Guidance on conducting energy consumption analysis

Created by the Behavioural Insights Team on behalf of the Department for Business, Energy and Industrial Strategy

July 2020

Contents

1. Purpose of this guidance

The Behavioural Insights Team (BIT) has developed this guidance for energy suppliers on behalf of the Department for Business, Energy and Industrial Strategy (BEIS).[1] It specifies key considerations for carrying out energy consumption analysis, BIT's recommended methodology, and alternative methodologies. Our intention is to provide suppliers and other organisations with an accessible, quality-assured method that can be tailored to their data and capacity.

This guidance focuses on suppliers' identification of the impact of smart meters on customers' energy consumption. Reduced energy consumption is a key anticipated benefit of the smart meter roll-out, and ongoing access to this evidence is valuable for helping the Smart Metering Implementation Programme (SMIP) monitor roll-out impacts.

However, the methodology we recommend may also be useful for analyses of other products, services, or interventions that energy suppliers offer customers where customers' energy consumption is an outcome of interest.

[1] The Behavioural Insights Team was founded in the Cabinet Office as the world's first organisation dedicated to the application of behavioural science to public policy and service design. We are now a social purpose consultancy, with expertise in the robust evaluation of policy interventions, products, and services.

Andrew Schein, Daniel Bogiatzis-Gibbons, and Tim Hardy wrote this guidance. Toby Park and James Lawrence provided quality assurance.

2. Analytical decisions

The below table outlines the decisions on analytical strategy that evaluators must make when examining the impact on energy consumption of smart meters, and our recommendations. It provides a high-level summary of the rest of the document's guidance.

Table 1: BIT's recommendations on key analytical decisions

Decision	BIT's recommendation
Creating a comparison group	Create a comparison group by matching on the following characteristics (listed in the order of the importance we ascribe them): 1. Previous energy consumption/year 2. Region 3. Other household characteristics, e.g. number of bedrooms The main analysis concerns dual-fuel credit customers. Other analyses should look at other customer types separately.
Post-installation period length	Use at least 12 months of post-installation consumption in calculating consumption/year and comparing between groups. Use the same time period (e.g. April 2020 through March 2021) for both groups.
Analysing a 'window' of installs	Use a 3-month window of smart-meter installations to construct your group of smart-metered households. *Use a longer installation window, if necessary to achieve sufficient sample size.*
Sample size	Ensure at least 10,000 customers for analyses (5,000 smart-metered households, and 5,000 households in the comparison group).
How to match on previous energy consumption/year	Use bands of ±50 kWh/year for electricity and ±200 kWh/year for gas analyses. *Widen these bands if sample size would otherwise be insufficient. We recommend using a maximum band size of 200 kWh/year for electricity and 800 kWh/year for gas*.
Region matching	Match on outer postcode (e.g. SW1H or AB12). *Use larger regions, such as Public Electricity Supplier area, instead, if sample size would otherwise be insufficient. We recommend prioritising a narrow match on consumption/year over having small regions.*
Matching on other characteristics	Match on further variables, such as number of bedrooms, if available and if sample size remains sufficient.
Outcome of interest	Calculate households' percent change in consumption (their post-installation minus their pre-installation consumption, divided by their prev-installation consumption).
Comparing consumption between the two groups	Compare smart-metered households' percent change to traditional-metered households' percent change using a paired t-test or regression analysis. Report the 95% confidence interval on the difference in percent change between the two groups.

3. Principles guiding the recommendations in this document

The critical step in any causal analysis is estimating the counterfactual.

- Hal Varian, Chief Economist, Google[2]

The challenge in identifying the impact of smart meters on customers' energy consumption is to compare customers' post-installation consumption to a valid counterfactual consumption. Estimating this counterfactual – the amount of energy the household would have consumed had they not received a smart meter – requires care to avoid confounding the smart meter's impact with other factors that may affect energy consumption. For example, a simple comparison of the consumption of households with and without smart meters could over- or underestimate the impact of smart meters should the two groups' baseline consumption levels differ.

If suppliers allocated smart meters randomly, this randomisation would create groups balanced on all factors other than meter type (on average).[3] The impact of the smart meter could then be identified by evaluating the difference in consumption between the two groups. In reality, smart meter allocation has been influenced by customer attitudes, eligibility criteria, and suppliers' marketing and targeting strategies. For this reason, evaluators must use quasi-experimental methods to identify the impact of smart meters.

What we recommend in this guidance is to 'build' the comparison group, comparing smart-metered customers to traditional-metered customers who have similar characteristics known to be associated with energy consumption. Specifically, we recommend matching on previous consumption, region, and number of bedrooms in the household. See Section 7 for further details on how we recommend creating this matched comparison group.

In addition, instead of calculating the difference between the two groups' post-installation consumption/year, we advocate calculating the difference between the groups' percent *change* in consumption/year. This 'difference in differences' approach is a further assurance that differences in baseline consumption do not confound identification of the impact of smart meters.

---

[2] Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, *113*(27), 7310-7315. https://www.pnas.org/content/113/27/7310

[3] BIT and BEIS are aware that some suppliers have waitlists for smart meters. Assuming that exit from waitlist to installation is truly random, this scenario sets up a 'natural experiment', and it *would* be valid to analyse the impact of smart meters on customers' energy consumption by comparing waitlisted customers who received a smart meter and those still on the waitlist. Indeed, such an analysis would have higher internal validity than any quasi-experimental method, including the methodology described in this document. That said, this natural experiment would likely identify only short-term impacts of smart meters given that customers presumably exit suppliers' waitlists relatively quickly, and results may not be generalisable to those who do not proactively join smart meter waitlists.

4. Creating the group of smart-metered households

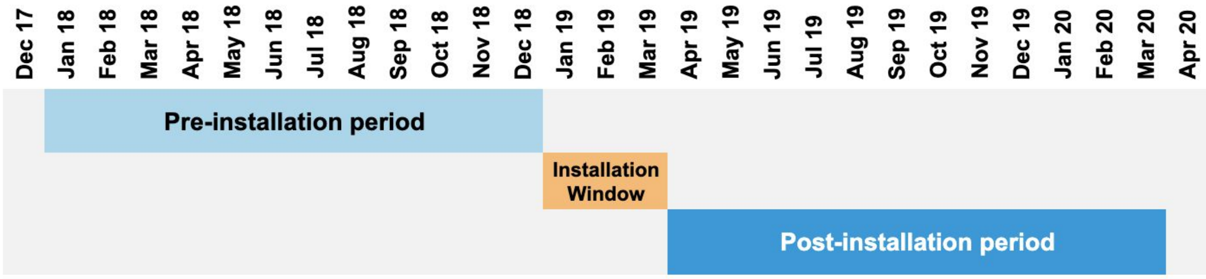
As a rule of thumb, we recommend analyses have at least 10,000 customers (5,000 smart-metered households, and 5,000 households in the comparison group). Energy consumption is influenced by many factors and varies significantly across the population over time. As a result, measurements of a population's consumption based on small samples are unreliable because there can be substantial differences in the consumption of two small samples drawn from the same population.

A larger sample will always be advantageous, and, where possible, we advise running power calculations using a sample of your data. However, by way of example, a supplier who shared their analysis with BEIS using consumption data from 10,000 customers calculated smart meter impacts with confidence intervals of ±0.3% at the 95% confidence level. Smaller sample sizes than this will generally have larger confidence intervals.

We recommend using a time-limited window of smart meter installations to create the group of smart-metered households. This can be a 3-month period, a 6-month period, a 12-month period, etc. We recommend using a 3-month period, if sample size permits. Shorter periods will allow suppliers to conduct analyses on a regular basis and analyse changes in impacts of smart meter installation across time.

For example, analysts working in June 2020 might choose January 2019 through March 2019 as their window of installations to determine the group of smart metered households.

Figure 1: A simple diagram showing an installation window January 2019 through March 2019.



To obtain an analysis with 5,000 smart-metered households, we recommend that suppliers ensure that the 'installation window' include approximately 9,000 or more smart-metered households because some of them will be excluded in later steps in the methodology we recommend in this document. If a 3-month installation window creates a group of smart-metered households that is too small for a precise analysis, we recommend increasing the length of the 'window' to increase the overall sample size available for analysis.[4]

[4] In addition to increasing the size of the group of smart-metered households in the smart-metered group, expanding the installation window affects the analysis in other ways:

1. It increases the ease of obtaining enough readings for customers by expanding the period during which a read can serve 'double duty' (a read in the installation window can serve as the

5. Use actual meter reads to calculate consumption

As Figure 1 implies, the approach we recommend involves matching on consumption/year using consumption during the *same* pre-installation period for all households (see section 7). It also involves calculating households' percent change in consumption between the post-installation and pre-installation years (see sections 8 and 10), where the post-installation year is also the *same* period for all households.

This approach simplifies much of the analysis. Because all households' pre-installation and post-installation periods are the same, evaluators do not need to correct for seasonal trends, use complex formulas to model consumption trends across time, or match on meter reading dates. However, to make this approach work, it is necessary to interpolate or extrapolate from actual meter readings to calculate households' consumption during the pre-installation and post-installation periods. We describe how to do so in this section.

5.1 Excluding customers without actual reads

In calculating consumption – both baseline consumption for matching (see Section 7), and post-installation consumption for comparison (see Sections 8 and 10) – a key question is how to treat estimated meter reads. We recommend using actual meter reads (and ignoring estimated reads) to calculate consumption (kWh/year) in the pre-installation and post-installation years.

We recommend suppliers have at least two actual meter reads at least nine months apart from each other to accurately calculate a customer's consumption in a year (and excluding customers who do not meet these criteria before building the matched comparison group or conducting further analysis). Small errors in estimation may be unproblematic for a monthly bill, especially as they are normally corrected when the supplier eventually obtains an actual read. However, they would be problematic and potentially bias this analysis, whose goal is the detection of small differences in consumption.

- In the group of households with traditional meters, this will mean including only households who submitted at least two actual meter reads, spaced at least nine months apart, in both the pre-installation *and* post-installation years. (As we note below, a read from the installation window can serve 'double duty' as a final read for the pre-installation period and a first read for the post-installation period.)
- In the group of households who received smart meters, this will mean excluding households who did not submit an actual meter read at least nine months before the

---

final read of pre-installation period *and* the first read of the post-installation period, as discussed later in this section).

2. It increases the length that a customer must remain on supply to be included in the analysis. The analysis described in this guidance requires 27 months of supply (12 months pre-installation-window, 3 months during the installation window, and 12 months post-installation window); increasing the length of the installation window will increase the total months involved. As an evaluator increases the length of continuous supply required, the composition of customers involved in the analysis will increasingly skew towards customers who do not frequently switch suppliers or move to a different property.

installation window. (We assume that the supplier normally obtains a closing read for the customer's traditional meter at the meter exchange and, in the post-installation year, automatically obtains at least monthly meter readings from these households.)

- As we note below, these reads can be used to extrapolate consumption, if they both fall within the year of interest, but they can also be used to interpolate consumption, if one or both reads fall outside the year of interest.

For example, imagine that the installation window was from 01 January 2019 through 31 March 2019 (as in Figure 1, above). In this case, we advise including only customers with sufficient actual reads as specified in Figure 2 (below).

Figure 2: Actual meter reads to calculate pre- and post-installation consumption (note: we show this figure, split in sections, in larger font in the next two figures)

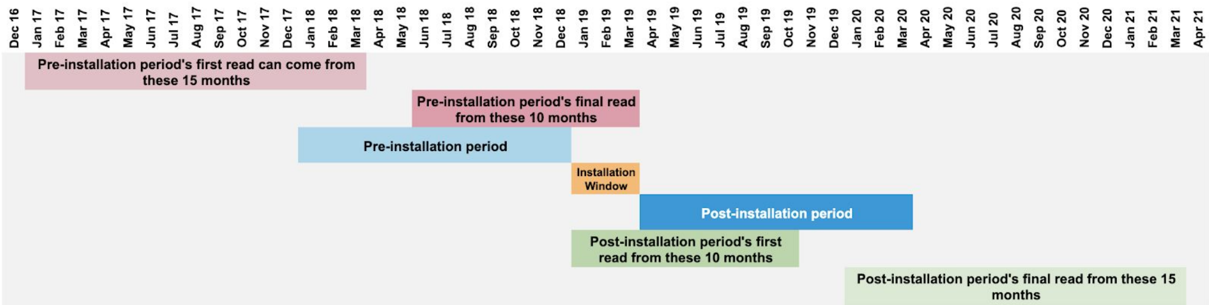
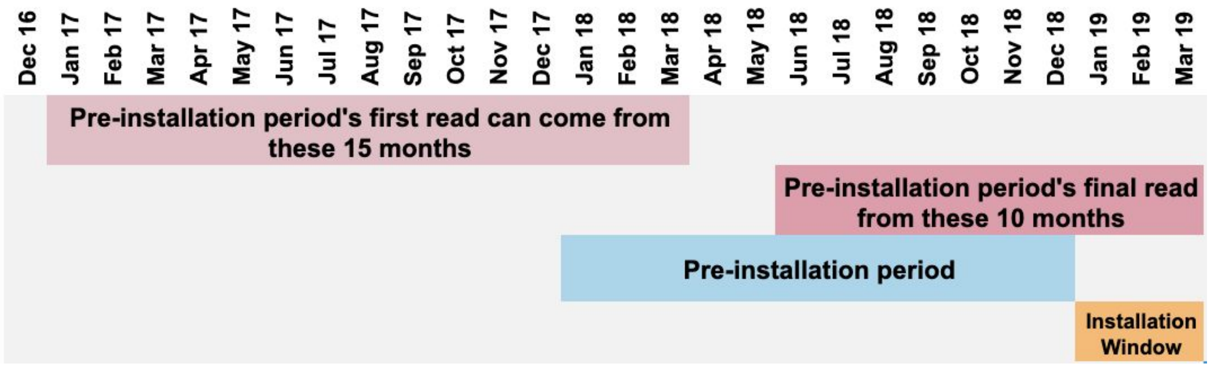


Figure 3: Calculating pre-installation consumption/year



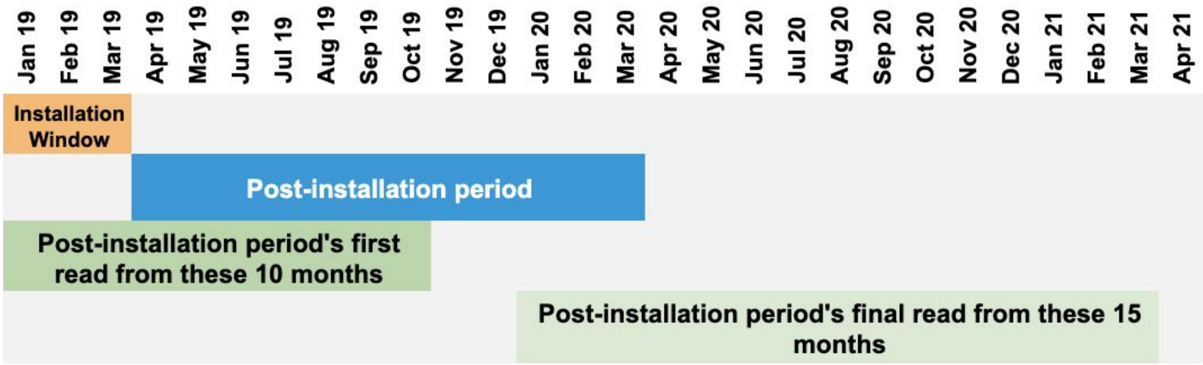
For the first read of the pre-installation period, evaluators must have at least one actual read from 12 months before to three months after the start of the period. For example, in calculating consumption from 01 January 2018 to 31 December 2018, it would be appropriate to use a read from 01 January 2017 through 31 March 2018 (the lighter pink rectangle in the images in this section).

For the final read of the pre-installation period (the dark pink rectangle in the images in this section), use a read from the 'installation window', if one exists.

- For smart-metered households, the meter exchange should provide such a read.
- For traditional-metered households who did not provide a meter read during the 'installation window', take the latest actual read in the pre-installation year. Ensure it is in the final seven months of the year (so, at least five months past the start of the pre-installation year). For example, in calculating consumption 01 January 2018 to 31 December 2018, it would be appropriate to use a read from 01 June 2018 through 31 March 2019 as the second read.

Note that the first and final reads should be spaced at least nine months apart (this condition is not captured in the above images.)

Figure 4: Calculating post-installation consumption/year (for traditional-metered households)



For smart-metered households, post-installation period consumption should come from the smart meter's reads.

For traditional-metered households, for the first read of the post-installation period, use a read from the 'installation window', if one exists. It can be the same read as the pre-installation period's final read. If such a read does not exist, take the earliest actual read in the post-installation year. Ensure it is in the first seven months of the period (so, at least five months before the end of the period). For example, in calculating consumption from 01 April 2019 through 31 March 2020, it would be appropriate to use the earliest read from the period 01 January 2019 through 31 October 2019 (the darker green rectangle in the images in this section).

For the final read of the post-installation period (the lighter green rectangle in the images in this section), use an actual read from any time from the three months before to 12 months after the end of the period. For example, in calculating consumption 01 April 2019 to 31 March 2020, it would be appropriate to use a read from 01 January 2020 through 31 March 2021 as the final read for the period.

Again, note that the first and final reads should be spaced at least nine months apart (this condition is not captured in the above images).

In summary:

- **Do not** use a read from the post-installation period to interpolate consumption in the pre-installation period.
- **Do not** use a read from the pre-installation period to interpolate consumption in the post-installation period.
- **Exclude from analysis** customers who do not have actual reads in line with the above criteria.

5.2 Interpolating or extrapolating from actual meter reads

Once a supplier has excluded customers with insufficient actual meter reads, what is the best approach to interpolate or extrapolate consumption for a given period? For example, if a supplier uses reads on 01 February 2019 and 30 May 2019 to calculate consumption during the period 01 April 2019 through 31 March 2020, what is the best way to interpolate consumption?[5]

Method 1: Using daily profile coefficients and LPA to extrapolate or interpolate consumption

We believe that the best method for extrapolating or interpolating consumption from actual reads is to imitate the method that industry bodies use to calculate Estimated Annual Consumption (EAC) for electricity and Annual Quantity (AQ) for gas.

For example, suppose a supplier has actual meter reads on 01 February 2019 and 30 May 2020 to calculate consumption for the post-installation period 01 April 2019 through 31 March 2020).

For gas:

- Multiply the household's actual gas consumption between the two reads by 365 (the number of days in the period of interest, in this case the post-installation period 01 April 2019 through 31 March 2020).
- Divide this product by the outturn LPA[6] between the reads.

$$\frac{(kWh\ actual\ consumption\ between\ reads) * 365}{(outturn\ LPA\ between\ reads)} \qquad \textit{(equation 1)}$$

- The result will be a year's worth of gas consumption, scaled by expected gas consumption load profile between the two reads from which the supplier is interpolating or extrapolating consumption.

For electricity:

- Sum the Period Profile Coefficients (PPCs) for a given customer over the period between the two actual meter reads.[7]
- Divide the household's actual electricity consumption between the two reads by this sum of PPCs.[8]

---

[5] We thank Manav Agarwal and Zoey White from British Gas, David Sykes from Octopus Energy, and Natalie Hockham and Katie Russell from Ovo Energy. They generously provided advice regarding the methods discussed in this section.

[6] Outturn LPA can be thought of as the amount of days of normal-year gas consumption that Xoserve expects the average household (in a given local distribution zone and end user category) to consume over a period. It may sum to more than 365 days per year in cold years, or to fewer than 365 days per year in warm years. Each local distribution zone and end user category has its own outturn LPA.

[7] PPCs can be thought of as the proportion of annual consumption that the average household is expected to consume in a given half hour.

[8] Note that there is no need to multiply by the number of days between the reads, as was necessary in the gas consumption calculation, because the profile coefficients sum to approximately 1, whereas the outturn LPA sums to approximately 365.

$$\frac{(kWh\ actual\ consumption\ between\ reads)}{(sum\ of\ profile\ coefficients\ in\ the\ period)}$$ *(equation 2)*

The result will be a year's worth of electricity consumption, scaled by expected electricity consumption load profile between the two reads from which the supplier is interpolating or extrapolating consumption.[9]

The key difference between these equations and the way that EACs and AQs are created is that, in this method, suppliers do not make use of previous EACs or AQs – instead, they calculate consumption solely from the actual reads.

If suppliers do not have access to the data required for this method (using profile coefficients and LPA to extrapolate or interpolate electricity and gas consumption), we suggest they use one of the following three alternative methods.

Method 2: Suppliers can spread consumption evenly across each day

The simplest method would be to take the total consumption from 01 February 2019 to 30 May 2019 – for the sake of this example, suppose this consumption totals 3,000 kWh across the 484 days. Suppliers could apportion 3,000 kWh / 484 days = 6.2 kWh/day to each day during that period. To interpolate consumption during the period 01 April 2019 through 31 March 2020 (365 days), suppliers multiply the number of days (365) by 6.2 kWh/day = 2,262 kWh in that period.

$$\frac{(kWh\ actual\ consumption\ between\ actual\ reads) * 365}{(Days\ between\ actual\ reads\ )}$$ *(equation 3)*

We think this is a valid approach, but it will underestimate gas consumption on cold days, when gas consumption is generally higher (and, similarly, underestimate electricity consumption on winter days with fewer daylight hours, when electricity consumption is generally higher).

Method 3: Suppliers can scale consumption according to HDDs or hours of darkness each day

A more accurate method involves taking account of weather conditions (for gas consumption analysis) or daylight hours (for electricity consumption analyses). A simple approach is to take note of the HDDs on each day during the period in the customer's HDD region and scale linearly according to these figures.

$$\frac{(kWh\ consumption\ between\ reads) * (HDDs\ in\ period\ of\ interest)}{(HDDs\ between\ reads)}$$ *(equation 4)*

[9] Note that this equation is essentially identical to Elexon's calculation of annualised advances. See Elexon. (2018). Load Profiles and their use in Electricity Settlement. Available at https://www.elexon.co.uk/documents/training-guidance/bsc-guidance-notes/load-profiles/ – in particular, see Section C, pages 17-19.

For example, suppose that there were 1,800 HDDs in the period 01 February 2019 to 30 May 2019, and the household faced 1,200 HDDs during the period 01 April 2019 through 31 March 2020. Then, the supplier would apportion:

$$\frac{(3{,}000\ kWh\ in\ period) * (1{,}200\ HDDs\ 01/04/19\ through\ 31/03/20)}{(1{,}800\ HDDs\ 01/02/19\ through\ 30/0519)} = 2{,}000 \text{ kWh}$$

They can do likewise with hours of darkness (24 minus daylight hours) for electricity consumption on each day during a given period.

We think this is a valid approach, and an improvement over method #2. However, it assumes no ‘baseload’ gas or electricity consumption – e.g. consumption that is not in response to cold weather or darkness. For this reason, it will slightly overestimate gas consumption on cold days (and, similarly, overestimate electricity consumption on short winter days) – in other words, overcorrecting for the problem associated with method #2.

Method 4: Suppliers can spread consumption according to HDDs or hours of darkness each day using slightly more nuanced methods than simply scaling linearly

Suppliers can improve on method #3 by conducting a hybrid of methods #2 and #3: apportioning 28% of gas consumption across each day, and then apportioning the other 72% according to the HDDs on each day.[10]

Another option is for suppliers to use more than two meter readings per period, which allows the ratio of kWh/HDD to differ somewhat by season (and likewise for kWh per daylight hours, in electricity analyses).

[10] The precise percentages will depend on the supplier’s assumption about how much of a household’s gas consumption is for heating. In our own work, we have estimated that 76% of households’ gas consumption is for heating; Ofgem has estimated this figure to be 72%. See our evaluation of the Nest Learning Thermostat for further details (https://www.bi.team/wp-content/uploads/2017/11/311013-Evaluating-Nest-BIT-Exec-Tech-Summaries.pdf).

6. Other notes on selecting the sample for this analysis

In addition to the criteria regarding meter read provision discussed in Section 5.1 (at least two meter reads, at least nine months apart, to determine consumption in both the pre-installation and post-installation years), we recommend excluding customers who:

- Were not supplied for at least 12 months prior and following the installation window (e.g. if there was a change of tenancy or supply), as presumably the supplier would not have sufficient meter readings for this customer
- Have post-installation consumption/year less than 50% or more than 150% of pre-installation consumption (regardless of whether they are in the smart-metered or traditional-metered group)
- Show a single day of electricity consumption ≥100 kWh/day
- Show a single day of gas consumption ≥ 500 kWh/day.

Suppliers should analyse dual-fuel credit customers on non-time-of-use electricity tariffs for their main analysis.

Separate analyses could use other samples:

- Dual-fuel customers on time-of-use electricity tariffs
- Customers on pay-as-you-go tariffs
- Single-fuel electricity customers on non-time-of-use electricity tariffs
- Single-fuel electricity customers on time-of-use electricity tariffs

The reason we suggest analysing these different categories of customers separately is to ensure ‘like for like’ comparisons, e.g. helping control for differences in heating sources.

7. Creating the comparison group of traditional-metered households

7.1 Trade-offs in matching

In 'building' the comparison group via matching, the evaluator must balance competing goals.

- Stricter matching criteria **reduce the chance of having differences between the groups** that might bias the estimation of the effect of smart meters
  - but -
- Stricter matching criteria can also **reduce sample sizes** by increasing the proportion of smart-metered households who have no match – e.g. where there are no traditional-metered households that match the smart-metered household on the criteria being used. In this instance, these smart-metered households with no traditional-metered household matches would be excluded from the analysis. As discussed previously, larger sample sizes allow evaluators to evaluate smart meter impacts with more precision.

Ultimately, the matching criteria chosen by each supplier should depend on their own sample size considerations and data availability.

7.2 Matching variables

BIT recommends using a variant of coarsened exact matching[11] to build the comparison group. The characteristics we recommend using for matching are (in order of the importance we ascribe them):

1. Previous energy consumption/year
2. Region
3. Other household characteristics, if known, such as number of bedrooms

1. Previous energy consumption is a strong predictor of future energy consumption. For example, BEIS analysis found that consumption recorded by gas meters in 2012 explained 80% of the variation in gas consumption between households in 2013.[12] By matching households with smart meters to households with traditional meters that have similar previous energy consumption, the evaluator avoids confounding the impact of smart meters with differences in baseline energy consumption between the two groups. Previous consumption also correlates with other variables that influence energy consumption, such as weather, location and property type, so matching on it can help control for their influence.

---

[11] Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, *20*(1), 1-24.

[12] BEIS (2016), *National Energy Efficiency Data-Framework (NEED) report: summary of analysis 2016, Annex C: Predicting Gas Consumption -* https://www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-report-summary-of-analysis-2016.

We recommend prioritising matches on pre-installation consumption/year, using tight matches if possible: ±50 kWh/year bands for electricity and ±200 kWh/year bands for gas (convert gas volume units into kWh according to the calorific value on the days in the period of consumption). Widen these bands if sample size would otherwise be insufficient. We recommend using a maximum band size of ±200 kWh/year for electricity and ±800 kWh/year for gas. (See Section 5 above for details on how to calculate consumption/year.)

Box 1: Methods for matching on pre-installation consumption

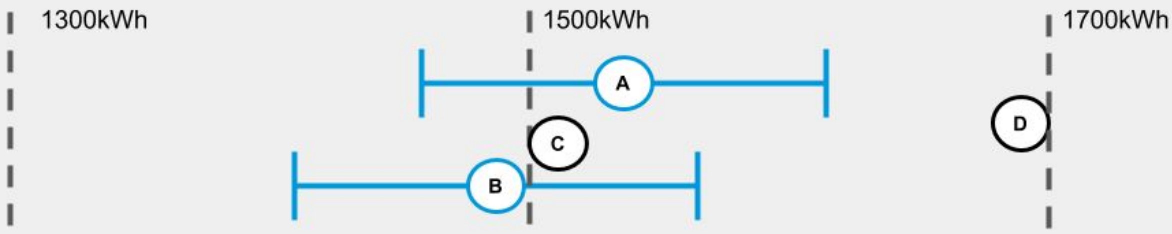
Caliper matching: In matching pre-installation consumption/year, we advocate a variant of matching in which suppliers perform one-to-one matching within a band for the previous year's consumption of ±50 kWh/year for electricity and ±200 kWh/year for gas. The formal term for this method is 'caliper' matching.

Coarsened exact matching: Another option is to create pre-installation consumption/year categories – for example, 1,100 kWh/year to 1,300 kWh/year, 1,300 kWh/year to 1,500 kWh/year, 1,500 kWh/year to 1,700 kWh/year, and so on – and then match customers on the category to which their pre-installation consumption belongs. This method, called Coarsened Exact Matching (CEM), is a valid approach.

Between caliper matching and CEM, we slightly prefer caliper matching because it avoids situations where customers are placed in different categories despite near-identical pre-installation consumption levels. For example, a customer with 1,499 kWh/year pre-installation consumption might be put in a different category from a customer with 1,501 kWh/year pre-installation consumption based on a 1,500 kWh/year category cut-off.

Using CEM to create matches in the following dataset, household A would match with households C and D, whereas household B would match to neither, despite household C being a close match with household B in terms of energy consumption.

Household	Meter	Previous energy consumption (kWh)
A	Smart	1541
B	Smart	1499
C	Traditional	1501
D	Traditional	1680



However, CEM may be easier to implement because there are 'off-the-shelf' resources to implement it in almost all major data analysis programmes – and, as we note, it is a perfectly valid approach.

Our recommendation on the band sizes is informed by the change in consumption (in kWh/year) that a 3% reduction would achieve for 'low', 'medium', and 'high' consumption households. This is an effect size we regard as realistic for smart meters, based on our and others' work in this area.[13] Matching on sufficiently small bands helps avoid differences in baseline consumption confounding the identification of the impact of the smart meter (as does the difference in differences methodology, discussed in Section 8).

Table 2: Examining 3% reductions in consumption/year, by Typical Domestic Consumption Value

	Ofgem TDCV category	Consumption/ year (kWh/year)	Consumption/ year with 3% decrease (kWh/year)	Difference (kWh/year)
Gas	Low	8,000	7,760	240
	Medium	12,000	11,640	360
	High	17,000	16,490	510
Electricity (profile class 1)	Low	1,900	1,843	57
	Medium	3,100	3,007	93
	High	4,600	4,462	138

2. Region is also a strong predictor of energy consumption. Region-specific events and trends – such as weather – affect energy consumption in that region. For this reason, even conditional on matching customers with similar pre-installation consumption/year, we expect that matching on region as well will produce a comparison group with consumption closer to smart-metered households' counterfactual consumption/year.

We advocate matching on small regions, such as outer postcode (e.g. SW1H or AB12). However, there may be insufficient sample size to use tight consumption/year matches *and* granular region matches. In this situation, our recommendation is that tighter matches on consumption/year are more important than more granular matches on region. Larger Public Electricity Supplier (PES) areas are sufficient for use as a matching variable in this case.

3. If suppliers do not match on region, or if they match on very large regions, we recommend matching on a variable that captures weather (for analyses of gas consumption) and daylight hours (for analyses of electricity consumption).

We recommend using **heating degree days (HDDs)** to capture weather effects on gas consumption,[14] though there are other good proxies that suppliers may use, such as the

[13] For example, see BEIS (2019), *Smart meter roll-out: cost-benefit analysis 2019 -* https://www.gov.uk/government/publications/smart-meter-roll-out-cost-benefit-analysis-2019.

[14] HDDs are the sum of the difference between the actual temperature in an 'HDD region' and a fixed 'baseline temperature' over a day. They are a measure of cold weather, and drive demand for energy,

composite weather variable (CWV). We would recommend matching customers on **HDDs in the pre-installation year** *and* **HDDs in the post-installation year**.

Our logic is that weather influences households' gas consumption, and hours of darkness influences their electricity consumption. These conditions (particularly weather) vary across geography. So, for suppliers who do not match on region, it is important to ensure households faced similar weather conditions, in both pre- and post-installation periods. (For suppliers that **do** match on region, we do not think it is necessary to additionally match on HDDs as well, given that households in the same region face approximately equivalent weather.)

HDDs/year are a continuous measure, much like consumption/year, so we recommend matching on HDDs using caliper matching (but CEM would also be appropriate). We recommend bands of ±50 HDDs/year for gas consumption. We do not have a specific band size recommendation in matching on daylight hours for electricity consumption analyses.

4. Finally, we recommend matching on **other household characteristics**, such as number of bedrooms, number of occupants, accommodation type, and accommodation size. Even conditional on matching customers with similar baseline consumption levels and identical regions, we expect a comparison group with consumption closer to smart-metered households' counterfactual when evaluators match on other characteristics about a household as well. However, if sample size constraints do not allow these additional matching criteria, we recommend forgoing them. If suppliers can only feasibly match on one further characteristic, **we recommend prioritising number of bedrooms** (treating 'unknown' number of bedrooms as its own category).

7.3 Multiple matches and non-matches

A smart-metered household may have **multiple matches** – for example, if multiple traditional-metered households have similar pre-installation consumption and are exact matches on the other matching variables. In this case, for simplicity, we recommend one-to-one matching, where suppliers randomly select one of the matching traditional-metered households for each smart-metered household. We recommend doing this 'without replacement' – this means that once a traditional-metered household has been selected as the match for a smart-metered household, it cannot serve as the match for a second smart-metered household.

A smart-metered household may have **no matches** – if no traditional-metered households have similar pre-installation consumption or are exact matches on the other matching variables. We recommend that suppliers report how many smart-metered households have no match. We counsel against proceeding if fewer than 70% of analysis-eligible smart-metered households have matches. In this case, we recommend relaxing matching criteria, such as by widening the consumption bands within which suppliers match households.

---

particularly gas, for heating purposes. They are publicly available from a range of data sources (e.g. https://www.degreedays.net/).

Note that a smart-metered household may be matched to different traditional-metered households for a supplier's electricity and gas analyses.

7.4 Many-to-many matching

If suppliers have the capability, they can significantly increase the precision of their analysis by using many-to-many matching. This allows a traditional-metered household to serve as the comparison household for multiple smart-metered households – and, similarly, allows multiple traditional-metered households to serve as a comparison for a single smart-metered household.

This many-to-many matching increases precision for two reasons.

1. Fewer households are unmatched, as a traditional-metered household can serve as a match for multiple smart-metered households.
2. For smart-metered households with multiple traditional-metered household matches, estimates will make use of these extra comparisons rather than randomly choosing just one traditional-metered household to serve as the match.

Note that, in many-to-many matching, the evaluator must apply weights to the households. If using Coarsened Exact Matching (CEM), we recommend using the weighting method specified by the developers of CEM, Gary King and co-authors.[15] If using caliper many-to-many matching, weighting is slightly more complicated. See the Appendix for further details on weighting when using many-to-many matching.

Also note that many-to-many matching is only relevant if suppliers use regression analysis (see section 10) rather than a paired t-test (see Section 8).

[15] Resources from Gary King and co-authors are available at: https://gking.harvard.edu/cem.

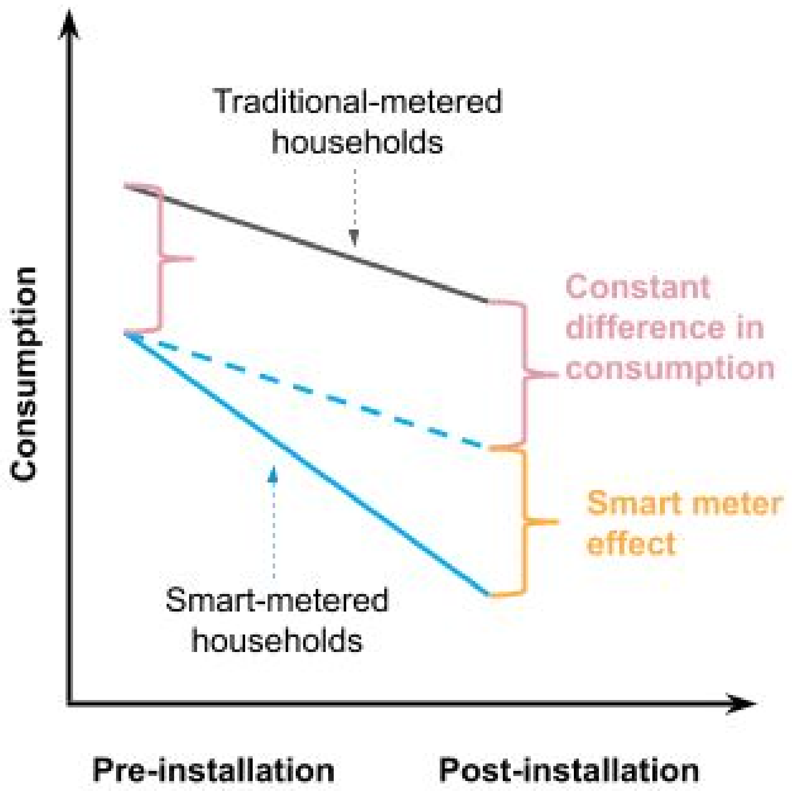
8. Comparing consumption in the two groups using a paired t-test

Once a supplier has a group of smart-metered households and a matched comparison group of households with traditional meters, they can compare the groups' consumption/year.

As implied in Figures 2-4, we recommend *not* analysing consumption during the installation window. For example, if the installation window were the three months from 01 January 2019 through 31 March 2019, we recommend analysing post-installation consumption/year using consumption from 01 April 2019 through 31 March 2020 (having matched on households' consumption/year from 01 January 2018 to 31 December 2018).

We recommend calculating the difference in households' percent change in consumption/year, by group. The formal name for this method is **difference in differences**: instead of calculating the difference between groups' post-installation consumption/year, we advocate calculating the difference between groups' *differences* in post- and pre-installation consumption/year. This is a further assurance (in addition to matching) that differences in baseline consumption do not confound identification of the impact of smart meters.

Figure 5: Calculating the difference in differences helps avoid confounding impact of smart meter installation with differences in pre-installation consumption[16]



[16] Diagram from *Health Policy Data Science*, Difference-in-Differences, by Bret Zeldow and Laura Hatfield (2019), available at https://diff.healthpolicydatascience.org/.

In some circumstances, the traditional-metered and smart-metered groups will be well-balanced on pre-installation consumption. In this situation, analysing the difference in differences will be equivalent to simply analysing post-installation consumption differences. Nevertheless, we recommend using the difference-in-differences approach, in case there are small differences between pre-installation consumption in the two groups.

For each household *i*, suppliers calculate:

$$D_i = (A_i - B_i) / B_i \qquad \textit{(equation 5)}$$

where:

- D_i is the percent change in consumption
- A_i is the post-installation consumption for the household (in kWh/year).
- B_i is the pre-installation consumption for the household (in kWh/year).

We recommend that suppliers compare the groups using a paired t-test. (The pair, in this case, would be the smart-metered household and its matched traditional-metered counterpart. For suppliers using an alternate methodology that does not involve pairing, we recommend a two-samples t-test.) So, for each pair, we calculate:

$$D_{i-s} - D_{i-t}$$

where *i-t* is the traditional-metered household that matched to the smart-metered household *i-s*. This difference is an estimated impact of smart meters on percent change in energy consumption. The paired t-test compares this difference to zero and provides the 95% confidence interval for this impact.

We recommend conducting separate tests for changes in gas consumption and changes in electricity consumption.

9. Summary of recommended methodology

<table>
<tr><th>Step</th><th>Description</th></tr>
<tr><td>1. Select an installation window</td><td> Select an installation window to create your group of smart-metered households. This can be a 3-month period, a 6-month period, a 12-month period, etc. We recommend using a 3-month period, if sample size permits. Shorter periods will create smaller sample sizes for analysis, but they also allow suppliers to conduct analyses on a regular basis and analyse changes in impacts of smart meter installation across time. Do not analyse consumption during the installation window. For example, if the installation window were the three months from 01 January 2019 through 31 March 2019, match on households’ pre-installation consumption/year from 01 January 2018 through 31 December 2018, and analyse post-installation consumption/year 01 April 2019 through 31 March 2020. An alternative method would be to analyse each smart-metered household’s post-installation consumption from the day of the smart meter installation – comparing it to its matched traditional-meter household's consumption over the same time period. We believe this is a valid approach, but involves more complicated data processing and analysis. </td></tr>
<tr><td>2. Create the group of smart-metered customers</td><td> Select households where a smart meter was installed during the installation window and who meet the following criteria: Supplied for at least 12 months prior to the installation window, and provided at least two actual meter readings during this 12-month period at least nine months apart (see Section 5.1 for details on actual meter reading criteria)
Supplied for at least 12 months following the installation window</td></tr>
<tr><td>3. Create the group of potential matches for smart-metered customers</td><td> Select households where a smart meter was not installed during the installation window, nor in the subsequent 12 months, and who meet the following criteria: Supplied for at least 12 months prior to the installation window, and provided at least two actual meter readings at least nine months apart to calculate consumption for this 12-month period
Supplied for at least 12 months following the installation window, and provided at least two actual meter readings at least nine months apart to calculate consumption for this 12-month period (see Section 5.1 for details on actual meter reading criteria)</td></tr>
<tr><td>4. Remove outliers</td><td> Remove households who meet any of the following conditions: Have post-installation consumption/year less than 50% or more than 150% of pre-installation consumption/year
Have a single day of electricity consumption ≥100 kWh
Have a single day of gas consumption ≥500 kWh.</td></tr>
</table>

5. Calculate consumption for all households	Calculate consumption from the available meter reads for all smart-metered and traditional-metered households. Use actual meter reads (ignore estimated reads) to calculate consumption (kWh/year) in the pre-installation and post-installation years (see Section 5.2).
6. Created the matched group of traditional-metered households	Create the matched traditional-metered households for the smart-metered households by matching on the following attributes, considering the trade-off between sample size and strictness of matching criteria: 1. Pre-installation consumption, within the chosen band for the fuel (ideally ±50 kWh/year for electricity consumption analyses and ±200 kWh/year for gas consumption analyses). 2. Region (e.g. outer postcode or PES area). 3. Other characteristic(s), such as number of bedrooms, number of occupants, house type, and house size. We believe number of bedrooms is the priority, here. Conduct matching separately for electricity and gas consumption analyses: a smart-metered household may have a different traditional-metered household match for each analysis. Report how many smart-metered households have no match, and what proportion of the total eligible smart-metered households these non-matched households represent. *Note: Although we advocate using 'caliper' matching, we believe Coarsened Exact Matching is also a valid approach, and it may be easier to implement given there are 'off-the-shelf' resources to implement it in most major data analysis programmes.* *If possible, use many-to-many matching to improve the precision of your analysis.*
7. Calculate percentage change for each household	For each household, compute the difference in consumption/year: post-installation consumption/year minus pre-installation consumption/year. Divide this difference by pre-installation consumption. The result D_i (in equation 5 in Section 8) is the percent change in consumption for each household. Suppliers should do this separately for electricity and gas consumption analyses.
8. Calculate 95% confidence interval for mean percent change	To check whether the traditional-metered group's mean percent change is statistically significantly different from the smart-metered group's mean percent change (whether the **difference in differences** is statistically significant), conduct a paired t-test of the two groups' percent-change statistics D_i. Report the 95% confidence interval.

10. Comparing consumption in the two groups using regression analysis

Where suppliers use software that allows regression analysis, we recommend identifying the impact of smart meters on consumption using a regression model, in addition to or instead of using the paired t-test method discussed above.[17]

We recommend using the following regression model predicting the scaled difference in energy consumption of household *i*:

$$D_i = \beta_1 + \beta_2 T_i + \beta_3 H1_i + \beta_4 H2_i + \beta_5 V_i + \beta_6 P_i + \varepsilon_i \qquad \textit{(equation 6)}$$

where

- D_i is as defined in equation 5 in Section 8 (post-installation consumption minus pre-installation consumption, divided by pre-installation consumption, for household *i*).
- T_i is 1 if the household received a smart meter, else 0.
- $H1_i$ is the HDDs in the pre-installation period, and $H2_i$ the HDDs in the post-installation period, in household *i*'s HDD region. Note that these parameters are only included if predicting gas consumption – we recommend dropping them from the regression predicting electricity consumption.
- V_i represents any other variables suppliers may hold about household *i*, such as bedrooms, number of occupants, house type, house size, channel through which the customer joined, tariff, etc. that may be associated with energy consumption. Adding variables that help predict energy consumption will increase the precision with which suppliers can identify the effect of the smart meter on customers' consumption (by explaining otherwise unexplained variance).
- P_i is a vector of indicator variables for each matched pair. The inclusion of these pair fixed effects renders the regression analytically identical (but for the covariates $H1_i$, $H2_i$, and V_i) to a paired t-test. Most statistical software packages have functions to execute this regression absorbing these fixed effects in the model rather than calculating and displaying each of their parameters (e.g. using `areg, absorb` in Stata, or `felm` in R).
- ε_i is an error term.

As discussed in Section 7.4, suppliers can significantly increase their analysis's precision by using many-to-many matching – using all traditional-metered households who match to a smart-metered household, and using a traditional-metered household to match to multiple smart-metered households. In this case, they would need to apply weights to the

[17] We would regard the regression result as primary in the case of disagreement between it and the t-test. However, such disagreement would indicate that the groups were imbalanced on the covariates in the regression and that this imbalance was correlated with energy consumption changes. We would recommend the evaluator investigate these issues.

traditional-metered households in their analysis, weighting traditional-metered households in line with the number of other traditional-metered households matched to the same smart-metered household (see Appendix for details on weighting). We would continue to recommend the inclusion of P_i in the regression – they would represent stratum fixed effects, rather than paired fixed effects.

The parameter of interest in equation 6 is β_2, which can be interpreted as the percent change in a household's energy consumption (D_i) caused by a smart meter. For example, if β_2 = -0.047 (95% CI -0.033 to -0.062), the supplier can say that smart meters cause a 4.7% reduction in consumption (95% CI 3.3% to 6.2%).

11. Threats to validity

There are two threats to the **internal validity** of the approach we recommend. In this context, internal validity is the extent to which an analysis identifies a *causal* relationship between smart meter installation and a customer's energy consumption reduction.

- In contrast to a randomised controlled trial, matching does not guarantee groups that are balanced, on average, between all observable and unobservable characteristics. Matching creates a comparison group that closely resembles the smart-metered group in terms of observable characteristics, and it assumes that unobservable characteristics are balanced across the groups. This assumption may be invalid. For example, customers who become interested in saving energy may be more likely to request a smart meter and follow through with the installation booking, and this self-selection could be correlated with energy efficiency behaviours independent of the smart meter installation. However, we expect this bias to be modest.
- Our recommended methodology involves excluding traditional-metered customers who do not give meter readings during the year in which their consumption is compared to smart-metered customers. This creates a subtle difference in the composition of the two groups. Customers in the smart-metered group may be composed of a greater proportion of customers who *would have* failed to give meter reads, whereas customers in the traditional-metered group are screened on actual meter read provision consistency. Insofar as changes in actual meter reading frequency are associated with changes in energy consumption, this difference in the composition of the smart-metered and traditional-metered groups could cause bias. However, we expect this bias to be modest as well.

Suppliers should also note threats to the **external validity** of the methodology we recommend. External validity is the extent to which findings of the analysis can be extrapolated to the wider population and other contexts.

- The methodology we recommend requires data from customers across at least 27 months – one year for matching smart-metered customers to traditional-metered customers with similar characteristics, and one year to compare the groups' consumption, separated by three months (the installation window). Our methodology also requires a minimum meter reading frequency. This means that suppliers will exclude from analysis customers who frequently switch suppliers and customers who rarely or never give actual meter reads. Although we have not seen evidence of this, it is possible that these customers may respond differently to obtaining a smart meter.
- The analysis results apply to the smart-metered customers in the installation window suppliers analyse – insofar as these customers or the installation window were atypical, their experience might not generalise to other customers' experience.

However, regarding the second point, our hope is that suppliers employ this method for each quarter of the year for each year that they have installed smart meters – or, if this were burdensome, to do so for one quarter each year. This analysis of different 'installation windows' would indicate if the savings from a smart meter have changed over time.

12. Robustness checks

12.1 Balance

We recommend suppliers analyse **balance** on key variables between the smart-metered household group and its matched comparison group, in particular:

- HDDs in the pre-installation period for each household, if evaluators are not matching on this measure (we assume that matching on region will ensure good balance on this measure, but the balance check would confirm this assumption)
- HDDs in the post-installation period for each household, if evaluators are not matching on this measure (again, we assume that matching on region will ensure good balance on this measure, but the balance check would confirm this assumption)
- Other variables about customers that suppliers did not match on, such as their tenure, tariff type, demographic information suppliers hold about the customer, etc. (If suppliers are using regression analysis recommended in Section 10, we recommend checking balance on any variable captured by V_i in equation 6.)

There will almost always be small differences between measures on which suppliers check balance. We recommend investigating cases where differences are large.

In any case, please do report the results of the balance checks – these results provide important context for interpreting the main results.[18]

12.2 Measurement error

Consumption extrapolation/interpolation from meters' actual reads will be associated with **measurement error**. There are a few robustness checks we recommend related to this.

- Pre-installation consumption/year is derived from two actual meter reads, in both the traditional-metered and smart-metered groups. For this reason, we do not expect the measurement error to affect one group more than the other. However, as a check against this, we suggest suppliers check balance on three aspects of these two reads, by group:
    1. Days between the first and last read
    2. Date of the first actual meter read
    3. Date of the final pre-installation period actual meter read.
- In the post-installation period, consumption/year is partially extrapolated or interpolated for the traditional-metered group, but not for the smart-metered group. It would be beneficial for suppliers to calculate and report the mean and standard deviation of this error.[19] If estimates are consistently higher than actual consumption,

[18] It is also valuable to compare balance between the two groups before and after matching. The evaluator should investigate if balance has *deteriorated* (e.g. differences between the groups has increased) after matching.

[19] Suppliers can calculate this error for customers who have provided three reads by extrapolating the customer's consumption/year from the first two reads and comparing this extrapolated consumption/year to the customer's actual consumption/year.

it could be that smart meters reduce *overestimation*, which might then be misinterpreted as smart meters reducing *consumption*. This possibility can be rejected – e.g. measurement error can be assumed to be immaterial to the results of the supplier's analysis – if the average measurement error is zero, or if it is small relative to the size of the impact of smart meters measured in the analysis.

12.3 Parallel trends

Finally, an important assumption underpinning the identification of the impact of smart meters from examining the difference in differences is that the smart-metered and traditional-metered groups would have had **parallel trends**. The 'parallel trends' assumption means that smart-metered customers would have had the same increase or decrease in consumption as traditional-metered customers, had they not received the smart meter. **This is an assumption that cannot be proved.** However, a common practice in analyses relying on a difference in differences is to check whether the groups *had parallel trends in the past*.

In the context of this energy consumption analysis, this would mean analysing percent consumption change between the pre-installation year and '**pre-**pre-installation year' – and showing its equivalence in the two groups. For example, if the pre-installation year is January 2019 through December 2019, suppliers should examine the percent consumption change amongst customers between:

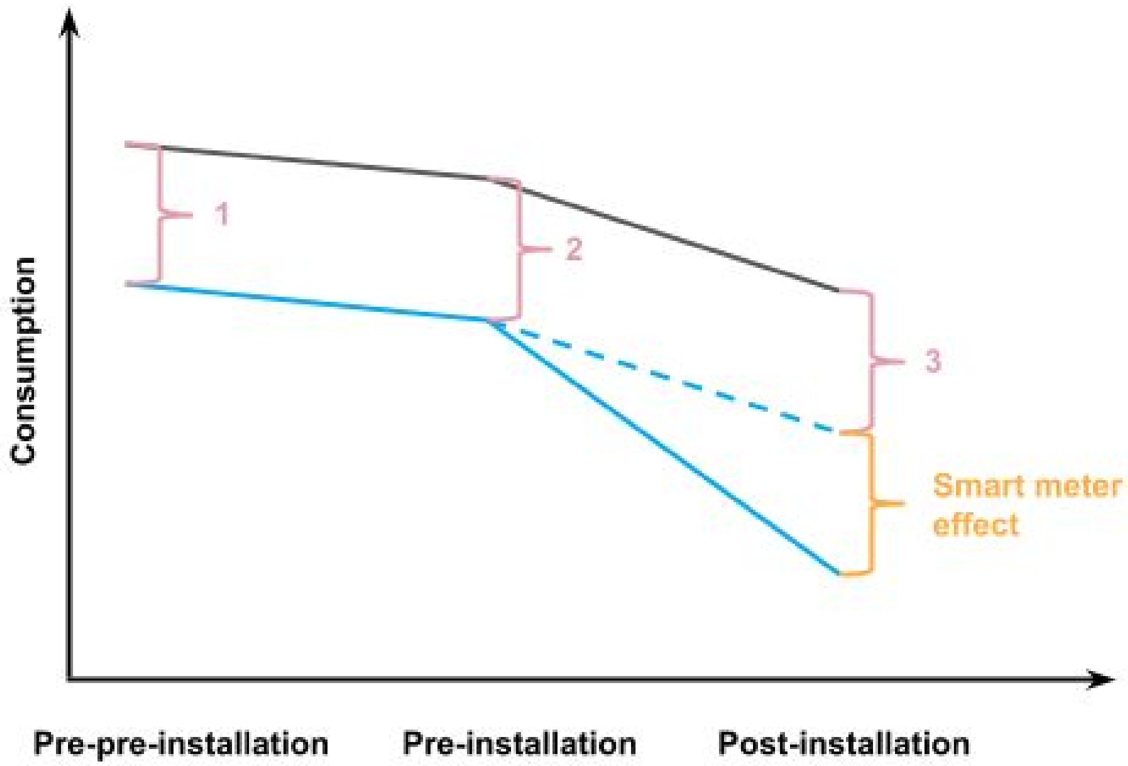
the '**pre-**pre-installation year': January 2018 through December 2018 (from 2 years before installation to the beginning of the pre-installation year)
- and -
the pre-installation year: January 2019 through December 2019 (from the beginning of the pre-installation year to the beginning of the installation window).

This analysis is identical to the analysis we described in Section 8, save that, instead of calculating percent change in consumption between the pre-installation year and the post-installation year, suppliers would calculate percent change in consumption between the 'pre-pre-installation year' and the pre-installation year. The evaluator should check that this difference in differences is zero.

We show this idea in Figure 2. The difference between groups' consumption should be the same in the '**pre-**pre-installation' period (pink bracket #1) as it is in the pre-installation period (pink bracket #2) – such that the difference in differences is zero.

This equivalence gives us greater confidence that the two groups' differences *would have been* equivalent between the pre-installation period (pink bracket #2) and post-installation period (pink bracket #3), were it not for the installation of the smart meter (which causes the difference in the post-installation period to be pink bracket #3 combined with the orange bracket).

Figure 6: Examining parallel trends



Note that this analysis will only be able to make use of households who were supplied during the 39 months represented by the '**pre-**pre-installation period' (12 months), the pre-installation period (12 months), the installation window (3 months), and the post-installation period (12 months). These data requirements mean that even more customers may be excluded from this check than were excluded from the main analysis.

Appendix: Weighting in many-to-many matching

In CEM[20], matched smart-metered households receive a weight of 1. Traditional-metered households receive a weight of:

*(n_S_stratum / n_T_stratum) * (n_T_matched / n_S_matched)* *(equation 7)*

where

- Each combination of matching variables forms its own 'stratum'.
- *n_S_stratum* is the number of smart-metered households in the stratum.
- *n_T_stratum* is the number of traditional-metered households in the stratum.
- *n_T_matched* is the number of matched traditional-metered households in the sample.
- *n_S_matched* is the number of matched smart-metered households in the sample.

In using caliper many-to-many matching, weighting is slightly more complicated. Multiple traditional-metered households might match with a smart-metered household, and one or more of those traditional-metered households might match with a different smart-metered household. For example, in the below dataset, assuming a band (or 'caliper') of ±50 kWh, smart-metered households 2 and 3 match with traditional-metered households 1 and 4. Smart-metered household 7 matches with traditional-metered households 4, 5, 6, and 8. Smart-metered household 10 matches with traditional-metered households 8 and 9.

Table 3: Example dataset to demonstrate caliper matching

Household id	Household type	Pre-installation consumption (kWh)
1	Traditional	970
2	Smart	990
3	Smart	990
4	Traditional	1030
5	Traditional	1100
6	Traditional	1090
7	Smart	1070
8	Traditional	1110
9	Traditional	1160
10	Smart	1150

[20] Resources from Gary King are available at: https://gking.harvard.edu/cem. In particular, a concise primer on weighting in CEM is available at https://docs.google.com/document/d/1xQwyLt_6EXdNpA685LjmhjO20y5pZDZYwe2qeNoI5dE.

In this method, matched smart-metered households continue to receive a weight of 1. Traditional-metered households receive a weight of:

$$(1 / n\_T\_stratum_1 + 1 / n\_T\_stratum_2 \ldots 1 / n\_T\_stratum_n) * (n\_T\_matched / n\_S\_matched) \qquad \textit{(equation 8)}$$

where

- Each smart-metered household forms its own 'stratum'. So, *n_S_stratum* is now always 1 – there is always just 1 smart-metered household per stratum.
- *n_T_stratum* is the number of traditional-metered households in a stratum.
- Because traditional-metered households can be matched to multiple smart-metered households, they can be in multiple strata. The evaluator sums (*1 / n_T_stratum)* for each stratum that the traditional-metered household occupies. Note that the strata listed in equation 8 ($stratum_1$ through $stratum_n$) are the strata that the matched traditional-metered household is **in**. They do not receive extra weight from strata they are not matched to.
- *n_T_matched* is the number of matched traditional-metered households in the sample.
- *n_S_matched* is the number of matched smart-metered households in the sample.

So, for example, in the dataset shown in Table 3, traditional-metered household 1 receives a weight of 1.5, traditional-metered household 4 receives a weight of 1.875, and traditional-metered household 3 receives a weight of 0.375.

Evaluators will use these weights in regression analysis to compare energy consumption changes between traditional-metered and smart-metered households. (See Section 10 for our recommended regression specifications.)

The intuition behind these weighting schemes is that each traditional-metered household receives a weight in proportion to the degree that it is acting as a counterfactual for smart-metered households.

- What this means in CEM is that, in strata with relatively more traditional-metered households per smart-metered households, evaluators reduce the weight a regression ascribes the traditional-metered households in that stratum. In strata with relatively fewer traditional-metered households per smart-metered households, evaluators increase the weight a regression ascribes the traditional-metered households in that stratum.
- Caliper matching weights uses this logic too. However, control units obtain extra weight if they belong to multiple strata (by virtue of matching to multiple treatment households).