THE
BEHAVIOURAL
INSIGHTS
TEAM

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

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

July 2019

Contents

Appendix 1: Foundations for rigorous evaluation

A1.1 The nature of this research

The research question at the heart of the derogation is:

> ***Do IHD-alternatives reduce customers' energy consumption as much or more than conventional IHDs?***

More specifically, three high-level research questions have been identified focusing on energy consumption, engagement and customer experience:

1. Does the app lead to reductions in energy consumption equal to or greater than the IHD?
2. How do levels of engagement compare between customers using an app and those using an IHD?
3. Does the app have a positive impact on customer experience and how does this compare to the IHD?

These research questions provide a clear starting point, though each leads to a number of sub-questions: Energy reduction for whom? How might different features of apps lead to varying impact? And given that real-world research is invariably complex, what kind of evidence should be deemed sufficient to change national policy? This Appendix, written at the beginning of the suppliers' trials, aims to overlay some structure to these questions and outline our approach to field trials.

A1.2 Types of evaluation

It is useful to categorise our research endeavours into four types of evaluation:

- impact evaluation (on primary and secondary outcomes);
- segment analysis;
- process evaluation; and
- exploratory analysis.

Each are described below.

Impact evaluation questions – research questions that aim to understand the impact of the intervention (the IHD or app) on our outcomes of interest.

1. **Relating to primary outcome measures.** In this case, the primary outcome measure is energy consumption (electricity and gas). Questions of customer experience and engagement may be considered primary or secondary research questions depending on whether they are viewed as important and worthwhile policy objectives in their own right or secondary to the question of energy savings.
2. **Relating to secondary outcome measures.** Secondary outcome measures are those that are either of subordinate or peripheral importance (e.g. customer retention), and/or those that add second-order detail to primary outcomes (e.g. if the primary outcome is energy savings, a secondary outcome might

be the persistence of energy savings over time, or maintenance of thermal comfort in the presence of energy savings).

Segment (subgroup) analysis questions – research questions that illuminate heterogeneity in impact, for example differences in energy consumption between different segments of the population. In this case relevant segment analysis is likely to include differences in energy savings and customer engagement by age, income, type of house, region, and payment type).

Process evaluation questions – research questions that aim to understand the 'how' and 'why' of the intervention's impact, evaluating the mechanisms that lead to energy savings or other outcomes. An impact evaluation may give us a robust estimate of the average or aggregate effect of an intervention in a particular context. However, it may tell us little about the mechanisms of impact. Process evaluations aim to 'open the black box'.

In this instance a process evaluation may seek to understand:

- the learning mechanisms through which energy feedback operates;
- the impact of different installation experiences on subsequent energy savings and customer engagement;
- the differences in feedback mechanisms between 'always-on' IHD feedback versus app feedback which requires proactive engagement;
- potential causes for a lack of impact (such as not downloading the app); and
- the level of engagement with different features of the feedback.

This understanding of mechanisms allows us to infer a deeper understanding of the intervention, and thus an understanding of when, for whom, and in what contexts the intervention will/will not be effective.

A process evaluation may also have its own segment analysis associated with it - i.e. does the mechanism differ for different segments of the population? Process evaluations are also integral to the development of a Theory of Change, which aims to illustrate the mechanisms and logic of an intervention's impact.

Exploratory Analysis – analysis of data with no a priori hypothesis in mind, instead looking for trends that emerge, and/or testing of non-causal relationships in the data. Exploratory analysis is inherently less rigorous due to the very high risk of drawing false-positive conclusions, which are based on patterns within data that will always exist by chance: if we look for them with no prior hypothesis, we will find them. Exploratory analysis is therefore best used to develop hypotheses for future research, or as heavily caveated supplementary detail where it adds contextual insight to pre-specified research questions. We would rarely accept exploratory findings as rigorous evidence in their own right.

Some research questions may fall across multiple categories. For example, if we are interested in customer satisfaction for its own sake, it would be part of an impact evaluation, but it may also be part of a process evaluation if we think customer satisfaction is an important mechanism through which energy savings emerge from IHD use. The fact that a research question serves two types of evaluation does not imply the distinction is false – each has its own clear purpose in mind. Moreover, we

should avoid conflating the type of research question, outlined above, with the methodology of enquiry. This may, for all types of research question, be qualitative or quantitative, and may draw upon any number of research designs and research tools including field trials, quasi-experiments, lab studies, customer surveys, data science, and ethnography.

A1.3 List of research questions

Based on the above categories of research, we identified the following research questions as relevant to the policy questions relevant to the IHD derogation. This list is not exhaustive, but it aims to capture all *types* of question such that additional variations not mentioned should be self-apparent and fit within this categorisation (for example, segment analysis could be undertaken on every conceivable household characteristic, and secondary impact evaluations could be undertaken on every conceivable peripheral outcome, but it would not be useful to list them all here). No exploratory analyses are listed, as by definition exploratory analysis is not pre-specified, though may still be a valid component of the suppliers' evidence.

Impact Evaluation Questions	
Primary	• What is the relative impact of IHDs versus apps on gas/electricity consumption?
Secondary	• What is the relative impact of IHDs versus apps on customer engagement? • What is the relative impact of IHDs versus apps on customer experience (satisfaction)? • Are impacts sustained over time (and do IHDs and apps differ in this regard)? • What is the impact of IHDs and apps on peripheral outcomes (including: propensity to switch tariff, subjective comfort, customer retention, purchase of other products, adoption of other pro-environmental behaviours, etc.)?

Segment Analysis Questions	
	• How does the impact of IHDs versus apps differ for different users based on: - Age - Household tenure - Attitudes to new technology

	- Socio-economic status - Household make-up (student, family, etc.) - Prior energy consumption (i.e. do those with the greatest potential savings save the most?) - Early/late adopters of smart meters - Prepayment versus credit accounts

Process Evaluation Questions (split into 4 key parts of the customer journey)	
Regarding the installation experience and suppliers' support	• To what extent is a positive installation process beneficial or necessary in leading to greater engagement with the feedback and ultimately changed energy consumption? • What is the impact of suppliers' post-installation support or advice? • Segment analysis of this process evaluation: Do certain demographics benefit more from greater support at installation or post-installation?
Regarding level of engagement with the feedback	• To what extent are customers engaging with the feedback provided by the IHD versus app? • To what extent do the devices elicit engagement from multiple residents? • How does this engagement change over time? • Are there certain events which disrupt engagement and which differ between devices? For example, if the IHD is unplugged and not plugged back in, or a change of phone for the app? • To what extent is engagement with, or use of, each technology a predictor of actually reducing energy consumption?
Regarding the learning and monitoring mechanisms	• How does the medium of feedback provision (e.g. utility bills, apps, IHDs) impact learning? • How important to driving energy savings is an IHD's 'always-on' feedback compared to the need to proactively engage with an app? • What novel features can 'power' the benefits of energy feedback (e.g. gamification, social norms/comparisons, disaggregation, frequency of feedback, alerts and prompts, colour-coding, budgeting tips etc.), and do these differ between apps and IHDs?

	• What specific behavioural barriers and motivators exist and how does the feedback address/harness them (e.g. forgetfulness, procrastination, lack of awareness)? • Are some impactful features inherently better or more feasible on apps than IHDs (and vice versa)? • Do certain features tend to lead to more sustained engagement, and more reliable behaviour change, over medium-long timeframes, and if so why? • Are there interaction effects between features, i.e. a combination of features is better than the sum of their individual impact? • To what extent might integration with other smart technology boost engagement and behaviour change? • Do IHDs and apps serve the same purpose in learning and monitoring, or are they providing distinct functions? Are they therefore interchangeable, or complementary to each other?
Regarding conversion of learning to actual behaviour change	• What are the most common behavioural changes being made? • What are the key differences between one-off purchasing decisions (e.g. buying a new boiler or insulation) and day-to-day habit changes (using heating more effectively or closing windows)? Does feedback tend to lead to one or the other? • Are there positive spillover effects to other behaviours, or negative rebound effects? • Do customers already know what to do to reduce their energy, or is IHD feedback insufficient without this additional information and guidance? • If customers are engaging, what factors determine whether behaviour change follows (e.g. know-how, busyness, motivation, savings potential, uncertainty, frictions and hassle, procrastination)? • Does additional support of any kind, from suppliers or elsewhere, help to overcome the intention-behaviour gap? • Are there particularly timely moments when behaviour changes? • To what extent does active engagement with smart metering encourage customers to seek out other sources of information and guidance? What are the principal sources? • What type of customer use is associated with energy reduction (e.g. more frequent engagement)? More sustained

	engagement? More informed engagement? ● What barriers do IHDs not overcome? Could these be overcome through IHDs/related technology with design changes?

Appendix 2: Interpreting the primary results

As discussed in Section 4.1, interpretation of the primary results from both trials is not straightforward. Any estimate of the difference in energy consumption between the two groups will contain a degree of uncertainty. In this case, since the differences we are trying to detect are small, and the confidence intervals wide, the range of possible 'true' values may span both sides of zero. For instance, the Supplier B electricity result suggests that the app users consumed fractionally more electricity, but the 95% confidence bands range from -0.18kWh/day to +0.20kWh/day (-1.9% to +2.3% of customers' average daily electricity consumption). Indeed this is the definition of statistical significance: a result would only be deemed significant (at the 5% level) if the 95% confidence intervals do not cross zero.

This is important because the conventional requirement of 95% confidence is deliberately burdensome: it puts a strong assumption on two populations (for example, those with the app and those with the IHD) consuming the same amount of energy, unless we are 95% confident that they are different. Under normal circumstances this is conservative (e.g. we conclude a drug does not work unless we find strong evidence of a difference in outcomes between the treatment and placebo groups). Thus conventional statistical tests are ordinarily employed to ascertain the veracity of a *difference* between populations or datasets. As such we conventionally start with the assumption (null hypothesis) that two datasets are the same, and only reject this null hypothesis if strong evidence is found to show that there is indeed a difference.

However, in this case, this convention is advantageous to demonstrating that an app is equivalent to an IHD, since this conclusion would align with the weight of presumption that the two populations' energy consumption is the same. Under these statistical tests, it would clearly be disingenuous to conclude that the two devices are definitely equivalent just because we failed to reject the null hypothesis. It is for this reason that null hypotheses should always reflect the opposite state to that which the analyst wishes to prove, or be confident about.

Our understanding of the derogation is that it puts a burden of proof on alternatives to IHDs to demonstrate 'equal or greater' energy savings to IHDs. We therefore need an approach that tests for *strong evidence of equivalence*, rather than testing for strong evidence of difference against an *assumption of equivalence*.

To summarise: 'no strong evidence that they are different' is not synonymous with 'strong evidence they are the same'. Convention is the former, but we are looking for the latter.

We therefore require a statistical test of equivalence in order to reject a null hypothesis that they are different. We outline a simple form of this analysis below.

A2.1 Testing for equivalence

A variety of statistical approaches can be employed for this purpose. For example, the two-one-sided-t-test (TOST) has been used to test for equivalence between

samples in a variety of contexts including pharmaceutical science, engineering, environmental science, psychology and medicine.[1] We do not advocate using t-tests in this analysis, as they do not enable the analyst to control for covariates or analyse panel data. Nonetheless the principles are valid: the null hypothesis is that two mean values are different, and the test attempts to demonstrate equivalence within a pre-set threshold of acceptability (the conceptual opposite to a normal two-sample t-test). In this way TOST and other similar methods appropriately guard against poor precision, noisy data, or small samples that would otherwise favour an analyst looking for equivalence.

There are two key steps to undertake when looking for evidence of equivalence. First, we must define an acceptable threshold of equivalence. By way of example, when undertaking this analysis BEIS could set a threshold of 0.5% of energy consumption. In other words, if the IHD saves 2% of energy, we could deem the app to be equivalent if its savings are at least 1.5%. We determined this to be a reasonable threshold; taking into account that the savings of IHDs are modest to begin with, a larger shortfall than 0.5% could render app-based IHD alternatives of limited benefit.

Second, we must set a probabilistic threshold, or confidence level. This is our required level of likelihood that the app's energy savings fall within the suggested margin of 0.5% of the IHD's energy consumption. For example, convention would set a 95% confidence level, i.e. we must be at least 95% sure that app users' energy consumption is not more than 0.5% higher than IHD users' energy consumption. Given that energy data tends to be noisy and has high variance, 95% is likely to be punitive for all but the most highly powered studies: in order to pass this test, we would either need an impractically large sample size, or the app would need to be substantially better than the IHD just to convincingly demonstrate that it is not worse. Again, this is a decision for the analyst based on the context. By way of example, in this analysis we use 80% as our required confidence level.

To summarise, these two thresholds should be set by completing the statement 'we require at least X% certainty (likelihood) that app users' energy consumption is within Y% of that of IHD users' energy consumption. In our example we set X at 80% and Y at 0.5%. Y could also be different for electricity and gas consumption.

Formally, this means our null hypothesis is that the app leads to energy consumption at least 0.5% higher than the IHD (H_0: $\beta > 0.5$), and our alternative hypothesis is that the app leads to energy consumption that is within 0.5% of the IHD or better (H_1: $\beta < 0.5$). We reject the null hypothesis at a significance level of α=0.2, i.e. an 80% confidence level.

[1] Limentani, G. B., Ringo, M. C., Ye, F., Bergquist, M. L., & McSorley, E. O. (2005). Beyond the t-test: statistical equivalence testing.

In undertaking this simple test of equivalence, if the point estimate savings figure from Supplier A's and Supplier B's regression analyses lie outside our 0.5% threshold, we can immediately conclude that the app does *not* pass the test of equivalence, since this implies the likelihood under H_0 is higher than the likelihood under H_1 (and since we are looking for at least 80% likelihood under H_1, this is a long way from adequate). This is the case for Supplier A's electricity consumption, since the point estimate of the app is 0.88% higher than the IHD (worse than 0.5%). In plain English: it is highly likely that the app is at least 0.5% worse than the IHD. Even with substantially more lenient thresholds of equivalence or of required confidence, we would lack strong evidence that the app is equivalent to the IHD on electricity consumption.

For Supplier A's gas consumption, our point estimate is 0.47%. This is within 0.5%, suggesting it is more likely than not that the app meets our required equivalence threshold. But *how* likely (noting that we require 80% confidence)? We can reject the null hypothesis and conclude the app is equivalent to the IHD *only if* {point estimate + 1.28* standard error of point estimate < 0.5}. In this case the app is still some way from passing this test. This should be intuitive since the point estimate is so close to the 0.5% threshold that there will be fractionally more than 50% likelihood under H_1, but not 80% as required. More lenient thresholds would paint a more positive picture for the app, but we would need to relax these thresholds considerably in order for the app to be deemed equivalent to the IHD on gas consumption (and even more so on electricity consumption).

The same process can be used for Supplier B's results. Here we see that gas consumption is significantly higher in the app group, and so unsurprisingly it fails the equivalence test. Supplier B's electricity consumption is the strongest of the four primary results, suggesting app users and IHD users are near identical in their consumption. However, the confidence bands are quite wide (95% confidence from 1.9% lower to 2.3% higher) and so there is still a possibility that app users' consumption is higher than the permissible 0.5% threshold. In this case, we can be *fairly* confident that the app consumption is within the 0.5% margin, but not 80% confident, so, again, it fails the test of equivalence. This raises the question of whether 80% is too burdensome - this could be relaxed. However, even then the other three results would likely fall short.

Synthesising evidence across the trials, we therefore conclude that apps and IHDs are *unlikely* to be equivalent, with apps *likely* to be inferior.

To conclude this section, we have presented an example analysis based on what we believe to be a reasonable interpretation of the derogation: apps should be deemed equivalent to IHDs only if we can be highly confident (say, 80%) that their energy consumption is close (say, within 0.5%). The evidence presented by Supplier A and Supplier B, though in 3 of 4 analyses failing to find a statistically significant difference between the consumption of app customers and IHD customers, fall short of this requirement. However, we note this interpretation is subject to the evidence requirements set.

Finally, we stress that this discussion pertains only to the interpretation of the primary results in Supplier A's and Supplier B's trials. There are other aspects of both sets of

results, discussed throughout the report, which provide further arguments both in favour of and against the use of apps.

Appendix 3: Technical details for Supplier A's trial

In this Appendix, we provide supplementary detail on Supplier A's derogation trial, in support of Section 2 of the main report. This is not a standalone summary of the trial, as it only provides *additional* detail to that summarised in Section 2.

A3.1 Balance Checks

Balance checks aim to determine if the two groups are equivalent on observable characteristics. Supplier A ran balance checks separately for the ITT groups, the ATT groups and for 'active device users', the latter used in exploratory analysis only. They ran Chi-squared tests on the following variables: number of rooms in property, tariff type (fixed, standard, or collective switch), paperless bill status, loyalty score, property council tax band, age, tenure with Supplier A, projected electricity consumption, projected gas consumption, and five marketing segmentations based on customer postcodes[2].

Supplier A found significant imbalance on the following variables:

ITT balance checks: Significant difference found on MPAN (geographic area), and tenure.

ATT balance checks (conducted every week to see how balance evolved as the sample changed due to attrition): Significant differences were found on number of rooms, MPAN (geographic area), and age. The average number of rooms was estimated from postcode data, so imbalances on this measure may be being driven by imbalances on MPAN. Age was particularly interesting, as no imbalance existed at the beginning of the trial. Imbalance on age only emerged through the course of the trial, with older customers more likely to attrite leaving a slightly younger sample in the app group. It is not clear what the cause of this was.

Active users balance checks: 'Active users' was defined differently in each group. For the app group, Supplier A defined them as having logged in at least once in the last month of the trial. For the IHD users, they were defined as having the IHD plugged in at least once over the course of three status checks, which occurred when a firmware update was sent (in February, August and November). Supplier A found significant differences in MPAN area, age, tenure with Supplier A, tariff type, paperless bills set up, and historical gas and electricity consumption. Given that these groups were defined in different ways, imbalance is to be expected (though still undermines our ability to compare the two groups).

A3.2 Customer surveys

[2] Variables such as age, tenure, projected electricity consumption, and projected gas consumption were put into bands in order to investigate balance using a Chi-squared test.

Four surveys were undertaken throughout the trial. The first three targeted a random subset of each group, and the fourth was sent to all participants remaining at the end of the trial. This received a ~20% response rate, yielding 540 responses in the app group and 544 in the IHD group. All survey responses were screened with an initial question to include only those who had the app or IHD installed. Survey responses were weighted based on energy consumption bands. Six IHD and six app users also participated in 60-minute, in-situ (at home) interviews, split across London and Scotland, aiming to capture a breadth of property types and demographics.

A3.3 Regression model

The available data for the analysis were weekly consumption data from smart meters, covering a period of one year after installation. Data prior to installation was from classic (dumb) meters. Covariates were included in the regression to control for past consumption, weather variation, and geographic and household characteristics.

The analysis estimated a random effects model Generalised Least Squares (GLS) estimator. This was used to account for the clustering in the errors within each individual consumer. In other words, while we have weekly observations for each customer, we cannot assume that these observations are independent of each other. The GLS estimator adjusts for this serial correlation within observations.

The adjustment in the analysis was done by estimating the error components for each of the participants, as well as the autocorrelation within each participant. The Supplier A team opted to not use standard statistical packages to run the analysis because of concerns over missing data. Because this approach was unconventional, Supplier A also ran robustness checks using standard statistical commands to ensure that the results they shared were not particularly sensitive to the choice of regression specification; they found negligible changes to the results.

To deal with outliers, the analysis tested a number of methods to remove outliers and each method produced similar results.

To estimate the effect of the treatment, Supplier A reported both an Intention to Treat (ITT) and an Average Treatment effect on the Treated (ATT). More precisely, Supplier A ran a Complier Average Causal Effect (CACE), rather than a true ATT (for simplicity, we have referred to this as ATT elsewhere in this report). The ATT analysis was used to address the issue of non-compliance with the treatment assignment, namely that some customers in the IHD group will install the app and vice versa.

To estimate the ATT, the analysis used an instrumental variables (IV) approach. This consisted of two stages:

- In the first stage the compliance with the treatment was regressed on the treatment assignment and a number of other available covariates.
- In the second stage, the outcome variable was regressed on the fitted values of compliance from the first stage.

In the standard IV approach, the same covariates are used in both the first and second stage. However, Supplier A used a different model in the second stage. We

would generally advise that the same model is used in both, though in this instance the ITT was the primary result, so the model specifications are not critical.

A3.4 Active versus non-active users (exploratory analysis)

In addition to the primary energy analysis across the full sample, Supplier A also undertook exploratory analysis looking at those who are 'active users' of the app and IHD, finding that active users appear to use less energy than non-active users.

Comparison	Gas	Gas p-value	Electricity	Electricity p-value
Active app user versus non-active app user	-1.78%	0.1808	-2.38%	0.0228
Active IHD user versus non-active IHD user	-4.13%	0.0004	-0.49%	0.5665
Active app user versus active IHD user	+0.53%	0.6908	+0.20%	0.8458
Non-active app user versus non-active IHD user	-1.82%	0.1594	+2.09%	0.0341
Active app user versus non-active IHD user	-3.60%	0.0130	-0.28%	0.7993

BIT evaluation of exploratory analysis findings

Supplier A's suggestion that active users use less energy than non-active users is a reasonable conclusion to draw from this exploratory analysis. However, the analysis should not be viewed as providing causal evidence since active-users and non-active users are fundamentally different, and there is no manipulation of an independent variable being applied in this analysis. This exploratory analysis therefore says nothing about the impact of each intervention on customers' engagement or degree of active use. We also note that 'active user' is defined very differently for each group due to different available data, so any comparison between IHD and app users along these boundaries is inappropriate.

A3.5 Customer engagement with the web portal

Supplier A measured the number of web portal logins for both groups before and after intervention, on the basis that 'reductions in web logins can be attributed to an increase in understanding of energy consumption from the allocated tracking device'. Supplier A observed that fewer app users log in to the web portal, and thus inferred high engagement with the app.

BIT evaluation of customer engagement with the web portal

It is important to clarify what can and cannot be inferred from this result. We understand that the app provides the same information as the web portal, while the IHD does not. We can therefore conclude that those provided with information on the app do not feel the need to log in to the web portal for the same information. Those who were not provided this information on the IHD, still feel the need to log in to the web portal for this information. This is a straightforward and unsurprising finding. As Supplier A remarked, this 'highlights the potential of the mobile app as a platform to provide an integrated domestic energy management experience'. We agree with this statement. However, we can infer nothing from this analysis about levels of 'engagement'. The IHD users, who are continuing to log in to the web portal, may be less engaged, equally engaged or more engaged with their device and with their energy consumption, with no evidence here supporting any of these conclusions.

A3.6 App engagement

The ability to set and use 'goals' was a feature launched on the app midway through the trial, with uptake rates of this feature recorded. Uptake remained quite high throughout the remaining trial period, at 40-50% of users. Broader app usage showed general decline over time. Supplier A noted a slight increase in engagement in November, again possibly reflecting a 'winter effect' as engagement increases with rising energy costs. However, this upward trend was slight, could be random fluctuation, and in any case was modest relative to the broader downward trend.

The proportion of app participants using the app also went down month on month, from around 60-65% at the beginning of the trial to 40-45% at the end of the trial, though Supplier A noted that this was still a high proportion of users. The average length of time spent per session on the app also declined from around 200 seconds at the beginning of the trial to 70 seconds in the late summer, 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 in diminishing numbers).

One of the clear benefits of an app is the ability to send personalised information, energy saving tips and push notifications. Supplier A's app had this functionality; opening rates of notifications averaged 38% across the trial (with this number on a

gradual downward trend as the trial progressed). The system was modified part way through and the process for recording this engagement was altered. In the second half of the trial, read rates of the messages typically varied between 10% and 40%, depending on the message, averaging at 26%.

Appendix 4. Technical details for Supplier B's trial

In this Appendix, we provide supplementary detail on Supplier B's derogation trial, in support of Section 3 of the main report. This is not a standalone summary of the trial, as it only provides *additional* detail to that summarised in Section 3.

Supplier B noted that, for the customers in the IHD group, it was not known what proportion took up the offer of a physical display when their smart meter was installed. Similarly, they did not exclude any customer in the app group who did not use the mobile phone app. We agree with this analytical approach. The appropriate research question concerns the effect of the devices' availability; in no realistic scenario would either technology be forced onto people. That said, as discussed in the main body, unlike Supplier A's primary analysis, Supplier B's analysis is not a true Intention-to-Treat (ITT). A majority of the customers whom Supplier B offered an app did not end up receiving one (through installer error, customer preference, or ineligibility); Supplier B excluded these customers from analysis. Put differently, Supplier B's analysis of the app's effect on consumption included only those customers who received their allocated device (regardless of how they used it).

A4.1 Matching

As discussed in the main body, Supplier B separated their customers into six 'residential segments'. Residential segments are a standard segmentation that Supplier B have created for marketing and communications purposes. The six segments were mostly based on age and estimated income. Within each residential segment, they matched app customers to their nearest neighbour by raw previous consumption, for electricity, or previous consumption divided by heating degree days (HDDs) in the relevant period, for gas. They used R's MatchIT package to conduct this matching process, using the Mahalanobis distance metric. They did this separately for electricity and gas, meaning that an app household could be paired with a different IHD household depending on the fuel being analysed.

Supplier B considered various matching strategies. They wrote in their final report, 'Initially we considered using Exact Matching but the number of exact matches would have been small. Coarsened Exact Matching was not used due to difficulties implementing that approach. Propensity scores were ruled out because we didn't believe that any model we built to predict whether a customer would choose to be in the app group would be sufficiently powerful.'

Supplier B checked balance after matching by conducting Chi-square tests on various demographic variables they had for customers. They did this separately for the gas consumption analysis groups and the electricity consumption analysis

groups.[3] In both cases, though balance had improved on most variables, statistically significant imbalances remained on tenure with Supplier B, number of bedrooms in property, age band, property type, council tax band, and tariff group. The groups had become balanced, insofar as differences that remained were not statistically significant, on number of adults in the household and residential segment.[4]

A4.2 Regression specifications

Supplier B used R's LMER function to conduct random-effects regressions on their customers' consumption data. They specified the following covariates:

- HDDs per day for each month of gas consumption, or a continuous measure of Average Daylight Hours (ADH) per day for each month of electricity consumption.
- Pre-installation consumption (for gas, divided by HDDs).
- Days since installation (and days since installation squared, in case learning effects were nonlinear).
- A binary variable for each month, and an interaction term for month and HDD/ADH during the month (because people's response to a given temperature may depend on the time of the year). The month used is the midpoint of the opening and closing read.
- Household characteristics: tenure with EON, number of adults in household, age, number of bedrooms, council tax band, tariff group, and property type.

For the gas analysis, Supplier B faced a choice between using unadjusted daily consumption (kWh/day), or using consumption per HDD as the dependent variable. BIT recommended using unadjusted daily consumption as the dependent variable, with average HDD/day as an explanatory variable. The reasoning was that this regression design better allows for the existence of a baseload of gas consumption not correlated with temperature. Supplier B agreed with this recommendation.

A4.3 Surveys and qualitative interviews

Supplier B contracted a market research agency to conduct their qualitative interviews and surveys. As discussed with Supplier B during their research, we treat the survey results with some caution due to low sample sizes and selection bias issues. For example, one concern that BIT had about Supplier A's surveys and interviews was that the app customers that Supplier A had interviewed and surveyed

[3] This was necessary because the matched IHD group differs in the two analyses (since app customers are matched with IHD customers separately for the gas and electricity analyses).

[4] Tenure, number of bedrooms, age, property type, council tax band, tariff group, and number of adults in household were all tested as categorical variable (where nominally continuous variables were put into bands in order to investigate balance using a Chi-squared test).

were a self-selected subset that had accepted and installed the app. This concern also applies to Supplier B's survey and qualitative research.

Another important piece of context for Supplier B's qualitative interviews is that various customers seemed to be confused about the functioning of their app or IHD (depending on the group to which they were allocated). For example, various app customers discuss trying to obtain information from their CAD, perhaps by analogy to their neighbours' IHDs, and they express frustration at the CAD failing to work as they expect it to. In discussing lessons learned about how the IHDs and apps affect customers' behaviour, Supplier B largely ignore these points of confusion. However, it is useful to bear in mind that at least some customers in both groups misunderstood how their devices worked.

Appendix 5. Building a Theory of Change

One of the key deliverables from this work is a deeper understanding of how customers use energy feedback, and the mechanisms through which behaviour change emerges. Developing this understanding is a process of theory-based evaluation and evidence synthesis. We aim to develop a Theory of Change to infer insight not only into whether, but also *how, why, and under what conditions* traditional IHDs and alternatives – here, we focus on app-based IHD alternatives – lead to changed energy behaviours. This requires us to combine robust impact evaluations (e.g. from randomised controlled trials) with supplementary quantitative and qualitative process evaluations, for example capturing intermediary measures of the experience of customers. This was the basis of the IHD derogation research guidance, which required suppliers not only to undertake robust impact evaluations, but also to explore customer engagement and experience to help 'open the black box' of energy feedback devices.

The below Theory of Change provides a mechanistic model of the impact of physical and virtual IHDs on energy consumption. It has been created by:

- reviewing all evidence summarised in this report, focusing on the impact of energy feedback and evidence of the processes through which it functions;
- drawing upon existing theories of behaviour change and the wider behavioural science literature, with a particular focus on the well-studied mechanisms of learning, prompts and feedback, and disruption to habit; and
- ongoing conversations and two workshops with BEIS.

As with any model, it is intended to be useful, though not necessarily perfectly accurate. Behaviour change is complex and cannot be fully described with a simple flowchart. Our model aims to approximate the mechanisms of behaviour change that consumers must pass through, and it highlights the key assumptions upon which these rest. It therefore provides an overarching framework through which we can interpret individual research findings, and provides a tool for identifying intervention ideas, developing new hypotheses to test, or diagnosing an intervention's success or failure. Several features of the Theory of Change are worth highlighting:

- The 'activities' (in grey) represent different features of the intervention. For example, the literature reviewed in this report highlights the importance of not only the device, but also of the installation visit and advice received at that point. Customers may also be influenced by peripheral marketing activity, for example that undertaken by Smart Energy GB (SEGB). These multiple aspects of the intervention are reflected in the four activities in the model.
- Moderating factors (in pale yellow) are all those external factors that impinge on the efficacy of the intervention, but which are not under our control and are generally static. We have presented these as personal and household factors. Not included here, but also potentially relevant, are wider socio-economic factors.
- All key steps a customer passes through before changing their energy consumption behaviour are indicated in blue. These are the 'mechanisms' of behaviour change. Mechanisms in purple are additional steps specific to

apps. Note there are only three additional steps: downloading the app to begin with; proactively using it in order to learn about and monitor energy consumption; and the existence of push notifications, which 'boost' the feedback loop involved in monitoring energy consumption. However, as revealed in the literature, these additional steps are very important: the first two are significant 'frictions', which reduce the device's energy saving potential, while Supplier A's and Supplier B's research suggests the third might be an advantage of apps, among those who use them.

- In green, assumptions are identified as they relate to the mechanisms above them. The boundary between what is a 'mechanism' and what is an 'assumption' is not always obvious, and interventions should consider both: behaviour change may fail to emerge either because a customer is not passing through a particular mechanism of change, and/or because a particular assumption is not being met.
- As identified by the ELP research, there are two distinct uses of an IHD. First, customers may explore their energy consumption and identify causes of high energy use through an 'initial learning period' (for example, turning appliances on and off and observing the impact they have). Secondly, with this knowledge in hand, customers may use their IHD as an 'ongoing monitoring' device, tracking daily, weekly, or monthly spend, and monitoring the improvements that come about as a result of behaviour change. These two mechanisms are reflected in the Theory of Change. The first initial learning period is an ongoing cycle of exploration, understanding the feedback in response to that exploration, and thus identifying sources of wasted energy. This process approximately mimics the Kolb learning cycle, a seminal theory of learning from the behavioural science literature.[5] The ongoing monitoring process is a longer-term activity of observing energy consumption in response to external events (e.g. changing weather), and to validate savings in response to changes in behaviour. The wider behavioural literature identifies this kind of validation as critical to, one, reinforce the sense of control and self-efficacy (a sense that our actions will generate the desired outcomes), and two, further motivate our intentions to do more. This is true, for example, with weight loss (which would be difficult without any feedback on progress, such as through a set of scales, or simply observing changes in our body). As such, in this model the monitoring process is represented by a positive feedback loop feeding into the mechanisms of control/self-efficacy, and intention formation. Only after monitoring the impact of our behaviour changes repeatedly, and gaining this sense of validation that our actions are worthwhile, might we begin to form new, lasting habits. Habit is based on repeated cues, actions, and rewards. The IHD provides the cue (we are

[5] Kolb, D. A. (2014). Experiential learning: Experience as the source of learning and development. FT press.

prompted to act when we see the red light, for example), we undertake an action (turning an appliance off), and the IHD provides the reward (the red light turns green, or monthly spend goes down).

- There are two distinct energy saving outcomes, displayed separately in the model. The first is changes to regular, ongoing behaviours such as turning off lights, putting less water in the kettle, turning the thermostat down, or taking shorter showers. The second is one-off decisions or investments, such as choosing to buy insulation or a smart thermostat. These two possible types of outcome are represented by two separate paths in the model. Each has its own mechanisms, and distinct behavioural and practical barriers to overcome.
- The entire model is predominantly concerned with deliberate, reasoned action (i.e. intention-driven behaviours). This follows the convention of many well-known models of behaviour change, including the theory of reasoned action and theory of planned behaviour. We note that behaviour is often driven by many factors not included in this model, including social norms, external motivators, such as price and choice architecture, and less cognisant mental processes, such as bias and emotion. These are beyond the scope of this model. Though they are highly relevant to the objective of reducing energy consumption, they are not central to the functioning of IHD feedback, which operates largely through these cognisant and intention-driven processes. In other words, this model describes how an IHD might lead to changes in energy consumption, and ignores other routes to the same end.