Testing the strength of impact evidence - applying contribution tracing within a realist evaluation of Transitional Arrangements for demand-side response

Mary Anderson, CAG Consultants, UK (ma@cagconsult.co.uk)
Tajbee Ahmed, Department of Business, Energy and Industrial Strategy, UK (tajbee.ahmed@beis.gov.uk)
Barbara Befani, University of Surrey, UK (befani@gmail.com)
Charles Michaelis, Strategy Development Solutions, UK (charles@camichaelis.com)

ABSTRACT

This paper focuses primarily on evaluation methodology rather than demand-side response. It presents a UK case study of rigorous testing of impact evidence within a realist evaluation of the 'Transitional Arrangements for demand-side response'. The Transitional Arrangements (TA) aimed to encourage commercial/industrial firms and aggregators to make more demand-side response (DSR) capacity available to the Capacity Market (CM) for electricity.

A realist approach was used because of the small sample of participants in the TA and the need to develop a deep understanding of the complex market for flexibility services in the UK. Competing theories about the causal influence of the TA were specified in terms of realist 'Context-Mechanism-Outcome' (C-M-O) hypotheses. The impacts of the TA and other influences were assessed using evidence from in-depth interviews and public statements by TA participants and non-participants, together with observed behaviour in TA and other CM auctions.

Statistical or experimental approaches to impact evaluation were not feasible, so process tracing with Bayesian updating was chosen to test the strength of evidence supporting different causal C-M-O hypotheses. The paper explores the issues involved in applying process/contribution tracing within a realist evaluation. It explores the degree to which evidence tests can realistically test individual components of the C-M-Os and can test the causal linkages between contexts, mechanisms and outcomes. It also suggests an approach to combine findings across multiple tests to provide an overall assessment of the strength of support for different contribution hypotheses.

Introduction

Background

The purpose of the paper is to examine some of the issues that arose from applying process tracing methods within a realist evaluation. Therefore, the focus of this paper is the evaluation methodology itself, rather than demand-side response results. These methodological issues are explored through a case study involving evaluation of the 'Transitional Arrangements for demand-side response' in the UK. This evaluation was undertaken by CAG Consultants, in partnership with Winning Moves, Verco, NERA Economic Consulting and Strategy Development Solutions, on behalf of the Department for Business, Energy and Industrial Strategy (BEIS) between 2016 and 2018. The Transitional Arrangements for demand-side response (TA) formed part of the Capacity Market (CM) for security of electricity supply, within the UK government's Electricity Market Reform (EMR) programme. The TA aimed to support BEIS's overall objectives of promoting growth and energy security, while ensuring affordability of the energy supply.

One challenge in this evaluation was that there were only nine participants in the second TA scheme – too small to allow quantitative analysis. A further challenge was the risk of TA participants using the evaluation to 'lobby' for more or continued Government subsidy for their operations.

To address these challenges, BEIS proposed a realist approach to the evaluation, which would explore in-depth the causality behind the TA scheme and the reasons why TA participants behaved as they did. A realist 'theory of change' was developed, in the form of 'Context-Mechanism-Outcome' (C-M-O) configurations. BEIS also encouraged the use of

generative causation methods to assess the influence of the TA scheme, to bring rigour to the realist analysis process and deal with potential 'lobbying bias'. Therefore, with encouragement from BEIS, and with expert support from Barbara Befani at the University of Surrey, process tracing was used to test the strength of the impact evidence. The team used the term 'contribution tracing' to indicate the use of process tracing tests in the context of additionality and contribution analysis.

Scope of the TA programme

The TA was a programme focused on the market for electricity capacity in Great Britain. Overall, the TA aimed to encourage the development of DSR to balance supply and demand in a decarbonized electricity grid (National Infrastructure Commission, 2016). This paper focuses on the second TA programme which had two main objectives:

- **Objective 1:** to contribute to the development of flexible capacity[1] for the future CM
- **Objective 2:** to encourage turn-down demand-side response[2] (DSR)

The TA was designed to be a stepping-stone for flexible capacity that might have difficulty in competing in the main CM. While the TA did not automatically lead to future CM participation, it aimed to build capacity and confidence so that providers of DSR were better placed to compete in future CM auctions.

The main CM auctions involved one-year ahead auctions (T-1) and four-year ahead auctions (T-4) which were open to all types of generating capacity as well as DSR. The TA auctions were one-year ahead, like the T-1 auctions, but they were restricted to specific types of capacity. The second TA scheme was designed as 'nursery' for turn-down DSR, involving slightly softer conditions than the main CM (e.g. lower minimum volume of capacity, lower credit cover).

Methodology

Research questions

This paper presents work undertaken to respond to two high-level questions (HLQs) posed by BEIS, examining outcomes related to the two objectives of the TA programme:

- **HLQ 1** - What outcomes can be attributed to the second TA and were they as intended by BEIS? What outcomes occurred for whom and under what circumstances?
- **HLQ 2** - Through what levers and causal mechanisms has the second TA contributed to these outcomes and the variation by group and circumstance?

This paper focuses on the methodologies used to respond to these research questions: these methodologies involved the application of 'process or contribution tracing' within the context of a 'realist' and 'theory-based' evaluation. These terms, and the rationale for choosing these methodological approaches, are explained below under the following headings:

- Outline of the methods used in this evaluation
- Our approach to applying these methods in this evaluation

Outline of the methods used in this evaluation

What is theory-based evaluation?

The major challenges in impact evaluation are, firstly, to assess whether the outcomes from an intervention have been observed as expected and, secondly, to assess whether the intervention has contributed to or caused the observed outcomes. One approach is to compare the outcomes that were observed under the intervention to what would have happened in the absence of the intervention (i.e. the counterfactual) using experimental approaches (such as random control trials (RCTs)) or quasi-experimental approaches (such as econometric or statistical analysis).

Theory-based evaluations provide an alternative approach when it is not practical or desirable to use an experimental or 'quasi-experimental' approach. Theory-based evaluation (Treasury Board of Canada Secretariat, 2009) involves the development of a 'theory of change' setting out how the intervention is expected to work and lead to its desired outcomes. This 'theory of change' is then tested against the evaluation evidence and refined by the evaluator to reflect their understanding of the causal influence of the intervention.

[1] Ofgem defines flexibility as 'modifying generation and/or consumption patterns in reaction to an external signal (such as a change in price) to provide a service within the energy system'.

[2] 'Turn-down' Demand-side response is temporary reduction in the electricity demand in response to signals from the GB grid.

A theory-based approach was chosen for the TA evaluation, partly because the number of participants in the second TA scheme was nine: six aggregators[3] and three direct participants[4]. This sample was too small to allow experimental or quasi-experimental analysis. A theory-based evaluation approach was also chosen because the research questions sought information about why and how the TA led to change, which could be explored through development and refinement of a 'theory of change'.

What is a realist approach to evaluation?

Realist evaluation is a specific type of theory-based evaluation. A realist approach to evaluation (Pawson and Tilly (1997), Pawson (2006)) emphasises the importance of understanding not only whether a policy contributes to observed outcomes but how, for whom and in what circumstances and why. For this case, the approach was chosen because it helped to respond to the second research question (HLQ2): i.e. understanding how the TA influenced outcomes, and how this varied between different groups and in different circumstances.

Realist approaches to evaluation attempt to identify the 'contexts' and 'mechanisms' that lead to a particular 'outcome', as defined here:

- **Context** - the circumstances which affect whether an intervention 'works' and for whom. Consideration of 'context' forms an important part of the realist approach.
- **Mechanism** - a change in people's reasoning, in response to the resources provided by an intervention, which leads to an outcome. Identification of causal 'mechanisms', which operate in particular 'contexts', forms an important part of realist approach.
- **Outcome** - a change in the state of the world, brought about as a result of an intervention or other influences.

Realist evaluation uses the idea of 'generative' causality. Rather than considering that a given outcome has a certain probability of happening, the realist approach aims to identify the combination of 'contexts' and 'mechanism' that will (nearly always) lead to that outcome. The development of the 'theory of change' behind an intervention, defined in terms of 'C-M-O' configurations, is central to a realist evaluation. Pawson (2006) describes a 'realist evaluation cycle' in which early theory (which we call 'candidate theory') is developed at the start of an evaluation cycle, tested against evaluation evidence and then refined to create revised theory which more closely represents the underlying causality. The C-M-O configurations in the candidate theory are effectively causal 'hypotheses' that are tested and refined during the evaluation.

What is process tracing (or 'contribution tracing')?

Process tracing (Collier, 2010) is a method that involves explicit testing of competing causal hypotheses, to assess which hypothesis is most likely to be true. The purpose of process tracing is to increase the transparency and replicability of qualitative analysis. The process tracing method involves the construction of evidence tests for each hypothesis and careful assessment of the conditional probabilities of observing different pieces of evidence, depending on whether a given hypothesis is or is not true.

Process tracing categorises evidence tests into four types, depending on these conditional probabilities (adapted from Befani, 2016):

- **Hoop tests** – necessary but not sufficient (this is a piece of evidence that we would 'expect to see' if the given hypothesis is true; hoop tests reject or weaken the hypothesis if not found but are not sufficient to confirm the hypothesis if found)
- **Double-decisive** – necessary and sufficient (this a piece of that is expected but is also confirmatory of the hypothesis; doubly-decisive tests confirm or strengthen the hypothesis if observed and if not observed the hypothesis is rejected or weakened).
- **Smoking gun** – sufficient but not necessary (this is a piece of evidence that we would 'like to see'; smoking gun tests confirm/strengthen the hypothesis if observed but do not reject/weaken the hypothesis if not observed)
- **Straw-in-the-Wind** – neither necessary nor sufficient (this is a piece of evidence that - if observed – would slightly strengthen but not confirm the hypothesis, and – if not observed- would slightly weaken but not reject the hypothesis).

In this evaluation, we chose to implement process tracing to mitigate the effect of potential lobbying by TA participants. It allowed us to consider the probability that TA participants might make a given statement, not because it was true but because it was in their interest to do so. For example, if we thought it fairly likely that TA participants would claim that the softer conditions of the TA were necessary as a precursor to participating in the main CM, even if they

[3] An aggregator is an intermediary organisation that provides a service of collating capacity (from generation and/or DSR) for National Grid balancing services or the Capacity Market (CM), from a range of other organisations (i.e. clients), in return for a share in the revenues generated.

[4] A direct participant is an organization that offers DSR capacity to the CM or National Grid in its right, rather than via an aggregator.

would actually have participated in the main CM straightaway if there had been no TA, we would specify such a test as a 'hoop test' (expect to see) or 'double-decisive' test for the additionality of the TA.

In a realist evaluation, the theory of change is defined in terms of C-M-O configurations (Context-Mechanism-Outcome), so each C-M-O can be treated as a causal 'hypothesis' to be tested using process tracing. The use of process tracing to test 'contribution' or 'additionality' hypotheses, is sometimes called 'contribution tracing' (Murray Brown, 2016).

Process tracing is often applied to a single case (e.g. an outcome for a single organisation), but it can also be applied to multiple cases. In this example, we treated each of the nine participants in the second TA scheme as a 'case'.

Process tracing can be combined with Bayesian updating, involving assignment of numerical probabilities to each hypothesis and evidence test, and updating of these probabilities based on observed evidence (Befani and Stedman-Bryce, 2016). For reasons explained below, we did not use Bayesian updating in our application of process tracing to the second TA scheme.

Our approach to applying these methods in this evaluation

How we developed a candidate 'realist theory of change' for the second TA

Our starting point for the realist theory of change for second TA scheme was the theory of change that we had already developed during our evaluation of the first TA scheme (BEIS, 2018). While the first TA scheme had included a range of DSR technologies, including back-up generation, the second TA scheme was more tightly focused on DSR provided by participants turning down electrical loads. Working closely with BEIS policy and evaluation staff, the evaluation team adjusted the theory of change to reflect the scope of the second TA scheme and to fit the research questions relevant to the second TA.

The theory of change for the second TA was set out in realist form, using 'Context-Mechanism-Outcome' configurations (Pawson and Tilly (1997), Pawson (2006)). These were realist hypotheses about how the policy was expected to work. The initial or 'candidate' theory of change was based around two high-level hypotheses about the additionality of the second TA scheme. There reflected the two objectives of the scheme introduced above:

- H1: The second TA leads to more and/or more competitive flexible capacity for the Capacity Market in 2018 - 2019 and subsequent years.
- H2: The second TA leads to wider encouragement of turn-down DSR.

For each of these hypotheses, the candidate theory presented several contexts and mechanisms which were expected to lead to different outcomes under these hypotheses - some of them additional, some non-additional. This candidate theory had been informed by our evaluation of the first TA scheme. For each high-level hypothesis, the theory defined detailed 'Context-Mechanism-Outcome' (C-M-O) configurations that explained how the objectives of the second TA might or might not be achieved. The C-M-Os for high-level hypothesis H1 are set out in Figure 1 below, while the C-M-O configurations for high-level hypothesis H2 are set out in Figure 2. Testing of whether these C-M-Os applied to a given participant in the second TA would effectively test for the outcome (i.e. the additionality or non-additionality of TA influence with respect to H1 or H2), and for the contexts and causal mechanisms that led to this outcome.

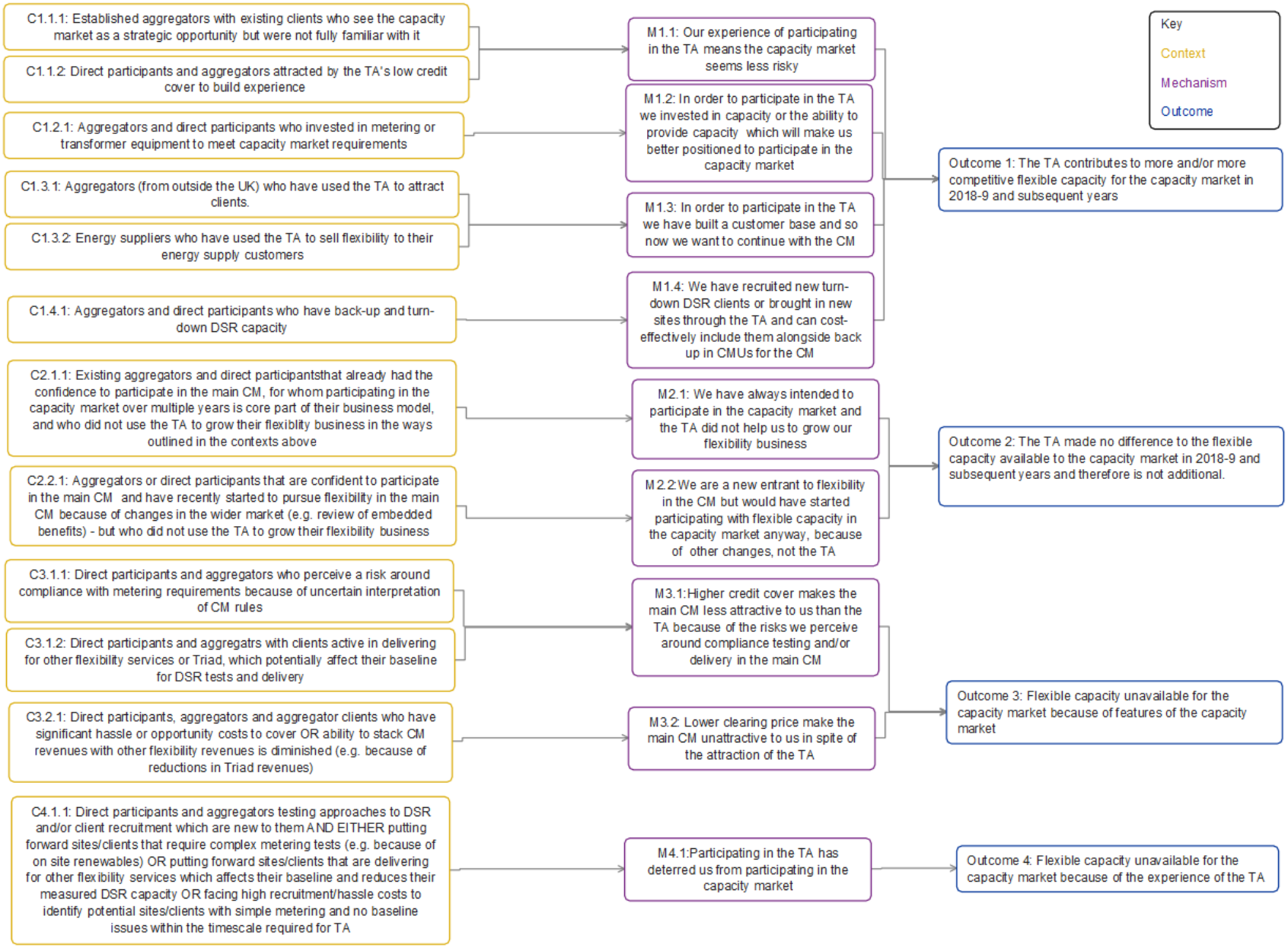

Figure 1: C-M-Os for hypothesis H1: the second TA contributes to more and/or more competitive flexible capacity for the Capacity market in 2018-19 and subsequent years

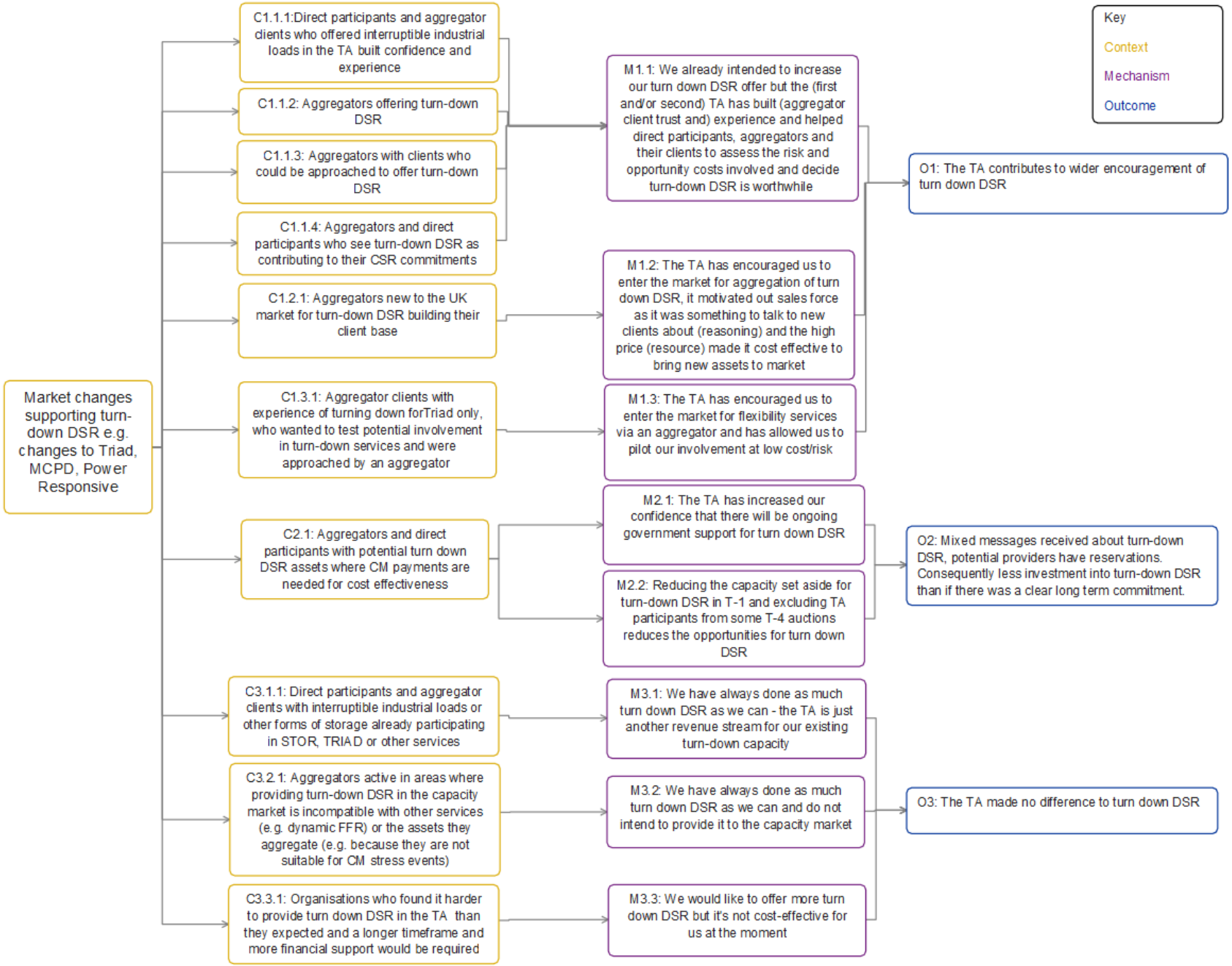

Figure 2: C-M-Os for hypothesis H2: the second TA leads to wider encouragement of turn-down DSR

How we collected evaluation evidence

The sources of evidence that were available to the evaluation, against which the C-M-Os could be tested, were as follows:

- Publicly available data (e.g. Capacity Market Register, published surveys and reports, aggregator and National Grid websites)
- In-depth interviews with all six aggregators participating in the second TA scheme (and access to notes from earlier interviews undertaken with the same organisations during our evaluation of the first TA scheme)
- In-depth interviews with a sample of eight organisations participating in the second TA scheme as clients of aggregators
- In-depth interviews with the two out of the three direct participants participating in the second TA scheme (and access to notes from earlier interviews undertaken with the same organisations during our evaluation of the first TA scheme)
- Email survey information from aggregators and clients, including confidential cost data and characterisation of capacity as new or existing, where available.

The in-depth interviews were structured using topic guides, agreed with BEIS, that fully explored the theory of change and the process tracing evidence tests. This involved some iteration between developing the topic guides and developing the process tracing method (see below) to ensure that the full range of evidence tests were covered. The interviews were recorded, transcribed and analysed in spreadsheet form, allowing analysis against different aspects of the theory of change and analysis against the process tracing evidence tests. The publicly available data sources and email survey information provided more objective sources of information that could be used to cross-check the subjective views put forward in interviews.

How we applied process/contribution tracing

We sought to use process tracing in a way that directly tested the C-M-O hypotheses in our candidate theory of change. This was because we wanted to integrate the process tracing into our assessment of the strength of evidence supporting each C-M-O and its constituent elements. We developed and applied evidence tests for each C-M-O hypothesis in the candidate additionality theory, based on process tracing concepts. And we then used process tracing to test the strength of evidence for competing additionality hypotheses (treating each C-M-O configuration in the candidate theoretical framework as a separate hypothesis). We applied process tracing on a case by case basis, which was consistent with our realist approach to analysis and synthesis. The steps in the process can be summarised as follows:

- We defined the competing hypotheses to be tested (i.e. realist C-M-Os), and the case or cases to be examined.
- We specified a set of evidence tests for each of the competing hypotheses, drawing on a range of available evidence sources (both objective and subjective). These were 'clues' to be looked for in the research evidence, to help distinguish between the competing hypotheses.
- We categorised each evidence test using the four process tracing categories ('hoop', 'double-decisive', 'smoking gun', and 'straw in the wind'). While we did not assign numerical probabilities to each evidence test, this involved broad assessment of the likelihood of observing the 'clue' if the relevant hypothesis was or was not true.
- We collected evidence as outlined above and reviewed all the available research evidence for each case, to assess whether each potential 'clue' has or has not been observed for a given case (i.e. whether each evidence test had or had not been passed for that case).
- We used this evidence to assess the relative merits of the competing hypotheses for that case (i.e. to assess which C-M-O best fitted each case).

We did not attempt to quantify probabilities and did not use Bayesian updating in this process. There were two reasons for this: firstly, there were 17 C-M-O hypotheses in the candidate additionality theory for H1 and H2. With (say) four independent tests per C-M-O there would have been over 50 evidence tests. We would have needed to estimate (or define ranges for) over 100 probabilities and we thought this was unmanageable. Secondly, it was problematic to define independent evidence tests for different aspects or elements of a C-M-O, because they were causally related. The Bayesian updating method requires that evidence tests are independent rather than causally related.

Our approach was therefore to categorise our evidence tests using the four categories of process tracing tests, without using Bayesian updating. We categorised the tests according to the rough likelihood of that piece of evidence being observed if the C-M-O was or was not true. We effectively used this categorisation to assess the weight that should be attached to a particular piece of evidence when considering whether a given case (i.e. organisation) exhibited a particular C-M-O.

In line with Pawson's 'realist evaluation cycle' (Pawson, 2006), we would have been prepared to refine or revise the theory and associated evidence tests, until we were confident that our refined theory was well supported by the evidence. However, in practice we found that each of the cases tested broadly supported one or more of the candidate C-M-Os so we did not need to refine or revise the theory to get meaningful results from process tracing. The reason for the accuracy of our candidate theory was that it was informed by extensive evidence collected during our evaluation of the first TA.

How we developed the evidence tests

We developed a set of tests that specified the evidence that we would expect or like to see if each of the C-M-O configurations in the H1 and H2 theory was true. While we tried to identify evidence tests that related specifically to the causal linkages between M-Os and C-Ms, these were in practice difficult to distinguish from the tests for Cs, Ms and Os. So, for each C-M-O hypothesis, we looked at all the tests for the constituent 'C', 'M' and 'O'. The outcome test provided evidence that the outcome had been observed, while the context and mechanism tests provided evidence of how and why the outcome occurred.

Table 1 shows example tests for the second C-M-O (CMO1.2) in the theory for H1. The tests for the outcome 'O' are presented first, followed by further tests for mechanisms 'M' and associated contexts 'C':

- **The main test for outcome 1** under H1 (increase in flexible capacity for future CMs) used objective evidence from the CM register to demonstrate that a TA participant had gone on to obtain a capacity agreement for flexible capacity in the main CM, after the second TA. This was a hoop test – without it, this C-M-O hypothesis would not apply to this TA participant. (There were other supplementary tests for outcome 1, not shown here).
- **The main test for mechanism 1.2** (one of the causal mechanisms hypothesized to lead to outcome 1) used evidence from interview statements with the TA participant to the effect that they made investments for the purposes of the second TA that would reduce their costs of participating in the main CM. This was classified as a 'hoop' test - it was deemed as essential for this C-M-O to apply. But it was not sufficient to prove that CMO1.2 applied, because TA participants might have invested time or money that would assist their participation in other flexibility services, rather than in the future CM. (There were similar tests for other mechanisms hypothesized to lead to outcome 1, but these are not shown here).
- **The tests for context 1.2.1** in Table 1 looked for evidence about the investment of money or time in meeting the metering requirements of the TA. The tests focused on metering because this was a specific requirement for the TA and CM which did not apply to other flexibility services. One of the tests (j) was classified as a 'smoking gun' (i.e. confirmatory of this C-M-O) because it provided externally verified evidence of metering tests, while the other evidence test (i) was interpreted as a weaker 'straw in the wind' because it was based on evidence provided by the participant. Neither of these contexts was strictly necessary for mechanism 1.2 to apply, since the mechanism 1.2 could have been based on investment in controls rather than metering. But test (j) would provide strong confirmatory evidence that the participant had undertaken investment/work for the TA that would reduce the cost of their future participation in the main CM.

There was considerable repetition in the evidence tests, so we assigned nicknames to the tests. We developed a set of tables, of which Table 1 is one example, indicating the source of evidence for each test, its categorisation using the four process tracing categories, the competing explanations for observing that evidence, and the rationale for classifying the test. The evidence tests were reviewed by two peer reviewers, a technical peer reviewer with expertise in DSR and by Dr. Barbara Befani, an expert in process tracing. We made some minor adjustments to the categorisation and wording of evidence tests during the testing process, to improve consistency across the tests.

Table 1: Example of evidence tests for the first C-M-O under H1.

Evidence tests for elements and linkages	Source of evidence	Type of test	Competing explanations	Rationale for classification of test
H1 – OUTCOME 1-test(a.1) Second TA participant obtains capacity agreements for flexible capacity in T-1 or T-4 auctions in 2018	CM registers for T-1 and T-4 held in Jan and Feb 2018.	Expect to see (hoop)	Necessary for O1. Could be observed for cases supporting Outcome 2 - flexible capacity put forward in CM but not attributable to TA	Evidence that this outcome applies (although there might be some external reason why they don't bid/clear in 2018/19). Could be observed even if TA had no influence on the flexible capacity they offer in the future CM.
H1 - M1.2 - test (h): Evidence of causal mechanism: Second TA participant saying in interview that they or their clients have developed or invested in assets (e.g. controls/metering) for the second TA that reduce costs of participation in future CM	In-depth interviews	Expect to see (hoop)	Necessary for M1.2. Could be observed for cases supporting Outcome 3/4 -they may have invested for the TA but may not go forward in the CM	Likely to see if the second TA has positively influenced the flexible capacity they offer to the future CM, and if this mechanism applies, but may also see if controls will really be used for other flexibility services, not the CM.
H1 – CONTEXT 1.2.1 - test (i): details of significant investment in metering or control assets by aggregator, direct participant or one of the aggregator's clients (for at least one of this participant's CMUs)	Email survey responses for TA participants and clients	Like to see (straw in wind)	Could be observed for cases supporting Outcome 3/4 -they may have invested for the TA but may not go forward in the CM	Specific details in email survey provide more confidence than test (h) but there's still a possibility that controls will really be used for other flexibility services, not the CM.
H1 – CONTEXT 1.2.1 -test (j): metering certificate or National Grid/Elexon statements indicate that meter testing has been completed for one or more components within this participant's CMUs (except if testing was only related to metering for onsite generation that could already have participated in wider CM)	Metering certificate (plus clarification on purpose of metering from interview data or National Grid/Elexon)	Like to see (smoking gun)	No significant competing explanations	Unlikely to see as most participants avoided meter testing through careful site selection. Undertaking metering testing was itself an investment of time and effort. Metering testing is specific to CM so very unlikely to invest in metering unless planning future CM involvement. Stronger test than test (i).

Legend:

- Dark grey row - evidence tests relating to outcomes 'O'
- Pale grey rows - evidence tests relating to mechanisms 'M'
- White rows - evidence tests relating to contexts 'C'

How we applied the evidence tests

As explained above, we applied the evidence tests to each potential C-M-O as a mini hypothesis. Given the number of evidence tests, and the limited resources available, we focused the testing on the most complex and important cases. The tests were therefore applied to the cases of the six aggregators that went forward to delivery in the second TA, taking into account evidence from these aggregators and from any of their clients that had been interviewed. In one case, we also took account of evidence from a sub-aggregator that had submitted capacity via one of the aggregators but was not a participant in the second TA.

There were two types of cases where we did not apply the tests. Firstly, we did not apply the evidence tests to direct participant cases, because it was already clear from the evidence that TA outcomes were not additional in these cases (i.e. that they would have offered the same flexible capacity to the Capacity Market, involving turning down of electrical loads, irrespective of the TA). Also, there were only three direct participants in the second TA, so the test findings were likely to be disclosive. Secondly, we did not apply the tests to two aggregators that initially participated in the second TA but then dropped out, because there was little additionality in these cases. This allowed us to focus on applying the evidence tests to six cases, namely, the six aggregators that went forward to delivery.

We streamlined the process and reduced duplication by only applying tests where relevant to a particular case. For example, where evidence tests for an outcome were failed, we did not test for the supporting mechanism and context. Similarly, where evidence tests for a mechanism were failed, we did not test for supporting contexts. The test findings therefore indicate those C-M-Os that are well supported by the evidence. Where there are competing mechanisms for the same outcome (e.g. one additional and one not), the evidence tests show the relative support for additional and non-additional C-M-Os in the theory.

In applying the tests, we synthesised evidence from a range of sources, including but going beyond self-reported evidence from the aggregators themselves. These included the sources of data outlined above:

- Publicly available data (e.g. Capacity Market Register, published documents, websites)
- Interviews with these aggregators
- Interviews with their clients
- Email survey information for aggregators and clients (where available)

We used a spreadsheet to code evidence for each case against the evidence tests. The evidence summaries and coding were prepared by one researcher and reviewed by another member of the project team. The detailed evidence and coding were also reviewed by Dr. Barbara Befani. Although this evidence was anonymised it was potentially disclosive because of the small number of TA participants. So, we prepared non-disclosive summaries of the results that could be shared with BEIS without breaching confidentiality that we had promised to interviewees. We created high-level summaries which combined test results using the following synthesis rules, to indicate the combined level of support for each C-M-O. These rules were developed by the project team but were peer reviewed by Dr. Barbara Befani.

Table 2: Rules for combining process tracing test results

Key:	Explanation	Process tracing concepts:
Strong support	Confirmatory evidence: at least one 'sufficient' or 'necessary and sufficient' test passed. No necessary tests failed. Allow failure of some tests which are 'not necessary or sufficient'.	Confirmatory evidence: at least one 'smoking gun' or 'double-decisive' test passed. No 'hoop tests' failed, but allow failure of some 'straw in the wind' tests.
Some support	No necessary tests failed. Allow failure of some tests which are 'not necessary or sufficient'. No 'sufficient' or 'necessary and sufficient' tests passed.	No 'hoop tests' failed, but allow failure of some 'straw in the wind' tests. No 'smoking gun' or 'double-decisive' test passed.
Mixed support	Apparently contradictory results - including at least one 'necessary' test being failed but also at least one 'sufficient' test being passed.	Mix of 'hoop' test failures and 'smoking gun' or 'double-decisive' tests being passed.
No support	At least one necessary test being failed, and no 'sufficient' tests being passed.	At least one 'hoop' test failed. No 'smoking gun' or 'double-decisive' tests passed.

Results

Findings on hypothesis 1 (the second TA contributed to more flexible capacity in the main CM)

Our findings on additionality of H1 are summarized in Table 3 below. These findings indicate strong support for Outcome 1 under hypothesis 1 (i.e. that the second TA contributed to more flexible capacity being brought forward to the main CM, for any type of DSR).

All the aggregators had gone ahead to participate in the main Capacity Market (CM) and all attributed some growth in their portfolios or knowledge to the second TA. The main causal mechanisms seemed to be that the second TA made participation in the main CM less risky, and that participants (or their clients) had invested time or money in

developing capacity that would make them better positioned to participate in the main CM. For aggregators new to the flexibility market in the UK, an additional causal mechanism was that the second TA had helped them to build a customer base for DSR. For aggregators already in the market, an additional causal mechanism was that the second TA had enabled them to bring new turn-down clients on board, positioning them to participate with higher volumes in the main CM.

There was very limited support for Outcome 2 under hypothesis 1 (i.e. that the second TA made no difference to the capacity available to the CM in 2018/19 and subsequent years and therefore was not additional). This was based on mixed evidence from two aggregators who commented in interview that they would have gone straight into the main CM even without the TA (although the scale of their portfolios might have been reduced).

While the participants did voice some comments and criticisms about the second TA, there process tracing tests found no support for Outcome 3 (i.e. that aggregators would not put forward capacity to the main CM, although they had been willing to do so for the second TA) or Outcome 4 (i.e. that aggregators were put off the CM as a whole because of their experiences of the second TA).

Findings on hypothesis 2 (the second TA contributed to encouragement of more turn-down DSR)

Our findings on additionality of H2 are summarized in Table 4 below. These findings indicate strong support for Outcome 1 under H2 (i.e. that the second TA contributed to encouragement of more turn-down DSR, both within and beyond the main CM). There was some support for Outcome 2 (i.e. that the additionality of the second TA could have been greater if the rules about participation in the main CM auctions had been different), but there was no support for Outcome 3 (that the second TA made no difference to turn-down DSR).

These H2 process tracing findings indicate strong support for additional outcomes (i.e. Outcome 1 and 2) at aggregator level, with aggregators developing their confidence and systems to aggregate turn-down DSR and/or building their portfolios of turn-down clients. There was less consistent support for the additionality CMO at client level (because some clients were already involved in turn-down DSR for Triad and other services, and were simply adding the TA as another revenue stream). The apparent inconsistency between these findings is explained by the fact that some clients changed aggregator as a result of the second TA: while these were perceived as 'new' to the aggregator they were not necessarily new to flexibility. Clients previously turning down their electrical loads solely to reduce demand during Triads[5] were counted as additional, provided that the TA was their first external contract for flexibility services.

The main mechanisms underlying these additionality outcomes were that:

- Existing aggregators already intended to increase our turn-down DSR offer but the second TA built aggregator client trust and experience and helped aggregators and their clients to assess the risk and opportunity costs involved.
- Aggregators new to the flexibility market in the UK were encouraged by the first and second TA to enter the market for aggregation of turn-down DSR. The TA gave them something to talk to new clients about and the high price (made it cost effective to bring new assets to market.
- For aggregator clients offering turn-down DSR, the second TA encouraged them to enter the market for flexibility services via an aggregator and has allowed them to pilot or increase their involvement at low cost or risk.

[5] At the time of this research, transmission charges were based on demand during three period of peak national demand (or 'Triads). Many major electricity consumers tried to anticipate the Triads and reduce their electricity demand during peak demand periods, in order to reduce their transmission charges.

Table 3: Process tracing findings for hypothesis H1 – the second TA contributed to more flexible capacity in the main CM

		Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Outcome 1: (additional)	**The second TA contributes to more and/or more competitive flexible capacity for the capacity market in 2018-19 and subsequent years**	Strong support	Strong support	Strong support	Strong support	Some support	Strong support
CMO 1.1	Our experience of participating in the second TA means the capacity market seems less risky	Strong support	Strong support	Strong support	No support	Strong support	No support
CMO 1.2	In order to participate in the second TA, we invested in capacity or the ability to provide capacity which will make us better positioned to participate in the main CM	Strong support	Strong support	Strong support	Some support	No support	Some support
CMO 1.3 (new entrants)	In order to participate in the second TA, we have built a customer base and so now we want to continue with the CM	Strong support	No support	No support	Strong support	Strong support	Strong support
CMO 1.4 (existing aggregators)	We have recruited new turn-down DSR clients or brought in new sites through the second TA and can cost-effectively include them alongside back-up in CMUs for the main CM	No support	Strong support	Strong support	No support	No support	No support
Outcome 2: (non-additional)	**The second TA made no difference to the capacity available to the CM in 2018/19 and subsequent years and therefore is not additional**	No support	No support	No support	No support	Some support	No support
CMO 2.1 (existing aggregators)	We have always intended to participate in the CM and the TA did not help us to grow our flexibility business.	Not relevant	No support	No support	Not relevant	Not relevant	Not relevant
CMO 2.2 (new entrants)	We are a new entrant to flexibility in the CM but would have started participating with flexible capacity in the CM at the same level anyway, because of other changes, not the TA	Not relevant	Not relevant	Not relevant	Mixed support	Mixed support	Not relevant

Table 4: **Process tracing findings for hypothesis H2 – the second TA contributes to encouraging more turn-down DSR**

		Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Outcome 1: (additional)	**The second TA contributes to wider encouragement of turn-down DSR**	Strong support	Strong support	Strong support	Strong support	Strong support	Strong support
CMO 1.1 (existing aggregators)	We ALREADY INTENDED to increase our turn-down DSR offer but the second TA has built (aggregator client trust and) experience and helped direct participants, aggregators and their clients to assess the risk and opportunity costs involved	No support	Strong support	Strong support	No support	No support	No support
CMO 1.2 (new aggregators)	The (first and) second TA has encouraged us to ENTER the market for aggregation of turn-down DSR, it gave us something to talk to new clients about and the high price made it cost effective to bring new assets to market	Strong support	No support	No support	Strong support	Strong support	Strong support
CMO 1.3 (clients only)	The TA has encouraged us to enter the market for flexibility services via an aggregator and has allowed us to pilot or increase our involvement at low cost/risk	Some support	Strong support	No support	Strong support	No evidence available	Strong support
Outcome 2: (additionality could have been greater)	**Mixed messages received about turn-down DSR, potential providers have reservations. Consequently less investment into turn-down DSR than if there was a clear long-term commitment.**	No support	No support	No support	Some support	No support	No support
CMO 2.1	Reducing the capacity set aside [for turn-down DSR/other capacity] in T-1 and excluding TA participants from T-4 reduces the opportunities for turn-down DSR	Not relevant	Not relevant	Not relevant	Some support	Not relevant	Not relevant

Reflections on this application of process tracing

The process tracing work was resource-intensive: 4-5 evidence tests were required to test each C-M-O fully, so the testing process involved nearly 80 tests across 17 C-M-Os (with some repetition of individual tests between different C-M-Os). The tests were kept manageable by focusing on only 6 cases where additionality was not straightforward.

One benefit of using process tracing was that the specification of tests encouraged explicit analysis of wider evidence (e.g. the extent to which participants had put forward capacity into other CM auctions). Another benefit was that the researchers could share high-level findings on the level of support for different C-M-Os with the client, without breaching anonymity for participants. Also, the testing process highlighted some scope for the simplification of theory by providing patterns of findings for new and existing aggregators.

Insights from this application of process tracing within a realist evaluation were:

- Treating each C-M-O as a causal hypothesis makes sense in a realist evaluation
- Ideally, we would have specified evidence tests for the causal linkages between C-M and M-O but this was problematic (e.g. all such tests would have been based on subjective interview statements about causality)
- Our tests helped us to confirm which Mechanisms triggered observed Outcomes, but were less useful in telling us which Contexts were important in triggering the Mechanisms
- Realist evaluation is an iterative approach, generally involving the development of C-M-Os until they accurately reflect all the available evidence. While we did not need to revise the C-M-Os during the testing process, possibly because they had already been refined during the earlier evaluation of the first TA scheme, it is possible that several iterations of theory refinement and process tracing might be required if this method was applied in another evaluation where the theory was less well developed.

Conclusions

Process tracing (or 'contribution tracing') provided strong evidence that the second TA scheme had supported the development of flexible capacity for the main CM, and that it had encouraged more turn-down DSR to come forward. This additionality was observed for aggregators and their clients rather than for direct participants. The second TA encouraged aggregators to find new clients offering turn-down DSR, reaching some firms that had not previously offered flexibility services (except, perhaps, internally - via Triad avoidance). However, the second TA attracted very few direct participants and no direct participants that were new to turn-down DSR. The few firms that were sufficiently experienced/confident, and had sufficiently large electrical loads, to offer load turn-down directly to the TA, would also have been confident enough to participate in the main CM without the stepping stone of the TA. So, aggregators were the route by which the second TA met its additionality objectives.

The value added by the process tracing to this evaluation was:

- Process tracing concepts were useful for assessing and weighing evidence according to its likely reliability (e.g. considering potential lobbying bias - "they would say that, wouldn't they?")
- The analysis process reminded us to consider alternative explanations for observed evidence
- Evidence test findings lent themselves to presentation in visual form and allowed non-disclosive presentation of findings about the strength of evidence
- Process tracing approach facilitated combination of evidence from different sources (e.g. objective and subjective, qualitative and quantitative)
- But the development of evidence tests in collaboration with policy/technical experts took considerable time and budget.

Developing and applying process tracing tests is time-consuming, particularly in a realist evaluation where there may be large numbers of C-M-Os and large numbers of cases. The application of process tracing to other realist evaluations is likely to be most practical where:

- There are just a few cases (or where causality is clear for all but a few cases)
- Evidence comes from a range of different sources, which need to be weighed against each other
- A few independent evidence tests can help to discriminate between these hypotheses
- There are adequate resources to support development of tests and assessment of probabilities – ideally this would be done participatively
- There are a relatively small number of competing causal C-M-O hypotheses.

Acknowledgements

CAG Consultants and BEIS are grateful to Barbara Befani and Charles Michaelis who advised on development of the process tracing method for this evaluation.

References

BEIS. 2018. Evaluation of the transitional arrangements for demand-side response – Phase 2. UK. Available at: https://www.gov.uk/government/publications/evaluation-of-the-transitional-arrangements-for-demand-side-response-phase-2 (Accessed 21/2/20)

Befani, B. 2016. Testing Contribution Claims with Bayesian Updating. CECAN Evaluation and Policy Note No.2.1 Available at: https://www.cecan.ac.uk/sites/default/files/2018-01/BARBARA%20v2.5.pdf (Accessed 21/2/20)

Befani, B. and Stedman-Bryce, G. 2016. Process Tracing and Bayesian updating for impact evaluation. Available at: https://journals.sagepub.com/doi/abs/10.1177/1356389016654584 (Accessed 1/12/2020)

Collier,D. 2010 'Process tracing: introduction and exercises', Department of Political Sciences, University of California, Berkeley.

Murray Brown, A. 2016. Contribution tracing – a brand new evaluation approach to improve your programme's impact. Available at: https://www.annmurraybrown.com/single-post/2016/06/20/contribution-tracing-a-brand-new-evaluation-approach-to-prove-your-programmes-impact

National Infrastructure Commission, 2016. Smart Power: A National Infrastructure Commission Report. Available at: https://www.gov.uk/government/publications/smart-power-a-national-infrastructure-commission-report. Accessed 21/11/2020

Pawson, R. and Tilley, N. (1997) Realistic Evaluation. London: SAGE Publications Ltd

Pawson, R. (2006) Evidence-Based Policy. London: SAGE Publications Ltd.

Treasury Board of Canada Secretariat (2009). Theory-Based Approaches to Evaluation: Concepts and Practices. Available at: http://www.tbs-sct.gc.ca/cee/tbae-aeat/tbae-aeattb-eng.asp. Accessed 21/11/20