Some 22% of our sample (510,000 households) were transferred to Octopus Energy Limited following the December 2022 government sale of their previous supplier, Bulb Energy, which had entered special administration in late 2021. These households were generally new to the program, having missed the bulk of the first-year events.

Across almost every other dimension – prior program experience, energy efficiency, dwelling type, tenure, metering, prior supplier, and geography – we found no meaningful heterogeneity (Table 2). Two nuances are worth flagging: prior participants were far more likely to re-enroll but reduced no more once enrolled, so experience shaped sign-up rather than response; and sign-up rates varied across regions while per-event reductions did not.

In summary, demand response engaged customers broadly. Maintaining program accessibility and appeal across customer segments, including those in rental housing or on prepayment meters, seemed to maximize overall participation and total response. At the same time, high-consumption and technology-enabled households delivered the largest per-household reductions, making them the natural focus where per-customer absolute impact is the priority.

3.3.2 Temperature and event conditions

The characteristics above were fixed household attributes. The treatment effect also varied with the *conditions* of each event – most visibly outdoor temperature. Unlike opt-in, these conditions were not randomized: each event bundled a date, weather, and remuneration rate together. The results below therefore describe *where* the randomized treatment effect was larger, not the causal effect of any one condition.

These conditions moved together over the trial. Earlier events were colder and paid more; later events were warmer and paid less, so temperature, remuneration, and event timing cannot be separated. A "horse race" interacting treatment with both temperature and remuneration (Appendix A9) leaves the temperature interactions mostly insignificant once price is included, while the remuneration interactions persist – but this only ranks two correlated conditions, neither of which was randomized. We therefore read the temperature and remuneration gradients as descriptive features of how response varied across events, not as separately identified effects.

3.4 Treatment persistence analysis

To evaluate whether treatment effects remained consistent across events, we re-estimated our main intent-to-treat and two-stage least squares models (Equations (3) to (5)) for each of the 13 individual events separately.[17] This allows us to examine whether the program's

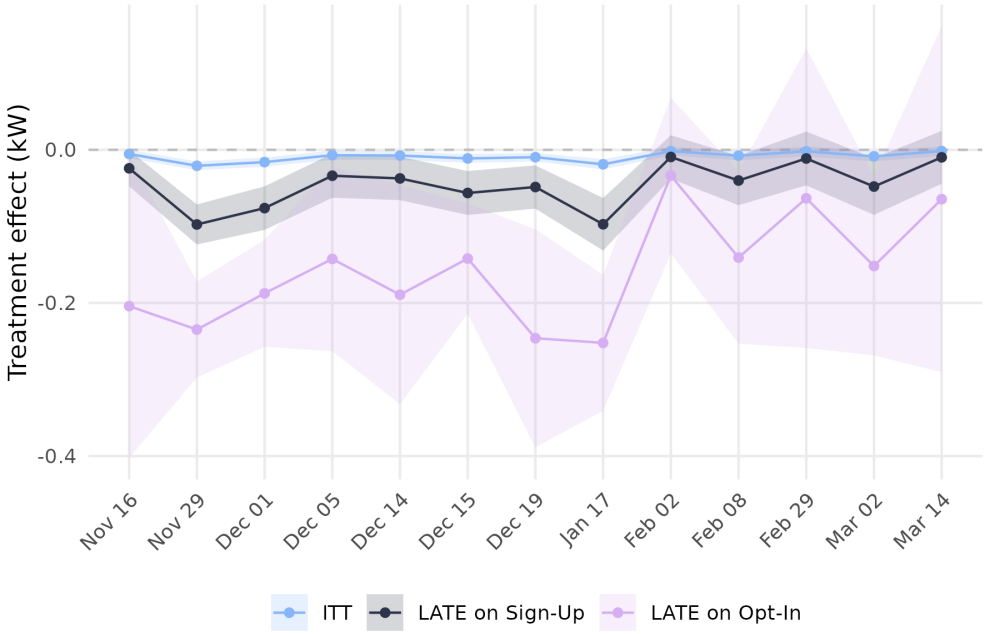
[17] Due to computational limitations, we did not estimate event-specific interacted two-stage least squares models; instead, we split the analysis into separate regressions per event.

effectiveness deteriorated, improved, or remained stable over time.

Electricity reductions varied substantially across events, with later events exhibiting slightly smaller reductions on average (see Figure 7). However, this pattern likely reflected differences in event characteristics, such as temperature and incentives, rather than participant fatigue or disengagement (Section 3.3.2 and Appendix A5.4).

Consistent with this interpretation, comparisons between households invited earlier and those invited later showed no decline in treatment response, although the estimates were imprecise (Appendix A5.2).

Figure 7: Impact of Being Invited, Sign-Up and Opt-In by Session



Notes: This figure reports the ITT and LATE estimates on event-period electricity consumption, estimated separately for each of the 13 events as part of the pre-specified event-by-event ITT/LATE. Estimates are in kW (kWh per hour) during the event window, with 95% confidence intervals. Cross-event variation largely reflects differences in event characteristics such as temperature, notice period and remuneration rather than participant fatigue.

3.5 Sub-trials and further quasi-experimental analyses

Beyond the core randomized encouragement design, we conducted a series of pre-specified sub-trials and exploratory quasi-experimental analyses to probe mechanisms, timing, external validity, and potential spillovers. We briefly summarize these analyses here and refer readers to the relevant Appendix sections for full details.

First, we implemented a late-season "chaser" email sent to households who had been invited but had not yet signed up, allowing us to test whether a reminder could meaningfully expand participation (Appendix A5.1). The chaser increased sign-up and opt-in rates by a small but non-trivial amount, but the number of newly induced participants was limited, resulting in imprecise estimates of consumption impacts. That said, the estimated complier-level effects for chaser-induced participants were similar in magnitude to those in the main trial, suggesting no obvious differences in underlying demand response.

Second, we randomized a late invitation to a subset of the original control group after most events had already occurred, enabling a comparison between early- and late-invited participants (Appendix A5.2). Late invitations generated substantial sign-up and opt-in, though at lower levels than early invitations. Consumption impacts during the final events were similar across early- and late-invited groups, providing little evidence of fatigue or declining effectiveness over the season; again, however, the ability to detect these differences was imprecise.

Third, we replicated the encouragement design among a sample of small commercial customers to assess external validity beyond domestic households (Appendix A5.3). While the invitation significantly increased participation, consumption effects were noisy and imprecisely estimated. Point estimates suggest potentially large reductions among compliers, but limited statistical power prevents strong conclusions.

Fourth, we tested the impact of a higher financial incentive per kWh of demand response during a single event, exploiting a randomized increase in the per-kWh reward offered to a subset of 5,000 already signed-up households (Appendix A5.4). The higher incentive substantially increased opt-in rates. We detected no corresponding increase in consumption reduction; however, given the small treated sample, these null results would also be consistent with economically meaningful effects below our minimum detectable effect size.

Finally, we examined whether participation affected longer-run tariff choices or utility switching (Appendix A6). We found no evidence that invited or participating households were more likely to adopt time-of-use tariffs or to remain with the utility, reinforcing the interpretation of the program as a targeted, short-run demand response tool rather than a catalyst for broader changes in customers' relationship with their energy provider.

Taken together, these additional analyses suggest that the estimated effects were broadly similar across different populations, timing regimes, and sources of quasi-random variation.

4 Welfare analysis

To estimate welfare impacts, we focused on demand reductions from the utility that delivered our randomized encouragement trial, for which we had high-frequency household data to credibly estimate realized demand reductions across the 12 events in which the implementing partner's bids were accepted and opt-in communications were successfully delivered to households.[18] We used LATE estimates of opt-in on electricity consumption, which provided an unbiased estimate of demand reduction per household in each half-hour for compliers. These LATE estimates were broadly similar to those from the standard industry baseline methodology, though with some event-to-event variation (Appendix A12.1). Every quantity in the welfare analysis is derived from this complier response: the realized demand reduction, and through it the CO_2e abatement, avoided balancing cost, and lost-load value in the numerator, together with the administrative cost and the per-MWh remuneration and break-even figures, are all measured on the compliers induced to participate by the encouragement.

The complier basis makes the MVPF internally coherent but conservative. Compliers turned down somewhat less than always-takers — roughly 0.18 kW versus 0.23 kW per opted-in household (Appendix A10). Costs and benefits are thus computed on a consistent complier basis; but the realized aggregate reduction is correspondingly smaller than the full campaign, which also mobilized the more-responsive always-takers, delivered. For context, we estimated that our compliers delivered 295.21 MWh of demand reduction. Including always-takers (see Table 1) – from both our treatment and control group – we estimated that total demand reduction, across our entire sample, was approximately 1,450 MWh.

4.1 Welfare analysis framework

We evaluated the welfare effects of the DFS using the Marginal Value of Public Funds (MVPF) framework (Finkelstein and Hendren, 2020, Hendren and Sprung-Keyser, 2020). The MVPF is the ratio of marginal social benefits to the net cost to the government. The relevant public expenditure is the system operator's payment for delivered demand response, net of balancing costs avoided because demand fell during the event:

[18] This excludes the two events previously omitted due to communication issues (12 December 2023 and 1 March 2024), as well as the 14 March 2024 event, which was delivered despite the bid being rejected by the system operator. In other words, we included the 14 March event in our demand analysis but excluded it from our welfare analysis.

$$\text{MVPF} = \frac{\text{Consumer surplus} + \text{Utility surplus} + \text{Externality benefits}}{\text{Costs to the system operator} - \text{Avoided balancing costs}} \tag{2}$$

Table 3: Welfare components: definitions and measurement. Benefits enter the MVPF numerator; costs enter the denominator.

Component	Construction
Benefits	
Consumer surplus	50% of household payments (80% of system operator outlay), assuming all participating households were marginal[a] (Appendix A12.3.1).
Utility surplus	50% of the utility's 20% share of system operator payments (Appendix A12.3.1).
CO_2e abatement	LATE-estimated reduction × marginal fuel emissions factor × social cost of carbon (Appendix A12.3.2).
Lost load avoidance	LATE-estimated reduction × £6,000/MWh UK regulatory Value of Lost Load (VoLL) (Appendix A12.3.3).
Costs	
Costs to the system operator	System operator payments for delivered reduction, from settlement records (Appendix A12.4.1).
Avoided balancing costs	Marginal balancing action price × LATE-estimated demand reduction; marginal unit identified as the highest-priced accepted Balancing Mechanism action ≥25 MW (Appendix A12.4.2).

[a] Under the standard Harberger triangle approximation with a linear demand curve, the average valuation of a marginal consumer equals one-half of the payment (Hahn et al., 2026, Harberger, 1964, Hendren and Sprung-Keyser, 2020). We treated all participating households as marginal, so consumer surplus equaled 50% of household payments. This analysis is exploratory.

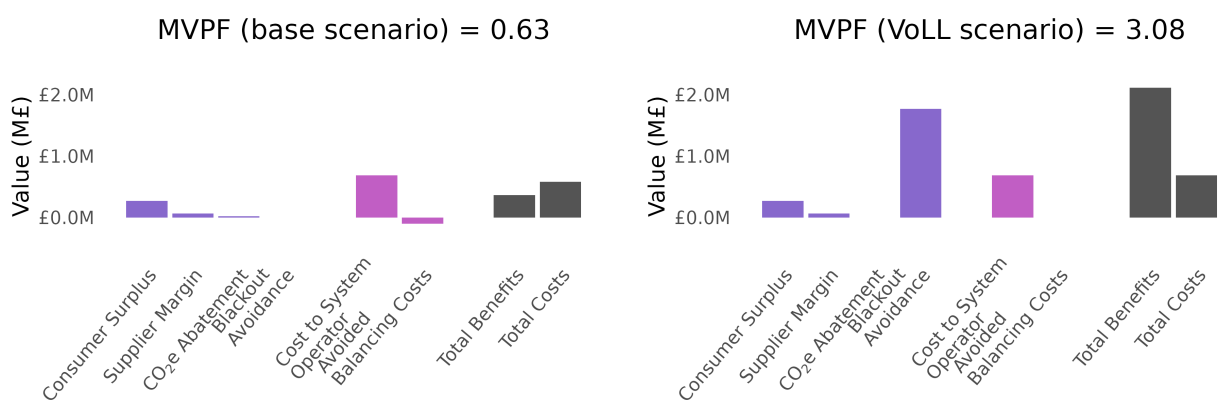
We computed two sets of MVPF estimates, corresponding to two views of what DFS substitutes for. In the *base scenario*, DFS is treated as a substitute for ordinary balancing actions: avoided balancing costs offset the denominator, and CO_2e abatement enters the numerator. In the *VoLL scenario*, DFS is instead treated as scarcity insurance deployed alongside, not in place of, ordinary balancing actions: the relevant counterfactual is involuntary disconnection, so avoided balancing costs and CO_2e abatement drop out and each MWh of realized reduction is valued at the GB regulatory VoLL of £6,000/MWh. Table 3 summarizes how each component was defined and measured; full details are provided in Appendix A12. Table A18 presents the event-level welfare inputs and resulting MVPFs; Figure 8 provides a visual summary of aggregate costs and benefits for the Winter 2023–24 campaign.

4.2 Marginal value of public funds

Base scenario: When treated as a substitute for other regular balancing actions, the program's MVPF was lower than 1 for all events but the one on 1 December 2023. At the campaign level, consumer surplus was £274,500, producer (utility) surplus £68,600, and carbon abatement value £24,500, against costs to the system operator of £686,400, yielding an overall MVPF of 0.63. Event-level estimates ranged from 0.51 to 1.48, with the highest value in the live event on 1 December 2023 – when alternative balancing actions were particularly expensive. The pattern is intuitive: the program appeared least attractive when household payments were high relative to the cost of alternative balancing actions, and most attractive when payments were low or the value of reductions was unusually high.

VoLL scenario: MVPFs were substantially higher when we treated DFS as scarcity insurance. At the program level, the VoLL MVPF was 3.08. All events equaled or exceeded 1.

Figure 8: Benefits, costs and overall MVPF



Notes: The figure decomposes the MVPF into its main benefit and cost components. This analysis is exploratory: welfare analysis was not part of the pre-analysis plan. Purple bars denote benefits, pink bars denote costs, and dark grey bars show total benefits and total costs. The left panel reports the base scenario, in which benefits consist of valued customer payments, utility surplus, and CO_2e abatement, while total costs are administrative expenditure net of avoided balancing costs. The right panel reports the VoLL scenario, in which the key additional benefit is avoidance of involuntary disconnection, valued at the GB regulatory VoLL; since DFS in this scenario is deployed alongside rather than in place of balancing actions, total costs are administrative expenditure alone.

4.3 Final welfare considerations

Break-even prices to reach average MVPF of 1: The base-case MVPF fell below 1 because household remuneration was high relative to the value of balancing actions displaced and benefits obtained through the demand reductions. A break-even analysis (Appendix A12.5) estimates the socially optimal remuneration and avoided balancing cost. The program would have had an MVPF of 1 if the balancing actions it displaced had cost around £1,080/MWh, roughly the level seen during acute system stress, or if households had been paid around £688/MWh rather than the average £1,860/MWh actually paid.

Both of these thresholds could shift in the future. On the cost side, the avoided balancing action price reflects the cost of gas-fired generation, which is volatile; periods of high gas prices or tightening carbon policy could push marginal balancing costs considerably higher. On the remuneration side, the high payments observed here partly reflect the early-stage nature of demand flexibility as a market: fixed costs of platform development, consumer recruitment, and learning are spread over a relatively small participant base, and per-unit costs are likely to fall as the market scales.

In addition, our MVPF estimates omit learning-by-doing benefits entirely. In the MVPF literature, spillovers from early-stage program deployment – i.e., reductions in future costs as technology matures, platforms improve, and consumer familiarity grows – are often among the largest components of social value (Hendren and Sprung-Keyser, 2020). We have not attempted to model these benefits, which means our base-case MVPF should be read as a lower bound rather than a point estimate of the program's true social value.

DFS as contingency insurance: The base and VoLL scenarios represent two extremes: DFS as a routine balancing substitute, and DFS as a last-resort measure to prevent involuntary lost load. In practice, the right framing may lie somewhere in between. System operators routinely pay substantial premia for contingency capacity that may never be called upon. For example, coal contingency contracts procured by NESO during Winter 2022–23 cost £340M to £395M despite being used only once.[19]

A further reference point is the Capacity Market, GB's principal instrument for procuring firm capacity for system adequacy. Taking the program's total cost (among *compliers*) of £686,400 over the 295.21 MWh of reductions delivered across the 12 events included in

[19] NESO procured these contracts "to ensure safe and secure operation of the electricity system throughout Winter" (National Grid, 2022), at a total capacity of 2.2 GW (John, 2023, National Grid, 2022, UK Parliament, Energy Security and Net Zero Committee, 2023). They were utilized once, on 7 March 2023, when two units at West Burton A delivered 2.5 GWh over seven hours.

our welfare analysis implies a cost of roughly £28/kW of demand reduction at the observed level of utilization.[20] The £28/kW figure falls within the range of recent Capacity Market clearing prices, though that range is wide. Four-year-ahead (T-4) auctions, which secure capacity for adequacy several years out, cleared between £65/kW for delivery year 2027–28 and £27/kW for 2029–30 (Deloitte LLP, 2024, 2026*b*); the most recent one-year-ahead (T-1) auction, which tops up near-term capacity, cleared far lower, at £5/kW for 2026–27 (Deloitte LLP, 2026*a*). The DFS figure is itself sensitive to how heavily the resource is used: because cost scales with delivered energy, fewer events would mechanically reduce the implied £/kW, so a resource reserved for genuine scarcity rather than the 12 events observed here would sit lower still.

The comparison should be read with care: the Capacity Market pays for *availability*, whereas DFS payments were contingent on *realized* reductions. Since no System Stress Event has been declared since the Capacity Market's inception in Great Britain, its non-delivery penalty has never bound; judged ex post on delivered hours, any availability-paid resource looks like deadweight. Valued ex ante against the reliability standard, the pertinent question is therefore not DFS's cost per kW but whether its availability would prove as dependable in a genuine scarcity event as contracted conventional capacity.

A resource that cleared that bar would also unlock a broader benefit that our analysis does not capture: by reliably displacing gas-fired peaking generation, demand flexibility could over time reduce the need to build or retain such capacity at all – a system-level decarbonization benefit that goes beyond the direct emissions avoided in any single event.

5 Conclusion

Our natural field experiment – to our knowledge, the largest natural field experiment of residential demand-side response ever conducted – demonstrated that a nationwide demand flexibility program could deliver substantial reductions in peak electricity usage and serve as a reliable grid resource when coordinated at scale. More broadly, the results suggest that residential demand flexibility can operate as a potentially important system-level resource in increasingly electrified and weather-dependent electricity systems.

[20] The implied annual capacity cost scales with the number of hours the resource is called: the unit cost of £2,325/MWh multiplied by H hours gives a figure in £/kW per year, or about £28/kW at the 12 event-hours observed here. This treats cost as proportional to delivered energy and is therefore illustrative; a portion reflects fixed platform, recruitment, and learning costs that do not scale with utilization, so the figure is best read as an order-of-magnitude comparison.

During Winter 2023–24, simply inviting households to participate via a targeted email nudge increased sign-up rates by 20 percentage points and reduced average peak-period consumption across all invited households by 1.4% (ITT). Among those who actively opted in, reductions were much larger – 23.1% during the peak window – highlighting the significant curtailment achievable when consumers engaged with the program. These results confirm that the observed demand reductions were driven by the intervention itself, rather than by low-consumption users self-selecting into participation or other confounding factors, and illustrate the potential of large-scale domestic demand flexibility to balance the grid in critical moments.

There was no strong evidence of a large "rebound" or load shifting. We found no statistically detectable increases in electricity usage in the hours before or after peak events, although small or diffuse effects cannot be ruled out given statistical power. There were no changes in overall consumption on non-event days, either. In other words, the demand response was concentrated during events.

We detected no fatigue or drop-off in performance as the number of events accumulated. Households invited later in the season responded similarly to those invited at the start, and the load reductions per participant remained fairly consistent across events. These findings suggested that short-notice, incentive-driven demand response can be deployed repeatedly over the course of a winter without detectable attenuation in response, at least at the frequency observed in our study (13 events over a four-month period).

In examining mediators and treatment effect heterogeneity, low-carbon technology adoption was the one factor that clearly shaped response. EV adoption modestly increased it, while adopting heat pumps or solar-and-battery roughly doubled it. By contrast, response was strikingly similar across household characteristics – no significant differences by tenure, dwelling type, or payment method, with prepayment and credit households delivering similar proportional reductions.

Our welfare analysis suggests that the value of demand flexibility hinges on the counterfactual it displaces. The incentives offered were large – in our trial, between £1.75 and £4 per kWh (£1,750 to £4,000 per MWh) for electricity not consumed during the peak events. While this figure is relatively high on a per-MWh basis, it must be weighed against the high costs of alternatives during rare supply crunches. In addition, if the demand reductions from programs like DFS contribute to avoiding lost load, the benefit-cost ratio can be well above 1. In less extreme scenarios, the intervention's benefit-cost ratio (0.63) shows that measured benefits did not fully offset measured costs under normal balancing conditions. This suggests that the value of demand flexibility depends importantly on procurement

design, remuneration levels, and the system conditions under which events are deployed.

Moving forward, our findings suggest that explicit demand flexibility services may be most valuable when deployed selectively during periods of genuine system stress, rather than as routine balancing mechanisms under normal conditions. In this sense, "critical peak rewards" may emerge as an important complement to wholesale price signals and traditional reserve capacity in increasingly renewable electricity systems. The challenge for system operators will be to balance offering attractive payments to ensure reliable performance with containing costs – but our findings give confidence that demand flexibility can be procured at a cost that is justified by the reliability and carbon benefits in critical periods.

References

Allcott, H., Collard-Wexler, A. and O'Connell, S. D. (2016), 'How do electricity shortages affect industry? Evidence from India', *American Economic Review* **106**(3), 587–624. 2, 6

Andersen, L. M., Hansen, L. G., Jensen, C. L. and Wolak, F. A. (2019), Can incentives to increase electricity use reduce the cost of integrating renewable resources, Technical report, National Bureau of Economic Research. 6

Bailey, M. R., Brown, D. P., Shaffer, B. C. and Wolak, F. A. (2025), Take the Load Off: Effort and Technology as Determinants of Electricity Demand Response, Working Paper 33836, National Bureau of Economic Research. Revised July 2025.
URL: *https://www.nber.org/papers/w33836* 7

Baldick, R., Kolos, S. and Tompaidis, S. (2006), 'Interruptible Electricity Contracts from an Electricity Retailer's Point of View: Valuation and Optimal Interruption', *Operations Research* **54**(4), 627–642. 6

Bergquist, M., Thiel, M., Goldberg, M. H. and Van Der Linden, S. (2023), 'Field Interventions for Climate Change Mitigation Behaviors: A Second-Order Meta-Analysis', *Proceedings of the National Academy of Sciences* **120**(13), e2214851120. 6

Bollinger, B. K. and Hartmann, W. R. (2020), 'Information vs. Automation and implications for dynamic pricing', *Management Science* **66**(1), 290–314. 6

Borenstein, S., Bushnell, J. and Mansur, E. (2023), 'The Economics of Electricity Reliability', *Journal of Economic Perspectives* **37**(4), 181–206. 2

Brandon, A., List, J. A., Metcalfe, R. D., Price, M. K. and Rundhammer, F. (2019), 'Testing for crowd out in social nudges: Evidence from a natural field experiment in the market for electricity', *Proceedings of the National Academy of Sciences* **116**(12), 5293–5298. 6

Brewer, D. and Crozier, J. (2025), 'Who heeds the call to conserve in an energy emergency? Evidence from smart thermostat data', *Journal of the Association of Environmental and Resource Economists* **12**(6), 1747–1789. 6

Burgess, R., Greenstone, M., Ryan, N. and Sudarshan, A. (2020), 'The consequences of treating electricity as a right', *Journal of Economic Perspectives* **34**(1), 145–169. 2

Burkhardt, J., Gillingham, K. T. and Kopalle, P. K. (2023), 'Field experimental evidence on the effect of pricing on residential electricity conservation', *Management Science* **69**(12), 7784–7798. 6

Burlando, A. (2014), 'Transitory shocks and birth weights: Evidence from a blackout in Zanzibar', *Journal of Development Economics* **108**, 154–168. 2

Callaway, B. and Sant'Anna, P. H. C. (2021), 'Difference-in-Differences with Multiple Time Periods', *Journal of Econometrics* **225**(2), 200–230. 61

Carreras, B. A., Colet, P., Reynolds-Barredo, J. M. and Gomila, D. (2021), 'Assessing Blackout Risk With High Penetration of Variable Renewable Energies', *IEEE Access* **9**, 132663–132674. 2

Caves, D. W. and Christensen, L. R. (1980), 'Econometric analysis of residential time-of-use electricity pricing experiments', *Journal of Econometrics* **14**(3), 287–306. 6

Caves, D. W., Christensen, L. R. and Herriges, J. A. (1984), 'Consistency of residential customer response in time-of-use electricity pricing experiments', *Journal of Econometrics* **26**(1-2), 179–203. 6

Cole, M. A., Elliott, R. J., Occhiali, G. and Strobl, E. (2018), 'Power outages and firm performance in Sub-Saharan Africa', *Journal of Development Economics* **134**, 150–159. 2

de Chaisemartin, C. and D'Haultfœuille, X. (2020), 'Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects', *American Economic Review* **110**(9), 2964–2996. 61

Deloitte LLP (2024), Capacity Market auction: Auction Monitor report for T-4 auction for 2027 to 2028, Technical report, Department for Energy Security and Net Zero. Auction concluded 27 February 2024; clearing price £65.00/kW/year. 33

Deloitte LLP (2026*a*), Capacity Market auction: Auction Monitor report for T-1 auction for 2026 to 2027, Technical report, Department for Energy Security and Net Zero. Auction concluded 4 March 2026; clearing price £5.00/kW/year. 33

Deloitte LLP (2026*b*), Capacity Market auction: Auction Monitor report for T-4 auction for 2029 to 2030, Technical report, Department for Energy Security and Net Zero. Auction concluded 11 March 2026; clearing price £27.10/kW/year. 33

Feng, K., Ouyang, M. and Lin, N. (2022), 'Tropical cyclone-blackout-heatwave compound hazard resilience in a changing climate', *Nature Communications* **13**(11), 4421. 2

Finkelstein, A. and Hendren, N. (2020), 'Welfare Analysis Meets Causal Inference', *Journal of Economic Perspectives* **34**(4), 146–167. 29

Fisher-Vanden, K., Mansur, E. T. and Wang, Q. J. (2015), 'Electricity shortages and firm productivity: evidence from China's industrial firms', *Journal of Development Economics* **114**, 172–188. 2

Fotis, G., Vita, V. and Maris, T. I. (2023), 'Risks in the European Transmission System and a Novel Restoration Strategy for a Power System after a Major Blackout', *Applied Sciences* **13**(11), 83. 2

Fowlie, M., Wolfram, C., Baylis, P., Spurlock, C. A., Todd-Blick, A. and Cappers, P. (2021), 'Default effects and follow-on behaviour: Evidence from an electricity pricing program', *The Review of Economic Studies* **88**(6), 2886–2934. 6, 7, 18, 65

Fried, S. and Lagakos, D. (2023), 'Electricity and firm productivity: A general-equilibrium approach', *American Economic Journal: Macroeconomics* **15**(4), 67–103. 2

Gertler, P. J., Shelef, O., Wolfram, C. D. and Fuchs, A. (2016), 'The demand for energy-using assets among the world's rising middle classes', *American Economic Review* **106**(6), 1366–1401. 2

Goodman-Bacon, A. (2021), 'Difference-in-Differences with Variation in Treatment Timing', *Journal of Econometrics* **225**(2), 254–277. 61

Gowrisankaran, G., Reynolds, S. S. and Samano, M. (2016), 'Intermittency and the value of renewable energy', *Journal of Political Economy* **124**(4), 1187–1234. 2

Hahn, R. W., Hendren, N., Metcalfe, R. D. and Sprung-Keyser, B. (2026), 'A welfare analysis of policies impacting climate change', *American Economic Review* **116**(7), 2368–2421. 30, 70

Harberger, A. C. (1964), 'The measurement of waste', *The American Economic Review* **54**(3), 58–76. 30, 70

Hendren, N. and Sprung-Keyser, B. (2020), 'A unified welfare analysis of government policies*', *The Quarterly Journal of Economics* **135**(3), 1209–1318. 29, 30, 32, 70

Houthakker, H. S. (1951), 'Electricity tariffs in theory and practice', *Economic Journal* **61**(241), 1–25. 6

Ito, K., Ida, T. and Tanaka, M. (2018), 'Moral suasion and economic incentives: Field experimental evidence from energy demand', *American Economic Journal: Economic Policy* **10**(1), 240–267. 6

Ito, K., Ida, T. and Tanaka, M. (2023), 'Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice', *American Economic Review* **113**(11), 2937–2973. 6

Jahn, W., Urban, J. L. and Rein, G. (2022), 'Powerlines and Wildfires: Overview, Perspectives, and Climate Change: Could There Be More Electricity Blackouts in the Future?', *IEEE Power and Energy Magazine* **20**(1), 16–27. 2

Jessoe, K. and Rapson, D. (2014), 'Knowledge is (less) power: Experimental evidence from residential energy use', *American Economic Review* **104**(4), 1417–1438. 6

John, A. (2023), 'ESO puts back-up coal plants to work for first time this winter'.
URL: *https://utilityweek.co.uk/eso-puts-back-up-coal-plants-to-work-for-first-time-this-winter/* 32

Joskow, P. L. (1976), 'Contributions to the theory of marginal cost pricing', *The Bell Journal of Economics* pp. 197–206. 6

Karaduman, Ö. (2021), *Large Scale Wind Power Investment's Impact on Wholesale Electricity Markets*, JSTOR. 2

List, J. A. (2025), *Experimental economics: Theory and practice*, University of Chicago Press. 7

Masood, N. A., Yan, R. and Kumar Saha, T. (2018), Cascading Contingencies in a Renewable Dominated Power System: Risk of Blackouts and Its Mitigation, *in* '2018 IEEE Power & Energy Society General Meeting (PESGM)', p. 1–5. 2

Moll, B., Schularick, M. and Zachmann, G. (2023), 'The power of substitution: The great German gas debate in retrospect', *Brookings Papers on Economic Activity* . 6

Morão, H. (2026), 'The 2025 Iberian Peninsula blackout: Lessons for modern power systems and policy implications', *Utilities Policy* **101**, 102196. 2

National Grid (2022), 'Winter Contingency Contracts'.
URL: *https://www.nationalgrideso.com/document/268126/download* 32

Ofgem (2023), 'Retail market indicators'.
URL: *https://www.ofgem.gov.uk/retail-market-indicators* 3

Panteli, M. and Mancarella, P. (2015), 'Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies', *Electric Power Systems Research* **127**, 259–270. 2

Tan, C.-W. and Varaiya, P. (1993), 'Interruptible electric power service contracts', *Journal of Economic Dynamics and Control* **17**(3), 495–517. 6

UK Parliament, Energy Security and Net Zero Committee (2023), 'Written evidence submitted to the Energy Security and Net Zero Committee'. Evidence on electricity system security and contingency arrangements.
URL: *https://committees.parliament.uk/writtenevidence/123822/html/* 32

Wolak, F. A. (2007), 'Residential customer response to real-time pricing: The Anaheim critical peak pricing experiment'. 6

Wolak, F. A. (2011), 'Do residential customers respond to hourly prices? Evidence from a dynamic pricing experiment', *American Economic Review* **101**(3), 83–87. 6

Wolak, F. A. (2022), 'Long-term resource adequacy in wholesale electricity markets with significant intermittent renewables', *Environmental and Energy Policy and the Economy* **3**(1), 155–220. 2

Yan, R., Masood, N.-A., Kumar Saha, T., Bai, F. and Gu, H. (2018), 'The Anatomy of the 2016 South Australia Blackout: A Catastrophic Event in a High Renewable Network', *IEEE Transactions on Power Systems* **33**(5), 5374–5388. 2

Złotecka, D. and Sroka, K. (2018), The characteristics and main causes of power system failures basing on the analysis of previous blackouts in the world, *in* '2018 International Interdisciplinary PhD Workshop (IIPhDW)', p. 257–262. 2

Appendix

A1 Balance tables

Table A1: Balance Table by Encouragement Group

	N		Means		P-values
Variable	Control	Invited	Control	Invited	T vs C
Dwelling type = House	100966	2162682	87.17	87.28	0.30
EPC	74104	1589701	65.39	65.37	0.71
Property Value	88484	1895179	371490.87	370392.65	0.22
Owner Occupied	100078	2142823	85.37	85.38	0.91
Urban	100987	2163125	77.85	77.76	0.49
IMD Score	88742	1899868	17.85	17.88	0.54
Main fuel = Electricity	107239	2296405	5.44	5.45	0.87
Has Export Meter	107239	2296405	4.97	5.01	0.59
Prepayment Customer	107239	2296405	1.57	1.57	0.82
EAC	107005	2291489	4130.08	4128.56	0.89
Is Bulb	107239	2296405	21.95	22.06	0.42
Is ToU	107218	2295963	8.20	8.20	0.98
Is Economy 7	107122	2293924	7.83	7.94	0.21
Returning Participant	107239	2296405	25.79	25.89	0.49

Note:
ToU (Time of Use) tariff excludes Economy 7 customers, who are reported as a separate category.

Table A3: Sorting on Invited, Sign-Up, and Opt-In

Dependent Variables:	Invited	Signed-Up	# Opt-Ins
Model:	(1)	(2)	(3)
Variables			
Dwelling type = House	0.0010*	0.0326***	0.3650***
	(0.0005)	(0.0014)	(0.0168)
EPC score	-1.6×10^{-6}	0.0004***	0.0030***
	(1.15×10^{-5})	(3.1×10^{-5})	(0.0004)
Property Value (£100k)	-9×10^{-5}	-0.0041***	-0.0443***
	(6.94×10^{-5})	(0.0002)	(0.0023)
Owner Occupier	7.16×10^{-5}	0.0243***	0.2292***
	(0.0004)	(0.0011)	(0.0129)
Urban	6.88×10^{-5}	-0.0024**	-0.1314***
	(0.0004)	(0.0010)	(0.0124)
IMD Score	9.01×10^{-6}	-0.0008***	-0.0088***
	(1.2×10^{-5})	(3.25×10^{-5})	(0.0004)
Main fuel = Electricity	0.0006	0.0171***	0.2817***
	(0.0006)	(0.0017)	(0.0204)
Has Export Meter	0.0012*	0.1161***	1.708***
	(0.0007)	(0.0017)	(0.0221)
Is Credit	-7.28×10^{-5}	0.1554***	1.645***
	(0.0011)	(0.0033)	(0.0403)
EAC in mWh	-1.86×10^{-5}	0.0006***	-0.0052***
	(4.52×10^{-5})	(0.0001)	(0.0016)
Is Bulb	0.0005	0.0228***	0.2079***
	(0.0004)	(0.0010)	(0.0125)
Is ToU	0.0004	0.1887***	2.608***
	(0.0006)	(0.0013)	(0.0179)
Returning Participant	0.0002	0.3854***	5.396***
	(0.0003)	(0.0008)	(0.0108)
Fixed-effects			
GSP Group	Yes	Yes	Yes
Fit statistics			
Observations	1,397,756	1,397,756	1,397,756
R^2	1.35×10^{-5}	0.14453	0.17834
F-test	0.72576	9,082.6	11,668.1

Heteroskedasticity-robust standard-errors in parentheses

Notes: The sample is restricted to customers for whom complete housing characteristics data are available. Column 1 confirms balance between treatment and control groups. Columns 2 and 3 reveal significant sorting: households with higher affluence (homeowners, high EPC) and those with prior program experience (participation in 2022-23) were more likely to engage. Increased participation among prepayment and Time-of-Use (ToU) customers suggests that cost-saving motives and familiarity with demand shifting are key drivers of program uptake. These patterns of self-selection underscore the necessity of the randomized encouragement design used in this study to isolate causal effects.

A2 Main analysis tables

We modeled the following ordinary least squares regression

$$Y_{it} = \alpha_0 + \alpha_1 Invited_i + \alpha_2 Temperature_{it} + X\alpha + \epsilon_{it} \tag{3}$$

Where Y_{it} represents electricity consumption (kWh) for household i at the half-hour level t during events. Results are presented as average power (kW) by scaling to hourly rates. We also ran a version with the logarithmic electricity consumption (excluding half-hours with null consumption) in Table A5. $Invited_i$ is a binary variable: 1 if the household i is part of the treatment group (received an email invitation encouraging sign-up to the demand flexibility program) and 0 if in the control group (we excluded late-invited customers from this analysis, and thus from the control group). α_1 represents the impact of being encouraged to sign up to the demand flexibility program. $Temperature_{it}$ is the temperature in household i's region at time t, rounded to the nearest degree and modeled as a categorical variable. X is a vector of household characteristics, including (i) the decile of pre-treatment electricity consumption, defined as the average half-hourly consumption during the peak period (16:30–18:00) over the Tuesday–Saturday preceding the start of the trial on 16 November 2023, and (ii) an indicator variable equal to one for credit customers and zero for pre-payment customers. ϵ_{it} denotes the error term.

To estimate the impact of actual program participation (signing up and opting in to events) for those who complied with the encouragement, we employed a two-stage least squares approach. In this instrumental variables analysis, assignment to the invitation serves as an instrument for program sign-up and event opt-in. This yields a Local Average Treatment Effect (LATE) for compliers (households who sign up or opt in because they were invited). The first stage assesses how the invitation influenced participation, and the second stage assesses how that induced participation affected consumption during events.

$$SignUp_i = \beta_0 + \beta_1 Invited_i + \beta_2 Temperature_{it} + X\beta + \nu_i \tag{4}$$

Where $SignUp_i$ is a binary variable indicating whether household i signed up to the program, $Invited_i$ is a binary variable equal to 1 if household i was part of the treatment group (received an invitation) and 0 otherwise, the rest of the variables are the same as Equation (3).

Table A4: Household Sign-Up and Opt-In Response to Invitation (First Stage)

Dependent Variables:	Signed-Up	Opted-In	Signed-Up	Opted-In
Sample	kW		Ln kW	
Model:	(1)	(2)	(3)	(4)
Variables				
Invited	0.2015***	0.0591***	0.2021***	0.0589***
	(0.0016)	(0.0009)	(0.0016)	(0.0009)
Group Average				
Control mean	0.264	0.115	0.261	0.112
Fixed-effects				
Temperature	Yes	Yes	Yes	Yes
Is Credit	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes
Fit statistics				
Observations	57,535,536	57,535,536	56,817,210	56,817,210
Number of Households	2,320,256	2,320,256	2,316,857	2,316,857
R^2	0.00650	0.01017	0.00671	0.00987
Wald (1st stage)	15,894.1	4,005.4	16,080.4	4,101.8

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Columns 1 and 3 report first-stage effects on sign-up, while Columns 2 and 4 report first-stage effects on opt-in. The kW and Ln kW headers refer to the second-stage specification paired with each first stage; the dependent variables in this table are participation indicators, so coefficients are percentage-point changes. Temperature: hourly average temperature in Celsius in the participant's region. Is Credit: binary indicator for credit customers. Pre-Trial Consumption: pre-treatment average half-hourly consumption during the peak period over the Tuesday-Saturday preceding the start of the trial on 16 November 2023. Pre-specified first stage.

In the second stage, we use the fitted values for $SignUp_i$ from the first stage to estimate the impact of sign-up on electricity consumption. The model is:

$$Y_{it} = \gamma_0 + \gamma_1 \widehat{SignUp}_i + \gamma_2 Temperature_{it} + X\gamma + \epsilon_{it} \tag{5}$$

Where Y_{it} represents electricity consumption (kWh) for household i at time t during events, $\widehat{SignUp}_i$ is the predicted probability of sign-up from the first-stage regression, $Temperature_{it}$ and X are as defined in Equation (3).

We also run an analogous analysis to Equations (4) and (5), replacing the sign-up indicator with an opt-in indicator. This estimates the effect of the encouragement on opt-in behavior.

The validity of this approach relies on the assumption that invitation status affects consumption only through its impact on signing up and opting in, and not through any other

channels (such as changes in tariffs or the purchase of low-carbon technologies).

Table A5: Household Consumption Response to Events

Dependent Variables:	kW			Ln kW		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Invited	-0.0105***			-0.0173***		
	(0.0017)			(0.0020)		
Signed-Up		-0.0520***			-0.0855***	
		(0.0085)			(0.0100)	
Opted-In			-0.1773***			-0.2930***
			(0.0287)			(0.0340)
Group Average						
Control mean	0.766	0.766	0.766	0.782	0.782	0.782
Fixed-effects						
Temperature	Yes	Yes	Yes	Yes	Yes	Yes
Is Credit	Yes	Yes	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	57,535,536	57,535,536	57,535,536	56,817,210	56,817,210	56,817,210
Number of Households	2,320,256	2,320,256	2,320,256	2,316,857	2,316,857	2,316,857
R^2	0.22827	0.22827	0.22827	0.31209	0.31209	0.31209
Wald (1st stage), Signed-Up		15,894.1			16,080.4	
Wald (1st stage), Opted-In			4,005.4			4,101.8

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Columns 1 and 4 report ITT estimates, Columns 2 and 5 report the LATE for sign-up compliers, and Columns 3 and 6 report the LATE for opt-in compliers. kW denotes kWh per hour during the event window. Temperature: hourly average temperature in Celsius in the participant's region, Is Credit: binary indicator for credit customers, Pre-Trial Consumption: pre-treatment average half-hourly consumption during the peak period over the Tuesday-Saturday preceding the start of the trial on 16 November 2023. Pre-specified, including pre-registered log-consumption variant.

Online Appendix

Contents

A1 Balance tables **41**

A2 Main analysis tables **43**

A3 Main specifications robustness checks **47**

A4 Habit formation **49**

A5 Sub-trials **49**

A5.1 Chaser email impact on program sign-up 50

A5.2 Did late joiners have a different demand response during events? 52

A5.3 Demand response from commercial customers 55

A5.4 Impact of higher incentive on demand response 56

A6 Adoption of new time-of-use tariffs **58**

A7 Heterogeneity analysis tables and figures **59**

A8 Low-carbon technology adoption analysis **60**

A9 Relationship between temperature and remuneration **63**

A10 Heterogeneous effect by compliance type **65**

A11 Comparison of field experimental coefficients with baseline estimates **66**

A12 Welfare analysis methodology **69**

A12.1 Demand reduction estimates for welfare analysis 69

A12.2 Marginal generator, price, and emissions. 69

A12.3 Benefits 69

A12.4 Costs 71

A12.5 Break-even prices 72

A13 Deviations from our pre-analysis plans **74**

A3 Main specifications robustness checks

Table A6: Impact on Consumption in kWh per Hour

Dependent Variables:	kW			Imported kW			Exported kW		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Invited	-0.0105***			-0.0101***			0.0117		
	(0.0017)			(0.0016)			(0.0097)		
Signed-Up		-0.0520***			-0.0502***			0.0597	
		(0.0085)			(0.0081)			(0.0495)	
Opted-In			-0.1773***			-0.1712***			0.1837
			(0.0287)			(0.0274)			(0.1490)
Group Average									
Control	0.7660	0.7660	0.7660	0.7731	0.7731	0.7731	0.1517	0.1517	0.1517
Fixed-effects									
Temperature	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Is Credit	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	57,535,536	57,535,536	57,535,536	57,535,536	57,535,536	57,535,536	2,714,736	2,714,736	2,714,736
Number of Households	2,320,256	2,320,256	2,320,256	2,320,256	2,320,256	2,320,256	116,801	116,801	116,801
R^2	0.22827	0.22970	0.23507	0.23039	0.23147	0.23574	0.03933	0.04505	0.07196
F-test (1st stage), Signed-Up		277,173.6			277,173.6			14,397.4	
F-test (1st stage), Opted-In			41,723.4			41,723.4			1,508.9

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Pre-specified primary analysis

Table A7 reports two robustness checks alongside the main specification. Columns 1–3 repeat the main analysis (Table A5) with no controls, on the same sample excluding the late-invitation group. Columns 4–6 restrict the main specification to participants below the notice account ID cut-offs: for some events, notifications were not sent to all households but only a portion of signed-up participants. As notifications are sent by account ID, we can find up to which accounts notifications were sent, and only include them in the analysis. We do not use actual notifications sent as some households are not opted-in to receive notifications, which would bias our results. Across both checks, we found similar coefficients for the impact of being invited, signing up and opting in.

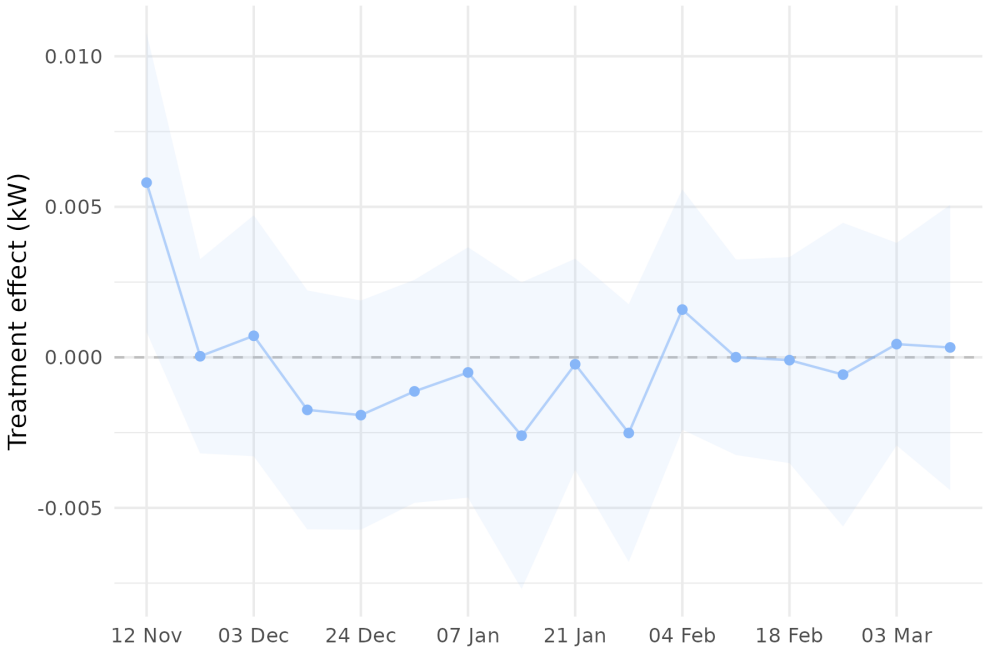
Table A7: Robustness: No Controls and Notice Cut-Off Restriction

Dependent Variable:	kW					
Sample	No Controls			Notice Cut-Off Restriction		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	0.7658***	0.7767***	0.7819***			
	(0.0025)	(0.0057)	(0.0073)			
Invited	-0.0082***			-0.0122***		
	(0.0025)			(0.0015)		
Signed-Up		-0.0412***			-0.0587***	
		(0.0124)			(0.0072)	
Opted-In			-0.1405***			-0.1746***
			(0.0423)			(0.0212)
Group Average						
Control mean	0.766	0.766	0.766	0.772	0.772	0.772
Fixed-effects						
Temperature				Yes	Yes	Yes
Is Credit				Yes	Yes	Yes
Pre-Trial Consumption				Yes	Yes	Yes
Fit statistics						
Observations	58,083,159	58,083,159	58,083,159	45,911,098	45,911,098	45,911,098
Number of Households	2,369,968	2,369,968	2,369,968	2,354,243	2,354,243	2,354,243
R^2	2.03×10^{-6}	2.03×10^{-6}	2.03×10^{-6}	0.22611	0.22611	0.22611
Wald (1st stage), Signed-Up		15,924.9			23,351.8	
Wald (1st stage), Opted-In			4,007.6			6,393.5

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Columns 1-3 use the main specification with no controls on the same sample as the main analysis (excluding the late-invitation group). Columns 4-6 restrict to participants below the notice account ID cut-offs, with full controls: for some events, notifications were not sent to all households but only a portion of signed-up participants. As notifications are sent by account ID, we can find up to which accounts notifications were sent, and only include them in the analysis. We do not use actual notifications sent as some households are not opted-in to receive notifications, which would bias our results. Columns 1-3 were pre-specified in the robustness checks; Columns 4-6 were not.

A4 Habit formation

Figure A1: Impact on Consumption Outside Event Days



Notes: This figure plots the estimated intention-to-treat (ITT) effect of the invitation on peak-window electricity consumption on non-event days, estimated separately for each calendar week of the trial as pre-specified non-event-day analysis. The non-event sample comprises half-hourly settlement periods between 16:30 and 18:30 (the window in which events were called) on Wednesdays through Sundays during the trial period, excluding every event delivery day together with a one-day buffer on either side of each event, so that anticipation before an event or rebound afterwards does not contaminate the non-event comparison. For each week, consumption is regressed on the treatment indicator with fixed effects for temperature, prepayment status, and pre-trial consumption decile, and standard errors are clustered at the household level. Points show the weekly ITT point estimate (kW) and the shaded band the corresponding 95% confidence interval.

A5 Sub-trials

In addition to the main trial, we ran several sub-trials. On 21 November 2023, 8,964 commercial customers were randomized: 7,136 customers (80%) received an invitation to join the demand flexibility program, while 1,828 customers (20%) received no communication. On 17 January 2024, among the 750,000 treatment group households who were subscribed to marketing emails, we randomly assigned 680,000 (approximately 90%) to receive a follow-up "chaser" email if they had not signed up initially, while the remaining 70,000 (approximately 10%) did not receive a reminder (the randomization was agnostic

to whether they had signed up already). On 8 February 2024, we introduced an incentive variation among 591,923 households who were signed up and likely to receive opt-in notices by email: 5,000 were randomly assigned to receive a £2.25/kWh incentive, compared to £1.75/kWh for the remaining 586,923 households. Lastly, on 27 February 2024, 40,000 households from the original control group were selected to receive a late invitation to join the demand response program (if they had not done so already), while the rest remained uncontacted (this randomization was, again, agnostic to whether they had signed up already).

A5.1 Chaser email impact on program sign-up

We implemented a "chaser" email sub-trial mid-way through the experiment: a reminder sent to treatment-group households (eligible for the original invitation on 24 October 2023) who had not yet signed up by mid-January 2024. On 17 January 2024, we randomly assigned 90% of households opted in to marketing communications to receive this chaser email, with the remaining 10% serving as a hold-out. We focused the outcome analysis on the final 6 events (after the chaser was sent), among households originally eligible to receive it.

The chaser had a notable impact on sign-ups: the chaser group's sign-up rate was about 4 percentage points higher than the hold-out group's by the end of the trial, and its opt-in rate about 1 percentage point higher (Table A8, Columns 7 and 9). Modest, but a single reminder clearly converted some hesitant households.

We then repeated the ITT/2SLS analysis from Equations (3) to (5), replacing $Invited_i$ with $Chaser_i$. For context, households outside the chaser sample (not opted in to marketing emails) showed similar ITT and LATE estimates over these last 6 events as the full-sample analysis (Table A8, Columns 1, 3 and 5). Within the chaser sub-trial itself, the chaser email produced no statistically significant reduction in consumption (Columns 6, 8 and 10): the ITT point estimate was a small additional 0.003 kWh (0.5% of consumption), indistinguishable from zero. This is plausibly a power issue – the control group was small and the first stage modest (a 4 percentage point sign-up increase) – since compliers induced by the chaser showed effects of similar magnitude to compliers from the original invitation (Columns 8 and 10 vs. 3 and 5), albeit likewise not significant.

In summary, the chaser was somewhat effective at boosting sign-ups, but the small number of newly recruited participants limited our power to detect consumption impacts.

Table A8: Household Consumption Response to Events (Chaser)

Dependent Variables:	kW	Signed-Up	kW	Opted-In	kW		Signed-Up	kW	Opted-In	kW
Sample	One email vs Control					Chased vs Not Chased				
IV stages		First	Second	First	Second		First	Second	First	Second
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Variables										
Invited	-0.0256***	0.3892***		0.1273***						
	(0.0017)	(0.0015)		(0.0009)						
Signed-Up			-0.0657***					-0.0977		
			(0.0044)					(0.0692)		
Opted-In					-0.2008***					-0.3096
					(0.0135)					(0.2191)
Chased						-0.0037	0.0378***		0.0119***	
						(0.0026)	(0.0007)		(0.0002)	
Fixed-effects										
Temperature	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Is Credit	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics										
Observations	14,678,411	14,678,411	14,678,411	14,678,411	14,678,411	6,307,709	6,307,709	6,307,709	6,307,709	6,307,709
Number of Households	1,640,803	1,640,803	1,640,803	1,640,803	1,640,803	704,775	704,775	704,775	704,775	704,775
R^2	0.20135	0.04216	0.20135	0.01487	0.20135	0.22506	0.00588	0.22506	0.00208	0.22506
Wald (1st stage)		69,993.4		21,257.9			2,728.1		2,484.0	
Wald (1st stage), Signed-Up			69,993.4					2,728.1		
Wald (1st stage), Opted-In					21,257.9					2,484.0
Control mean	0.767	0.309	0.767	0.124	0.767	0.811	0.027	0.811	0.005	0.811

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Columns 1 to 5 compare households that received one invitation email to control: Column 1 reports the ITT effect on consumption, Columns 2 and 3 report the first stage and LATE for sign-up compliers, and Columns 4 and 5 report the first stage and LATE for opt-in compliers. Columns 6 to 10 make the same comparison for chased versus not-chased households (all of whom received the marketing email). kW denotes kWh per hour during the event window, and the dependent variables in the first-stage columns are participation indicators, so those coefficients are percentage-point changes. Pre-specified chaser sub-trial analysis.

A5.2 Did late joiners have a different demand response during events?

In a second sub-trial, we examined the effect of encouragement timing by introducing a late invitation for a subset of households who had initially been in the control group. On 27 February 2024 (after 10 of the 13 events had already occurred), a random portion of the original control group – the late invitation group – was encouraged to join the demand flexibility program. This let us test whether being invited so late, with only the final 3 events left, led to different outcomes than being invited early: essentially, a test for fatigue (early-invited households may have tired after many events, or instead grown more efficient) versus novelty (late-invited households coming to the program "fresh").

The late invitation had a large impact on sign-up: the late-invited group's sign-up rate rose 9 percentage points above the never-invited group, though it stayed 9 points below the early-invited group. Opt-in followed the same pattern – up 2 points on never-invited, down 2 points on early-invited – so once signed up, late-invited households opted in at about the same rate as everyone else.

We modeled the impact on consumption with an OLS regression analogous to Equation (3), replacing $Invited_i$ with a late-invitation indicator, restricting to the last three events (after the late invitations were sent) to compare three groups against the main trial: early-invited vs. never-invited, late-invited vs. never-invited, and late-invited vs. early-invited.

Table A9 Columns 1 to 3 re-estimate the main trial's early-vs-never comparison over just the final three events: the ITT and LATEs are about half the main trial's, significant at the 10% level. Rather than fatigue, we attribute this to milder weather during these later events, when demand reduction is generally smaller (Section 3.3.2).

Columns 4 to 9 report the late-invitation comparisons. Magnitudes are similar to the early-vs-never comparison, though not statistically significant, which we read as suggestive evidence against strong fatigue effects: LATEs on sign-up and opt-in are of similar magnitude whether compliers were induced by early or late invitation.[21]

[21] Recall that sign-up and opt-in rates were higher among the early invited. As a result, the LATEs for these participation outcomes identify marginal compliers who participated due to early invitation but would not have done so if invited later.

Table A9: Household Consumption Response to Events (Late Invitation)

Dependent Variable:	kW								
Sample	Early vs Never Invited			Late vs Never Invited			Late vs Early		
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Variables									
Invited	-0.0041*			-0.0018			0.0023		
	(0.0023)			(0.0040)			(0.0033)		
Signed-Up		-0.0229*			-0.0195			-0.0272	
		(0.0129)			(0.0431)			(0.0381)	
Opted-In			-0.1057*			-0.0896			-0.1261
			(0.0591)			(0.1970)			(0.1760)
Group Average									
Control mean	0.720	0.720	0.720	0.720	0.720	0.720	0.718	0.718	0.718
Fixed-effects									
Temperature	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Is Credit	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	6,883,997	6,883,997	6,883,997	311,719	311,719	311,719	6,780,136	6,780,136	6,780,136
Number of Households	2,302,342	2,302,342	2,302,342	104,261	104,261	104,261	2,267,599	2,267,599	2,267,599
R^2	0.19118	0.19118	0.19118	0.19362	0.19362	0.19362	0.19090	0.19090	0.19090
Wald (1st stage), Signed-Up		9,653.0			841.26			1,058.7	
Wald (1st stage), Opted-In			1,564.9			135.47			165.01

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Columns 1 to 3 re-estimate the main trial using only the final three events and compare early-invited households to never-invited households. Columns 4 to 6 compare late-invited households to never-invited households. Columns 7 to 9 compare late-invited households to households invited at the start of the season. kW denotes kWh per hour during the event window. Pre-specified late-invitation sub-trial analysis.

Table A10: Household Sign-Up and Opt-In Response to Late Invitation (First Stage)

Dependent Variables:	Signed-Up	Opted-In	Signed-Up	Opted-In	Signed-Up	Opted-In
Sample	Early vs Never Invited		Late vs Early		Late vs Never Invited	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Invited	0.1777***	0.0385***	-0.0859***	-0.0185***	0.0918***	0.0200***
	(0.0018)	(0.0010)	(0.0026)	(0.0014)	(0.0032)	(0.0017)
Fixed-effects						
Temperature	Yes	Yes	Yes	Yes	Yes	Yes
Is Credit	Yes	Yes	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	6,883,997	6,883,997	6,780,136	6,780,136	311,719	311,719
Number of Households	2,302,342	2,302,342	2,267,599	2,267,599	104,261	104,261
R^2	0.00452	0.00865	0.00128	0.00842	0.00971	0.00665
Wald (1st stage)	9,653.0	1,564.9	1,058.7	165.01	841.26	135.47
Control mean	0.332	0.103	0.509	0.142	0.332	0.103

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Columns 1 and 2 re-estimate the early-versus-never comparison using only the final three events. Columns 3 and 4 compare late-invited households to households invited at the start of the season. Columns 5 and 6 compare late-invited households to never-invited households. The dependent variables are sign-up and opt-in indicators, so coefficients are percentage-point changes. Pre-specified late-invitation sub-trial analysis.

A5.3 Demand response from commercial customers

In this sub-trial, we applied the same randomized encouragement design used for domestic customers to a sample of small commercial customers. On 21 November 2023, we randomly assigned 8,964 eligible business accounts to receive an invitation to join the demand flexibility program. Of these, 7,136 customers (80%) were assigned to the treatment group and received the email, while the remaining 1,828 customers (20%) formed the control group and received no communication. The invitation raised commercial customers' sign-up rate by 10 percentage points relative to the control group (Table A11), confirming meaningful variation in program participation. We estimated treatment effects using the same ITT/2SLS framework as the main analysis (Section 3.1.1): Equation (3) for the ITT effect of the invitation, Equation (4) for the first stage (effect of treatment on sign-up or opt-in), and Equation (5) for the resulting LATEs among compliers.

Results are not statistically significant, but the point estimates suggest potentially large effects (Table A11). The ITT estimate implies invited commercial customers reduced consumption by 0.0763 kWh during events (5.4% relative to the control group's baseline), while the LATEs are much larger in magnitude: 0.6812 kWh (48.3%) for sign-up compliers and 2.405 kWh (170.4%) for opt-in compliers. This last figure exceeding 100% is implausible and likely reflects estimate noise rather than a real effect, especially given only 3% of this sample has an export meter.

In summary, the encouragement induced a non-trivial number of commercial customers to sign up, and some of them did participate in events and appear to reduce their consumption. This confirms the feasibility of recruiting commercial customers into an event-based distributed demand response program. However, due to insufficient statistical power and potentially small effect sizes, we are unable to make strong claims about the average or marginal treatment effects in this subpopulation.

Table A11: Firm Consumption Response to Events

Dependent Variables:	kW	Signed-Up	kW	Opted-In	kW
IV stages		First	Second	First	Second
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Invited	-0.0763	0.1120***		0.0317***	
	(0.0550)	(0.0063)		(0.0026)	
Signed-Up			-0.6812		
			(0.4926)		
Opted-In					-2.405
					(1.740)
Group Average					
Control mean	1.411	0.055	1.411	0.014	1.411
Fixed-effects					
Temperature	Yes	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes	Yes
GSP Group	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	197,689	197,689	197,689	197,689	197,689
Number of Households	8,726	8,726	8,726	8,726	8,726
R^2	0.15270	0.02228	0.15270	0.00938	0.15270
Wald (1st stage)		316.33		152.87	
Wald (1st stage), Signed-Up			316.33		
Wald (1st stage), Opted-In					152.87

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
The sample consists of small commercial customers. Column 1 reports the ITT effect of invitation. Columns 2 and 3 report the first stage and LATE for sign-up compliers; Columns 4 and 5 report the first stage and LATE for opt-in compliers. kW denotes kWh per hour during the event window, and the dependent variables in the first-stage columns are participation indicators, so those coefficients are percentage-point changes. Temperature: hourly average temperature in Celsius in the participant's GSP region. Pre-Trial Consumption: indicator for whether pre-treatment peak consumption was above or below the sample median. Pre-specified business-customer sub-trial analysis.

A5.4 Impact of higher incentive on demand response

Finally, in a separately pre-registered sub-trial, we tested how a higher financial incentive affected participation and performance during a single event, among the 591,923 (of 1,318,441) signed-up domestic participants without push notifications enabled as of 8 February 2024 (restricted since our treatment depended on customers reading the email notice). Within this group, we randomly selected 5,000 households to receive £2.25 per kWh saved for an upcoming event, rather than the standard £1.75/kWh, and estimated the ITT effect on opt-in and consumption, then the LATE of opt-in among compliers using

a two-stage approach.

Households offered £2.25/kWh were significantly more likely to opt in (Table A12): about 8 percentage points higher than under the standard incentive (36.7% vs 27.9%, a 31.5% increase for a 28.6% increase in incentive, or an "opt-in elasticity" of approximately 1.10) – striking given the treatment's subtlety, communicated identically apart from the reward amount.

Despite this, we found no significant difference in demand, by ITT or LATE, possibly suggesting the extra opt-ins were opportunistic rather than genuine commitments. But we caution against over-interpreting this null: with only 5,000 treated households, the minimum detectable LATE (0.42 kW) is more than double the main trial's estimated LATE (0.18 kW, Figure 4), so a null result here is unsurprising regardless of the true effect.

Table A12: Household Consumption Response to Extra Incentive

Dependent Variables:	kW	Opted-In	kW
IV stages		First	Second
Model:	(1)	(2)	(3)
Variables			
Extra Incentive	0.0024	0.0871***	
	(0.0131)	(0.0070)	
Opted-In			0.0270
			(0.1503)
Group Average			
Control mean	0.784	0.279	0.784
Fixed-effects			
Half Hour	Yes	Yes	Yes
Fit statistics			
Observations	1,129,410	1,129,410	1,129,410
Number of Households	564,708	564,708	564,708
R^2	0.00010	0.00032	0.00010
Wald (1st stage)		155.16	
Wald (1st stage), Opted-In			155.16

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
This sub-trial uses a single event and only households without push notifications enabled. Column 1 reports the ITT effect of offering a higher reward on event-period consumption. Columns 2 and 3 report the first stage and the corresponding LATE for opt-in compliers. kW denotes kWh per hour during the event window. Pre-specified higher-incentive sub-trial.

A6 Adoption of new time-of-use tariffs

Lastly, in a pre-specified analysis, we assessed whether participation led to longer-term changes in tariff or utility switching – specifically, whether invited households were more likely to move onto a time-of-use (ToU) tariff (one that varies by hour of day, rather than flat) or switch to a new energy utility. We compared each household's tariff and utility status on the day the trial began to their status on the day it ended, using a two-period panel regression with period fixed effects, with the treatment indicator switched on only for treated households in the second period.

Table A13 shows no evidence that treatment group households were more likely to adopt a ToU tariff or switch utility, by ITT or by LATE among sign-up compliers – counter to our expectation that program experience would increase ToU interest and reduce switching, but still informative: it helps rule out these channels as drivers of sustained behavioral change, reinforcing our interpretation of the program as a time-specific engagement tool rather than a driver of long-run demand changes.

Table A13: Pre-Post Analysis on Adopting ToU Tariff and Switching Provider

Dependent Variables:	Is ToU		Switch Energy Provider	
Model:	(1)	(2)	(3)	(4)
Variables				
Invited	6.64×10^{-6}		3.66×10^{-7}	
	(0.0009)		(4.96×10^{-5})	
Signed-Up		3.64×10^{-5}		2.04×10^{-5}
		(0.0047)		(0.0001)
Group Average				
Control mean	0.088	0.088	0.000	0.000
Fixed-effects				
Period	Yes	Yes	Yes	Yes
Fit statistics				
Observations	4,805,807	4,805,807	4,807,288	4,806,362
Number of Households	2,403,181	2,403,181	2,403,644	2,403,181
R^2	0.00046	0.00046	4.33×10^{-5}	0.00012
Wald (1st stage), Signed-Up		15,400.6		15,399.3

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

A7 Heterogeneity analysis tables and figures

The pre-specified heterogeneity analysis list is: EPC, IMD, smart tariff, previous program participation (2022-23), Octopus-before-Bulb, Bulb-acquired, EAC $\geq$ 2900 kWh, prepayment, region, domestic/non-domestic.

Table A14 reports the estimated percent-change effects (ITT and LATE on sign-up and opt-in) for every characteristic in Table 2, by subgroup level, grouped by theme.

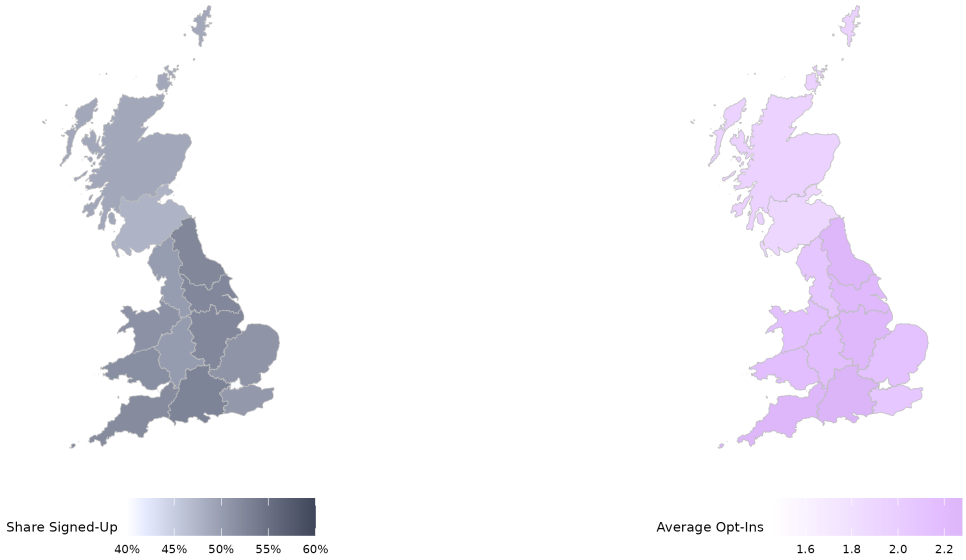
Table A14: Summary of treatment-effect heterogeneity (percent change in consumption)

Level	ITT	LATE (Signed-Up)	LATE (Opted-In)
Home size and consumption			
EAC			
Up to 1,980 kWh	-1.3%**	-7.0%**	-21.8%**
Up to 2,786 kWh	-1.4%***	-6.6%***	-22.0%***
Up to 3,747 kWh	-1.2%***	-6.0%***	-20.1%***
Up to 5,471 kWh	-0.9%**	-4.3%**	-15.3%**
Above 5,471 kWh	-1.7%***	-8.3%***	-30.0%***
Property Value			
Up to £197k	-1.2%**	-7.0%**	-25.2%**
Up to £271k	-1.4%***	-7.0%***	-23.9%***
Up to £356k	-0.7%	-3.6%	-11.8%
Up to £494k	-1.8%***	-8.3%***	-26.9%***
Above £494k	-1.8%***	-8.3%***	-28.5%***
Floor Area (sq m)			
9-76	-1.5%**	-7.0%***	-22.6%***
76-98	-1.6%***	-7.8%***	-26.0%***
98-126	-1.8%***	-8.7%***	-30.5%***
126-180	-0.9%	-4.2%	-14.3%
180+	-1.6%**	-7.9%**	-26.1%**
Prior program experience			
Returning Participant			
No	-1.1%***	-5.8%***	-22.8%***
Yes	-2.3%***	-8.8%***	-24.1%***
Housing and geography			
EPC			
A	3.6%	22.2%	52.1%
B	-0.8%	-4.6%	-15.3%
C	-1.1%**	-6.0%**	-20.2%**
D	-1.3%***	-6.7%***	-23.5%***
E	-1.1%	-5.6%	-20.0%
F	-0.5%	-2.4%	-8.4%
G	-4.0%	-18.3%	-51.4%
Dwelling type			
Flat	-1.2%*	-7.3%*	-24.2%*
House	-1.4%***	-6.9%***	-23.4%***
Occupancy Tenure (Own/Rent)			
Owner Occupied	-1.3%***	-6.2%***	-20.7%***
Private Rented	-1.7%**	-10.5%**	-36.6%**
Social Housing	-1.4%*	-8.2%*	-34.2%*

Level	ITT	LATE (Signed-Up)	LATE (Opted-In)
Housing and geography			
Urban			
No	-1.3%**	-5.7%**	-19.5%**
Yes	-1.4%***	-7.1%***	-24.2%***
IMD Score			
1-7	-1.8%***	-8.1%***	-26.3%***
7-12	-0.7%	-3.4%	-11.2%
12-17	-2.4%***	-11.2%***	-36.5%***
17-27	-0.9%*	-4.7%*	-17.0%*
27+	-1.3%***	-7.8%***	-29.1%***
Total annual income (£)			
GBP 22k-GBP 38k	-0.7%	-3.5%	-12.1%
GBP 38k-GBP 44k	-2.6%***	-12.4%***	-41.1%***
GBP 44k-GBP 49k	-1.1%*	-5.2%*	-17.5%*
GBP 49k-GBP 56k	-1.5%**	-6.8%**	-22.4%**
GBP 56k+	-1.3%**	-6.0%**	-20.0%**
Main Fuel = Electricity			
No	-1.4%***	-6.9%***	-23.5%***
Yes	-1.3%	-7.2%	-22.8%
Supplier and tariff			
Is ToU			
No	-1.3%***	-6.7%***	-23.1%***
Yes	-1.6%	-8.1%	-23.8%
Economy 7			
No	-1.3%***	-6.3%***	-21.7%***
Yes	-2.4%***	-10.9%***	-36.1%***
Credit Customers			
No	-1.4%***	-6.8%***	-23.0%***
Yes	-1.1%	-7.6%	-38.3%
Bulb Customers			
No	-1.4%***	-6.8%***	-21.8%***
Yes	-1.3%***	-7.0%***	-30.3%***
Intelligent Octopus Go			
Exc. IO Go	-1.4%***	-6.9%***	-23.3%***
IO Go	-1.0%	-5.5%	-19.3%
Tenure with Energy Provider			
0 to <1 yrs	-0.8%	-4.6%	-19.2%
1 to <2 yrs	-1.9%***	-9.7%***	-29.6%***
2 to <3 yrs	-1.7%***	-8.2%***	-25.4%***
3 to <4 yrs	-0.9%*	-4.3%*	-14.4%*

Notes: Cells report the estimated percent change in event-period consumption for the invitation (ITT) or for compliers (LATE), by subgroup level, from the same specifications as the individual heterogeneity regressions. Significance stars are from clustered standard errors (Signif. Codes: ***: 0.01, **: 0.05, *: 0.1).

Figure A2: Regional Heterogeneity in Sign-Up and Opt-In



(**a**) Share Signed-Up by Region

(**b**) Average Opt-Ins by Region

Notes: Panel (a) shows the share of invited households who signed up in each region, and Panel (b) shows the average number of event opt-ins among those who signed up. Some regions exhibited higher sign-up and opt-in rates than others, reflecting variation in household engagement. These differences in participation, however, did not correspond to meaningful differences in per-event load reductions.

A8 Low-carbon technology adoption analysis

This exploratory analysis focuses on the $17{,}085$ households that signed up to the program and subsequently installed a low-carbon technology after the first event, excluding households with a pre-existing technology (since there are no pre-adoption observations to test parallel trends).

We use a two-way fixed effects difference-in-differences specification exploiting within-household variation in installation timing, interacting event-time indicators (event time -1 omitted) with both installation status and opt-in, with household, event-day, and temperature fixed effects and household-clustered standard errors. Pre-adoption coefficients are statistically indistinguishable from zero for all three technologies, supporting the parallel-trends assumption (Figure A3). Post-adoption direct effects are clearly identified: heat pump installation sharply raises event-period consumption before attenuating, EV installation has no detectable direct effect, and solar gradually lowers net consumption.

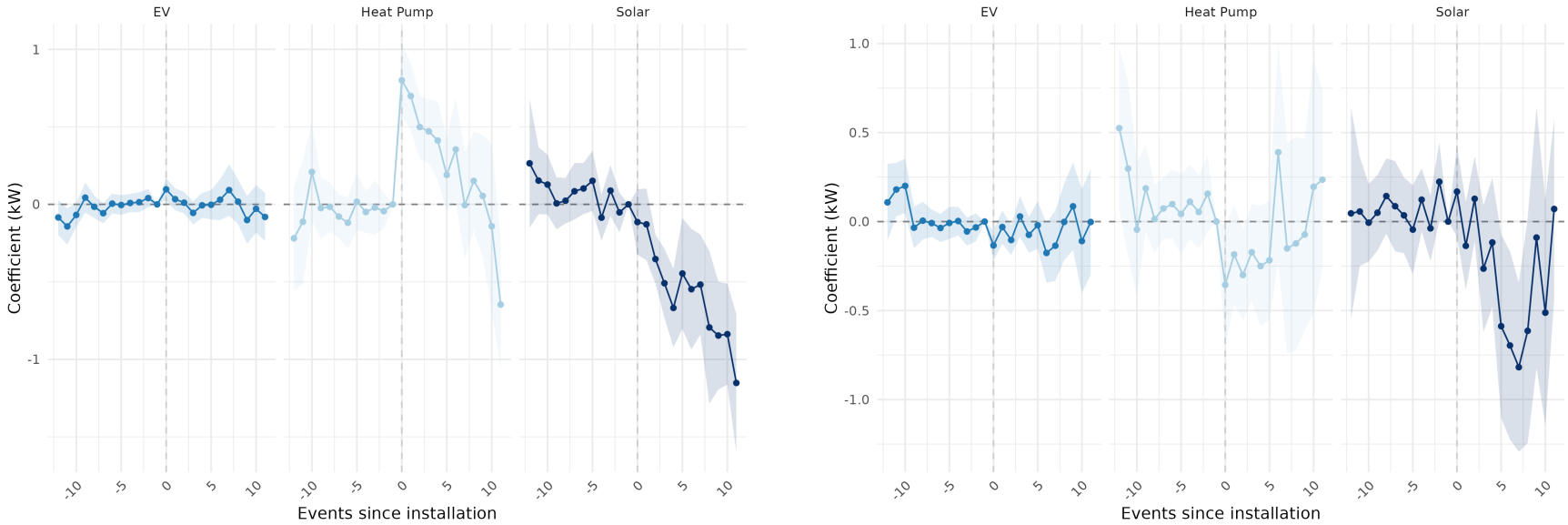
The dynamic opt-in interaction estimates – our coefficient of primary interest – are noisier, since identifying them event by event requires opt-in variation within each post-adoption bin, and few adopters opt in to any single event. The pooled specification in the main text, averaging across event-time bins, has substantially more power and yields the precisely estimated effects in Figure 6.

TWFE can be biased under staggered adoption when treatment effects are heterogeneous across cohorts or event time (Callaway and Sant'Anna, 2021, de Chaisemartin and D'Haultfœuille, 2020, Goodman-Bacon, 2021). Here, though, the demand response to LCT adoption is largely mechanical (a smart charger pauses, a heat pump switches off, a battery discharges), so the effect of installation should be roughly constant in event time – under which TWFE recovers the same estimand as cohort-robust estimators like Callaway and Sant'Anna (2021).

Table A15: Balance Table for Early vs Late Adopters

	N		Means		P-values
Variable	Late	Early	Late	Early	Early vs Late
EV					
EAC	7716	2088	4070.34	4229.12	0.01
Property Value	7523	2042	423231.95	445849.71	0.00
EPC	5637	1618	67.20	67.32	0.75
Baseline	6595	1804	0.84	0.87	0.17
Consumption	6551	1792	0.82	0.82	0.75
Reduction	6548	1791	0.03	0.04	0.43
Heatpump					
EAC	3529	321	3670.16	3592.19	0.58
Property Value	3582	327	376350.77	416045.70	0.00
EPC	3375	324	75.72	74.19	0.03
Baseline	3253	305	0.57	0.58	0.92
Consumption	3244	304	0.48	0.42	0.28
Reduction	3237	304	0.09	0.16	0.20
Solar					
EAC	1679	272	4357.06	4393.68	0.85
Property Value	1705	273	429873.60	448664.09	0.22
EPC	1150	168	65.56	66.08	0.59
Baseline	1463	230	0.89	0.91	0.69
Consumption	1455	231	0.81	0.86	0.41
Reduction	1455	230	0.08	0.05	0.55

Figure A3: Dynamic impact of low-carbon technology adoption



(**a**) Direct effect of installation

(**b**) Additional effect when opting in

Notes: This figure shows event-study estimates from a two-way fixed effects specification, estimated separately for EV chargers, heat pumps, and solar PV with battery storage. Panel (a) plots the $\mathrm{EventTime}_{it}$ main effect: the change in event-period consumption following installation with no opt-in. Panel (b) plots the $\mathrm{OptedIn}_{it} \times \mathrm{EventTime}_{it}$ interaction: the additional consumption response when opting in at each event after installation. Coefficients are measured relative to the reference period at event time -1, the last event prior to the technology installation. Negative event-time values denote events before installation; positive values denote events afterward. Household, event-day, and temperature fixed effects are included, with standard errors clustered at the household level. Shaded bands are 95% confidence intervals. This analysis was not pre-specified.

Table A16: Impact of Low-carbon Technology Adoption on Consumption During Events via Opt-In

Dependent Variable:		Actual kW	
LCT	EV	Heat Pump	Solar
Model:	(1)	(2)	(3)
Variables			
Installed	0.0064	0.6788***	-0.3805***
	(0.0191)	(0.0610)	(0.0560)
Opted-In	-0.1530***	-0.2309***	-0.2160***
	(0.0074)	(0.0140)	(0.0154)
Installed × Opted-In	-0.0567**	-0.2918***	-0.1966**
	(0.0244)	(0.0799)	(0.0851)
Fixed-effects			
Day	Yes	Yes	Yes
Household	Yes	Yes	Yes
Temperature	Yes	Yes	Yes
Fit statistics			
Observations	112,757	48,510	22,966
Number of Households	8,993	3,929	1,859
R^2	0.51311	0.60297	0.47146

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Outcome is realized event-period consumption in kWh. 'Installed × Opted-In' captures the additional event response of households after installing the relevant low-carbon technology, relative to otherwise similar households that have not yet installed. Temperature: average temperature in Celsius in the participant's GSP region during the Saving Sessions. Is Installed: binary indicator for whether or not the LCT is installed.

A9 Relationship between temperature and remuneration

In this section, we explore the relationship between temperature, remuneration and flexibility.

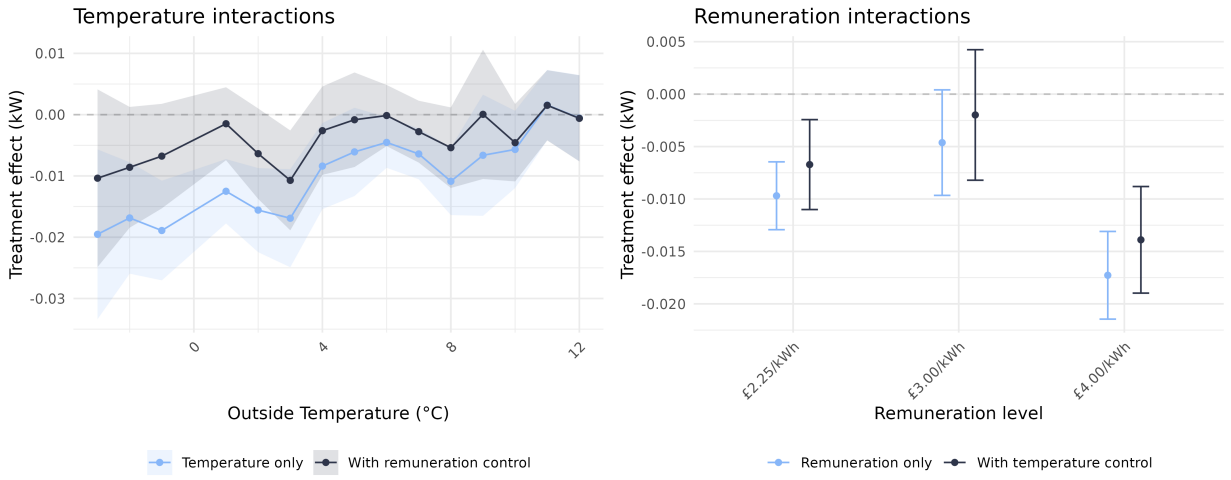
Earlier events tended to coincide with lower outdoor temperatures and higher remuneration per kWh, whereas later events occurred on warmer days and offered lower remuneration. This produces a strong negative correlation between temperature and remuneration per kWh, which raises the possibility that the two are confounded in their influence on participation.

To disentangle them, we run a "horse race" in Figure A4, estimating three ITT specifications that interact treatment with remuneration and with binned outdoor temperature. Temperature is binned in 1°C intervals (finer disaggregation was computationally infea-

sible on our dataset), with 0°C as the omitted category; the omitted remuneration level is £1.75/kWh. All models control for temperature, day fixed effects, pre-payment status, and pre-trial average consumption. Column (2) interacts treatment with temperature only, column (3) interacts treatment with remuneration only, and Column (1) includes both sets of interactions to assess their roles jointly.

The horse race indicates that the apparent temperature gradient is largely driven by remuneration. In the temperature-only model (Column 2), the temperature interactions are sizeable and highly significant across most bins. Once remuneration is added (Column 1), however, these coefficients are sharply attenuated and become statistically insignificant for the large majority of bins. By contrast, the remuneration interactions remain economically and statistically significant in the combined model (Column 1), changing only modestly relative to the remuneration-only specification (Column 3). Taken together, this suggests that the earlier negative association between temperature and treatment response is mostly an artifact of the temperature–remuneration confound: because later, warmer events also paid less, temperature proxies for remuneration. After accounting for remuneration, temperature retains little independent explanatory power, and remuneration emerges as the dominant driver of heterogeneity in response.

Figure A4: Horse race: Temperature vs Remuneration



A10 Heterogeneous effect by compliance type

The LATE estimates in the main analysis identify the consumption response for compliers – households that signed up, or opted in, as a direct consequence of the invitation. A natural question is whether this population responds differently from always-takers, the households that enroll without any encouragement and are observed directly in the control arm: if the invitation draws in households with weaker underlying flexibility, the complier response will understate what a more motivated population could achieve, whereas if the two groups respond similarly, the LATE generalizes more readily. This comparison parallels Fowlie et al. (2021), who study a residential time-varying pricing experiment in which one group was invited to opt in and another was defaulted into the program with the option to opt out.

Table A17 presents the comparison across both participation margins. For the sign-up margin, compliers reduce consumption by 0.052 kW, against 0.102 kW for always-takers; for the opt-in margin, compliers reduce by 0.177 kW, against 0.233 kW for always-takers. In both cases always-takers turn down more, consistent with the Fowlie et al. (2021) finding that more motivated participants respond more strongly. The gap is proportionally larger on the sign-up margin, which is intuitive: opting in is itself an active choice, so opt-in compliers are already a relatively engaged subset, and the always-taker premium is correspondingly smaller.

Building on this contrast, we construct a population-level estimate of the demand response by reweighting the complier and always-taker effects according to their shares in a fully invited population. Letting p_0 and p_1 denote the participation rates in the control and treatment arms (Table A4), the always-taker share among participants is p_0/p_1 and the complier share is $1-p_0/p_1$. On the sign-up margin, $p_0 = 0.261$ and $p_1 = 0.463$, giving shares of 0.56 and 0.44; on the opt-in margin, $p_0 = 0.116$ and $p_1 = 0.174$, giving shares of 0.67 and 0.33. The resulting participation-weighted average reduction is $0.56 \times 0.102 + 0.44 \times 0.052 = 0.080$ kW per signed-up household and $0.67 \times 0.233 + 0.33 \times 0.177 = 0.215$ kW per opted-in household-event – the average turn-down per participating household, accounting for the invited population's mix of always-takers and marginal compliers.

Table A17: Household Consumption Response to Events (Compliers vs Always Takers)

Dependent Variable:	kW				
	ITT	Sign-Up		Opt-In	
	Encouraged	Compliers	Always Takers	Compliers	Always Takers
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Invited	-0.0105***				
	(0.0017)				
Signed-Up		-0.0520***	-0.1021***		
		(0.0085)	(0.0039)		
Opted-In				-0.1773***	-0.2328***
				(0.0287)	(0.0054)
Fixed-effects					
Temperature	Yes	Yes	Yes	Yes	Yes
Is Credit	Yes	Yes	Yes	Yes	Yes
Pre-Trial Consumption	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	57,535,536	57,535,536	1,737,166	57,535,536	1,737,166
Number of Households	2,320,256	2,320,256	70,055	2,320,256	70,055
R^2	0.22827	0.22827	0.23759	0.22827	0.24111
Wald (1st stage), Signed-Up		15,894.1			
Wald (1st stage), Opted-In				4,005.4	

Clustered (Household) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
This analysis is exploratory.

A11 Comparison of field experimental coefficients with baseline estimates

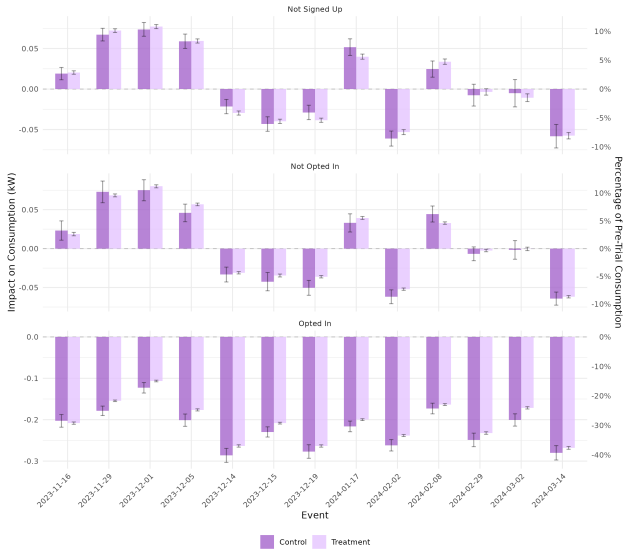
In this section, we compare the event-by-event LATE estimates from the main analysis to per-event reductions computed under the P376 baselining methodology. We begin by computing baselined reductions, event by event, in three groups and both arms: customers not yet signed up, customers signed up but not opted into the event, and customers who have opted in. The not-yet-signed-up group has no incentive to reduce demand and should, by construction, show no turn-down.

Figure A5, row 1, shows the baselined "reduction" for this group averages just 0.005 kW across the 13 events – about 2.5% of the opted-in baselined reduction (0.2 kW), and negligible at the seasonal level. Event by event, the picture is noisier: a slight increase in early events gives way to a decrease in later ones, more consistent with a systematic baselining

bias – likely driven by warmer-than-usual weather before early events and colder-than-usual weather before later ones – than with a behavioral response. Row 2 repeats this for signed-up-but-not-opted-in customers and shows the same pattern, reassuring us that neither group differs systematically from the other (e.g. via unusually high event consumption), and confirming that the baseline under-estimates demand early in the trial and over-estimates it later, with control and treatment moving in tandem within each group.

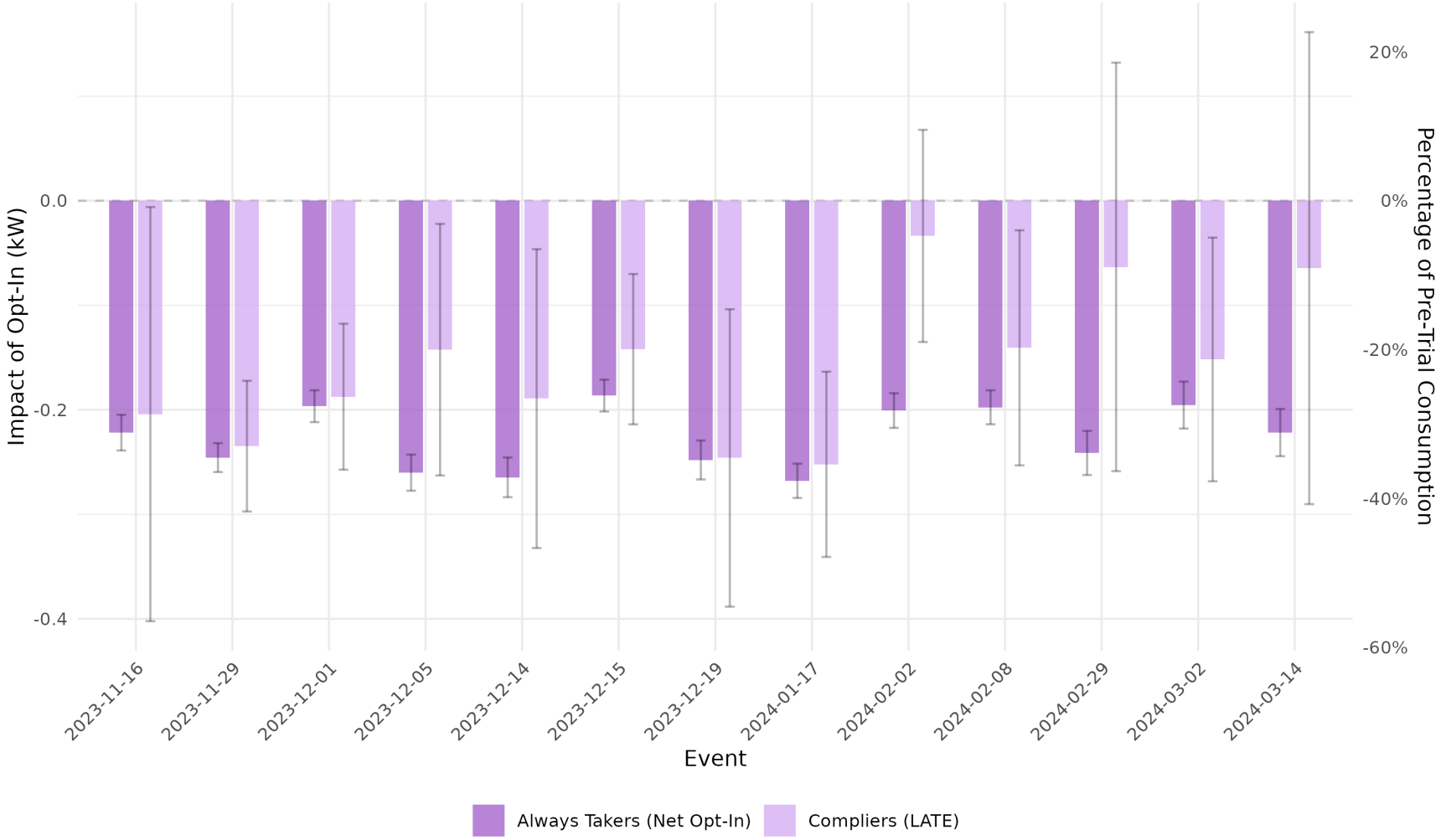
Motivated by these findings, we construct a net baselined reduction for opted-in customers by subtracting, for each event, the bias estimated from the not-yet-signed-up group. Figure A6 compares this net baselined reduction for always-takers in the control arm to the LATE estimate for treatment compliers: the two lie within each other's 95% confidence intervals for almost all events (the exception is 2 February, where the LATE is markedly smaller). Across the 13 events, always-takers turn down 0.22 kW on average versus 0.18 kW for compliers – about 22% larger, though with large confidence intervals. Reassuringly, the season-level baselined reduction for always-takers (0.22 kW) closely matches the OLS estimate for the opt-in always-taker group in Table A17 (0.23 kW), confirming that the P376 baselining methodology and the regression-based approach recover essentially the same demand response despite resting on different identifying assumptions.

Figure A5: Comparison of baselined results by enrollment status



Notes: Each panel plots the P376-baselined demand reduction, event by event, in both trial arms. Row 1: customers not yet signed up. Row 2: signed-up customers who did not opt in to the event. Row 3: opted-in customers. The near-zero reduction for the first two groups, with opposite signs early and late in the season, indicates a weather-driven baselining bias rather than a behavioral response. This analysis is exploratory.

Figure A6: Impact of Opt-in by Session



Notes: This figure compares, event by event, the net baselined reduction for always-takers in the control arm (P376 reduction with the not-yet-signed-up bias netted out) against the LATE on opt-in for treatment compliers. 95% confidence intervals shown. The impact of opt-in by event was pre-specified but the comparison with baselined reduction is exploratory.

A12 Welfare analysis methodology

This appendix provides full details of the welfare framework in Section 4.

A12.1 Demand reduction estimates for welfare analysis

We calculated a separate LATE for each half-hour during events – the aggregate reduction in consumption per settlement period caused by opt-in – and multiplied each half-hour-specific LATE (in kWh per half-hour per household) by the cumulative number of households opted into the event to estimate total demand reduction in that event.

This approach assumes our LATE of opt-in captures the impact on the average opt-in, potentially understating or overstating the always-takers' impact. The industry-standard alternative – a modified pre-post approach that "clips" wrong-direction demand response to zero – yielded similar estimates on average, slightly larger, though not uniformly so across events (Section 3.4; Figures A5 and A6). We used LATE-based estimates throughout as a conservative measure of program benefits.

A12.2 Marginal generator, price, and emissions.

To value the balancing actions displaced by DFS, we assigned each settlement period a marginal generator: the highest-priced accepted Balancing Mechanism action that would have been required absent the demand reduction, restricted to actions managing the overall supply-demand balance (excluding those resolving local network constraints) with an accepted volume of at least 25 MW – avoiding price spikes from actions too small to represent what DFS would realistically displace. Each settlement period was thereby assigned a marginal price, fuel type, and emissions factor. Implied marginal prices ranged from about £58/MWh to £900/MWh, with the highest values concentrated in live November–December events.

A12.3 Benefits

A12.3.1 Payments to consumers and the utility.

While transfers from the system operator to private agents do not create social value mechanically, the MVPF numerator captures the net willingness to pay of beneficiaries. Treat-

ing all participating households as marginal, the standard Harberger triangle approximation with a linear demand curve implies consumer surplus equal to 50% of household payments (Hahn et al., 2026, Harberger, 1964, Hendren and Sprung-Keyser, 2020). We did not treat bill savings from reduced consumption as a separate benefit: by the envelope theorem, participants would not have reduced consumption without compensation, so the payment itself captures the marginal value of foregone consumption. Of system operator payments, 80% were allocated to households and 20% retained by the utility, split evenly between administrative costs and utility surplus:

$$\begin{aligned}\text{Private Benefit}_t = &\underbrace{0.50 \times 0.80 \times \text{System Operator payment}_t}_{\text{Consumer Surplus}} \\ &+ \underbrace{0.50 \times 0.20 \times \text{System Operator payment}_t}_{\text{Utility Surplus}}.\end{aligned} \tag{6}$$

A12.3.2 CO_2-equivalent abatement.

Reducing demand at peak times avoided greenhouse gas emissions by displacing fossil generation at the margin, valued using the marginal fuel type's emissions factor and the social cost of carbon:

$$\text{CO}_2\text{e Abatement Benefit}_t = \text{Demand Reduction}_t^{\text{LATE}} \times \text{Emissions Factor}_t \times \text{Social Cost of Carbon}. \tag{7}$$

Some events displaced zero-emissions generation while others displaced carbon-intensive units (Table A18); the resulting carbon value was economically meaningful but modest relative to transfers and reliability value.

A12.3.3 Reliability and avoided lost load.

One of the most important benefits of demand flexibility is its role in reducing the risk of lost load. In an extreme scarcity scenario, the relevant counterfactual is not the marginal generator but involuntary disconnection: each MWh of realized reduction was instead valued at the GB regulatory Value of Lost Load (VoLL) of £6,000/MWh:

$$\text{Lost Load Avoidance Benefit}_t = \text{Demand Reduction}_t^{\text{LATE}} \times \text{VoLL}. \tag{8}$$

In the VoLL scenario, avoided balancing costs were set to zero, since the system operator would have relied on load curtailment rather than cheaper supply-side actions. The

resulting MVPF was:

$$\text{MVPF}_t^{\text{VoLL}} = \frac{\text{Consumer surplus}_t + \text{Utility surplus}_t + \text{Lost Load Avoidance Benefit}_t}{\text{Administrative cost}_t}. \tag{9}$$

A12.4 Costs

A12.4.1 Costs to the system operator.

The cost to the system operator was the payment made to the flexibility provider for delivered demand reduction, measured directly from settlement records at the half-hourly settlement-period level and aggregated to the event level for reporting. Denoting the DFS remuneration in settlement period t by P_t^{DFS}:

$$\text{Costs to the system operator}_t = \text{Delivered reduction}_t^{\text{P376}} \times P_t^{DFS}. \tag{10}$$

Administrative cost varied substantially across events (Table A18), from roughly £18,100 for the small 2 March 2024 test event to about £118,000 for the large live event on 29 November 2023.

A12.4.2 Avoided balancing costs.

Avoided balancing costs were the cost of the marginal generator that would have been needed without the realized demand reduction: the LATE-estimated reduction multiplied by the marginal price for each event:

$$\text{Net Cost}_t = \text{Administrative Cost}_t - \text{Demand Reduction}_t^{\text{LATE}} \times \text{Marginal Price}_t. \tag{11}$$

For example, on 1 December 2023 the marginal price was about £900/MWh, compared with household remuneration of approximately £4,000/MWh – break-even prices were typically far below actual remuneration, which explains why the base-case MVPF remained below 1 across events.

A12.4.3 Fiscal externalities.

The program's VAT, corporate tax, and utility revenue effects were small relative to the system operator's direct expenditure, so we treated the payment net of avoided balancing

cost as the relevant government cost.

A12.5 Break-even prices

A useful way to interpret the welfare results is to ask what prices would be consistent with an MVPF of 1 under the base case: holding remuneration fixed, what avoided balancing cost per MWh would do so, and, holding balancing costs fixed, what remuneration level would. In our implementation, 80% of the system operator payment was passed through to households (valued at 50%, all consumers assumed marginal), with the remaining 20% accruing to the utility (half retained as profit). At MVPF = 1, the avoided balancing cost must therefore cover the share of administrative cost not already returned as private surplus, net of carbon benefits. Solving for the break-even marginal price in event e gives:

$$P_e^{BE,\text{marg}} = \frac{\text{Admin cost}_e - 0.50 \times \text{Consumer payment}_e - \text{Utility surplus}_e - \text{CO}_2\text{ value}_e}{\text{Delivered reduction}_e^{\text{LATE}}}. \tag{12}$$

The resulting break-even avoided balancing cost (reported in the main text) is well within actual GB market conditions: marginal prices, almost always set by gas generation, exceeded £500/MWh during the 2021–2023 energy crisis and reached nearly £1,400/MWh in January 2025 – so DFS at observed remuneration would have justified itself on routine balancing grounds only during acute system stress, not a typical winter peak. The complementary benchmark – holding observed marginal prices fixed, what remuneration would have been consistent with MVPF = 1 – is reported in the notes to Table A18. Both figures are accounting benchmarks rather than directly implementable recommendations: if participation is price-elastic, the welfare-optimal remuneration would need to be determined jointly with the behavioral response of participants.

Table A18: Event-level costs, benefits and MVPF

	Prices and Incentives			MVPF Components						MVPFs	
Date	Marginal Price (£/MWh)	HH Remuneration (£/MWh)	Break Even Marginal Price (£/MWh)	Total Turn Down Late (MWh)	Total CO_2e (t)	Consumer Surplus (k£)	Surplus to Supplier (k£)	CO_2e Value (k£)	Cost to System Operator (k£)	MVPF	MVPF (VoLL)
2023-11-16	341	3000	1450	12.44	4.4	15.3	3.8	1.1	38.3	0.59	2.45
2023-11-29	429	4000	737	71.46	25.0	47.2	11.8	6.3	118.0	0.75	4.13
2023-12-01	900	4000	637	54.85	19.2	31.8	8.0	4.8	79.6	1.48	4.63
2023-12-05	200	2250	1171	16.05	5.6	16.2	4.0	1.4	40.4	0.58	2.88
2023-12-14	195	2250	1497	17.17	6.0	21.8	5.4	1.5	54.4	0.56	2.39
2023-12-15	125	2250	1618	25.70	5.6	34.4	8.6	1.4	86.0	0.54	2.29
2023-12-19	113	2250	1237	22.23	7.8	23.6	5.9	2.0	58.9	0.56	2.76
2024-01-17	58	2250	839	43.76	15.3	32.5	8.1	3.9	81.2	0.57	3.74
2024-02-02	110	1750	5952	4.12	1.4	19.9	5.0	0.4	49.8	0.51	1.00
2024-02-08	100	1750	1000	17.36	6.1	15.1	3.8	1.5	37.8	0.57	3.26
2024-02-29	97	1750	5937	1.98	0.7	9.6	2.4	0.2	23.9	0.51	1.00
2024-03-02	114	1750	1119	8.07	0.0	7.2	1.8	0.0	18.1	0.53	3.18
Total				295.21	97.1	274.5	68.6	24.5	686.4	0.63	3.08

Notes: Marginal price is the cost of the marginal balancing action during the event, averaged across settlement periods for events longer than half an hour. Remuneration to households is the remuneration per MWh of settled demand reduction – customers were remunerated only if they had opted in to the event and had actual demand lower than their personalized baseline demand. Total demand reduction is from our randomized encouragement, among compliers in the treatment group: it is the LATE on opt-in multiplied by the number of customers who opted in to a given event. Total CO_2e reduction is the total demand reduction multiplied by the estimated carbon intensity per MWh of the generator providing the marginal balancing action. Consumer surplus is estimated to be 50% of total payments to customers; payments to customers are settled on their *baseline*-based demand reduction. Producer surplus is estimated to be 10% of the total bid, where we estimate the retailer keeps 20% of NESO payments to DFS providers, using half for the costs of logistical costs and overheads. CO_2e value is the monetized value of the CO_2e reduction at HM Treasury's specified valuation, approximately £250/t CO_2e at the time of the experiment. Admin cost is the cost to NESO of procuring the demand reduction – where, as noted above, we assume the retailer spent 80% on remuneration to households, kept 10% as profit, and used 10% to cover logistical costs and overheads. Volume-weighted average marginal price: £347/MWh. Average household remuneration: £1,860/MWh. Average cost to the system operator: £2,325/MWh. Break-even household remuneration: £688/MWh. Break-even marginal price: £1,080/MWh.

A13 Deviations from our pre-analysis plans

Notice Period: Due to technical difficulties, we could not randomize notice periods and thus did not perform the planned notice-period analysis, which would have regressed event-period consumption on the amount of advance notice a household received before an event (a categorical variable), controlling for household characteristics and heating degree days. Our original plan was to randomize the order in which opt-in emails were sent. Because sending notifications to hundreds of thousands of households can take several hours, this would have generated quasi-random variation in notice length across customers within the same event. For example, if the first opt-in emails were sent between 16:00 and 17:00 on the day before an event, the full mailing could still take several hours to complete, creating multiple notice-length categories.

Event Study: We planned to estimate an event-study specification interacting treatment with time since the last event, grouped into 4-hour blocks (with the half-hours during an event itself blocked together), estimating leads and lags up to 24 hours around each event. The specification also included a separate indicator for the half-hour when households first began receiving the opt-in notice email, plus time-since-event fixed effects and controls for heating degree days and other covariates.

However, due to computational limitations, we instead estimated hourly treatment effects over a symmetric 48-hour window around each event (24 hours before, the event hour(s), and 24 hours after), restricting the sample to the seven events whose windows did not overlap with another event (Figure 5).

Placebo events: Our pre-analysis plan specified placebo tests using non-event "placebo" events – three dated before any OctoPlus invitation emails were sent (4–6 October 2023) and three after (24–26 October 2023) – to verify that the treatment group exhibited no consumption response on days without a genuine event. We did not report these specific placebo dates. The pre-specified non-event-day analysis (Figure A1) serves the same function across the full trial: it estimates the invitation's effect on peak-window consumption on all non-event weekdays and finds no systematic difference between arms.

Intelligent Octopus exclusion: Our pre-analysis plan specified a robustness check that excluded customers on the Intelligent Octopus tariff for the entire trial period and re-ran

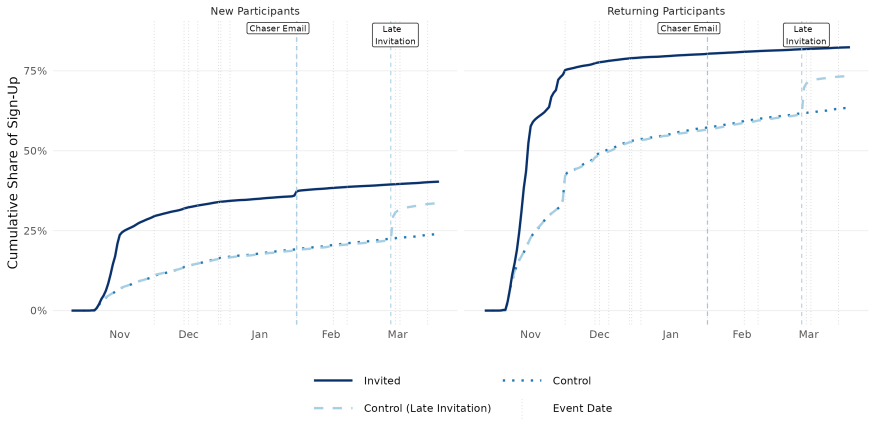
the analysis, to confirm that results were not driven by households with automated (e.g. EV-charging) load-shifting. We did not report this exclusion separately.

Email Sent to Control Group from Previous Year Participants: In Figure A7, we present sign-up rates separately for new participants and returning participants — households who had taken part in the previous 2022–23 iteration of the program. As shown in the balance tables, returning participants are substantially more likely to sign up, but otherwise follow similar trends to new participants.

The exception is mid-November, where we observe a spike in sign-ups among returning participants. This was caused by a technical error: the system sent opt-in reminder notifications for a mid-November event to all households who had been signed up in 2022–23, regardless of whether they had voluntarily re-enrolled in 2023–24 or had been assigned to the control group. As a result, some returning participants in the control group signed up just before that event.

In our heterogeneity analysis, we show that our main results are not affected by previous participation status.

Figure A7: Sign-Up Rates by Trial Arm for New and Returning Participants



Notes: This figure shows the cumulative share of households that signed up for the demand flexibility program, split by whether they participated in the first (2022–23) iteration of the program, across the three primary experimental arms: the invited (treatment) group in solid dark blue, the original control group in dotted medium blue, and the late invitation subset of the control group in dashed light blue. The dashed vertical lines indicate the timing of specific sub-trials – the chaser email campaign in mid-January and the late invitations issued in late February – and the dotted light grey vertical lines indicate individual demand flexibility events. Returning participants sign up at much higher levels but follow similar trends. First-stage outcome of the pre-specified RED analysis.