THE
BEHAVIOURAL
INSIGHTS
TEAM

Impacts of alternatives to In-Home Displays on customers' energy consumption

A report from the Behavioural Insights Team for the Department for Business, Energy and Industrial Strategy

July 2019

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Business, Energy and Industrial Strategy (BEIS).

Any enquiries regarding this publication should be sent to us at:
enquiries@beis.gov.uk

Contents

Executive summary

Scope and purpose of this report

The government is committed to ensuring that smart meters are offered to every home and small business by the end of 2020. Smart meters are the next generation of gas and electricity meters and are expected to deliver a digital transformation of Great Britain's energy system.

As part of the installation of smart meters, energy suppliers are currently obligated to offer households a free in-home display (IHD), which can provide near real-time energy consumption feedback to help them reduce their energy use and bills. However, since the introduction of this mandate, several new energy feedback technologies or methods have emerged that could act as alternatives to IHDs and may be able to deliver similar or better impacts. They may also bring other advantages, such as greater potential for ongoing innovation and subsequently a greater long-term impact. This includes, for example, real-time feedback delivered through smartphone apps. However, there is a lack of robust, independent and f evidence on the efficacy of these alternatives.

The Department for Business, Energy and Industrial Strategy[1] (BEIS) therefore issued derogations in 2016 allowing suppliers to offer customers alternatives to physical IHDs within the confines of approved, rigorous research to ascertain their energy saving potential. Two suppliers, Supplier A and Supplier B[2], undertook such research, each testing smartphone apps that provided near real-time consumption data, feedback, notifications, and past consumption history, and comparing them to conventional IHDs. A third supplier, Supplier C, received a derogation, developed an app, and prepared a trial; however, due to installation and participant recruitment challenges, they discontinued their trial.

The Behavioural Insights Team (BIT) was commissioned to provide guidance and quality assurance on appropriate research methods, in conversation with BEIS and the suppliers themselves, and to interpret suppliers' trial results and analysis. The second of the two trials concluded in early 2019. This report reviews the evidence presented from the trials and synthesises this with existing evidence.

In appendices to this report, we also:

- identify the research questions which the derogation set out to answer and summarise our suggested approach to research of this nature;
- provide further methodological details of the suppliers' trials; and
- develop and validate a Theory of Change describing the mechanisms through which IHDs and app-based alternatives help provide the means, the motivation, and the opportunity to save energy.

[1] The Department of Energy and Climate Change (DECC) merged with the Department for Business, Innovation and Skills (BIS) to form BEIS in July 2016. We refer to the Department for Business, Energy and Industrial Strategy (BEIS) throughout this report, including where the actions being referenced were undertaken by DECC.
[2] Suppliers have been anonymised to protect commercially sensitive information.

Background context and previous evidence

In 2015 the UK government published the Smart Metering Early Learning Project (ELP). The ELP was the culmination of several strands of primary research, including an evaluation of the impact of smart meters early in the rollout across Great Britain. The ELP also involved a synthesis of this new research with existing evidence at the time on consumer engagement and the energy saving impacts of smart meters and in-home displays. Notably, this included the findings of the Energy Demand Research Project (EDRP), a series of trials evaluating energy consumption feedback interventions across Great Britain that began in 2007 and finished in 2010 that confirmed the value of real-time feedback delivered through In-Home Displays in British households.

The ELP and subsequent follow-up work undertaken by BEIS indicated that customer satisfaction and engagement with IHDs is high and found that customers who had early smart-type meters and IHDs installed in 2011 consumed 1.5% less gas (with a 95% confidence interval of 0.9 - 2.1%) and 2.2% less electricity (with a 95% confidence interval of 1.6 - 2.8%) than those with legacy meters and no IHDs. Considering these savings were observed with only a nascent consumer engagement policy framework (before the requirement for installers to offer tailored energy saving advice, for example), and the international evidence on the impact of real-time energy feedback, the ELP synthesis concluded that it would be 'realistic to expect durable savings of 3%'[3] of customers' energy consumption in the rollout of smart meters across Great Britain.

There is relatively little publicly available new evidence of substance from which we can update the conclusions of the ELP. As such, we believe the ELP still provides a reasonable account of the likely energy savings and customer experience.

The particular focus of this project is not on the energy saving potential of IHDs, but on the *relative* energy saving potential of alternative mechanisms of feedback, since the pertinent policy question is whether there is evidence to support the relaxing of the mandate to offer IHDs in favour of alternatives. On this question, there are no robust impact evaluations outside of the two trials undertaken through this derogation.

Testing of IHD alternatives, including apps, was of interest because of their potential to integrate behavioural nudges and other features to promote energy savings – though it should be noted that conventional IHDs could be designed or updated so as to integrate these features. The wider research on behaviour change and energy savings suggests that social norms and social comparisons (comparing a customer's energy consumption to that of their more efficient neighbours), gamification, personalisation and loss aversion (framing energy use as a financial loss) are all features that may be effective in causing reductions in households' energy consumption. Yet, prior to the completion of the suppliers' derogation trials, there was no known evidence comparing apps to conventional IHDs.

[3] Department of Energy & Climate Change. (2015). Smart Metering Early Learning Project: Synthesis report. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/407568/8_Synthesis_FINAL_25feb15.pdf

Supplier A's trial

Supplier A undertook a randomised controlled trial with 7,903 customers, of whom 5,519 were still active at the end of the trial. They allocated roughly half of the customers to receive apps with their smart meters, and roughly half to receive IHDs with their smart meters. They captured weekly smart meter gas and electricity consumption data for one year after installation of smart meters.

Supplier A's IHD showed real-time energy consumption, in kWh or pounds and pence, allowed exploration of past energy consumption history through usage charts, and included the ability to set targets for energy use, with green/amber/red 'traffic light' style feedback. Their app had broadly the same functionality as the IHD, with a number of additional features, including the ability to view energy use while away from home, the provision of energy saving tips, customer messaging functionality, and the ability to capture household information and thus tailor feedback and other features.

To compare the groups' consumption, Supplier A conducted an Intention-to-Treat (ITT) analysis, which compares the whole cohort offered the IHD to the whole cohort offered the app. They also undertook an Average Treatment effect on the Treated (ATT) analysis to provide an approximate estimate of the impact of the app among those who adopted it.

The ITT analysis is the primary analysis because it is not biased by self-selection into the app group. In the context of Supplier A's trial, around 22% of those allocated to the app group did not receive the app, either because of installer error or because they opted to forgo the app.[4] Among the 78% who received the app, 21% never downloaded it. In contrast, only 7% of those offered the IHD declined one, and the remaining 93% all had an IHD installed. This may lead to bias, since a simple comparison between app users and IHD users would be a comparison between a self-selected and relatively engaged 62%[5] of the app group, and a much broader 93% of the IHD group. The ITT analysis addresses this bias by comparing the two groups as originally (randomly) allocated, before self-selection occurs.

In comparing consumption between the app group and the IHD group, Supplier A found that app users consumed more energy than IHD users – 0.47% more gas and 0.88% more electricity. These estimates were not statistically significantly different from zero at a 95% confidence level. However, our conclusion is based on the weight of probability, which suggests the app was likely (but not undoubtedly) to lead to lower energy savings.

Supplier A also carried out survey work with samples of trial participants and undertook detailed qualitative interviews with 12 customers. The findings showed greater engagement, likelihood of recommending the device, and a greater sense of control over consumption levels among the app group. However, the survey was only carried out with those who had either device installed. With 22% of the app group declining the app at the offer (compared to 7% of the IHD group), and a further 21% failing to install the app (compared to no customers failing to have the IHD installed in the IHD group), it is likely that the survey presents a biased account in which the most engaged 62% of the app group are compared to a much broader 93% of the IHD group. Though this is not easily avoided given the nature of the research, wherein the customer' journeys were quite different for customers in the two groups, this bias must be

[4] It is not known what fraction of this 22% were mistakenly offered the IHD, but not the app, by the installer, versus actively opting to forgo the app.
[5] Non-compliance at the point of installation = 22%; non-compliance at the point of customers downloading the app = 21%. 0.78*0.79 = 0.616.

acknowledged when interpreting the results, as it may inflate estimates of engagement in the case of the app, but not the IHD.

In summary, we conclude that Supplier A's app was *likely* to be slightly less effective at saving consumers energy than their IHD.

Supplier B's trial

Supplier B's trial compared the energy consumption and levels of engagement between 2,717 customers using Supplier B's traditional IHDs and 1,550 customers using Supplier B's smartphone app-based alternative (hereafter 'app').

Supplier B's IHD showed real-time energy consumption, presented information about customers' CO2 footprint, allowed customers to explore past energy consumption history charts, and gave customers the opportunity to create target budgets. Their app provided broadly the same functionality as their IHD, including a traffic light system showing energy usage per hour and per day, as well as a number of additional features. These additional features included the ability to view energy use while away from home and weekly notifications about the household's energy expenditure.

Supplier B installed smart meters and IHDs or apps from April 2017 to December 2017. Their primary analysis analysed customers' consumption data from the date of smart meter installation to December 2018.

Due to logistical constraints during the trial, there were some differences between the customers in the IHD and app groups: in their appetite for an app, and in their installer, region, weather, and time of year of installation. This created potential for significant bias between the two groups. To mitigate this bias, Supplier B adopted a matched-samples approach, creating roughly 1,000 matched pairs of customers from the available sample.[6]

In comparing the energy consumption during the trial period of the app customers to the matched sample of IHD customers, Supplier B found that customers from the two groups consumed near-identical amounts of electricity, while app users consumed 2.6% more gas than IHD users. The difference in gas consumption was statistically significant at 95% confidence levels.

In addition to the collection of energy consumption data, three surveys were completed over the course of the trial, and in-depth interviews were conducted with roughly 50 customers in the app and IHD groups (by telephone, online, and in person). Within the app group, Supplier B surveyed customers who were using the app, so, as with supplier A, there is selection bias, which limits the ability to make direct comparisons in attitudes between customers using the two devices. Nonetheless, the responses provide insight into the use cases for the technologies and mechanisms through which customers may save energy. The findings inform many elements of the Theory of Change (in Appendix 5) that BIT posits for how IHDs and app-based alternatives impact behaviour. For instance, a learning period appears to be an important mechanism for behaviour change for both IHDs and apps, after which customers

[6] This quasi-experimental approach is an established evaluation technique. However, it is a compromise made necessary by logistical constraints, as it relies on stronger assumptions than a randomised control trial. A randomised controlled trial creates groups that are on average balanced on observable and unobservable characteristics. The matched samples approach creates a control group that closely resembles the treatment group in terms of observable characteristics only, and it assumes that unobservable characteristics are balanced across samples.

mostly use the devices to monitor against unexpected energy use. Supplier B's trial revealed that, among app users, push notifications seem to be particularly beneficial for prompting ongoing monitoring of energy consumption.

On balance, we note that the quantitative and qualitative research suggested a number of positive features of Supplier B's app. However, as with supplier A, this is not reconciled against the energy consumption analysis, which suggested that app-using customers consumed more gas while consuming roughly equivalent amounts of electricity.

Supplier C's trial

Supplier C received a derogation to undertake a randomised controlled trial comparing their app to a conventional IHD. However, they had problems recruiting a large enough sample of customers to robustly compare consumption between the groups. They also had challenges training installers to install app-based IHD alternatives consistently. Many installers ended up installing IHDs in most of the houses that had been allocated to the app groups. Due to these logistical challenges, Supplier C discontinued their trial.

Energy saving mechanisms of IHDs and apps

Suppliers A and B's trials suggest a number of mechanisms for how their IHDs and apps may have influenced their customers' energy consumption.

Box 1: Comparing the devices' mechanisms of behaviour change

Energy saving mechanisms of the IHDs	Energy saving mechanisms of the apps
• The IHDs' always-on, ambient feedback was likely to engage customers, acting as a constant reminder and prompt. The apps were less able to exploit this feature by virtue of requiring customers to proactively open the app to see traffic-light style feedback on their energy consumption. • The IHDs seem to be superior at delivering 'initial learning' functionality, perhaps because of the always-on, ambient feedback, which may be an important part of the overall energy saving journey. • Like the apps, the IHDs allowed customers to examine historical data and set budgets. • No proactive engagement was required to receive a functioning IHD, and very little proactive engagement was required to get some value out of the IHD once installed. In contrast, apps needed to be proactively downloaded before they could be used. It is well established that these small frictions and hassles can be large barriers to behaviour change. • Evidence suggests IHDs were significantly more likely to be used by multiple members of a household.	• Like IHDs, the apps showed real-time feedback on customers' energy consumption, but customers needed to have the app open to see this information. • The apps provided more detailed energy saving tips and personalised feedback. • Like the IHDs, the apps allowed customers to examine historical data and set budgets. • Once installed, apps were likely to remain available (as mobile devices are generally used every day), whereas IHDs could be moved to less prominent locations (to stay plugged in) or run out of charge. • More knowledgeable and engaged customers may have benefited from the apps' ability to provide sophisticated information about their energy use patterns. In Supplier A's trial, for example, a subset of app users made use of a new feature allowing them to set personalised goals. • Push notifications encouraged ongoing monitoring of weekly/monthly spend and may have partly addressed the challenge of the apps requiring more proactive engagement.

We conclude that the two technologies are not direct substitutes: each has certain advantages and disadvantages, functioning through slightly different behavioural mechanisms.

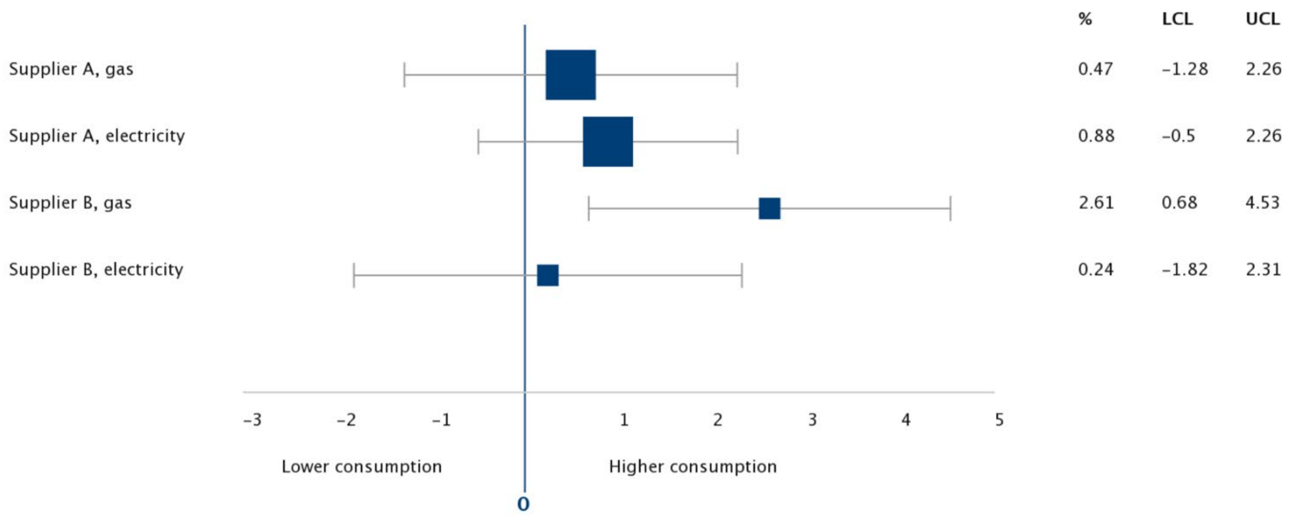
BIT's conclusion

The weight of available evidence suggests that the conventional IHDs were more effective at reducing customers' energy consumption than the suppliers' apps:

- Supplier A's trial shows that while the energy saving potential of their app was not statistically significantly less than that of their IHD, their app was nonetheless *likely* to result in higher electricity and gas consumption.
- In Supplier B's trial, the app was roughly equivalent in its energy saving ability for electricity, but resulted in statistically significantly higher gas consumption relative to the IHD.

The four results are summarised in the plot below, with 95% confidence intervals shown. A result is statistically significant where the line does not cross zero.

Figure 1: Electricity and gas consumption of app customers compared to IHD customers (in both suppliers' analyses)[7]



Though a degree of statistical uncertainty remains around these results, **we conclude that Supplier A's and B's apps were on average likely to save slightly less energy than their IHDs**.

A number of insights from the trials may help to explain why this was the case. The trials' IHDs seemed to be more widely suitable across the suppliers' population, they tended to be used by more members of the household, and the provision of always-present feedback may have been more likely to prompt widespread behaviour change (as opposed to only impacting the more engaged and proactive users of an app), particularly during the 'learning' period immediately after smart meter installation. Though apps may have brought benefits to a segment of more engaged consumers who valued their features, and their push notifications

[7] The boxes representing Supplier A's trial results are larger than the boxes representing Supplier B's trial results in order to reflect the larger sample size in Supplier A's trial. Graph generated by DistillerSR Forest Plot Generator from Evidence Partners.

helped customers monitor their energy consumption, they also seemed to require greater proactive engagement in accessing feedback.

1. Introduction

The government is committed to ensuring that smart meters are offered to every home and small business by the end of 2020. Smart meters are the next generation of gas and electricity meters and are expected to deliver a digital transformation of Great Britain's energy system.

As part of the smart meter installation, energy suppliers are currently obligated to offer households a free in-home display (IHD), which can provide near real-time energy consumption feedback to help them reduce their energy use and bills. However, since the introduction of this mandate, several new technologies or methods that could act as alternatives to IHDs have emerged that may be able to deliver similar or better impacts. They may also bring other advantages, such as greater potential for ongoing innovation. This includes, for example, energy consumption feedback delivered through smartphone apps; however, there is a lack of robust, independent evidence from Great Britain on the efficacy of these alternatives.

The Department for Business, Energy and Industrial Strategy[8] (BEIS) therefore published a derogation in April 2016 allowing suppliers to offer customers alternatives to physical IHDs within the confines of approved, rigorous research to ascertain their energy saving potential. This research, undertaken by energy suppliers, aims to help BEIS assess whether the IHD mandate remains the most appropriate policy.

BEIS commissioned the Behavioural Insights Team (BIT) to advise on managing the derogation and assist with the associated research. To date, BIT's role has been to:

- support BEIS in developing the application guidance for suppliers wishing to undertake research under the derogation;
- develop a framework of assessment criteria under which suppliers' applications were evaluated;
- provide ongoing support and guidance to suppliers during their research design and implementation to maintain high standards of rigour; and
- evaluate the evidence emerging from these supplier trials and synthesise this evidence with other existing evidence.

Three suppliers were granted derogations and undertook research to compare IHDs to apps in the form of mobile phone apps with varying functionality: Suppliers A, B and C.[9] The research proposals from each supplier varied in approach, but they universally aimed to quantify the difference in energy consumption between customers using an IHD and customers using an app by drawing on electricity and gas billing data. Each trial also included a variety of quantitative and qualitative research elements to understand consumer engagement with the devices.

[8] The Department of Energy and Climate Change (DECC) merged with the Department for Business, Innovation and Skills (BIS) to form the Department for Business, Energy and Industrial Strategy (BEIS) in July 2016. We refer to BEIS throughout this report, including where the actions being referenced were undertaken by DECC.
[9] Suppliers have been anonymised to protect commercially sensitive information.

1.1 Purpose of this report

The purpose of this report is to:

- review the evidence generated by the suppliers' trials; and
- draw conclusions from these trials with a view to synthesising this with other available evidence from academic, industry and policy research.

In appendices, we also:

- identify the research questions that the derogation set out to answer and summarise our suggested approach to research of this nature;
- provide further methodological details of the suppliers' trials; and
- develop and validate a Theory of Change describing the mechanisms through which IHDs and app-based alternatives help provide the means, the motivation, and the opportunity to save energy.

1.2 Existing evidence

In reviewing the available evidence related to the energy saving potential of IHDs and app-based alternatives, we drew three key observations, which are relevant to the derogation.

First, there is an absence of trials that directly address our research question: How do IHDs compare to alternatives? While there is some evidence on the energy saving potential of physical and virtual devices, we are not aware of any high-quality impact evaluations that compare the two, nor is it possible to compare the results of separate studies (since methodologies and samples vary). The supplier trials within this derogation are therefore critical in answering the question of interest.

Second, research into energy saving products and services varies significantly in methodological rigour. Small sample sizes are common, energy savings estimates are often implausibly large, and outcome measures sometimes unreliable (e.g. where researchers rely on self-reported behaviour change, rather than analysing actual energy consumption).

The number of high-quality studies is therefore limited. Among those that are available, the Energy Demand Research Project (EDRP), a series of trials by energy suppliers that began in 2007 and finished in 2010, remains among the most informative. The government's Smart Metering Early Learning Project (ELP), published in 2015, drew on the EDRP trials and provides insight into customer use of IHDs and the mechanisms of learning and monitoring that drive energy savings (which are reflected in our Theory of Change, in Appendix 5). This research highlighted the importance of always-on, ambient feedback - a key consideration for the question of devices' impact on energy consumption, since app-based feedback requires more proactive engagement.

The ELP analysed the consumption data of customers that had early installations of smart type meters in 2011 and found that the installation of a smart meter with an IHD resulted in average reductions of 2.2% in customers' electricity consumption (with a 95% confidence interval of 1.6 - 2.8%) and 1.5% in their gas consumption (with a 95% confidence interval of 0.9 - 2.1%).

Considering these savings were observed with only a nascent consumer engagement policy framework (before the requirement for installers to offer tailored energy saving advice, for example), and the international evidence on the impact of real-time energy feedback, the ELP synthesis concluded that it would be 'realistic to expect durable savings of 3%'[10] of customers' energy consumption in the rollout of smart meters across Great Britain.

For further information about the evidence underpinning the IHD mandate, see BEIS's consultation response on in-home display licence conditions.[11]

Our third observation is that while there is no reliable evidence comparing apps or other alternatives to IHDs (besides these trials), the wider academic literature on energy saving feedback and behavioural science provides indirect insight into the potential that apps have to support households in reducing their energy consumption. This is simply because their versatility as a platform means they can more feasibly utilise a wide range of behaviourally-informed techniques compared to IHDs. For instance, social comparisons, gamification, tailored advice, and prompts and reminders are all well-evidenced techniques for nudging behaviour (there is some evidence on energy saving behaviours in each case and plentiful evidence from other contexts). While these features are not impossible to provide through IHDs, it is clear that physical hardware presents a trade-off on innovation, given that households would not be expected to replace their IHD regularly and firmware updates operate within the fixed constraints of IHDs' hardware. In addition to apps being able to support more novel features, the flexibility inherent to the mobile platform is itself a feature, allowing iteration and ongoing evaluation such that apps can continually evolve and provide a platform for ongoing supplier innovation.

Of course, apps require customers to have a smartphone, download the app, set it up, and use it. The question is whether the possible advantages of apps discussed immediately above are sufficient to counter their lower accessibility and higher barriers of engagement.

1.3 Progress of the three suppliers

Three suppliers initially set out to undertake trials in accordance with the derogation guidance.

- Supplier A completed a randomised controlled trial.
- Supplier B used a quasi-experimental design to analyse energy consumption differences between their app-using and IHD-using customers. Specifically, they matched app-using customers to IHD-using customers who had the most similar previous consumption. This was necessary because allocation into the app and IHD groups varied by region and time of year and was partly driven by customers' self-selection into the app group.
- Supplier C received a derogation to undertake a randomised controlled trial comparing their app to a conventional IHD. However, they had problems recruiting a large enough sample of customers to robustly compare consumption between the groups. They also had challenges training installers to install app-based alternatives to IHDs consistently.

---

[10] Department of Energy & Climate Change. (2015). Smart Metering Early Learning Project: Synthesis report. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/407568/8_Synthesis_FINAL_25feb15.pdf

[11] Department of Energy & Climate Change. (2016). In-Home Display Licence Conditions: Consultation response. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/497078/IHD_Policy_Framework_Licence_Conditions_Post_Consultation_Decisions_Final_for_Publication.pdf

Many installers ended up installing IHDs in most of the houses that had been allocated to the app groups. Due to these logistical challenges, Supplier C discontinued their trial.

We summarise the methodologies and results of Supplier A's and Supplier B's trials in Sections 2 and 3 respectively. We synthesise the evidence in Section 4 and conclude in Section 5.

2. Supplier A’s derogation trial

2.1 Summary of trial

Supplier A’s IHDs showed real-time energy consumption, in kWh or pounds and pence, allowed exploration of past energy consumption history through usage charts, and included the ability to set targets for energy use, with green/amber/red ‘traffic light’ style feedback. Their app-based alternative (hereafter ‘app’) had broadly the same functionality as their IHDs (enabled by a separate consumer access device (CAD)), with a number of additional features, including the ability to view energy use while away from home, the provision of energy saving tips, customer messaging functionality, and the ability to capture household information and thus tailor feedback and other features.

The trial was carried out between September 2016 and December 2017 with a total of 7,903 customers, of whom 5,519 were still active at the end of the trial (2,753 with an IHD and 2,766 with an app). Supplier A analysed weekly consumption of trial participants from October 2016 through 8th December 2017. In addition to the collection of energy consumption data, four surveys were completed over the course of the trial. The first three surveys were conducted on a sub-sample of the participants, and the final survey was sent to all remaining participants (with a 20% response rate). In-depth interviews were conducted with 12 customers (six with the IHD and six with the app) in their homes.

Supplier A sought to answer three key questions associated with the derogation, namely:

- Does the app lead to a reduction in energy consumption that is equal to or greater than the IHD?
- How do levels of engagement vary between the app and the IHD?
- Does the app have a positive impact on customer experience compared to the IHD?

2.2 Supplier A’s methodology and implementation

Trial design and analytical strategy

The research design was a randomised controlled trial with two arms. Supplier A initially intended to have an equal split between customers with their app and customers with their IHD. However, they subsequently altered recruitment to 60% allocation to the app, 40% allocation to the IHD. This was to overcome differential rates of attrition occurring at later stages of the recruitment and installation process, as fewer apps were accepted and installed.

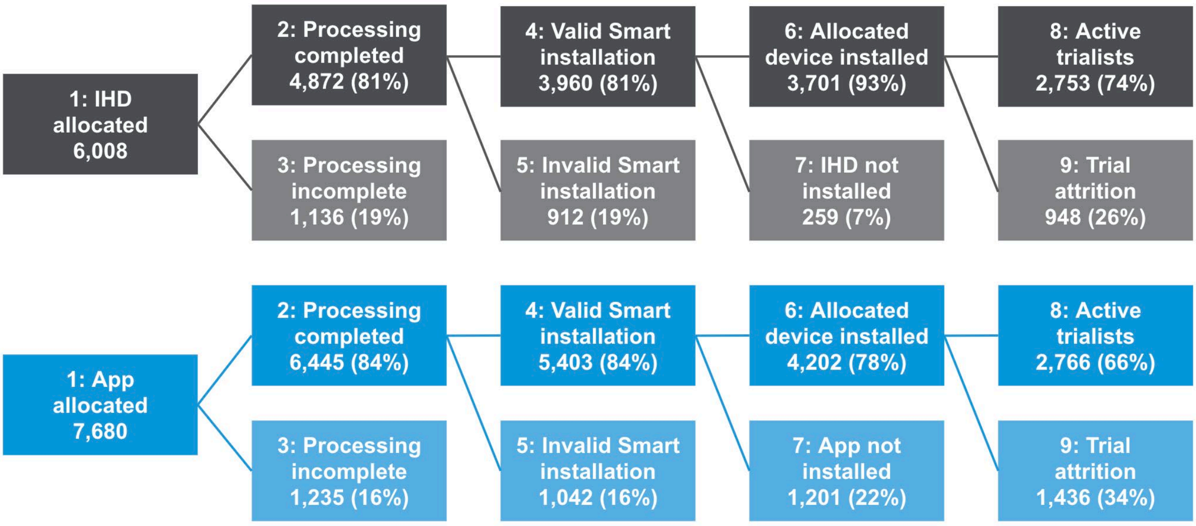
Supplier A used an Intention-to-Treat analysis (ITT) as their primary analysis strategy. The ITT is a comparison of the energy consumption of the two groups based on their allocation to treatment, rather than their acceptance of treatment. This comparison therefore includes all those offered the IHD, against all those offered the app, regardless of whether they received, downloaded or used either device.

The ITT approach was adopted because of selection bias causing non-compliance, predominantly in the app group. That is, around 22% of those allocated to the app group did not receive the app, either because of installer error or because they opted to forgo the app[12]. Among the 78% who accepted the app, around 21% never downloaded it. In contrast, only 7% of those offered the IHD declined one, and the remaining 93% all had an IHD installed.

This may lead to bias, since a simple comparison between app users and IHD users would be a comparison between a self-selected and relatively engaged 62%[13] of the app group, and a much broader 93% of the IHD group. The ITT analysis addresses this bias by comparing the two groups as originally (randomly) allocated, before self-selection occurs. The trade-off is that any difference in energy consumption between the two groups (whether positive or negative) will be somewhat diluted, since not all members of the groups have the intended device.

The flowchart below summarises the recruitment numbers at each step.[14]

Figure 2: Trial recruitment for the Supplier A trial, showing the various points of non-compliance and attrition



From left to right:

- Points 3 and 5 were the first and second points of non-compliance. They reflected administrative processes and failed smart meter installations respectively, which may slightly diminish the representativeness of the trial (e.g. properties unable to have a smart meter installed, or households where the bill payer is more likely to be out, are no longer part of the trial). However, attrition was balanced across the groups, and so not of major concern.
- Point 7 represented those given the wrong device or no device. This was a combination of selection bias (e.g. customers not receiving the device that Supplier A originally allocated to them) and installer error; 22% of the app group did not receive an app, whereas only 7% of the IHD group did not receive an IHD.
- Among those who received the intended device, further attrition (point 9) occurred throughout the trial, primarily caused by people moving house or changing supplier. It is not clear why there is greater attrition in the app group than in the IHD group, though

[12] It is not known what fraction of this 22% were mistakenly offered the IHD, but not the app, by the installer, versus actively opting to forgo the app.
[13] Non-compliance at the point of installation = 22%, non-compliance at the point of customers downloading the app = 21%. 0.78*0.79 = 0.616.
[14] Figures from Supplier A's final report.

this is of some concern as it suggests further bias stemming from selection on some (potentially unobservable) characteristics.[15]

In understanding why the two groups' device allocation rates were different (at Point 7 in the flowchart), it is important to note the two groups' different customer journeys.

- The IHD group had the IHD installed and demonstrated, and were immediately able to use it.
- If those in the app group accepted the app offer, the following working day the supplier shipped them a CAD that connected to their smart meter to enable their app to provide real-time information from their smart meter. These customers also received an email with instructions for downloading the app. The user then had to install the CAD, download the app, and log in to it before they could begin to use it. (Supplier A estimated that having customers carry out this part of the installation themselves saves the smart meter installation engineer 10-15 minutes per installation.)

Within the app group, of those who accepted the offer, 21& never logged in to the app (some of these never installed the CAD, while others did install the CAD but didn't download the app). A 21% attrition rate is actually quite low (i.e. 79% of those accepting the app completed the process) considering the customer needed to complete many steps themselves. Nonetheless, from a methodological standpoint, this does represent a further source of non-compliance in addition to those described above. As such, only 62% of those allocated to the app group ultimately accepted the offer, installed the CAD, and downloaded the app, thus being fully 'treated'.

As noted above, the ITT analysis is the appropriate method to address this bias, though it 'dilutes' the effect of the app (whether positive or negative, relative to the IHD). The ITT analysis also answers a slightly different research question, comparing energy consumption between those offered (rather than between those necessarily using) the two devices.

However, it is also useful to know how effective the app was compared to the IHD among those who actually used it. To answer this question, Supplier A also conducted an average treatment effect on the treated (ATT) analysis, which is a weighting of the ITT analysis using an instrumental variables (IV) approach to account for the dilution of the effect through non-compliance.

Customer surveys and interviews

Four surveys were undertaken throughout the trial. The first three targeted a random subset of each trial group, and the fourth was sent to all participants remaining at the end of the trial. The fourth survey received an approximately 20% response rate, yielding 540 responses in the app group and 544 in the IHD group. Only those who had received the device they had been assigned (the app or the IHD) answered the surveys. Survey responses were weighted based on energy consumption bands. In addition, interviews were conducted with six IHD users and six app users. Each interview was 60 minutes long and was conducted in-person at the user's home. Interviews were split across London and Scotland to capture a breadth of property types and demographics.

[15] For example, app users may have been generally more engaged customers and thus more likely to switch supplier. This explanation is also supported by the fact that app users were significantly more likely to have been with Supplier A for a shorter period (see 'balance checks' in Appendix 3). However, we give this explanation by way of an example of the kind of bias that may be emerging – there is no strong evidence for this specific explanation.

2.3 Supplier A's results

Energy consumption

The ITT analysis, in blue, shows that energy use in the app group was slightly higher than the IHD group for both gas and electricity, though the differences are not statistically significant. Supplier A also reported energy consumption for active versus inactive users, detailed in Appendix 3.

Model	Gas	Gas 95% confidence interval	Gas p-value	Elec	Elec 95% confidence interval	Elec p-value
ITT	+0.47%	-1.29% to +2.26%	0.5994	+0.88%	-0.50% to +2.26%	0.2118
ATT	+0.61%	-1.58% to +2.85%	0.5858	+1.16%	-0.54% to +2.89%	0.1825

Engagement with IHD and app

The metrics used to track engagement differ between the app and IHD. For this reason, we cannot meaningfully compare engagement between the two products. However, some broad patterns emerge. First, app and IHD engagement were both highest at the beginning of the trial, and then steadily declined. Both devices showed slightly higher engagement as the winter approached, as would be expected given the rising cost of heating the home during this period. However, the winter recovery in engagement is modest across all metrics, and much smaller than the prior decline.

Not all app participants used the app at the beginning of the trial, whereas nearly all IHD participants seem to have had their IHD plugged in. As discussed above, the groups' different installation customer journeys may drive this difference. Engineers set-up the IHD for IHD-allocated customers, but app-allocated customers were required to set up the CAD and install and log into the app themselves. In February 2017 (a few months into the trial), around three quarters of IHD-allocated customers had their IHDs turned on, compared to about 60% of app customers using the app at least once in the month. Though engagement among app customers was lower at the beginning of the trial, this engagement was somewhat more sustained for the app than the IHD, in that a greater proportion of the customers who used the device at the beginning were still using it at the end of the trial.

Midway through the trial, Supplier A launched the ability to set and use 'goals' as a new feature on the app; 40-50% of users set goals throughout the remaining trial period.

The average length of time spent per session on the app also generally declined throughout the trial, again with a slight increase in the winter. Reduced time on the app might suggest less interest in its various features, or may just be a sign of habituation as customers become more familiar with it, and consult it for their priority needs without protracted exploration. As such we would not necessarily view a drop in the time spent per session as a negative result (within limits) provided users are still using it at all (which, as noted previously, many are, but also in diminishing numbers). One important feature of Supplier A's app was the ability to send personalised information, energy saving tips and push notifications, and most customers who used the app opened these notifications – approximately 75% of app customers received the notifications and approximately 50% of app-using survey respondents reported actively responding to the notifications.

Survey and interview results

Supplier A sent a survey to all trial participants remaining at the end of the trial who had installed the device that Supplier A had allocated to them. The response rate was approximately 20%.

When interpreting these survey findings, it must be noted that there is potential for bias; as we discuss below (in Section 2.4) this is because only those with the allocated device installed were able to provide answers, with app users being a more strongly self-selected group than IHD users. For this reason, the survey responses may show a bias towards more engagement with the app, making it impossible to directly compare responses between the two groups of respondents. Furthermore, though a 20% response rate is relatively high for a survey, we nonetheless are observing the experiences of a more engaged subset of each group who may not fully represent the wider sample of participants.

Consumption:

- **Both groups claimed to have used less energy as a result of their device**: 49% of app-using customers and 46% of IHD-using customers reported that they thought they consumed either 'a little less' or 'a lot less' electricity due to the installation of the smart meter and the feedback from the app/IHD; 40% of app-using customers and 37% of IHD-using customers felt this way about their gas consumption. When asked about financial savings, 25% of app-using customers were 'sure' they had saved money; 21% of IHD-using customers felt this way. (Another 32% of each group thought they had saved money but could not be 'sure'.)
- **IHD-using customers were slightly more likely to report that they rarely or never changed their behaviour** to reduce their consumption after looking at their device (38% versus 31% for app-using customers).
- Few of these differences were statistically significant, but the pattern appears to be that **app-using customers were slightly more likely to claim they saved energy than IHD-using customers**. This is not evidenced in the energy data analysis, possibly reflecting measurement error, selection bias, or the fact that this may be true for the most-engaged 20% who answered the survey, but not on average for the whole ITT group.

Engagement:

- **IHD-using customers reported engaging with their device more frequently**. This was found to be true throughout all surveys. However, app-using customers reported being more likely to check their device at least once per week.
- **Those with the IHD were significantly more likely to report that the device was used by other members of the household**. Among the app group, just 15% reported that it was also used by a partner (compared to 27% in the IHD); 5% of IHD-using customers also indicated their children used the device (none in the app group reported this).

Sense of understanding of, and control over, energy consumption:

- **App users were more likely to feel informed about their spend**, and slightly more likely to agree that it is easier to control spend, than IHD users are. App-using customers were also slightly more likely to say they felt more in control of their energy consumption thanks to their device (65% of app-using respondents reported this, compared to 55% of IHD-using respondents). Relatedly, IHD-using customers' sense of control over their energy consumption declined more than the app group's over the course of the trial.

- **Younger respondents were more likely to manage energy for cost savings**, and older people were more likely to manage energy for comfort.

Customer satisfaction:

- **Overall customer satisfaction was higher with the app**, and this difference increased over the duration of the trial. At the end of the trial, satisfaction with the device continued to be greater among app users (66% rated their device as 'excellent' or 'very good', compared to 50% of IHD users). Charging of the IHD was cited as a frustration for some IHD users.
- **App-using customers reported their device was useful at higher rates than IHD-using customers** (42% of app-using respondents said the device was 'very useful', compared to 30% of IHD-using respondents). This gap was highest at the end of the trial. This may reflect the fact that the app can continue to provide novel information.
- **App-using customers reported their device was easy to use at higher rates than IHD-using customers** (68% of app-using respondents said the device was 'very easy' to use, compared to 43% of IHD-using respondents).
- **Significantly more customers were likely to recommend the app** (on a 0-10 scale, where 0 represented 'not likely at all' to recommend and 10 represented 'extremely likely' to recommend, 36% of app-using respondents reported an 8, 9, or 10, compared to 27% of IHD-using respondents).

2.4 BIT evaluation of Supplier A's evidence

BIT evaluation of primary results

The Intention-to-Treat (ITT) analysis was the appropriate way to address possible selection bias, as it remains unbiased. We therefore view the ITT analysis as the primary result and as a robust approach. It is also worth noting that the main source of bias[16] is at least partly driven by the difference in customer journey: whereas installation engineers carried out the full set-up for IHD customers, app customers received a CAD in the post and installed it themselves. Results are therefore not a like-for-like comparison between two technologies, but rather a comparison between two different technologies *and* customer journeys.

The primary result suggests customers in the app group consumed slightly more energy than customers in the IHD group. This difference was not statistically significant, which means there was a fair chance the difference was observed only by chance. However, the correct interpretation of this finding depends upon where the burden of proof lies, and how strong the evidence is required to be.

Our understanding is that in order to deviate from the current mandate of offering an IHD, alternatives (in this case apps) must convincingly demonstrate that they provide *equivalent or better* energy savings. In this light, we conclude that *strong evidence of equivalence* between the devices is lacking. Though not definitive, Supplier A's results suggested that apps are slightly inferior, with app users *likely* to have used more electricity than the IHD users, and that

[16] That is, 38% of the app group not being 'treated', compared to 7% of the IHD group, as discussed in section 2.2.

the same was *somewhat likely* to be true for gas consumption.[17] We discuss this further in Section 4.1.

BIT evaluation of surveys and interviews

As discussed above, all survey respondents faced a screening question that removed those who did not have the allocated device installed. This allows us to understand the effect of the devices among those who used it, but it is also a source of bias that undermines a direct comparison between the two technologies: due to different customer journeys and self-selection, 22% of app participants did not receive the app, and a further 21% did not install their CAD or download the app. In contrast, only 7% of the IHD group did not receive an IHD.

While the ITT analysis addressed this bias in the primary energy analysis, it is not practical to adopt an equivalent strategy to compare the survey responses between groups. The survey results therefore compared the 62% most engaged in the app group to a more representative 93% of the IHD group (or, more specifically, responses from 20% of each of these cohorts).

This selection bias reflects genuine challenges in designing and implementing this kind of research, where customer journeys differed and customers in the app group were required to take more proactive steps (causing self-selection). Nonetheless, we must acknowledge the survey results are likely to skew towards higher levels of engagement in the app group, and this undermines our ability to confidently compare the two groups.

For this reason, it is not possible to determine the extent to which these results, including the higher customer satisfaction and sense of control that app-using customers reported, were driven by genuine benefits of the app, or by the biased sample. For example, the survey suggested that, though app users tended to use their device less than IHD users, they were (according to self-reports) significantly more likely to actually change their behaviour. This finding is not easily reconciled with the primary finding that app users consumed slightly more energy: a plausible explanation is that it is only true among the more engaged app users who were eligible for, and responded to, the survey.

The bias in survey samples and responses may be revealed most clearly in one particular question, which asked 'To assist in our future development, which of the following would you like to be able to do?' IHD respondents were more than twice as likely to respond, 'none of the above - I would prefer less technology', indicating a sample that is significantly less technology-savvy. The app group were significantly more likely to highlight advanced features as appealing, including 'know the energy use of individual appliances', 'automatically control heating and water', 'change tariff through an integrated app', and 'automatically control lights'. Though it is possible that use of the app caused these differences in preferences for future features, it is more plausible that these results are driven by (and revealing of) an inherent bias in the survey sample with app respondents being more engaged and knowledgeable.

Supplier A attempted to overcome this bias by weighting the survey responses by energy consumption band. Supplier A shared further information on the balance of the two survey groups on demographics (for both the weighted and un-weighted samples). The differences are not large, though we note that the app survey responders were slightly younger and slightly more likely to have children at home, characteristics that may impact energy consumption and/or customers' attitudes and responses to feedback about their energy consumption. The

[17] For the avoidance of doubt, this does **not** constitute unambiguously strong or statistically significant evidence that the app uses *more* energy than the IHD - rather it refutes the opposite conclusion that they are unambiguously *equivalent*.

groups were balanced on council tax band and gender (though notably there were more men than women in both groups). However, we would not expect energy consumption to adequately correlate with (and thus control for) the many unobservable characteristics that would skew these results (namely, interest in technology, interest in and knowledge of energy saving behaviours etc.). For this reason, we are limited in the conclusions that can reliably be inferred from comparisons between the two groups.

Finally, it is worth noting that the results may not be representative due to the self-selecting nature of responding to a survey (not an easily avoided problem) and the fact that it was administered online. This is the case with both groups, so does not worsen the validity of comparisons between groups, but we would urge against taking these findings as wholly representative of the wider population. For example, levels of engagement might, in general, be somewhat lower than indicated in the survey for both groups.

All of this said, Supplier A's survey research provides a great deal of useful evidence on overall use and customer experience, data on engagement with the technologies within the groups, and long-term engagement trends.

BIT conclusions on Supplier A's trial

We acknowledge that much of the energy feedback functionality offered by the app was valued by the app-using customers, particularly by those who are technologically savvy and well suited to this technology.

However, we conclude that the presented evidence does not constitute strong evidence of Supplier A's app being 'as good as or better than' their IHD in reducing consumer energy consumption.

Some statistical uncertainty remains around the results, and as such it is imperative that this conclusion is understood alongside a consideration by BEIS of how strong the evidence must be to conclude equivalence. We discuss this further in section 4.1.

3. Supplier B's derogation trial

3.1 Summary of trial

Supplier B's IHD showed real-time energy consumption, presented information about customers' CO2 footprint, allowed customers to explore past energy consumption history charts, and gave customers the opportunity to create target budgets. Their app-based IHD alternative (hereafter 'app') provided broadly the same functionality as their IHD (using data provided by a consumer access device (CAD)), including a traffic light system showing energy usage per hour and per day calibrated to Ofgem's typical domestic consumption values for households[18], as well as a number of additional features. These additional features included the ability to view energy use while away from home and weekly notifications about the household's energy expenditure.

Supplier B's primary research question was: Does providing customers with an app cause them to reduce their consumption by at least as much as giving them an IHD? Supplier B also investigated customer engagement and customer experience/satisfaction.

Supplier B installed smart meters with apps in 1,550 customers' homes and installed smart meters with IHDs in 2,717 customers' homes, with the installations occurring between April 2017 and December 2017. Due to unforeseen logistical constraints, these groups were not randomly allocated. As discussed in more depth below, the groups differed on age, geographic region, and time of year of installation; furthermore, allocation into the groups was partly driven by self-selection, and, as with Supplier A, the two groups experienced different customer journeys. Because of this non-random allocation, Supplier B matched customers in the app group to customers in the IHD group according to previous energy consumption as well as a coarse income/age segmentation.

Because Supplier B non-randomly allocated apps to their trial participants, their trial answers a slightly different question than Supplier A's trial. As discussed in Section 2, Supplier A's analysis estimated the Intention-to-Treat effect (ITT), comparing all those offered the IHD against all those offered the app, regardless of whether a customer actually received the respective treatment. In contrast, Supplier B's trial estimates the effect of the app compared to the IHD among customers who successfully received the technologies.

In addition to the collection of energy consumption data, Supplier B conducted three surveys over the course of the trial with roughly 200 customers, and they conducted in-depth interviews with roughly 50 customers, drawing roughly evenly from the app and IHD groups. The findings from this research support many elements of the Theory of Change for IHDs and alternatives described in Appendix 5.

[18] 'Low' consumption is classified as 8000 kWh annual gas consumption and 2,000 kWh annual electricity consumption. 'Medium' consumption is classified as 12,500 kWh annual gas consumption and 3,100 kWh annual electricity consumption. 'High' consumption is classified as 18,000 kWh annual gas consumption and 4,600 kWh annual electricity consumption.

3.2 Supplier B's methodology and implementation

Supplier B's primary consumption analysis used a random-effects model, where the unit of observation was monthly gas and electricity consumption in each household. They analysed the consumption data from the date of smart meter installation (April 2017 to December 2017, depending on the household's installation date) through December 2018.

Supplier B installed smart meters with apps in 1,550 customers' homes and installed smart meters with IHDs in 2,717 customers' homes. As noted above these groups were not randomly allocated. We summarise below the four principal ways in which allocation was non-randomised:

1. Relatively early in the trial, Supplier B limited their offer of apps to customers who were younger than 65 years old. They made this decision due to recruitment problems that had occurred in the first wave of installations. For example, in some instances, installers were offering the app to those in homes with no home broadband (and thus were ineligible for the trial), and they found that these installation failures were more prevalent among customers aged 65 and older.
2. Due to installation engineer training constraints, it was necessary for certain installers to focus on one technology or the other, creating an imbalance between groups in their installer experience (which is recognised as an important part of the intervention itself) and in the regions in which these installers operated. Overall, customers in the app group were more likely than IHD customers to be in North East England, and IHD customers were more likely than app customers to be in North West England and the Midlands.
3. Due to supply-chain issues with the model of smart meter with which the app was integrated, the installation months are systematically different by group. Overall, these supply chain issues caused fewer app installations in the summer than in other months, whereas installations were relatively constant across the year in the IHD group.
4. App recruitment and allocation was separate from IHD allocation. App customers proactively responded to email or mail campaigns in their regions, affirmatively requesting that Supplier B install a smart meter with an integrated app (with the implication that this would be instead of an IHD). This is inherently a process of self-selection. Further selection occurred after setting up an installation appointment because only 40-50% of scheduled app installations resulted in a successfully installed smart meter and app.[19] Reasons for unsuccessful app installation include installer error, lack of broadband internet at the house (broadband internet is necessary for the app's CAD to function), and/or customer preference for an IHD instead of an app.

For these reasons, the app group in Supplier B's trial consisted of internet-connected and self-selected customers, whereas the participants in the IHD group were more representative of the general smart-meter-accepting population.

To create an unbiased comparison between app and IHD customers' energy consumption, Supplier B created a matched IHD group for the app customers. They matched each app customer with a customer from the IHD group who 1) was in the same 'residential segment',[20]

[19] We do not know what percentage of the IHD customers had successful smart meter and IHD installations.
[20] Residential segment is a six-segment categorisation Supplier B have created for marketing and customer communication purposes. It is based mostly on income and age.

and 2) had the closest daily consumption (weather-adjusted, for gas, and raw, for electricity) before the installation of the smart meter. They did this separately for electricity and gas, and, therefore, an app customer may be paired with one household for gas analysis and another for electricity analysis.

Creating matched samples invariably involves a trade-off: the more variables one attempts to match on, the cruder the matches will be. Supplier B struck a balance by matching on just two variables, recognising that matching on past consumption was most critical. After matching, the groups were still imbalanced on certain customer characteristics, such as tenure with Supplier B and number of bedrooms in the property. In order to control for any bias resulting from these imbalances, Supplier B included these characteristics as covariates in the regression analysis.[21]

In total, Supplier B's matching process created 1,069 pairs of gas customers and 984 pairs of electricity customers. These sample sizes differ because Supplier B removed customers from analysis if their consumption was extremely high or extremely low (because it indicates they are an outlier or another factor was substantially affecting their consumption), or Supplier B lacked actual consumption data for a given month for any reason. Electricity consumption data necessitated this trimming more frequently than gas consumption data.

Supplier B compared the app customers' energy consumption to the matched IHD customers' energy consumption using a random effects regression on monthly consumption. They specified as covariates a set of customer characteristics, previous consumption, and weather/daylight hours. Note that comparing the app group and its matched IHD sample theoretically provides an unbiased estimate of the effect of the app on energy consumption; by including the covariates in their regressions, Supplier B controlled for imbalances between groups that exist despite the matching and increased the precision in their estimate of the effect of the app. However, a matched-samples design is always a trade-off relative to a pure randomised trial, the implications of which are discussed further in section 3.4.

3.3 Supplier B's results

Energy consumption

Supplier B found that customers who received apps consumed 0.93kWh (or 2.6%) more gas than the matched group of customers who received IHDs (with the 95% confidence interval spanning 0.24kWh to 1.62kWh higher gas consumption, or 0.68% to 4.53%). This result is statistically significant.

Electricity consumption was almost identical between groups, with app users' consumption being just 0.2% higher (with the 95% confidence interval spanning -0.18kWh lower to 0.20kWh higher electricity consumption, or 1.82% to 2.31%). This result is not statistically significant.

Gas	Gas 95% confidence interval	Gas p-value	Elec	Elec 95% confidence interval	Elec p-value
+2.61%	+0.68% to +4.53%	0.0082	+0.24%	-1.82% to +2.31%	0.8185

[21] See Appendix 4 for full covariate set and regression specifications.

Supplier B also conducted exploratory analysis looking at whether the effect of the app versus IHD depends on the amount of days since the smart meter (and IHD or app) was installed.[22]

- For gas, the exercise indicated that the app's disadvantage with respect to the IHD is highest immediately after installation, perhaps indicating a disadvantage in the 'learning period'. The app then partially 'catches up' to the IHD in terms of its consumption-reducing effect.
- The effect of days since installation interacted with treatment is small for electricity. The difference between the electricity consumption of app customers and IHD customers appears very close to zero, no matter how many days since installation have passed.

However, we caution against drawing inferences from these trends with high confidence. The results are sensitive to the way BIT and Supplier B constructed the regression specification.[23] Moreover, a larger sample size would be desirable to precisely identify interaction effects such as these.

Qualitative interview and survey results

Supplier B conducted two interview rounds, with 53 customers (split roughly evenly between the app and IHD groups), through multiple media (pop-up online communities, telephone interviews, and in-home interviews) and three rounds of online surveys with roughly 200 customers (again split roughly evenly between the app and IHD groups). Their findings informed many elements of our Theory of Change for how IHDs and apps impact behaviour (see Appendix 5). A learning period appears to be an important mechanism for behaviour change for both IHD types, after which customers mostly use the devices to budget and monitor against unexpected energy use. Supplier B noted that IHDs tend to be used by multiple members of the family, which is an important hypothesised advantage of the devices vis-à-vis apps – though there is also some evidence that the app may be downloaded by multiple members of the same household.

Notifications: Supplier B's customers noted that the weekly and monthly energy spend summaries were useful, suggesting that notifications are an important mechanism through which customers engage with the app, especially during the longer-term monitoring of energy consumption. Supplier B's login and usage data indicated that activity peaked on Mondays, around the hour when Supplier B sent the expenditure notifications, suggesting that the notifications motivated active use of the app. These analytics also indicated relatively stable engagement. Roughly 25% of the weekly notifications that Supplier B sent appear to be seen by customers, even in the final few months of the trial, 12-18 months after the smart meter and app installation.

Current energy use feature: Supplier B noted that customers made use of the app's current energy use feature to see how much energy (gas and/or electricity) their household uses at specific moments. Supplier B's app usage and login data indicated that the 'current energy use' screen is the most popular screen, though this is likely to be in part because it is the

---

[22] To do this, they interacted the indicator variable for app with days since smart meter installation (as a quadratic). This attempts to quantify any changes in the difference between the devices as time passes since installation.

[23] Specifically, Supplier B's exploratory analysis looked at days since installation as a quadratic (days since installation, and the square of days since installation). However, it is possible that the effects are more complicated than can be captured by this specification. By way of two examples, the energy consumption reductions that one or both devices cause could ramp up and down multiple times in a complicated cycle, or the 'learning effects' could vary with weather, time of year, or region.

default customers see first when they open the app. This feature's 'traffic light' system was similar to the supplier's IHDs' feedback in some respects:

- As with their IHD, Supplier B's app was calibrated to Ofgem average usages[24], where 'low' usage creates a small green energy pulse, 'medium' creates a medium-sized yellow pulse, and 'high' usage creates a large red pulse.
- This created potential calibration issues, in that a household's typical baseline consumption did not influence the app/IHD's feedback (this is not necessarily the case for all IHDs, but our understanding is that it was the case with Supplier B's IHD and app).
- As with IHDs, this calibration issue may have had different effects on energy consumption for gas versus electricity. Supplier B pointed out that both devices had access to electricity consumption multiple times per minute, whereas they had access to gas consumption every half hour. The use of an electrical appliance like a kettle will often be electricity-intensive enough, for a minute or so, to almost certainly count as 'high' usage, potentially giving customers the spurious belief that short-use, high-current appliances like kettles are more problematic than background use such as refrigerators. Gas consumption and feedback will be smoother in both devices' feedback.

That said, the 'current energy use feature was different from IHDs' traffic-light feedback in a few respects.

- The app's default screen showed a sum of gas and electricity use, displayed as pounds/pence used per hour. In contrast, IHDs showed electricity use (not electricity and gas) as the default screen.
- Customers had to have the app open to see energy use changing. In contrast, IHDs sometimes turn 'red' in response to an action, and occupants may passively notice this.

Because of this second difference, the app likely demanded more active customer engagement (at least in the learning phase) in order to have the same impact on customers' behaviour as IHDs. For ongoing monitoring of energy consumption, the app's push notifications alerting customers to expenditure may help overcome a lack of proactiveness by their users.

Energy history feature: Supplier B's app allowed customers to inspect their energy consumption over the past day, week, month, or year. Various customers described the usefulness of the energy history graphs to help them learn about and monitor their usage patterns, explaining that they used the feature proactively and in response to the weekly and monthly spend summary notifications.

Involvement of multiple members of household: Customers noted that they used the app as a tool to start conversations about energy use and energy efficiency with other members of the household. Supplier B's login data also indicated that multiple members of a household may have downloaded the app.[25]

---

[24] As explained in a previous footnote: 'Low' consumption is classified as 8000 kWh annual gas consumption and 2,000 kWh annual electricity consumption. 'Medium' consumption is classified as 12,500 kWh annual gas consumption and 3,100 kWh annual electricity consumption. 'High' consumption is classified as 18,000 kWh annual gas consumption and 4,600 kWh annual electricity consumption.

[25] The evidence here is not definite. Logins outnumber trial participants, at least for some months, but these numbers are not solely due to multiple users downloading the app. A user who has the app on multiple devices, or purchases a new phone and re-downloads the app, would be double-counted, for example.

Monitoring energy use: Supplier B’s app- and IHD-using customers said they referred to consumption information frequently after installation. For both devices, customers commented in surveys and interviews that it would be useful to receive advice rather than just information from the device. Supplier B noted that after the first six months with the smart meter and app, customers logged in and used the app less frequently, though as with Supplier A, colder weather seemed to increase engagement temporarily. Customers reduced their time spent per session on the app over the course of the trial, which again echoes Supplier A’s findings. This reduction in time spent on the app is not necessarily negative, as we would expect users to get quicker at using the app as they become accustomed to it and know what information they want to access.

Supplier B's app hub: Supplier B’s app operated through a ‘hub’, a CAD that synced with the smart meter, which itself had a display with basic information on it. According to the qualitative data, this seemed to lead to some confusion, with some customers thinking this device was an IHD, despite Supplier B’s assurances that this hub should not be displaying any data beyond home broadband connectivity status.

3.4 BIT evaluation of Supplier B's evidence

BIT evaluation of primary results

Supplier B’s primary results suggest that app users use more gas. We can be reasonably confident of this conclusion given the result is statistically significant. Electricity consumption looks to be nearly identical between the groups, suggesting the two devices may be equivalent – though, with wide confidence intervals, this is still not certain, and there is a fair chance that the app is worse (or indeed better). Again, the interpretation of these findings therefore requires consideration on how strong the evidence must be, and we return to this issue in Section 4.1.

These results need to be treated with some caution due to the methodology adopted by Supplier B. For reasons discussed above, Supplier B’s allocation of customers to the app and IHD groups varied by region, installer, time of year, weather, age, and self-selection towards apps. To mitigate these issues, Supplier B matched customers from the app group with households in the IHD group based on previous consumption and their age/income segment. BIT regards this quasi-experimental analysis strategy as the best option available, and it is generally a robust approach. Nonetheless, the matching may not have been sufficient to make the groups equivalent on all characteristics that may affect energy consumption after the installation of a smart meter and provision of an IHD.

By matching groups on historic consumption, the impact of unobservable but time-invariant characteristics, which impact energy consumption consistently over time (such as the thermal properties of the home), should be reasonably well balanced given the balance on past consumption. Age was also relatively well balanced after matching, as was the number of adults in the household. The main risk, therefore, is that there remains some imbalance on attitudinal variables, which mean one group responds to the intervention differently to the other group, and/or on other characteristics that impact energy consumption that vary with time. For example, it is possible that one group has on average already undergone a higher level of energy saving home modifications or habit changes, and thus has less room for improvement in response to the IHD or app. There is little that can be done to either identify, quantify, or control for these possible sources of bias, other than to caveat the findings accordingly.

With these caveats in mind, it is useful to put these methodological limits into perspective. In comparing Supplier A and Supplier B's trial methodologies, our conclusion is that Supplier A's trial methodology is superior given their ability to maintain a true randomised controlled trial. Though non-compliance was an issue for Supplier A, the ITT analysis remains robust to selection bias by investigating the groups as allocated *before* selection bias created concerns about imbalance. Supplier B's matched sample approach was the correct method, but is invariably a compromise. This being said, we believe the results of Supplier B's trial are robust enough to draw meaningful conclusions.

BIT evaluation of secondary results

As with Supplier A's survey results, Supplier B's surveys suffer from selection bias because only customers who successfully received the IHD or app could respond. These two groups experienced different customer journeys, with the app-using customers needing to take proactive steps, which the IHD customers did not. Moreover, because the matching exercise for consumption analysis occurred after the trial finished, all surveys were conducted on the pre-matched samples. We therefore cannot make direct comparisons between the two groups on survey responses.

That said, there are some important points worth highlighting.

- **Engagement of multiple household members**: It is an important qualitative finding that some of Supplier B's app-using customers report that the app sparked conversations about energy consumption with other members of the household. A key feature of IHDs is that they are accessible to all members of the household. Supplier B's qualitative interviews and app analytics data are a reminder that apps may also be used by multiple family members, either through multiple members of a household downloading the app or because notifications and other features of the app spark conversations between members of the household.

- **Future improvements to the app**: Some customers reported that they wanted notifications to suggest specific and immediate actions, give more targeted and personalised recommendations, alert to unusual or high usage, alert to specific devices' consumption (i.e. disaggregating customers' energy consumption data) and give personalised budgets and thresholds. These comments echoed findings in the wider behavioural literature showing the value of customisation.

- **Stable residual engagement**: Though engagement with the app decreased over the course of the trial, comments by customers in qualitative interviews, survey results, and analytics data all point to at least a portion of customers — perhaps 30 to 50% of the overall app group — remaining engaged, i.e. checking energy consumption graphs from time to time and opening at least some energy consumption summary notifications.

BIT conclusions on Supplier B's trial

Despite concerns about whether Supplier B's analysis strategies sufficiently controlled for differences between the app and IHD groups, we regard Supplier B's trial as an important piece of evidence in evaluating the question of whether apps are good substitutes for IHDs.

On balance, our conclusions are similar to that for Supplier A's trials. A degree of uncertainty remains around the comparative energy saving of the two technologies – in this case not only because confidence intervals are wide, but also because for one fuel the app appears to be significantly inferior to the IHD, while for the other fuel there is no notable difference. We do not have a clear explanation for why the result is different for the two fuels. A number of possible explanations exist; however, we caution against over-interpreting the findings with post-hoc narratives. We therefore conclude that the app is *likely* to save less energy than the IHD, and, certainly, the trial falls short of providing compelling evidence that the app is unambiguously equivalent or better than the IHD. Again, we discuss the rationale behind this conclusion further in Section 4.1.

4. Evidence synthesis

This report and appendices present a wide range of evidence, in particular reviewing the results from Suppliers A and B's derogation trials. Our objective is now to synthesise this evidence to help BEIS consider whether the IHD mandate remains appropriate considering this new evidence on the impact of alternatives.

Supplementary to this chapter, we also outline a Theory of Change, describing the mechanisms through which IHD and app feedback lead to energy savings, included in Appendix 5.

4.1 Energy savings

The EDRP trials, ELP research, subsequent follow-up work by BEIS, and recent academic research provide a reasonable amount of evidence of IHDs' energy savings, but almost no evidence on the energy saving potential of app-based energy feedback relative to those of IHDs. We therefore rely heavily on the trials undertaken by Supplier A and Supplier B in answering this question. Though this evidence is not definitive, on balance we conclude that there is insufficient evidence to be confident that the suppliers' apps caused reductions in customers' energy consumption equal to or greater than the suppliers' IHDs.

However, there is some statistical uncertainty around these findings, with wide confidence intervals and three of the four differences not found to be statistically significant. So, a proper interpretation of the findings requires some consideration of where the burden of proof lies, and how definitive the evidence is required to be.

Burden of proof

Conventional statistical tests would assume equivalence between two populations (i.e. between those using the app and the IHD) unless there is 95% certainty that a difference in energy consumption exists. This approach is clearly not suited to the policy question at hand because it puts the weight of assumption strongly in favour of apps being equivalent to IHDs, until categorically proven to be inferior. Indeed, given the wide confidence intervals around all results, this approach would carry significant risk of falsely finding a negative result (a result of no difference) where there is actually a potentially meaningful difference between the technologies. This approach would also favour noisier data or smaller samples (i.e. weaker research) because these methodological shortcomings would widen the confidence intervals and decrease the likelihood of detecting a significant difference between the two devices.

To avoid this trap, the appropriate approach is to assume an app's consumption impact is *not* equivalent to an IHD's consumption impact unless the evidence convincingly shows that the apps are equivalent or superior to the IHD. Taking this approach, in order to draw reasonable conclusions from the results of the two trials, we must first answer two questions[26]:

1. How do we define 'equivalence'? For example, must the energy consumption of app users be within 0.1% of IHD users? Or 0.5%? Or 1%?

[26] We discuss this in some detail within Appendix 2.

2. What is the strength of evidence we would require in order to conclude that this standard of equivalence has been met? Must we be 95% sure? Or 90% sure? Or just 50% sure?

These are not analytical questions, but rather judgments that should reflect the needs of the smart meter rollout. For the sake of following this logic to its conclusion, we presume the following:

1. The app users' consumption must be no more than 0.5% higher than that of the IHD users. Given a conservative assumption that smart meters with IHDs cause a 2% reduction in electricity and gas consumption, a greater penalty than 0.5% may be deemed too large because it would remove a quarter of the energy savings resulting from the rollout.
2. We must be at least 80% confident that this standard is being met. Statistical convention would be 95%, but we recognise this is a burdensome requirement, which even highly effective apps might fail to meet. At the opposite extreme, 50% confidence would provide no better insight than a guess into whether the app meets the standard described in point (1) (for example, one has 50% confidence in the outcome of a coin toss). Requiring 80% confidence is therefore reasonable.

On this basis, we conclude from both trials that the apps fail to demonstrate equivalence to the IHD. Specifically, for each of the four results:

- Supplier A's gas result suggested app users use 0.47% more gas. This is within our threshold of 0.5%, but the confidence intervals are wide, so we cannot be very sure the difference is not bigger than that. In fact, we can only be slightly better than 50% sure (close to chance) that the difference between the IHD and app is less than 0.5%. This would not meet the 80% confidence threshold.
- Supplier A's electricity result falls short. The results suggested app users use 0.88% more electricity, beyond our threshold of 0.5%. Again, the confidence intervals are wide, so there is a chance that the app is actually within the 0.5% threshold, but this is a slim chance - very far from 80% certainty.
- Supplier B's gas result falls very short. The results suggested app users use 2.6% more gas, with the 95% confidence interval spanning +0.68% to +4.53% additional energy consumption. In other words, we can be very confident that the app fails to meet our standards of equivalence.
- Supplier B's electricity result is the strongest of the four. The result suggested app users have near-identical electricity consumption to IHD users. Furthermore, looking at the margins of confidence around this result, we can be fairly confident that the true difference falls within our 0.5% tolerance – but still not the required 80% confidence, due to the wide confidence intervals in calculating the app's effect on consumption compared to the IHD. We might argue that 80% confidence is still too burdensome and could be relaxed further, but recall that the other three results would still fall short.

To conclude, we are not suggesting smart meters with apps do not reduce energy consumption. We are also not stating that they *definitely* cause smaller reductions in energy consumption than smart meters with IHDs. Instead, the evidence is ambiguous but suggests apps are on average *likely* to be inferior in their energy savings (and from some of the results, highly likely). In other words, **the apps tested by Supplier A and Supplier B are not convincingly shown to reduce energy consumption by as much as or more than conventional IHDs**.

4.2 Energy saving mechanisms of IHDs and apps

The body of existing evidence, including the findings presented by Suppliers A and B, provides a number of possible mechanisms for how IHDs and apps may influence customers' energy consumption. The results of the suppliers' trials comparing IHDs to an app indicated that, on balance, the IHDs' energy saving mechanisms caused a higher reduction in energy consumption than the apps' energy saving mechanisms. Below, we summarise the energy saving mechanisms that differ between the two devices.

Box 2: Comparing the devices' mechanisms of behaviour change

Energy saving mechanisms of the IHDs	Energy saving mechanisms of the apps
• The IHDs' always-on, ambient feedback was likely to engage customers, acting as a constant reminder and prompt. The apps were less able to exploit this feature by virtue of requiring customers to proactively open the app to see traffic-light style feedback on their energy consumption. • The IHDs seem to be superior at delivering 'initial learning' functionality, perhaps because of the always-on, ambient feedback, which may be an important part of the overall energy saving journey. • Like the apps, the IHDs allowed customers to examine historical data and set budgets. • No proactive engagement was required to receive a functioning IHD, and very little proactive engagement was required to get some value out of the IHD once installed. In contrast, apps needed to be proactively downloaded before they could be used. It is well established that these small frictions and hassles can be large barriers to behaviour change. • Evidence suggests IHDs were significantly more likely to be used by multiple members of a household.	• Like IHDs, the apps showed real-time feedback on customers' energy consumption, but customers needed to have the app open to see this information. • The apps provided more detailed energy saving tips and personalised feedback. • Like the IHDs, the apps allowed customers to examine historical data and set budgets. • Once installed, apps were likely to remain available (as mobile devices are generally used every day), whereas IHDs could be moved to less prominent locations (to stay plugged in) or run out of charge. • More knowledgeable and engaged customers may have benefited from the apps' ability to provide sophisticated information about their energy use patterns. In Supplier A's trial, for example, a subset of app users made use of a new feature allowing them to set personalised goals. • Push notifications encouraged ongoing monitoring of weekly/monthly spend and may have partly addressed the challenge of the apps requiring more proactive engagement.

4.3 Suitability

In the UK, 78% of adults 16 years old or older have a smartphone.[27] The majority of the population would thus be able to use the suppliers' apps evaluated in these trials, but this does not mean app-based alternatives to IHDs are universally suitable. The qualitative research that both suppliers conducted indicates that even among the customers who accepted the app offer, there was a continuum of technological know-how. Qualitative evidence from both trials suggested that some people had difficulty using both devices, and, in some cases, there was significant confusion. We cannot quantify the frequency of these issues from the qualitative interviews, and the quantitative surveys did not allow for robust comparisons between the two

[27] Ofcom. (2018). Communications Market Report 2018. Available at: www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf

devices. Supplier B's recruitment difficulties also showed many customers over 65 years old to be without home broadband and thus an unsuitable group to be using their app, which relied on connecting a CAD to the home broadband. This echoes findings from BEIS' 2017 Customer Experience Study, which found that app engagement was lower among those over 60. In contrast, IHDs appear to be relatively universal in their suitability.

Even notwithstanding households without internet, we also find that the three suppliers undertaking trials within the derogation had varying degrees of difficulty in recruiting participants, largely due to difficulties in recruiting sufficient customers into the app arm of their trials. This may suggest that apps are not well suited to everybody, though other factors are likely to be present - for instance, the wider public may be relatively well aware of their entitlement to a physical IHD, whereas apps have not benefitted from public awareness efforts. Supplier C had similar difficulties, which ended up being sufficiently severe that they discontinued the trial.

Once customers were successfully recruited into the app groups, the trials suggest there were still some frictions to getting apps into some customers' hands. Among Supplier A's customers, we see that among those who accepted the offer, a sizeable minority did not receive/download the app – this is partly due to some customers (21%) choosing not to install their CAD or download their app. In the context of our analysis, these are factors that contribute to the lower energy savings of the app, though we should be cautious not to conflate a higher-friction customer journey with an inferior product. Put differently, it is questionable whether these issues are a fundamental downside to the technology; if apps became widespread alternatives to IHDs, an 'app customer journey' could more closely reflect an 'IHD customer journey' or become generally lower-friction.[28]

Supplier B had a similar experience, estimating that 50-60% of the app-allocated customers ended up taking an IHD instead of an app, again due to a mix of home-compatibility issues, installer errors, and customer preferences. These numbers are substantial and have a major impact – if 40% or more of the population is not getting the product, the benefits (or detriments) it provides across a population are immediately diluted. This does not necessarily mean IHDs are themselves suitable for everybody (or, indeed, the first preference for everybody), though the fact that installations remain above 90% in the trials is encouraging.

It could be argued that customers could be given the choice of having an app or IHD, allowing them to take whichever technology is better for them. Both suppliers' survey responses and qualitative interviews suggested that some customers enjoyed the app and wanted even more functionality, while others did not. Given this variety of preferences, this approach certainly looks to have some merit. However, we urge some caution when considering this option: it is not certain that customers would choose the best option for themselves. The wider behavioural science evidence highlights our propensity to make suboptimal choices, and to fail to follow through on our good intentions. The trial evidence reinforces this: a sizeable minority who willingly took the app did not download it. The evidence on the behavioural mechanisms of the two devices also shows that each technology has certain distinct advantages, and customers may not fully recognise these subtle differences in behavioural mechanisms in making a choice between devices.

[28] For the avoidance of doubt, we are not suggesting the likely inferiority of apps is solely due to an inferior customer journey leading to diminished uptake – there is also evidence from these trials that apps are likely to be inferior among those who actually used them.

4.4 Customer engagement and behavioural mechanisms

There is a modest body of existing literature researching the impact of different behavioural interventions delivered within energy feedback. For example, framing energy expenditure as a loss harnesses our loss aversion (we are more sensitive to losses than to equivalent gains) and thus has the potential to create stronger motivation to save energy; introducing games into the customer experience may motivate learning; and drawing comparisons with more efficient neighbours has been robustly shown over multiple studies to lead to energy reductions of a few percentage points.

These findings are relevant because they are features that may start to appear in all energy feedback devices, but apps may offer more potential for these kinds of innovations than physical IHDs. Though IHDs do receive occasional firmware updates, their functionality will be somewhat limited by the fixed constraints of their hardware compared to the 'blank slate' of a smartphone.

Supplier A's and Supplier B's research begins to highlight the benefits of some of these kinds of features. Specifically, the two suppliers identify the advantages of being able to send push notifications and ongoing energy saving tips to customers to prompt action. This feature seemed to be effective in maintaining long-term engagement with feedback among app users.[29] It is possible that a simple push notification requires *less* engagement than turning towards the relevant feature on the IHD on a regular basis to monitor ongoing energy use.

Supplier A's trial also highlights a valid concern that IHDs may be moved away from the most prominent position in the home due to the need to plug it in, or the device may remain unused for long periods of time when the battery is dead. In contrast, smartphones will generally remain charged and on-hand to the customer.

Despite Supplier A's and B's apps possessing some beneficial mechanisms of behaviour change, the primary energy consumption analysis suggests their energy saving potential was likely to be slightly inferior to the suppliers' IHDs. This suggests that these benefits are not, at present, sufficient to overcome their more fundamental drawbacks.

Principally this relates to apps' greater dependence on proactive engagement, both in downloading and installing them to begin with, and subsequently choosing to use them. In contrast, IHDs are installed without the need for customers to proactively engage and are immediately visible to the whole household. They provide always-on, ambient feedback and are accessible to anyone who happens to be in the room. The benefits of these features are confirmed by the suppliers' qualitative and survey research. Supplier A's research shows that partners of the bill-payer are nearly twice as likely to engage with the IHD as with the app, and children engage with the IHD in 5% of homes compared to no homes in the app sample. Supplier B's qualitative interviews and survey research indicated a similar dynamic. This also strongly echoes earlier findings in the ELP and follow-up work by BEIS.

In Appendix 5 we propose a Theory of Change for IHDs and apps, highlighting some of the commonalities and differences in the energy saving mechanisms.

[29] It should be noted that CAD-enabled IHDs could also make use of push notifications, for example triggering notifications about consumption that suppliers send through their standard account management apps.

5. Conclusion

Though there is much overlap in their function, the IHDs in the suppliers' trials appeared to:

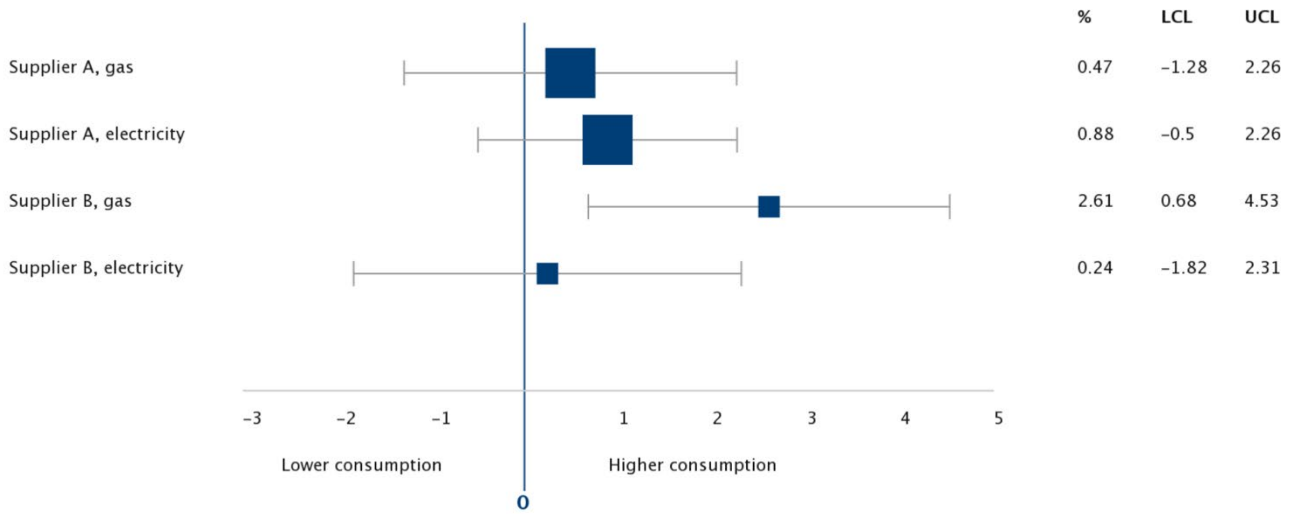
- be more readily adopted (fewer customers declined them, and they did not require downloading), in part due to different customer journeys;
- be suited to a wider demographic;
- be better at engaging the whole household; and
- have advantages at the beginning of the customer journey when members of the household were learning about their energy consumption.

This means the IHDs may have been effective at prompting engagement and creating motivation, even among relatively disinterested customers. In comparison, the apps in these trials may have:

- been better at providing tailored tips;
- had greater potential for novel features and ongoing innovation; and
- benefited in the long-term from push notifications, which may contribute to more sustained engagement and help encourage ongoing monitoring of energy expenditure.

However, despite these advantages, **the energy consumption data suggested that Supplier A's and B's apps were on average likely to save slightly less energy than their IHDs**. In Supplier A's trial, the differences between the groups' gas and electricity consumption were not statistically significant. However, the weight of evidence indicated higher consumption in the app group. In Supplier B's trial, the groups consumed roughly similar amounts of electricity, but the app group consumed significantly greater gas. The four results are summarised in Figure 3 below, with 95% confidence intervals indicated.

Figure 3: Electricity and gas consumption of app customers compared to IHD customers (in both suppliers' analyses)[30]



[30] The boxes representing Supplier A's trial results are larger than the boxes representing Supplier B's trial results in order to reflect the larger sample size in Supplier A's trial. Graph generated by DistillerSR Forest Plot Generator from Evidence Partners.

Given the differences in behavioural mechanisms, discussed above, the two technologies may complement each other well, but the apps did not, on average, appear to be good substitutes for IHDs for the general population.

These findings are based on, and fully supported by, the empirical evidence summarised in this report, though our interpretation is also informed by a wider understanding of behavioural science theory reflected in the Theory of Change in Appendix 5.

Finally, we reiterate that these conclusions are not definitive. The evidence is limited, and the differences in energy savings are small. Moreover, because there is uncertainty around the results, where the confidence intervals in three of the four cases indicate the app-using customers *may* have consumed the same or less energy than the IHD-using customers, the interpretation of the results depends on the burden of proof one places on the apps to demonstrate equivalence to the IHDs, as discussed in Section 4.1.

We also stress that this report focuses solely on the evidence for energy savings. We recognise there are many other considerations not discussed in this report that may ultimately feed into the policy decision.

This publication is available from: www.gov.uk/beis

If you need a version of this document in a more accessible format, please email enquiries@beis.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.