2019

Be the Business Evaluation Framework

Table of Contents

Executive Summary

The UK has a longstanding productivity challenge, with large productivity gaps between the UK and other advanced economies such as Germany, France and the United States. The Government's modern Industrial Strategy sets out a long-term plan to boost the productivity and earning power of the UK by focusing on five foundations of productivity: Ideas, People, Infrastructure, Business Environment, and Places.

In July 2015, some of Britain's most senior business leaders came together to identify practical steps to raise productivity among British businesses under the banner of the Productivity Leadership Group (PLG)[1]. In the Autumn Statement 2016, the Chancellor announced £13m in funding over three years to support this work. Be the Business (BtB) was created to help UK businesses to take practical steps to improve their performance. It is the movement to drive up UK productivity through inspiring and helping every firm in the country to improve their own performance, and the performance of those they work with.

This framework is intended to assist key stakeholders of BtB, senior leadership, programme teams and delivery bodies to understand the approach BtB takes towards evaluation and what standards are expected in the form of robust evaluations to inform future decision making. Robust evaluation helps key stakeholders understand impact, informs spending reviews and fiscal events. Programmes vary in nature and as a result this framework will not set out one particular way of conducting evaluation. Instead it will highlight different approaches that can be taken, what data and governance requirements there will be and provide some examples of evaluation approaches taken so far. Evidence generated from these evaluations will help build up a picture of the relative impact and cost-effectiveness of these programmes and BtB overall which will inform the future allocation of resources.

The BEIS 2019 evaluation framework and the Green Book underpins this framework (BEIS, 2019; HM Treasury, 2018). All evaluations will be conducted in line with Green Book principles and should use both this framework and the BEIS evaluation framework as the starting points for the work completed.

Be the Business Programmes

The BtB portfolio involves several programmes at different stages of development. The framework focuses specifically on the programme activity of BtB rather than the wider campaign and movement (effects which will be explored further as the organisation evolves).

Programmes are typically delivered by internal programme leads and a mix of regional delivery partners such as Universities, Growth Hubs and LEPs. Programmes are funded by both BtB and private sector partners. This diverse stakeholder network will have potentially different requirements for data collection (both process and governance), evaluation design and ask different questions about impacts and outcomes. Therefore, the evaluation strategy must be sufficiently detailed and flexible enough to accommodate a mix of objectives. For example, programme leads will

[1] Members of the Productivity Leadership Group include Sir Charlie Mayfield (John Lewis Partnership), David Abraham, Tera Allas (McKinsey & Co), Jeremy Anderson (KPMG), Sir Roger Carr (BAE Systems), Roger Connor (GlaxoSmithKline), Ian Davis (Rolls-Royce), Carolyn Fairbairn (CBI), Doug Gurr (Amazon), Christopher Handscomb (McKinsey & Co), Lady Barbara Judge (IoD), Dame Fiona Kendrick (Nestle), Sir Richard Lambert (British Museum), Juergen Maier (Siemens), Sir Mike Rake (BT Group), Phil Smith (Cisco) and Nigel Whitehead (BAE Systems).

benefit from early outcome evidence to help design and improve their interventions; whereas Government would like to see evidence of longer-term impacts on overall business productivity.

BtB is currently at the stage of testing and learning what approaches and programmes work best to provide productivity impacts. This means programmes may change and develop overtime and the evaluation team will have to be engaged throughout this process. The portfolio approach may mean different programmes and methods might be deployed to work out the most effective ways to engage firms.

Some of the current programmes are outlined below.

The **benchmarking tool** serves as a diagnostic, that helps address the performance issues most businesses face. It highlights strengths and weaknesses in different thematic areas of management practice and provides a customised action report that features guidelines and “how to” on how and why to address a weakness.

There are two **Network** interventions. The Collaborative Networks Programme for Hospitality promotes business improvement in Cornish hospitality firms. Launched in early 2018, the network gives business leaders the time and tools to work on their business together. The network in the North West focuses on family businesses. There, the scope of the network is to complement existing family business membership bodies, focusing on support to improve management practice.

Mentoring for Growth facilitates collaboration between mentors from big corporations and key decision makers, owners and Managing Directors from SME organisations. An open and honest exchange of ideas, expertise and experiences helps shape the future direction an SME might take.

The **Productivity through People** programme provides leadership training with a strong focus on people management (rather than finance or marketing) and is run through business schools.

Productivity and the long-term impacts of programmes

In recent years, productivity literature has grown at the firm level, offering a more granular insight into productivity performance. Analysis undertaken by the Productivity Leadership Group, with support from McKinsey, shows that whilst the UK has many innovative businesses, the country also has a long tail of businesses whose productivity is below its potential (McKinsey, 2017). This problem is seen across the spectrum: in small, medium and large businesses, and in all sectors of the economy. Two thirds of employees in the UK (66%) work for under-performing firms, a figure eleven percentage points higher than in Germany.

Evidence points to a strong link between better managerial skills and formal management practices (e.g. HRM, standards and certifications, accounting, etc.) on the one hand and productivity growth on the other. A relatively new and growing strand of research in this area finds a correlation between the prevalence of structured management practices and firm performance, including productivity (Bloom and others 2013, ONS 2017a, Bender and others 2016, Broszeit and others 2016).

Recent evidence has also shown that UK businesses underperform on the adoption of effective management practices, relative to top performing countries. In a survey of global management practices, the UK scored 3.02 out of 5 for management best practice, behind the US, Japan and Germany (Bloom et al., 2012).

Chapter one: Evaluation

This chapter describes the process of evaluation. As BtB has been operational for almost two years, the framework has been developed to build on and improve the structures already in place. The chapter highlights key evaluation challenges and theories and describes how data collection can be structured in BtB programmes to enable evaluations.

This chapter starts with the basics for implementing a robust evaluation. The description of the framework is at a high level, establishing principles for evaluation and then the approaches that might be used to support the development of a plan for each evaluation.

What is Evaluation?

High quality evaluation enables an understanding of what works, how, and for whom, and whether programmes have met their objectives. It can help answer key questions about the value for money of interventions, looking at how effective, efficient and sustainable they are. Evaluation should be built into programmes from the beginning, as this will help to prove what has been achieved and improve on-going activities. Evaluation should also be part of decision making processes with regards to the continuation or scale up of programmes.

To put it simply, it examines the implementation and impacts of a policy to assess whether the anticipated effects, costs and benefits were realised.

The purpose of this evaluation framework is to formalise our approach to capturing evidence for the impact and outcomes from productivity interventions. BtB programmes focus on the productivity challenge and the aim is to evaluate how effective those programmes are at improving productivity for supported firms and value for money.

Different types of Evaluation

There are different types of evaluation that can be completed, and they should be chosen based on the current stage of the programme. Some of these are outlined below:

- **Process evaluations** complement impact evaluations by capturing the experiences of end-users and focus on the implementation and delivery of a programme, identifying which components were particularly effective and ineffective (BIT, 2017). The Green Book defines process evaluation as supporting an understanding of internal processes used to deliver outputs, alongside what was actually delivered and when (2018).
- **Impact evaluations** quantify the effect of a given activity or set of activities (an intervention) and the extent to which they can be attributed to a programme by comparing those that receive the intervention to others that do not (BIT, 2017). These will be completed in line with green book principles including full cost/benefit analysis where possible.

An evaluation plan should be put in place alongside the development of a programme, the first stage of this is developing a theory of change underpinning the intervention.

Theories of Change

The theory of change should outline the overall programme aims and how they will be achieved through a set of activities. They should also highlight the underlying assumptions which form the basis for the change. The theory of change is a working document which should evolve as the programme develops over time, and as data is collected. BtB is currently at a stage where the organisation is testing and learning from its programmes so it is likely the theories of change will be adapted as the programmes evolve into a scalable intervention. The theory of change can help an organisation:

- Understand how specific elements of a programme are thought to bring about change;
- Outline milestones which are believed to be essential to the overall goal, as well as identify key stakeholders involved;
- Detect potential risks and barriers by outlining key assumptions about what is needed to achieve change across all stages of the programme; and,
- Develop a shared understanding and consensus on their overarching aim and purpose, keeping a focus on the end goal throughout.

When a specific element of the programme or discrete intervention is implemented, the theory of change can be translated into a logic model (see figure 1) to illustrate a specific pathway within the theory of change that leads from activities to outcomes. It is important in logic models that the impacts relate back to the initial market gap / failure that exists in the theory of change, otherwise it will not be clear how the programme will go about achieving the initial aims and solving the failure.

The design of the evaluation, including what data to collect, flows directly from the identified driver(s) of change in the logic model. These include short, medium, and long-term measures so that the programme can identify at an early stage whether the intervention is working and then develop a narrative about impacts as they emerge.

BtB has developed theories of change for each of the programmes. These theories of change are usually reviewed and updated as and when new implementation evidence emerges or if there are changes to programme delivery. BtB should ensure that any new programme it develops has in place a theory of change since it supports the monitoring of programme performance (i.e. outputs) as well as giving early evidence for how the intervention is working in practice (i.e. outcomes).

Figure 1: Logic Model

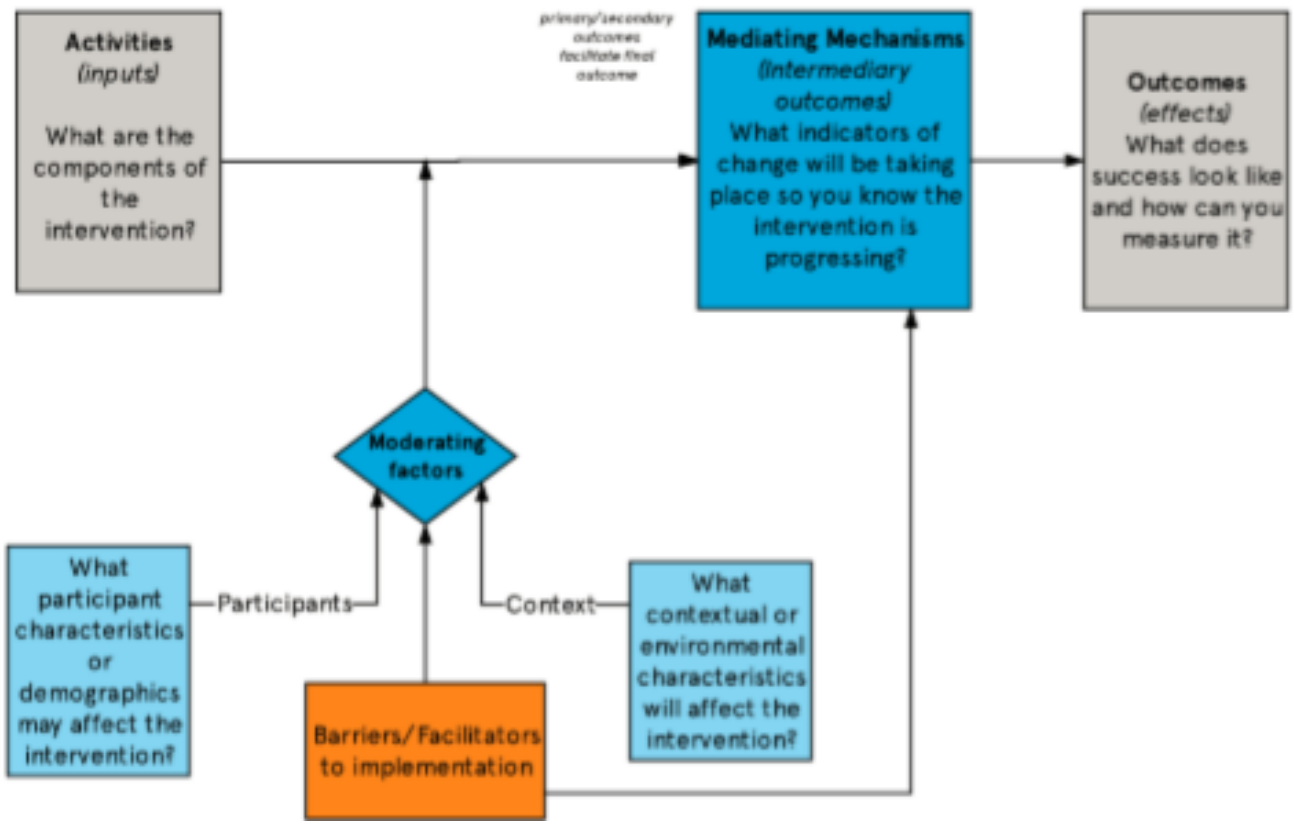


Figure 2 below synthesises across programmes to provide a general description also taking into account the timings of impacts and outcomes. This serves as useful context for the decision on questionnaire design as well as phasing of evaluation.

To the right of Figure 2 are the outcomes and impacts expected in the three periods after support that underpin the evaluation framework: immediately, medium-term and long-term (first six months, one-two years and after two years respectively). Broadly, the figure indicates that very soon after a support measure, in the first six months, the outcomes and impacts are changes in behaviours, knowledge and awareness, and related outcomes, such as confidence to take an action. After a year, the change in an SME leader's confidence should have resulted in outcomes improving management practices, such as new systems being put in place. The final set of impacts are focused on productivity, looking at improvements in business performance resulting from support.

The timeframe for changes in productivity can be up to seven years, with measuring the interim outcomes being important both because these are targeted by the programmes and because it is viewed as evidence that longer-term outcomes and impacts are on track.

Post-intervention, the expectation is to see similar impacts for all programmes. This will include, but is not limited to, sustained learning and adoption of good management practices and the increased uptake of other business support programmes as leaders and their firms become more engaged.

Figure 2: Programme Impacts and Outcomes

Inputs	Activities	Outputs	Outcomes	Impacts
IP Human resources Infrastructure and Facilities	BtB Programmes Productivity through People Mentoring for Growth Networks in Cornwall & the Northwest Benchmarking …and more to come	X number of individuals completing the Y programme	**Immediate: <6 months** **Individual:** Change in Knowledge, Awareness of Leadership Practices **Firm:** Improved understanding of benefits of support **Medium Term: 1-2 years** **Individual:** Increased self confidence **Firm:** take up of further support. Action towards better business practices Reduced wastage Increased staff engagement, skills and retention Improved quality of goods and services Increased margins, or sales Adoption of new to firm technology	**Medium term: 1-2 years** Improved management and leadership skills for labour force (industry volunteers and SME business leaders) **Long Term: >2 years** Increased productivity of SME participants measured by GVA/worker which is sales minus depreciation and costs of goods and services bought.

The next stage of evaluation planning is to design an approach to data collection and measurement. This allows the correct identification of correlations between programme activities and measures of success. However, without measuring the change in beneficiaries compared to a similar group of businesses that did not have access to the programme it cannot be concluded that the cause of change was the intervention, since there might have been an alternative explanation. The next section discusses different methods for constructing counterfactuals to ensure that there is no alternative explanation of the observed change.

Counterfactuals and Robustness

The purpose of constructing a counterfactual in evaluation is to understand what would have happened to businesses (or any other intended beneficiary of a programme) in the absence of an intervention. It is not possible to observe the outcomes for an individual or business both receiving and simultaneously not receiving support, so data should be collected for a group of almost identical subjects who did not get the support. Since the businesses are similar in every way, whether they got support or not, the only difference between the two groups (those receiving the programme and those not) is the programme itself. The quality of the tools used to reach this counterfactual analysis is determined by factors such as the design of recruitment to the programme, the data collected, and the analysis undertaken to check comparability. This is sometimes known as robustness. In evaluation methodological robustness is measured in terms of the Scientific Maryland Scale (SMS).

Identifying a suitable comparison group, alongside collecting before/after data for the supported and the comparison group will be key to all BtB evaluations. The comparability of these businesses determines the robustness of the

evaluation, especially in terms of reaching higher SMS scores. Box 1, on the next page shows the What Works Centre's scoring of evaluation robustness which is based on the SMS. BtB will always aim for the most robust form of impact evaluation; however, the key will be to ensure that every evaluation is appropriate and proportionate to the type of programme, its characteristics and potential impact. The next section discusses some of the challenges to the optimal evaluation approach.

Box 1: Levels of Evaluation Robustness

The figure summarises the Scientific Maryland Scale, scoring the level of evaluation robustness from one to five. Level 5 is reserved for research designs that involve explicit randomisation into treatment and control groups, with Randomised Control Trials (RCTs) providing the definitive example. At the other extremes are levels 1 (a cross-sectional comparison of treated groups with untreated groups, or a before-and-after comparison of treated group, without an untreated comparison group) and level 2, which improves on level 1 by using control variables in statistical analysis to adjust for differences between treated and untreated groups or periods.

Challenges for Evaluation

There are several challenges that evaluations of productivity initiatives will present, and these are outlined below.

Selection bias

Selection bias is a challenge for evaluators to overcome because the businesses coming forward for BtB programmes could be inherently better performing - for example, more ambitious, or capable. When identifying a comparison group for evaluation, the difference between firms that choose to sign up for programmes, and those that do not must be taken into account, as it is likely the result of unobservable characteristics (ambition or capability). This can be achieved in at least one of four ways:

1) Identification of other observable characteristics or variables that predict differences in uptake of a programme and correct for this in some way. An example of this approach is propensity score matching which could measure knowledge of productivity and desire to improve. If this approach is able to predict programme participation accurately, using the propensity score, the change in outcomes is determined to be a result of these

characteristics rather than due to higher capability for programme participants when compared to non-programme participants. One of the challenges to overcome here is getting enough insightful data on businesses to enable the identification of a strong matching counterfactual.

2) Identification of other observable characteristics or variables that lead to variation in whether a firm applies for the programme that are uncorrelated with the decision to apply. Examples of these approaches are known as instrumental variable analysis.
3) The use of “near misses” or businesses that were interested in programmes, but unable to take it up, as they will likely have similar motivations to the treated group. This is a form of regression discontinuity or kinked design and uses observable characteristics of a firm around an eligibility criterion to analyse what are, in essence, the same business type but so happen to be on two sides of a cut off for eligibility. For example, BtB could impose a strict lower threshold for its programmes at 10 FTEs. Firms with 9 or 11 FTEs might be otherwise identical but those with fewer than 10 staff could not be eligible for the programmes. The difference between firms on two sides of the cut off is measured to understand the effect of the Programme.
4) The gold standard for evaluation is the randomised control design (RCT). RCTs remove most selection bias through random assignment of firms to treatment and control because the uptake of a programme is no longer connected to individual characteristics; rather, the toss of a coin. However, careful attention must also be paid to the point at which randomisation takes place: if randomisation takes place too late, other forms of selection bias will emerge and so many RCTs randomise early to avoid measuring the effects of motivation to participate or experimenter rather than actual programme effects. Where this is a concern, evaluators typically apply either two stage randomisation or an intent to treat analysis.

Lagged effects and duration

Evaluation planning should always consider when impacts become observable. Productivity impacts may not be observable until 7 years after the intervention has been completed, which means it could be beneficial to wait until several years after the programme to complete the evaluation. However, this presents a trade-off between timeliness of evaluation reporting and certainty that any expected time lag in programme impact has translated into observable outcomes.

A further consideration is respondent recall that could be jeopardised by significant time between programme delivery and follow up survey. Indeed, contact details still need to be viable and a long follow up inhibits accurate tracking of beneficiaries. Therefore, planning the correct timing of evaluation depends, on the one hand, on ensuring the respondent has an accurate recollection of the support received and, on the other, sufficient time having passed to observe the actual effects of the intervention. The section on theory of change (above) provides some analysis of when to expect effects from the programme to show up and should serve as a guide to decisions on what to ask in surveys and when. When an evaluation plan is drawn up this will be taken into account. Interviews will be conducted in a timely manner and external data source will be used to gather information on the longer-term effects.

Statistical power

BtB interventions can be quite light touch, so small scale observable impacts might be expected. Light touch interventions can present a challenge for evaluation design if the sample size has not been planned to account for the expected magnitude of change of a programme on the outcome of interest. In other words, small effects are more difficult to detect so you need more data to ensure that what you observed is not random but results from the programme. However, a statistically desirable sample size might conflict with programme constraints such as time, cost and recruiting sufficient numbers of programme participants. The section below considers some options for mitigating some of these programmatic concerns whilst retaining the ability to detect small effects such as stepped wedge designs.

Attribution error

There could be concern that the reason for success might be some other support that the beneficiaries received at the same time as the programme, this is known as attribution error. The fundamental concern is that programme impacts will not be isolated from other background effects. BtB builds close relationships and creates programmes which integrate with support provided by the current network of business support organisations with a particular focus on improving productivity. Integration with the wider support network and business support programmes enables BtB to make limited resources go a lot further and benefit from the relationships that already exist and the knowledge of those partners. The challenge here comes with attributing any effects from that support and ensuring the impacts of the programmes are not double counted.

Evaluation should always include survey questions to ask what other support programmes or initiatives a firm has been involved with during the intervention period or before the follow up has happened to as best as possible avoid over or underestimating the impact of our programmes.

Evaluation Methods

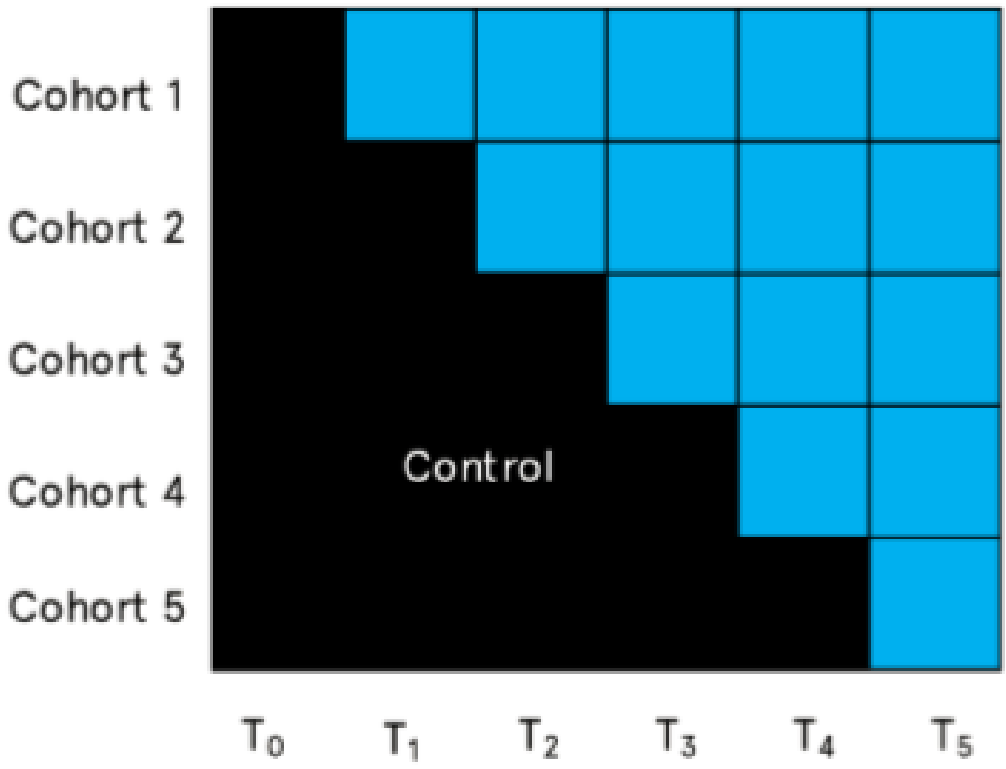
The different methods that can be used for impact evaluation are outlined in more detail below. When using these methods, the above challenges need to be considered to ensure the results of the evaluation remain robust.

As highlighted above, randomised control trials (RCTs) are widely considered to be the “gold standard” of evaluation methodology. Although there are several ways to construct a counterfactual, all other methods in some way face difficulties in ensuring that programme participants are not different to a comparison group in either observable or unobservable characteristics. By randomly allocating a group of firms who sign up for a programme either to take part in the programme or not, it can be ensured that the control group has similar characteristics to the treated group of firms since the only reason for whether a firm is in the programme is the toss of a coin. Given a large enough sample size, randomising participants into groups will produce, on average, groups that are similar to each other in terms of the things that can be observed (e.g: firm size, sector) as well as things that cannot be observed easily with businesses (e.g. staff motivation).

Randomised Control Trials

One implication of an RCT is the need for a group of firms that do not participate in a programme. Without this untreated group of firms, the counterfactual would be lost as there could not be a comparison between the impact of receiving support and not receiving support at the same point in time for a similar group of firms. This might be deemed unsuitable for certain programmes or organisations that are uneasy about the prospect of surveying firms who do not get access to a programme or could in some way discourage future participation in their activities. This must also be considered in light of the challenge of recruiting sufficient numbers to meet programme delivery targets. A way to overcome this could be a technique called the “stepped wedge RCT”, resulting in all participants that take part in the evaluation gradually being treated. The support would be rolled out sequentially over a specific time period, such that at the beginning of the trial no participants are treated, and by the end of the trial all participants are treated. Over the time frame, firms from the target group will be randomly assigned treatment at specific intervals. The diagram in Figure 3 opposite illustrates this point.

Figure 3: Stepped Wedge RCT



Randomised Encouragement Design & Dose Response

Other randomisation methods, designed to avoid a pure control group have been developed such as randomising the communication channels for potential recruits or varying the level of support received by beneficiaries. Randomised encouragement design can be used where BtB is unable to control who has access to the intervention. This could happen in the case of the BtB Benchmarking tool which is open to everyone via the web. Encouragement designs work by promoting an intervention to a specific group (the treatment group) more often over a defined period of time. Any difference in effects between the target group that has received encouragement, and others that use the tool, can be observed.

This is similar to a “dose response” method which actively varies the level of support for two groups in order to observe differences in impacts. By giving different ‘doses’ of an intervention, such as frequency in mentoring relationships or action learning sets, evaluation can help to determine the optimal intervention design required to have the desired effects. For example, in a mentoring relationship would six sessions be enough, or would 12 sessions be more effective?

Quasi-Experimental Methods

Quasi-experimental approaches may be possible if it is deemed that an RCT approach is not feasible. These would be the next most robust methods, but do not require randomisation of participants or businesses, hence, no control group is required. These methods use variation that naturally exists in the world in order to simulate or approximate random assignment.

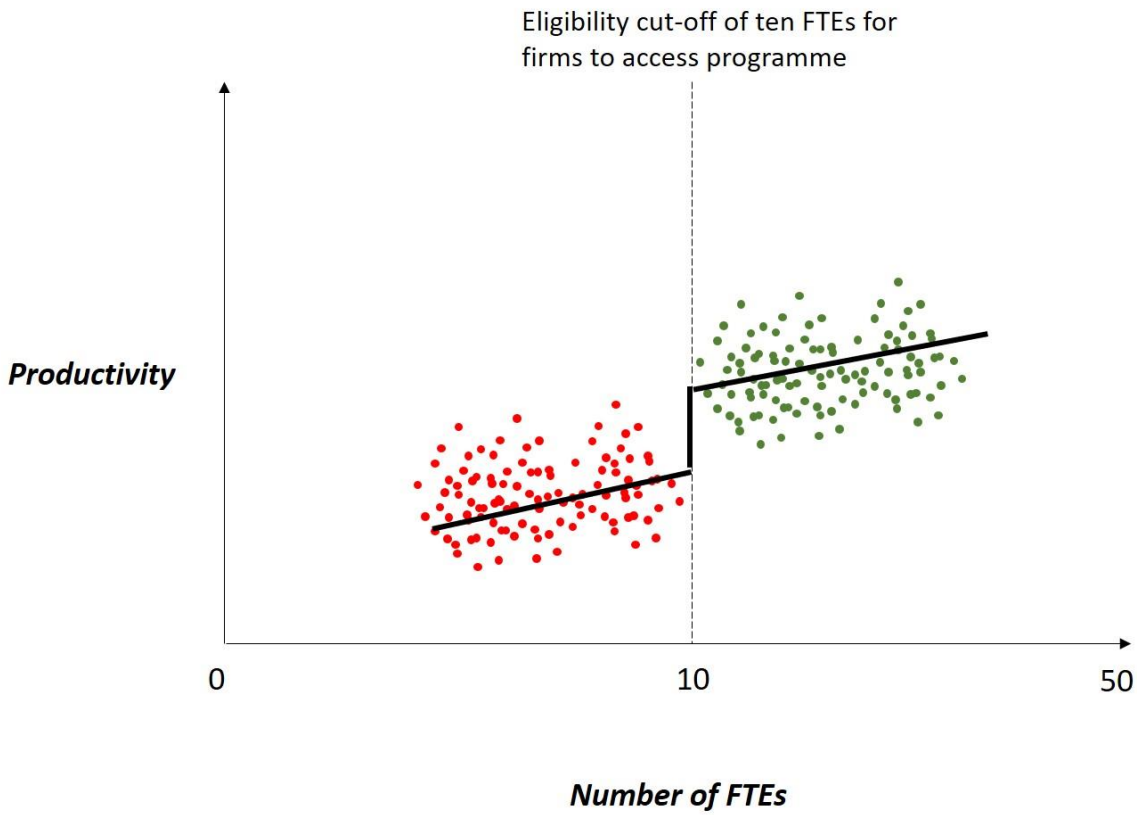
Regression Discontinuity Design

Regression Discontinuity Design (RDD) is a method which will provide a high level of robustness if performed correctly. It requires that firms or individuals that have signed up for a programme are ordered by some form of scoring method with a threshold (or cut-off) that determines whether they are eligible for support. An example might be number of employees or turnover. It can then be assumed that those just above the threshold are very similar to those just below, except for not receiving the support, and as a result, any difference in performance can be argued as being a result of the support rather than firmographic differences. The near misses can be a robust comparison group for the nearly rejected beneficiaries. The idea is that those near the threshold businesses are almost identical and that allocation of support is nearly random. While RDD is a robust method, it does have design features which limit its usefulness for some types of intervention.

These include:

- Firms well above the threshold for support cannot be included in the evaluation sample meaning the impacts from support could be understated (or overstated).
- It requires a sufficient sample of firms around the cut-off to be able to see a trend since evaluation only estimates a local average treatment effect.
- Finally, programmes that do not impose a strict inclusion criterion for participation in the programme are not suitable for this kind of methodology.

Figure 4: Regression Discontinuity Design



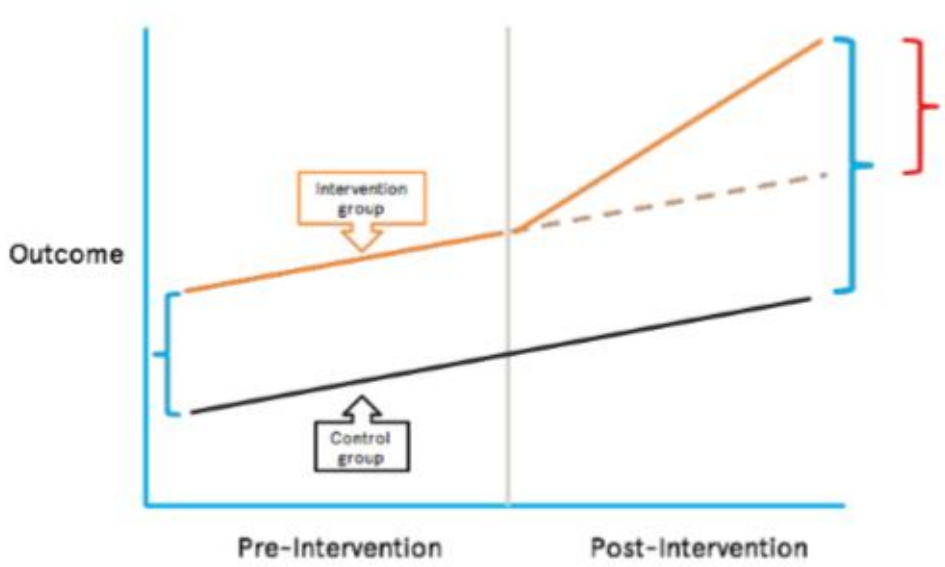
Difference in Difference

Difference in Difference (DiD) is a quasi-experimental approach that can be used to measure the impact of an intervention by finding a comparable set of control firms to the treated firms and comparing the outcomes before and after the introduction of the programme. This might be used, for instance, to compare performance of firms in one region, where an intervention was being rolled out, to the performance of firms in a similar region where it was not being rolled out.

A simplified version of the DiD method is displayed in figure 5 below. The analysis involves comparing two differences. The first difference is calculated before the programme is implemented and it measures any underlying differences between the two regions that are unrelated to the treatments. The second difference captures any underlying difference between the two treatments as well as the impact of the programme. By taking the difference of the two differences (hence the term “difference in difference”) the net impact of the programme can be worked out (signified by the red brackets in figure 5 below).

While this seems like a simpler method to implement, it is important to highlight that it is not as robust as an RCT and it relies on a number of assumptions. The key requirement of this approach is to find suitable comparators to the treatment group. These comparator firms would need to be sufficiently similar to the treatment firms and exhibit similar trends in productivity before the implementation of the programme such that, in the absence of the intervention, the treatment group would have undergone the same change in productivity as the control group. This is depicted as the dashed orange line in figure 5. Furthermore, firms would need to be found that have not benefitted from other concurrent programmes which aimed to improve their productivity. The intervention would also need to be designed in such a way as to avoid spillovers between the treatment and control group, as well as avoiding cases where firms in the control group are participating in the programme.

Figure 5: Difference in Difference



Matching

Matching is another quasi-experimental statistical method that attempts to address the issue of selection bias without conducting an RCT. With matching, businesses that take up the intervention (the treatment group) are matched with close comparator businesses that did not take up the intervention (the control group), in terms of the observable features that predict participation in the programme such as firm size, sector or turnover. In other words, the technique tries to predict a business’s propensity to be part of a programme. Outcomes are then measured across the two groups

for similar businesses that had a similar propensity to choose the programme. There are a variety of matching methods available such as propensity score or regression, but each method matches units on the basis of the observable characteristics (i.e. business traits that are measurable and contained in the available data). So, if the supported businesses tend to be large and in certain industries, these approaches will statistically create sets of comparable businesses that mimic those characteristics.

One downside of matching methods is that unobserved characteristics may make it difficult to robustly match businesses that take up the intervention to suitable comparator businesses. For example, while characteristics like region, industry, number of employees, and revenues are observable, other characteristics like awareness of productivity issues and management ability are harder to observe and may be associated with the propensity of businesses to seek support and high growth performance. Therefore, there should also be a consideration of what information could be collected as a proxy for motivation to sign up for a programme to tackle the selection issues of any programmes. Statistical modelling estimates the chance of support given the known characteristics of the business – estimating a propensity score – and then businesses are selected from the unsupported pool that have similar scores to the supported businesses.

Other Considerations

Evaluations need to consider how to address effects of interventions on different groups in society and opportunity costs and should justify any adjustments made based on existing literature on business support, and following guidance laid out in the Green Book, which provides these explanations:

- Deadweight: The outcomes that would have occurred anyway without the intervention.
- Displacement and diversion: The extent to which an increase in economic activity promoted by an intervention is offset by reductions in economic activity elsewhere.
- Substitution: Where firms or consumers substitute one activity for another as a result of intervention.
- Spillovers: the extent to which firms benefit from treatment (through diffusion or learning) even if they were not the intended targets of an intervention or where an unintended consequence is observed which may seem unrelated to the intervention.
- Leakage: The extent to which effects “leak out” of a target area into others.
- Opportunity Cost: The costs of using assets and resources are defined by the value which reflects the best alternative use a good or service could be put to.

A mixed methods approach will be considered where suitable in our evaluations in which researchers include both quantitative and qualitative research to measure the impacts of an intervention. As outlined in BEIS (2019), qualitative research can help in answering the how and why questions, which may not be captured in quantitative analysis and using both techniques can help to communicate the results in a way that appeals to a wider audience using both narrative as well as interviews, ethnographic surveys, case studies and charts.

Chapter two: Data and Governance

Good data underpins any evaluation and will be required throughout the evaluation process. As such, the governance and management of that data is also key. This chapter will cover general points about data, including what needs to be collected and when for evaluation purposes, but will not go into detail on full data governance. A detailed description of data governance and templates for data sharing agreements are located in BtB's sharepoint folders.

For the purpose of evaluation, emphasis should be with collecting high quality baseline data and contact information about the businesses participating in BtB and putting in place steps to centralise data such as appropriate permissions as well as track businesses through the support lifecycle.

The Magenta book outlines the following questions that should be asked around identifying data requirements (The Magenta Book, 2011) for evaluation.

- What data is required?
- What is already being collected / available?
- What additional data needs to be collected?
- If the evaluation is assessing impact, at what point in time should the impact be measured?
- Who will be responsible for data collection and what processes need to be set up?
- What data transfer and data security considerations are there?

These will be covered below along with more in-depth sections on the key parts of the process where data is required, how data might be collected and what data might be collected.

Baseline data

As covered previously, evaluation is trying to understand what has happened as a result of the intervention. To do that, it is important to understand what was happening before the intervention and then after the intervention so these can be compared. This means some form of baseline data collection is required.

At a basic level this should include any qualifying criteria for the programme and business characteristics (contact details, firmographics, the programme provided) so it can be understood who the individuals are and the characteristics of the types of businesses taking part in the programme. As outlined below, the baseline data could be collected through some form of expression of interest or application form. It would also be useful to collect contact details and identifiers of those who applied for the programme unsuccessfully. Near-misses, businesses that sought but did not receive support, are an important potential comparison group.

BtB's programmes vary in their approaches to firm recruitment, delivery and learning model hence careful thought is needed for how to sign up and track participants. Individuals can access BtB services online through the benchmarking tool; they are also likely to be recruited through partner networks. Both Productivity through People (PTP) and Mentoring for Growth (MFG) use partner networks (university or Growth Hubs) to target and find SME managers for the programmes. Further, the Growth Hubs will administer all surveys and data collection before and after the interventions they support. Whatever method is used, it will be important that the questions asked are suitable for the programme requirements, but also that it is accessible to the evaluation team. Some examples of the data it would be useful to collect are shown in Appendix C.

Ongoing Monitoring

Throughout a programme, it can be useful to collect ongoing monitoring data. For an evaluation team, this regular data can add value as it helps to understand how activities are engaged with by participants through a mix of interview, observation and data collection. Gathering opinions or follow up actions after a touch point, can help evaluators analyse change over time as well as the most effective delivery mechanism.

This will again differ for each programme. For example, the MFG programme might collect data after each mentoring session, providing both mentees and mentors the opportunity to comment on the session. Networks conduct a diverse set of activities and businesses are funnelled from the light touch (such as newsletters) through to more in depth interactions such as a lecture series (masterclass) or group learning activities (peer groups). There is likely to be a programme specific data capture as the intervention is provided and this could be useful for the evaluation.

Some examples of the data that it would be useful to collect for each programme are shown in Appendix C along with the examples of the baseline data.

Post Intervention

The next stage of data collection will begin once the beneficiary has completed the programme. This will provide an updated view of the firm after the intervention has occurred. There are three main ways this can be done and any one or a mix of them can be used depending on the type of evaluation and level of robustness required.

- Internal assessment form completed by every subject after taking part in a programme.
- Survey of subjects conducted by BtB or external consultants using a range of survey methods.
- Data linking with external datasets.

An internal assessment form at the end of the programme can provide some useful information for programme leads and anecdotal evidence from the respondents however, these will not be applicable for all BtB programmes as for some there is not a hard end date. These internal assessment questions ask mostly for self-reported data and are not always completed with accurate information. As such, they are not generally suitable on their own for a robust evaluation.

Surveys for impact and outcome measures can be used to understand productivity impacts. The surveys should be short and focused, and usually concentrate on actions taken by individuals following the programme. If possible, surveys could cover no-shows or those that expressed an interest but then did not go to the event.

Typically, the programme beneficiary in a BtB programme is usually the leader of a business. This means that surveys can contact an individual who is both the BtB participant and the individual responsible for (and able to comment on) business level impacts. The topics covered in surveys should be:

- Background characteristics of firm including age, size, sector, growth orientation, technology focus, innovation activity etc.;
- The performance outcomes the business is seeking to achieve;
- Type and intensity of support received: advice and guidance, training, networking, cost, hours, etc.
- How they heard about the programme and views on the programme;
- Any other support the business received;
- Attitudinal and confidence of business to take actions;
- Management practices and actions taken;
- Changes in business performance over time and contribution made by the intervention.

The use of business surveys in impact evaluations is common and there have been initiatives to bring a degree of standardisation to the questions used in the surveys. Generally, questions assessing the nature of the support and the views about its quality can be specific to the intervention. However, there may be advantages to using standard survey questions for business performance (such as management practices) and firmographic data, especially as this can allow comparisons with the data collected in public surveys. For example, using the same questions on employment change as the Longitudinal Small Business Survey (LSBS) and then having similar questions characterising the business will mean a comparison group could be drawn from the LSBS. There has also been a standardisation of questions about the adoption of best practice or new ideas by businesses in public surveys.

Some other examples of the questions to be asked can be found in Appendix B.

Data linking with external datasets is most likely to be effective when completed post intervention as the datasets will take time to record changes due to an intervention. However, there may be some instances where this can be done throughout the programme if the evaluation team has access to real-time data.

Any data sharing conditions should allow linking across different incidences of support and datasets. A business should be tracked through support pathways as this is useful in understanding selection processes and establishing the outcome of an intervention.

A key dimension to tracking the long-term productivity impacts is whether interventions are high intensity and applied to large enough numbers of businesses to be detectable. Low intensity interventions applied to small numbers of businesses are less likely to produce observable impacts in the administrative data.

To enable businesses to be linked to other datasets it is vital to collect at some point a form of business identifier. The most useful and common examples of these would be the full and correct registered company name and addresses as recorded at Companies House, a Companies House number and / or a VAT number (if they are registered for VAT).

Timing of Activities

The programmes that operate under the BtB umbrella differ considerably in the way they are delivered, the amount of support that is provided, how selection takes place and when impacts might occur. There is a complexity then in the pathways of support and the way in which businesses address the productivity challenges they face.

This results in support and impact timing being quite different for different interventions. For example, in the Networks, there may be an expectation of a build-up in the level of contact, with an initial introductory session translating over months into a more intensive form of support. It is important to have a governance structure that allows for the timing differing across interventions, while not overly complicating the way BtB's evaluation team interacts with the programmes.

Programme teams will work with the evaluation team to identify a key milestone around which a beneficiary becomes a contact for evaluation. At this milestone, programme teams would facilitate sharing data and enabling evaluation to take place.

This handover point will vary from programme to programme. For some, the initial application to a BtB programme will be a sufficiently significant interaction – such as applying for the Productivity through People programme. For the Mentoring for Growth programme, the Growth Hubs may be the initiator of connecting businesses to the programme and BtB would be provided with the details about this.

Identifying the milestone for each programme, and the steps the BtB evaluation team takes at the milestone, can be done by looking at the processes being followed by each programme and elaborated on in any process evaluation.

Evaluation built into Programme design

The BtB evaluation team will engage with programme teams from the outset of any intervention. This is to facilitate a smooth process for analysis and reporting in order to prove the success of a programme. It also ensures the data required for evaluation is collected and can be linked to administrative data by collecting the relevant baseline information (e.g. Companies House reference number) as well as ensuring the feasibility of robust evaluation methods such as RCTs are discussed and assistance provided for implementation of any evaluation.

BtB may consider the implementation of a data management system and customer relationship management system (CRM). It is important that these solutions are designed with consideration of evaluation requirements. Having the data in a centralised system can also enable cross-programme evaluations to take place.

Responsibility for Data Collection

Data collection is a joint responsibility between the programme and evaluation teams. Management reporting and programme monitoring falls under the programme team's remit and so it is expected that registration, beneficiary experience and monitoring of programme attendance will continue as their responsibility. The evaluation team will support the design of questions, data collection and analysis of this data as necessary.

For impact evaluations, data collection is the sole responsibility of the evaluation team (or through independent evaluators where this is commissioned). The evaluation team will continue to work with the programmes to identify the right balance between capturing as much data as necessary to answer intervention effectiveness without surveys becoming overly burdensome or costly to undertake.

Independence

For an effective and robust evaluation to take place, independent evaluation organisations can be an important part in that process. Not only do they provide valuable resource to conduct the evaluations, they also provide objectivity to the evaluation process. As it is usual to conduct surveys under research protocols, respondents can give views about the programmes knowing there is no risk of disclosure and they can be more open and honest about their experience. This can add greater credibility to evaluation outcomes.

Measures of productivity

A key dimension for the evaluation framework is whether productivity impacts can be measured. BEIS (2019) sets out measures of productivity impact, asking that all evaluations should collect the gross value added (GVA) and GVA per hour worked or GVA per worker measures.

However, it is recognised that this is a departure from the proxy measures using sales or turnover as it is not possible to calculate GVA unless a lot of financial data is collected from businesses. Even if this can be collected, unless the comparison businesses are surveyed, there will be no administrative source or public survey reporting GVA reliably over time for even a sample of SMEs. (The ONS Annual Business Survey refreshes its sample of small businesses each year meaning that changes in GVA would not be possible; full company accounts report GVA but SMEs are not required to complete these).

For SMEs, reviewing studies conducted in the UK, the proxy measure for productivity, turnover per employee, is feasible and the dataset is very rich. In the short and medium-term, evaluation outcome measures would be based on self-reporting/survey evidence. In the longer-term, administrative data is increasingly being used, particularly the ONS Business Structure Database, which tracks turnover and employment as reported to HMRC through VAT and PAYE returns.

The overarching aim of BtB is to improve productivity in the UK. All evaluations will assess how the programmes in the BtB portfolio have contributed towards that aim.

Chapter three: Applying the Framework

This chapter provides more advice on implementing the framework, applying it to four of the interventions that are currently (June 2019) in the BtB portfolio. The programmes differ in terms of the number of businesses that might be touched by the programme, the scale of impact an individual business would expect, and the size and timing of outcomes and impacts expected.

Evaluation approaches for the Productivity through People Programme

Productivity through People (PtP) is one of the original BtB programmes. As it is currently being evaluated, the focus in this section will be to outline what it is involved in the impact evaluation. The evaluation will use quasi-experimental methods following up on both the beneficiary and counterfactual groups.

Telephone surveys will be conducted with both the beneficiary group and three control groups. The three control groups are formed of a range of different firms who could be deemed similar to the beneficiary firms. These are:

- A group of firms that could be regarded as 'near misses'. Those who were approached to go on the programme, showed interest, but in the end were unable to attend.
- A group of firms that have been on a similar but lighter touch programme which did not provide the same level of support intensity.
- A control group who are being followed up from a previous survey. This group has been picked based on having similar firm motivations as the beneficiary firms in key areas. The proxy for this is based on questions from a previous questionnaire relating to 'plans to invest in L&D training' and 'being interested in support'.

There are also plans to undertake data linking with administrative datasets to observe the future performance of both the beneficiary and control groups. This entails matching businesses from the ONS Business Structure Database (BSD) to current delegates on the PtP programme to capture data on selected variables including: turnover, employment, etc. This aims to use the BSD snapshots for consecutive years from 2015 to 2019 and beyond. A control group may also be identified from the BSD data matched to the PtP group based on similar firm characteristics such as size, sector etc. Outcomes in terms of employment and turnover of both groups will therefore be tracked from the start of the programme, at the end and after completion. The evaluation will also collect historical impact data to establish any longer running trend in performance.

The evaluation has started to explore how variation exists within the delivery of the programme and how this could help BtB understand a difference in outcomes. However, due to beneficiary participant numbers this will be unlikely to provide detailed analysis of the variation and contribution to effects.

Evaluation approaches for the Mentoring through Growth Programme

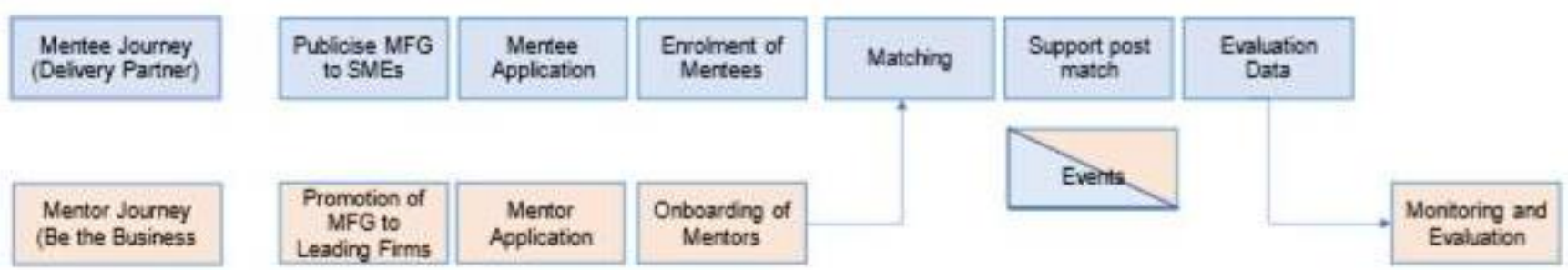
The Mentoring for Growth programme is currently aiming to provide benefits for both individual Mentees and Mentors and their organisations. The programme brings together key decision makers from SMEs and senior individuals from large firms, with the senior individual from the large firm providing advice to the SME leader.

The journey for the mentee and mentor is set out below. Going forward, this process and the data collection will incorporate a core system called iMentor. Mentee and Mentor application details will be collected in the system and will provide a user-friendly journey.

The Growth Hub advisor will use the system to help ensure they are making an appropriate match and monitor the overall progress of the mentoring relationship. The data collected in the system will be used for any future evaluations on impact.

The matching is one of the key stages of the process. Mentee's will be asked questions about their visions and aims of the company, future productivity plans, what support they need and the qualities they are looking for in a mentor.

Mentoring for Growth Mentee and Mentor Journey's



The logic model (see figure 1) is the first place where it is set out how the inputs will translate into the outcomes for these groups. This will be updated in the process evaluation of the programme to ensure any impact evaluation has a clear model to follow.

The programme has some clear inclusion criteria for the Mentees. These are:

- They need to have at least £2 million in turnover
- They need to have more than 10, but less than 250 employees
- The mentee must be a key decision maker / owner

These are some useful criteria which could enable a quasi-experimental method to be used such as Regression Discontinuity Design or Difference in Difference. However, the possibility of using one of these methods based on the inclusion criteria may depend on how rigorously that criteria is applied or how likely a firm is to apply if they don't meet the criteria. The level of rigour used in applying the criteria may depend on the number of applications for the programme.

Other potential options might include:

- To consider whether comparing early cohorts of the programme can be compared to those that receive mentoring afterwards, e.g. those businesses where leaders are mentored in 2018 are compared with those receiving support after 2020. This could be deployed where there aren't enough mentors immediately available,

so a queue is then formed. The baseline would be taken at the same time as those receiving support. Evaluation could then compare those receiving support against those waiting for support.

- The evaluation approach could be randomised with this programme, if application levels exceed capacity. Alternatively, the rejected applicants may be used as a control group.

During the early stages of the programme BtB will complete a process evaluation which will reflect on the selection process of mentees along with a range of other parts of the delivery and implementation. The outcomes of this process will inform the evaluation methods used for any future impact evaluation.

Evaluation approaches for the Networks Programme

The Networks Programme is designed as a multidimensional intervention that includes peer group learning, action learning and business advisory initiatives with the aim of drawing similar businesses together into networks that support one another to solve individual and collective productivity challenges.

The programme operates through multiple channels that each address a different aspect of the theory of change for how productivity initiatives improve management practices. At least four mechanisms have been identified that translate activities into outcomes: knowledge, aspirations, competition and collaboration:

1) Closing a knowledge gap will teach SMEs how to improve management practices;
2) Raising aspirations will drive a willingness to commit to behaviour change and adopt new practices;
3) Becoming more aware of competitor activities in the network will compel firms to consider how management practices can help them improve; and,
4) More connections with other businesses lead to higher levels of trust which drives collaboration, new ideas or mimicking the leading firms.

This complex web of change mechanisms leads the programme to focus on a multidimensional approach to growing networks as discussed above. Therefore, not all evaluation approaches will be appropriate for Networks. The programme is a package of activities that are designed to draw businesses to greater use of the network to learn through peer effects, to collaborate and to engender competition in working to improve productivity through adoption of new management practices. As a result, a mixed methods approach is desirable since it is expected that the programme operates through multiple channels which may be better understood if impact evaluation is combined with stakeholder interviews and feedback from beneficiaries on the most effective features of the programme.

There is no proposal at this stage for strict inclusion criteria for programme participation, hence, a discontinuity design is not considered appropriate.

A strict RCT is unlikely to be appropriate either given considerations for recruiting a control sample, contamination and spillovers. A larger sample is required to establish a pure control, but it is likely that firms may respond negatively to being denied access to the network. Further, there will be spillovers to firms in the control group who will likely be treated through the diffusion of ideas across networks even if they were not the intended beneficiaries of support.

In addition, there could be a displacement effect from supported firms drawing staff, customers and resources away from a control group. This might result from treated firms attracting business away from unsupported firms. In this case the intervention itself is not raising the overall productivity of the region but merely displacing business activity away from control firms and the wider unsupported firm population. Careful attention should be paid to not miss this phenomenon as it would not represent a net positive impact.

Therefore, the recommendation for evaluating networks is a mixed methods approach that couples a quasi-experimental or experimental design that does not require a pure control group with qualitative interviews.

The appendix highlights that this means some standardised baseline data is needed to track the support provided. The networks provide multiple incidences of support and the programme collects event related data. At present, the programme team collects behavioural data such as click through rates on any email contact or event feedback forms. This would need to be linked together to understand the pathways through successive incidences of support from the networks and the team is exploring how to track beneficiaries throughout their journey for this purpose.

One example of an evaluation design is a mixed methods evaluation of PLATO, a Belgian initiative to build networks, which found that the programme had a positive impact on productivity. The Before/after surveys studied impacts on attitudes, confidence and behaviour change (Van Cauwenberge, 2012). This was combined with quasi-experimental analysis using administrative data to track the performance of supported and unsupported businesses to conclude that the intervention was likely responsible for behaviour change amongst beneficiaries.

Evaluation approaches for the Benchmarking & Assessment Tool

Evaluation will proceed in two phases: first, a pilot stage which will test the impact of different messages to encourage uptake of the tool; and, second, a phase which will evaluate the impact of the intervention on productivity.

As an initial pilot phase, randomisation will be used to analyse the impact on behaviours of different messages or prompts by allocating different options randomly to users and then tracking further interaction by the user as the next actions are taken. These actions might be click through to the advice provided on the website or to further support measures.

As an objective of the tool is for firms to continue engagement with management practices, results will be accessible, and users will be encouraged to use other BtB services after undergoing a benchmarking process. Therefore, the messaging trial will be used also to test the optimal path to repeat interactions and follow up by the user.

Following the pilot phase, the intention is to run an impact evaluation of the tool. The hypothesis being tested through the Benchmarking & Assessment tool is that encouraging firms to focus on management practices, using a web-based self-assessment of those practices, leads to higher firm-level productivity.

Assessing the impact of the intervention on productivity will rely on being able to track users via firm level identifiers such as VAT numbers or Companies House registration. The evaluation is still being designed and three options are under consideration:

1) Tracking tool users with firm-level data from publicly available sources and the ONS Secure Research Service with an appropriate matched dataset of non-users constructed from the Annual Business Survey and Companies House data.
2) A Randomised Encouragement Design drawing from the pool included in the messaging pilot trial whereby those firms that received the invitation to participate are tracked against other firms who go through the tool by finding out about it in other ways. The difference in uptake rates between the tool respondents from the invited group to the background group of uninvited respondents allows the estimation of the Local Average Treatment Effect of the intervention for the complier pool that has been randomised into receiving an encouragement but who would not have tried the tool if not invited to do so.
3) A mixed methods evaluation which incorporates elements from each of the approaches above but makes an attempt to understand, via participant interviews, the changes and investment in management practices made by tool users from the point of benchmarking to the assessment of impacts.

Recognising the light touch nature of the intervention, it may be that for most beneficiaries the tool's purpose is solely as a recruitment method for other BtB programmes or other activities that lead to an increase in management capabilities. Therefore, the chain of results that begins with benchmarking could run through multiple management initiatives, only made possible through the tool, that in time generate improvements in firm level productivity.

Careful attention must be taken when designing evaluation to properly account for the role of benchmarking in driving behaviour change whilst recognising that the attribution of productivity benefits may not be directly related back to the tool.

Recommendations

This framework outlines the following recommendations based on a review of the current state of BtB.

Area	Key recommendations
Using random allocation	Whether a programme should introduce some randomisation to provide a control group could be tested as plans are made about the marketing of BtB programmes or – for randomised testing focused on behaviours – when programmes are being improved requiring systems development. The tests centre on whether the number of applicants is forecast to be high and the impacts of the support are large enough to be discernible statistically (see Appendix B, approach 3).
Plan the counterfactual before evaluation	Collecting contact details about potential comparison businesses that did not receive support is important. This might include collecting data on unsupported applicants from partner delivery bodies that manage selection. Further, to increase the quality

	of evaluation, one approach is to link into administrative datasets any firm-level variables or data that might correlate with support seeking by businesses.
More standardised baseline data needs to be collected across programmes	Baseline data collection should closely align across programmes and some key details should be collected such as the contact details of the supported individual (name, phone or email) and business details, with identifiers. Data should be collected about those who were rejected support or who contacted programmes.
Governance of evaluation and data collection	Governance can be quite light touch but some structure involving programme leads and BtB evaluators to own this framework should be put in place. Areas to cover: • Programme Leads overseeing the data collections for the baseline data and the data collected before and after the programme. • Common approaches to allow data sharing and linking across programmes. BtB should centralise data through the use of a data management system across the programmes which may also include a Customer Relationship Management system.
Establish a system for monitoring progress towards the implementation of evaluation recommendations for programmes	Ongoing monitoring, process evaluations and impact evaluations are a vital source of evidence for making incremental improvements to programme delivery as well as for deeper aspects of the BtB operating model such as beneficiary targeting and recruitment. BtB should think about how it acts to improve the delivery model of programmes so that it is held accountable for following through with recommendations emerging from evaluation. BtB should look to establish a regular forum (such as at programme board or equivalent) in which progress against implementing recommendations is reported on periodically. The reporting framework should be standardised and as light touch as possible to ensure ease of use for all programme teams.

References

BBB (2016) “Evaluation of Start-Up Loans: Year 1 Report”. Report to the British Business Bank by SQW Ltd and Policy Research Group at Durham University with BMG Research.

BEIS (2019) Business Support Evaluation Framework, Department for Business, Energy and Industrial Strategy, January.

Bender, Stefan, Nicholas Bloom, David Card, John Van Reenen and Stefanie Wolter (2016) "Management Practices, Workforce Selection and Productivity," NBER Working Papers 22101, National Bureau of Economic Research, Inc.

BIT (2013) “Applying Behavioural Insights to Organ Donation: preliminary results from a randomised controlled trial”. Report by the Cabinet Office Behavioural Insights Team to Department of Health.

BIT (2017) “ Evaluation and monitoring: Report by the Behavioural Insights Team for the Productivity Group”, September, unpublished.

Bloom, Nicholas, Christos Genakos, Raffaella Sadun, John Van Reenen (2012) ‘Management Practices Across Firms and Countries’, Feb, NBER Working Paper.

Bratberg, Espen, Astrid Grasdal and Alf Risa (2002) “Evaluating Social Policy by Experimental and Nonexperimental Methods”. Scandinavian Journal of Economics, Vol. 104, No. 1, pp. 147-171, March.

Broszeit S, Fritsch U, Görg H and Marie-Christine Liable (2016) “Management practices and productivity in Germany”, IAB Discussion Paper No. 32/2016, Nuremberg: Institute for Employment Research

Criscuolo, Chiara, Ralf Martin, Henry G. Overman and John Van Reenen (2019) “Some Causal Effects of an Industrial Policy”, American Economic Review, 109(1): 48-85

Frontier (2017) “The impact of public support for innovation on firm outcomes” BEIS Research Paper Number 3.

Goldman Sachs (2018) “10KSB: Building Small Business Britain”. Impact Report to Goldman Sachs.

Guest, David, Jaap Paauwe and Patrick Wright (2012) *HRM and performance: Achievements and challenges*. London: John Wiley HM Treasury (2011) “The Magenta Book – Guidance for Evaluation”, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/220542/magenta_book_combined.pdf

HM Treasury (2018) “The Green Book – Central Government Guidance on Appraisal and Evaluation”, https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-governent

McKinsey Global Institute (2017) Unpublished report to the Productivity Leadership Group.

Mentoring Plus (2004) “Mentoring disaffected young people An evaluation of Mentoring Plus”. Report by Michael Shiner, Tara Young, Tim Newburn and Sylvie Groben to the Joseph Rowntree Foundation.

Phipps, James (2017) “Taking the first steps in business policy experimentation”, Nesta Innovation Growth Lab blog, February.

Potter, Jonathan and David Storey (2007) “OECD framework for the evaluation of SME and entrepreneurship policies and programmes“, Paris: OECD.

Roper, Stephen (2018) “Using RCTs as a research method for SME policy research: The UK experience”. ERC Research Paper 66, Warwich Business School.

Van Cauwenberge, Philippe, Heidi Vander Bauwhede, Bilitis Schoonjans (2013) “An evaluation of public spending: the effectiveness of a government-supported networking program in Flanders”, Environment and Planning C: Government and Policy, 31: 24 – 38.

Appendix A: Reviewing Theories of Change

Theories of change are expected as a prerequisite for deciding whether to commence a new BtB programme. These theories of change are updated as and when new implementation evidence emerges or if there are changes to programme delivery.

Each of the programmes currently being run by BtB have theories of change underpinning them. These were reviewed by Belmana in spring 2019. The aim was to understand the logic of each intervention and then to synthesise across the interventions to establish a common set of outcomes and impacts.

Box 2 presents the detail about the early parts of intervention logic, focusing on the inputs, activities and outputs of Mentoring for Growth. Similar analyses have been conducted for each programme.

Programme intensity of interventions varies from a year commitment (PTP) to 6-8 minutes on the website (Benchmarking). For Mentoring for Growth a suggested commitment of 6-8 sessions between mentees and mentors is spread over a maximum of a year. The Networks programme involves an array of activities and interventions ranging from one Meet the Owner to intensive year long peer group activities. There is less of a clear expectation of intense involvement compared to PTP or MFG since beneficiaries choose their level of commitment from the varied activities all under the network membership. Benchmarking is considered an introductory tool to bring businesses into the website and then potentially attract them to other programmes.

The theories of change then consider outcomes. Changing business practices is the common theme, with increased confidence in making changes a frequent interim outcome for business leaders. The specifics can range from hiring new people, expanding the load of an SME leader, to new HR strategies learned from peers. The time-frame for changes in productivity is up to seven years, with measuring the intermediate outcomes being important both because these are targeted by the programmes and because it is viewed as evidence that longer-term outcomes and impacts are on track. For example, the Networks programme advance core leadership and management practices to improve business resilience.

BOX 2: Mentoring for Growth inputs, activities and outputs (Reviewed Spring 2019)

The *Mentoring for Growth* pilot worked with over 40 SMEs predominantly in manufacturing, pairing up business leaders with mentors drawn from large businesses. This pilot ended in September 2018 and was delivered through the Manchester and Birmingham Growth Hubs. Initial results in terms of both mentee and mentor satisfaction and impact have been extremely positive, with over 90% benefiting from the relationship according to the monthly surveys.

The objectives for the programme in year 2 were to: roll out the programme at greater scale; expand delivery to two new Growth Hubs (The North East and London); and, to develop and strengthen the evaluation. The scaled programme would provide 100 mentors to work with business across these four regions.

The core emphasis of the programme is:

- increase in SME awareness of technology and/or management practices;
- change in attitude towards use of technology and/or management practices;
- increase in SME ability to introduce change to organisations; and,
- successful adoption of new technologies and/or management practices.

Inputs (pre Oct 2018) *Tools or resources*	**Activities (Oct 2018-2019)** *Ways tools/resources used*	**Outputs (Oct 2019)** *Tangible services or products*

Intellectual property
• Design: GSK/BEIS/BAE
• Skills matrix - matching & code of conduct (Manchester Growth Hub)
Human resources
• 3 FTE across the East Midlands with extra support for admin, marketing, management, matching and sourcing
• 2 FTE from GSK for coordination and lead and sourcing mentors
• 120 Mentors from PLCs/PLG
Infrastructure and facilities
• Provided by growth hubs and BtB
• Partners (PLC and PLGs)
	• **Mentor recruitment and training** and support
	• **Mentee identification** target 120-200 mentees.
	• **Matching** process.
	• 3-4 **sessions** of 1-2 hours
	• **Support**
	  • Mentor induction prior to start of programme.
	  • Monthly connect and collaborate calls where all the mentors come together.
		SMEs / Mentees
		1. 200 completed relationships (0% attrition)
		2. >75% satisfaction for mentors
		3. >75% satisfaction for mentees.

Appendix B: Evaluation Approaches

Appendix B considers approaches to the evaluation of both immediate impacts of BtB programmes and longer-term outcomes and impacts. The appendix is structured bearing in mind how the timing of outcomes and impacts influences data collection approaches at the outset of the programme, moving towards approaches to assessing impacts and outcomes in the medium and long-term. It also references how different evaluation approaches will result in higher levels on the Scientific Maryland Scale highlighted in Box 1 of the framework as well as providing recommendations for data collection. For each of the five categories:

- A description of the approach and recent examples;
- The steps to ensure the highest quality level of evaluation for the approach; and
- Some indicative timings for the steps where the approach is implemented.

Approach 0: Baseline data collection

At the first interaction with either BtB or delivery partners the nature of support, contact details and firmographics are collected for supported individuals or businesses. Sometimes, baseline data collection will be staged, so that a first set of data would be collected at first contact and more data collected as a business applies to a programme.

The data can be split between the basic recruitment questions, covering the name of the contact in the SME and their email, phone number, role or job title and then details about the SME (entity name, address). In an evaluation, the business is likely to be linked to administrative firm-level data, which is made considerably easier if a business identifier is provided by the contact.

Typically, Companies House Number is the easiest to collect, though VAT, DUNS or PAYE registration numbers may also be known. If these are not known, it is also sensible to ask whether an entity is registered at Companies House, or for VAT/PAYE, as with name and address it is relatively straightforward for an evaluator to link the entity to its register entry if the name has been collected accurately. Beyond identifying the business and the contact point in the business, the baseline characterises the business in ways relevant to the programmes, such as asking whether a business is family owned, or a multinational business which has units in other countries.

Generally, the baseline evidence is collected by programmes and can form the basis for achieving a level 1 evaluation if the data collected is high quality using the nature of the businesses being supported to provide some insight about the businesses receiving support.

Approach 1: Before/after analysis

The before/after design is the most important non-experimental design. Although it suffers from many threats to internal validity, it can provide preliminary evidence for intervention effectiveness, especially when supplemented with complementary information.

The SMEs that interact with BtB are the focus of before/after analysis. Data collection seeks to track the progress of the businesses, providing the main evidence of what happens to the supported businesses. The key feature of before/after analysis is then to relate the timings of any changes in attitudes, behaviours or any actions taken to when

BtB support was provided. A key dimension to this analysis is to have good data about the timing and nature of the support provided, recognising that this may be over multiple discrete interactions with the support provided maturing over time.

BtB does start in a good position with regard to baseline data:

- Programmes typically collect benchmarking or diagnostic data at the start of the programme to assist in action planning; and
- The Networks programme, which does not baseline management practices or benchmark performance as standard, does collect behavioural data such as click through rates on email contact or event evaluation forms.

The benchmark questions on leadership are in the Box 3 below for the BtB web tool, covering the awareness, attitudes and behaviours of business leaders. The questions ask about the advice seeking behaviours of businesses. When BtB Network programmes hold events, the evaluation forms focus on similar issues, exploring the motivation for action of attendees. This covers key issues in networking, such as the number of connections (Pre and Post), the confidence levels of participants and other networking activities. An evaluation form then asks whether this has changed, as well as conversations about the issues raised in events.

Improving the SMS level to 2

Before/after analysis depends on the before and after measurements being carried out using the same methodology. A programme's satisfaction survey may ask individuals to rate key aspects of an event or engagement on scores from 1-10, with net promotion then calculated using the share with high scores adjusted for the low scores. If this is administered at an event that is imparting some skills, then the scoring may cover the change in awareness and confidence to use a new skill and the intention to do this.

There are some general guidelines about interpreting analysis, such as to identify outliers and contextualising outcomes by monitoring any natural changes in the population over time which could obscure the effect of the intervention. This may even require possible statistical corrections for their effect during the statistical analysis. Finally, it is sensible to identify the effects of intervention participants dropping out and allow for this in the analysis.

Satisfaction data provides an early indicator of outcomes and support improvements to the delivery and content of the support provided. However, to gauge impacts relies on collecting performance data, especially where an incidence of support is designed to improve business performance.

One way to facilitate these improvements is to ensure – where appropriate – that the data collected during benchmarking or as support is delivered is as standardised as possible. Box 3 indicates one possible improvement in the benchmark: it may be sensible to ask the question about whether a business received support from a mentor or advisor in a manner that is comparable to the Small Business Survey conducted annually by BEIS, where a similar question is "Whether the business used information or advice in the last year", allowing the SBS results to be used to establish if there are background trends.

Box 3

Survey questions on business awareness, attitudes and behaviours
Which of the following traits are your greatest leadership strengths? (BtB - Leadership)
How many times in the last year did you ask for support from a mentor or advisor? (BtB - Leadership)
Which of the following characteristics are most important to running your business? (BtB - Leadership)
Thinking back over the past year, what best describes the degree to which data was available to support decision-making in your business? (BtB - Leadership)
Whether you used information or advice in the last 12 months. (SBS)
Whether you used information or advice. (SBS)
Reason for using information/advice. (SBS)
Who you received information/advice from (SBS)
Awareness of business support organisations (SBS-A)

Box 4 focuses on questions about management practices, taken from the recent ONS Management and Expectations Survey, building on the Management Practices Survey of 2016.

The MES questionnaire covered a broader and slightly modified set of questions than the MPS, bringing it in closer alignment to the Management and Organisational Practice Survey (MOPS) conducted by the US Census Bureau. The MES survey attempts to measure four aspects of firms' management practices:

- continuous improvement practices – how well does the firm monitor its operations and use this information for continuous improvement?
- key performance indicators (KPIs) – how many KPIs the firm has and how often they are reviewed.
- targets – are the firm's targets stretching, tracked and appropriately reviewed?
- employment practices – is the firm promoting and rewarding employees based on performance, managing employee underperformance and providing adequate training opportunities?
- The mandatory questions were Question 6 - "How many key performance indicators were monitored with this business?" and Question 8 - "In 2016, which one of the following best describes the main time frames for achieving production/services targets within this business?".

The MPS questions are in Box 4 with comparable questions from the BtB benchmarking tool and the BEIS SBS. As with the behaviour questions above, there would be value in aligning questions to the public surveys so that the before/after analysis can use the public surveys to understand wider trends and improve the robustness of the data collected by programmes.

Box 4

Survey questions about the management practices in a business

In 2015, what generally best describes what happened at this business when a production problem arose? (MPS)
In 2015, how many key performance indicators were monitored at this business? (MPS)
In 2015, how frequently were the key performance indicators reviewed at this business? (MPS)
In 2015, what best describes the time frame of production targets at this business? (MPS)
In 2015, how easy or difficult was it for this business to achieve its production targets? (MPS)
In 2015, how were employees usually promoted at this business? (MPS)
In 2015, when was an under-performing employee moved from their current role? (MPS)
In 2015, who made decisions over the hiring of permanent full-time employees? (MPS)
In the past year, how frequently did you discuss your company's vision with your staff and/or the senior team? (BtB)
In the past year, how frequently did you discuss your company's values with your staff and/or the senior team? (BtB)
Do you make online purchases for your business? (BtB - Digital)
Can your customers complete orders and transactions with your company online? (BtB - Digital)
Do you make use of collaborative software for sharing documents, conferencing and communicating as a team or remote working? (BtB - Digital)
Do you use customer relationship management (CRM) software to collect, store and share customer information and data insights with other internal business functions? (BtB - Digital)
Do you use sensors, automation or Internet of Things (IoT) as part of your business operations, e.g. asset tracking, automation in production, machine learning? (BtB - Digital)
Do you use enterprise resource planning (ERP) or process management software to integrate and automate back office functions, e.g. those related to technology, services and human resources? (BtB - Digital)
Do you use a digital platform or application to automate and collaborate across your entire supply chain? (BtB - Digital)
Do you have a formal written business plan? (SBS)
In which of these ways does your business keep records for VAT?(SBS-C)
In which of these ways does your business keep records for income tax self-assessment? Cohort C only. (SBS-C)
In which of these ways does your business keep records for company tax? Cohort C only. (SBS-C)
Do you use any technologies or web-based software to sell to customers, or for use in the management of your business? (SBS-C)
Which of the following do you use? Cohort C only. (SBS-C)
Does your business have any of the following business and management practices?(SBS-C)

Timings for this approach

Data collection should occur immediately before and after an intervention with follow up survey data collected to asses impacts. The approach might be designed as follows:

- **Before support:** collect benchmark evidence repeating questions in the public surveys where practical.
- **After support:** establish any immediate changes in attitudes, confidence or intentions.
- **Follow up:** 6 months or so after the end of an intervention, a survey of actions taken would establish any immediate outcomes, such as whether advice covered in an event has been used. An e-Survey may be quick and reliable for this.

Approach 2: Quasi-experimental approaches using data-linking and surveys

Quasi-experimental approaches have been used in past programme evaluations of business support in the UK. These typically use company accounts data combined with business surveys or using financial data derived from the administrative data held at the ONS Secure Research Service (SRS). For the former, London Economics (2012a, 2012b) evaluated the beneficiaries of various export support interventions provided by UK Trade and Investment linking the supported businesses to the FAME/ORBIS database using Companies House identifiers.

The shift to using ONS data, especially in evaluating SMEs, was primarily because of the increased accessibility of the datasets and because the accounting databases did not contain accounts from small businesses. Reporting requirements are modest for small and micro businesses. The ONS data is based on tax administrative datasets, especially the VAT return and the PAYE accounts used to track employment and turnover for all businesses on the two

tax registers. This covers all SMEs of economic significance. In evaluating the Innovate UK support for R&D, beneficiaries were linked in the ONS SRS and propensity score matching (PSM) was used to identify a comparison group. Then the employment and turnover performance of both the supported and comparator businesses could be tracked (Frontier, 2017) (SMS level 3).

The evaluation of the Goldman Sachs 10k Small Businesses intervention is a recent example of a quasi-experimental approach in a policy evaluation (Goldman Sachs, 2018). To strengthen the quality of evaluation, the propensity score matching approach using the administrative data was mixed with a survey. The latter instrument allows the behaviours, actions and motives of the beneficiaries to be collected, alongside the economic impacts through the experimental approach on administrative data. It can also collect changes in management practices.

Survey questions about the performance of a business in the Small Business Survey
How many employees did the business have on the payroll 12 months ago across all UK sites (still excluding owners and partners)? (SBS)
How many employees do you expect the business to have on the payroll in the UK in twelve month's time (excluding owners and partners)? (SBS)
Whether export goods or services (SBS)
JHas your business introduced any new or significantly improved processes for producing or supplying goods or services in the last three years? (SBS)
Whether processes new to the business. (SBS)
Annual turnover. (SBS)
Compared with the previous 12 months, has your turnover in the past 12 months increased, decreased or stayed roughly the same? (SBS)
Percentage turnover increase (SBS)
Percentage turnover decrease (SBS)
Expectations of turnover growth in next 12 months. (SBS)
Percentage turnover increase (SBS)
Percentage turnover decrease (SBS)
Taking into account all sources of income in the last financial year, did you generate a profit or surplus? (SBS)

Surveys have sometimes designed questions to allow the results to be compared to a public survey. The evaluation of the Start-up Loan programme by SQW and Aston Business School utilised a constructed control group from the Global Entrepreneurship Monitor (GEM) Survey and other screened individuals who did not obtain the Start-Up loans but were similar in characteristics (BBB, 2016). The survey of beneficiaries replicated the questions from the GEM Survey. The evaluation of a leadership training programme in Northern Ireland constructed a control group from the non-supported Enterprise Ireland clients who fit certain clear characteristics – being within the same sector and size as the supported businesses – and using data on turnover, exports and employment (NI 2015. PLM Leadership 4 growth programme).

Generally, these studies receive an SMS level 3 assessment, because they collect before/after evidence for the beneficiaries and take steps to match the supported businesses to a comparison group based on observed characteristics.

Improving the SMS level to 4

Quasi-experimental approaches can become level 4 in quality if there is sufficient evidence that the statistical matching is robust, especially where it can show that some factor has been included which correlates perfectly with the chance of receiving support, but the business does not receive support.

A common way to achieve this is to look at those that were nearly selected for support. Whilst not a business support intervention, an evaluation of Mentoring Plus (2004) interviewed 'near-misses', people who had expressed interest, but not gone forward with the programme.

In business support evaluations, there are very few level 4 studies. One case is where an instrumental variable was identified. In evaluating the Regional Selective Assistance, the changes in the criteria by which areas were deemed eligible for support in England changed in a specific year. Criscuolo et al (2018) exploit changes in area-specific eligibility criteria to create instruments for programme participation

Timings for this approach

This is a longer-term evaluation approach since management behaviours and productivity impacts take time to materialise. The approach to data collection might proceed as follows:

- **Year 0:** collect beneficiary details and the support provided.
- **Year 2:** use linked administrative data to conduct early impact study.
- **Year 2:** conduct surveys of beneficiaries and – if possible – near misses or rejected applicants. Understand management impacts using survey evidence drawing any counterfactual from non-beneficiaries surveyed
- **Year 4:** full economic impact evaluation.

Approach 3: Programmatic Randomised Control Trials

Programme RCTs involve subjects that are facing some common productivity problem requesting BtB support but are then randomly allocated to a treatment and control group. Differences in outcomes between the treatment and control groups are then attributed to the effect of the policy intervention. In terms of industrial policy, however, experimental policy evaluation approaches remain marginal, with non-experimental, ex post policy evaluations remaining the norm. In the context of small business policy evaluation, Potter and Storey (2007), for example, provide an extensive review of best practice in OECD countries without any mention of either the application or potential for experimental methods.

Where an RCT has been used in an evaluation, the evidence has been considered very high quality, achieving a level 5 assessment. A study looking at the impact of consulting services on SMEs in Mexico used an RCT. The randomised control trial with 432 small and medium enterprises shows positive impacts on total factor productivity and return on assets following access to a year of management consulting services. Owners also had an increase in "entrepreneurial spirit" (an index that measures entrepreneurial confidence and goal setting). These results used surveys and administrative data with Mexican social security data used and finding a persistent large increase (about 50 percent) in the number of employees and total wage bill even five years after the programme.

Nesta instigated a series of RCTs with their 2015 RCT evaluation of the Creative Credits programme in Manchester. Using mixed methods, they measured immediate output additionality, behavioural additionality and network additionality after the programme, and 12 months later. For output additionality they looked at sales growth, and new process or product innovations. This did not look at long-term impacts in terms of productivity. Overall, this was an example of capturing why and how programmes lead to impacts for a business. In Northern Ireland an evaluation of a leadership training programme for managers and CEOs looked both at attitudinal changes, changes in processes or management, and then longer-term impact changes in turnover per employee.

Improving the SMS level to 5

Experimental approaches are level 5 in quality. However, there are conditions that have had to be met for this by programmes. In particular, the programmes must have been:

- designed to include randomisation in the selection process;
- successfully recruited enough participants for the control to be large enough for statistical tests to be robust;
- delivered the support as planned to the selected participants; and,
- tracked the businesses randomised out of support and those receiving support using surveys or in administrative data.

The second condition has been quantified. In small samples of less than 300 randomisation may be ineffective at ensuring the comparability of control and treatment groups (Bratberg, Grasdal, and Risa 2002), especially if the intervention's impact is likely to be modest. For smaller groups, experiments can still be run if the intervention is sufficiently powerful (Bloom et al. 2011).

If significant numbers of businesses that were randomly allocated into support subsequently did not make use of it, then impacts may be underestimated.

Timings for this approach

The core focus for an RCT at programme level is to integrate randomisation at an early stage and make the decision to do this based on a firm expectation that the quality conditions can be met. Data collection might proceed as follows:

- **Year 0:** develop a marketing and selection process for an RCT (explain the RCT to applicants and ensure recruitment is sufficient to accommodate attrition and non-compliance). Collect beneficiary details and the support provided.
- **Year 1:** consider options for interim data collection, or link to administrative data.
- **Year 3:** conduct surveys of beneficiaries and non-beneficiaries; link to administrative data for analysis.

Approach 4: Randomised Tests

The spread of behavioural insights has helped build the capacity of a wide range of public sector bodies to think about their work through a more behavioural, or 'human-centric', way (BIT, 2017). This has occurred alongside a growing evidence base for what works in the application of behavioural science to practice; the growing capability of public sector bodies to apply behavioural approaches to simpler policy problems; and the growing confidence of policymakers to view policy through a behavioural lens.

The focus has been behavioural approaches affecting one-off decisions, such as whether to pay a tax bill on time. These have often been in relation to interactions with policy delivery either through messaging with clients, especially in online tools associated with policy delivery. Randomisation is used to analyse the impact on behaviours of different messages or prompts by allocating different options randomly to users and then tracking further use by the client as the next actions are taken.

For example, BIT (2013) looked at options to encourage organ donation, especially through prompting sign-up after a citizen had either renewed their vehicle tax or registered for a driving licence online. It tested this approach on these websites through different messages and pictures to work out which increased organ donation registration rates the

most. The trial ran for five weeks, during which time over one million people saw one of the eight variants (over 135,000 for each).

Roper (2017) describes how behavioural insights were used to improve completion of a business support intervention as part of a wider effort to promote business mentoring. BEIS (then BIS) launched the 'Get Mentoring' campaign, which had the aim of training 15,000 new volunteer mentors by December 2012.

Despite lots of people registering to become mentors, relatively few were completing the necessary training. To rectify this, BEIS decided to test whether behavioural insights could be applied to email reminders sent out to volunteers to 'nudge' them into completing their training (Phipps, 2017). For the first time in this policy area, RCTs were used to learn what worked best. Content which was emotionally engaging was found to have a significant impact with important gender differences also identified in individuals' responses. The process of testing and improving proved to be instrumental in the programme meeting its goal. By the end of December 2012, 15,305 people had completed their Get Mentoring training. The trials directly generated an additional 778 mentors and perhaps more than double that, once the wider application of lessons learnt are included.

This research found that asking businesses to reflect on their growth ambitions, and then telling them that mentors help to spot growth opportunities, was the most effective approach - increasing both their attitude toward mentoring and motivation to get a mentor. A series of trials tested simple tweaks to improve engagement with an email newsletter on available support. For example, opening rates increased by 3.9 per cent just by inviting businesses to reflect on their growth ambitions in the subject line ("Realise your hopes and aspirations"), but highlighting that support was "free" completely undermined this. Possibly due to a lack of perceived value from free support or perhaps simply as the term triggered spam filters (human or automated).

Approach 5: Qualitative approaches

In evaluating the educational outcomes for a school refurbishment project, BEIS (2019) recognises that there is clearly an argument for the inclusion of some case studies to explore a range of qualitative aspects of the impacts of a sample of projects. Equally, the Magenta Book details some of the specific approaches used in qualitatively assessing outcomes and impacts, often complementing the evidence available in quantitative methods.

HMT (2011) highlights two sets of approaches:

- **Theory-based evaluation** Theory-based evaluation approaches involve understanding, systematically testing and refining the assumed connection (i.e. the theory) between an intervention and the anticipated impacts. These connections can be explored using qualitative findings to verify any empirical impact evaluation.
- **Meta-evaluations** can use quantitative or qualitative techniques to bring together several related evaluations to derive an overview or summary conclusion from their results.

This appendix is not seeking to describe the many approaches used. However, in assessing the BtB programmes, the use of case studies and the filling of evidence gaps using meta-evaluations can be important.

A first area is the question of whether behavioural effects of BtB programmes are sustained and whether they reach beyond the specific target behaviour to deeper, long-term consequences. In many interventions, the full impacts are likely to be long-term and the tracking of beneficiaries is not practical for a long enough period. Meta-evaluations can

often show – to a relatively robust level – that the changed behaviour observed early after an intervention correlates or predicts the long-term impacts. These predictions will take account of the likely loss of impact, such as an intention after the policy to take an action may not be followed through. In management practices, the implementation of HR systems is correlated with later productivity improvements (Guest and others, 2012), so that behaviour change may be assessed directly, and then long-term impacts estimated.

Related to this qualitative approach are the theory-based evaluation methods. They provide a systematic and cumulative study of the links between activities, outcomes, and context of a policy intervention. It involves the specification of an explicit theory of "how" and "why" a policy might cause an effect which is used to guide the evaluation. It does this by investigating the causal relationships between context-input-output-outcomes-impact to understand the combination of factors that has led to the intended or unintended outcomes and impacts. Theory of Change analysis therefore normally develops and tests the implementation theory of the policy and allows this to be modified or refined through the evaluation process.

Raising the robustness of qualitative evidence

The Magenta Book highlights how a systematic review differs from literature reviews, in having a focused review of the evaluation evidence that is available in past studies. The review is guided by a clearly stated set of objectives with pre-defined eligibility criteria for studies that would be used in the meta-evaluation. Also, there is an explicit, reproducible methodology, including structure around the searches to identify all studies that meet the eligibility criteria (else there may be a bias to finding only confirmatory evidence). There is a formal assessment of the validity of the findings of the included studies.

The qualitative approaches can validate the assumptions behind a theory of change. Approaches taken may be case studies workshops, in-depth interviews etc, where multiple participants, taken from different groups of stakeholders, are used to confirm the assumptions. The robustness is enhanced when multiple views of the same link in the theory can be contrasted. This can often be informed by the quantitative evidence. For example, after the analysis of evidence after an intervention, the evaluator might focus on those participants who viewed the support as successful or unsuccessful. Then, this may be complemented by delivery partners or others involved in the programme.

Timings for this approach

These approaches are usually complementary to the other approaches. Timings therefore link to the design of evaluation approaches and the development of the theory of change for a programme. This approach might proceed as follows:

- **Year 0:** identify the (mainly behavioural) aspects of the programmes where immediate effects will be assessed relatively robustly but the long-term impacts are harder to quantify; design behavioural surveys to focus on early outcomes that correlate with later impacts.
- **Year 1:** complete meta-evaluation and confirm that evidence gathering – both qualitative and quantitative – can validate (i.e. confirm the assumptions of) the theory of change.

Appendix C: Baseline Questions

Baseline Requirements			
Data required	**Importance**	**Description**	**Reason**
Unique BTB Number	High	A systematically generated unique ID anonymised	To track the business anonymously throughout the evaluation stages
Company Name	High	Name in Companies House	To identify company from administrative datasets
Company Number CRN/ VAT number	High	Company number in Companies House or VAT number on HMRC	To identify company from administrative datasets
Contact details (telephone, email)	High	Name of contact, email, telephone, address	For tracking for survey responses
Size – Turnover – if using the Benchmarking tool or accessing high intensity activity	Medium	Total operating revenue (£)	To help build representative sample of business types and for measuring impacts
Size - Employee	Medium	Full Time Equivalent	To help build representative sample of business types and for measuring impacts
If they have received any other support - government	Medium	List of other programmes or funding received before, during and after the intervention	A way to check for confounding factors that might be the reason for an observed change rather than the intervention
Industry	Low	SIC code	To help build representative sample of business types
Location – Postcode	Low	Postcode	Background data to monitor spillovers, confounders, etc.
Location – Region (Nuts1)	Low	Region name/number	Background data to monitor spillovers, confounders, etc.
Cost of Sales	Low	Cost of sales (£)	To help build representative sample of

			business types and for measuring impacts
Behavioural questions	Low	Aspirations, attitudes to risk, attitudes to management, ambition, locus of control.	Building a picture about the types of business leaders that benefit most from support and whether different behavioural traits change the impact of an intervention
Qualifying criteria	Optional	If appropriate	The qualifying criteria for programmes which determine who receives support