Department for Education

Understanding the impact of Taking Teaching Further: A feasibility study

Research report

November 2022

Author: Stephen Morris, Manchester Metropolitan University

GSR
GOVERNMENT SOCIAL RESEARCH
Social Science in Government

Contents

List of figures

List of tables

Executive Summary

Introduction

The Further Education (FE) sector has experienced considerable difficulty in recruiting and retaining suitably qualified teaching staff, particularly those with prior commercial or industrial experience. The Taking Teaching Further (TTF) programme aims to address this problem as part of a suite of policy measures. Overseen by the Education Training Foundation (ETF) on behalf of the Department for Education (DfE), the programme has two Strands:

- **Strand 1** provides financial support for up to 150 industry experts to become FE teachers, covering the course costs of teacher training as well as support and mentoring.
- **Strand 2** supports up to 40 projects that help develop local partnerships and collaborations between FE and industry.

The focus of this report is the feasibility of conducting a rigorous and credible impact evaluation of future rounds of Strand 1 of TTF.

Launched in 2018, the TTF programme has been piloted across two Rounds to date. Round 1 took place in the 2018/19 academic year and Round 2 took place in the 2019/20 academic year. Each Round of TTF comprised of the two Strands described above. Participating providers could apply to participate in one or both Strands in an individual Round. They could also apply to participate in both Rounds of the TTF pilot.

Both Strands in each Round have been subject to a process evaluation. The process evaluation, together with this feasibility study, forms a programme of activities designed to understand how TTF has operated in practice from the perspectives of both providers and newly recruited teachers (IFF Research, 2019).

Approach to the impact evaluation

This report commences through examining the UK government's guidance on evaluation - the Magenta Book.[1] It covers what is meant by 'impact evaluation' and sets out a provisional high-level discussion of relevant approaches consistent with the Magenta Book. The report concludes that any future proposed evaluation should focus on the average causal effect of TTF on the providers, teachers, and (by extension) learners that participate in the programme. In this situation, the average causal effect of TTF is the average outcome for those participating in TTF (individuals and providers), minus the average outcome had those that participated not been exposed to TTF.

The intervention being evaluated

In impact evaluation, the evaluand refers to the subject of the evaluation. This is usually a project, programme, or intervention. The evaluand discussed in this report is Strand 1 of the TTF programme. The nature of the evaluand has important implications for the design of an evaluation. As such, this report discusses the important features of Strand 1 of TTF and describes patterns of participation of Rounds 1 and 2. These are used as the best guide as to what a future version of Strand 1 of TTF might look like, which helps to formulate views about any future proposed evaluation.

Target population, level of analysis and outcome indicators

The prior implementation of TTF is also examined from the perspective of identifying the fundamental unit of analysis that might be chosen for any future impact evaluation. In particular, the possibilities of estimating the causal effects of TTF at the levels of the provider, teacher and learner are examined. This discussion is supplemented by further considerations around what these various levels might mean for the definition of outcomes and indicators. For example, if the evaluation proceeded based on the provider or college as the fundamental unit of analysis then outcomes and indicators would primarily be defined at the level of the provider or college.

[1] HM Treasury (2020), The Magenta Book

Sources of data

A counterfactual approach to impact evaluation (please refer to the Glossary for definitions of counterfactual and impact evaluation), consistent with the requirements of the Magenta Book, is reliant on the availability and quality of data that:

- outcome indicators can be derived from;
- captures the target populations eligible to participate;
- enables the analyst to distinguish between that portion of the target population that participates or is exposed to TTF and that which is not exposed; and
- enables 'control variables' to be identified that permit statistical adjustments in analysis, particularly in relation to Level 3 approaches to evaluation.

This report reviews a range of existing data sources and assesses their usefulness from this perspective.

Proposed approaches

This report uses the Maryland Scale of Scientific Methods[2] to examine the prospects for implementing either a Level 5 impact evaluation (i.e. a randomised controlled trial (RCT) design) or a Level 3 design based on a non-random comparison group design. It looks at what it might take to implement such designs practically and how effective TTF would have to be for the likely samples available to be large enough to detect any effect (should one exist).

The report demonstrates that impact evaluation designs at Levels 3 and / or 5 of the Maryland Scale are practically achievable. This is the case if appropriate primary data is collected. These include individual unit level data, data that record individual unit level outcomes for both exposed and unexposed cases and data that indicate which cases are exposed and which cases, at the point in time outcomes were measured, remain unexposed. It also assumes that enough time is available before future Rounds commence for the necessary data collection processes to be put in place. However, the future effects of TTF Strand 1 are likely to be modest and therefore difficult to identify statistically, particularly given the likely size of the samples available to the evaluation. An RCT (i.e. Level 5 evaluation) is also practically viable and could be designed to consider outcomes by provider and/or by declared vacancy. However, results are likely to be inconclusive in statistical terms due to a) the likely modest scale of any impact; and b)

[2] https://whatworksgrowth.org/resources/the-scientific-maryland-scale/

samples that are relatively small (this assumption is based on the number of colleges recruited during Rounds 1 and 2 of the TTF programme).

This does not mean results will be inconclusive, just that the risk that they will be so is higher than would be generally acceptable at the commencement of most evaluations. Such results could not be interpreted as revealing that TTF did not work, but instead that the data were not consistent with a strength of effect that might have reached statistical significance. A Level 3 design is discussed as an alternative. Broadly, the challenges that face a Level 3 design are similar and relate to the limited samples available leading to a high chance of inconclusive findings if TTF produces modest impacts.

Conclusions and recommendations

This report recommends proceeding with a counterfactual impact evaluation only if substantial numbers of providers other than General FE colleges could be attracted to the programme. This is because the total population of General FE colleges is smaller than the total number of providers that would need to take part in TTF, to minimise the risk of the counterfactual impact evaluation producing inconclusive findings. Approximately 350 providers would need to take part in total in TTF, over two future Rounds, before the risk of inconclusive findings would stand at levels generally accepted at the planning stage of most evaluations.

If participation of this order of magnitude is not felt feasible, then it is recommended that other forms of evaluation that attempt to shed light on impact be considered. These approaches are non-statistical and aim to provide the best and most plausible explanation for the effects of interventions derived from evidence that comes from mixed-methods evaluation designs. Such methods include realist evaluation (Pawson & Tilley, 1997), theories of change (Funnell & Rogers, 2011), contribution analysis (Mayne, 2012) and possibly sophisticated case study approaches such as qualitative comparative analysis (Schneider & Wagemann, 2013).

Glossary

Average causal effect	The average of the effect of the intervention (in this case, TTF) across all units (colleges or teachers for example) in the population.
Average causal effect on the treated	The average of the effect of the intervention (TTF) across all treated units (colleges or teachers for example). In this report we refer to treated units as colleges, teachers or learners taking part in the TTF.
Bias	The extent to which sample estimates, given a particular estimator, differ systematically across repeated samples from the true or expected value of the quantity being estimated.
Causal attribution	The process of attributing an effect to a particular cause. In the case of an evaluation, this is usually the process of assessing whether an observed effect can be attributed to an intervention (or evaluand – see below). So, in this case, causal attribution refers to the extent to which observed effects identified via the evaluation are caused by the TTF.
Causal pathway	A series of distinct and related intermediate processes that bring about the causal effect of the intervention and which lie on a pathway between the intervention itself (i.e. the TTF) and outcomes of interest (i.e. positive impacts on colleges, teachers or learners).
Comparison group	A group of units from which an evaluation can calculate estimates of the average outcomes that would have happened, for those treated or exposed to an intervention, in the absence of that intervention. For example, a comparison group can be used to measure the same outcomes among a group of providers that did not take part in the TTF, to compare with those that did, in order to assess its impact. A comparison group, in contrast to a control group,

	is not formed through randomisation (see below).
Control group	A group of units formed at random that remain unexposed to the intervention (the evaluand) and from which estimates of average outcomes that would have occurred for the intervention group (see below) had they remained unexposed are obtained. So, for example, a randomly selected group of providers who did not take part in the TTF, used to measure the same outcomes as those who did, to assess its impact.
Control variables	Used in this report to refer to pre-intervention variables that are statistically associated with the decision of teachers or colleges to participate in the intervention (the TTF) and/or that are correlated with outcomes. For example, if large colleges were systematically more likely to take part in the TTF than small ones, size of college would need to be controlled for in the analysis.
Counterfactual	The counterfactual is a measure of “what would have happened” to the units (e.g. providers / teachers) which took part in TTF, had they not participated or been exposed to it (and all else being equal).
Difference-in-differences	A process whereby the difference in outcomes before and after an intervention for a comparison group (see above) is subtracted from the before and after difference in outcomes for an intervention group (see below) in order to obtain an estimate of the average causal effect (see above).
Effect size	Generally an effect size quantifies the size and direction of a difference between two groups or the strength of an association between two variables (Durlak, 2009). The effect size can be defined in a number of different ways. In this report the standardised difference in means (the

	difference in mean outcomes between the intervention and control or comparison group, divided by the pool standard deviation) is the effect size referred to.
Estimand	The target of estimation or the quantity to be estimated, as distinct from the estimator that is the statistical procedure used to obtain an estimate.
Evaluand	The subject of an evaluation, usually a project, programme or intervention. In this report the evaluand is the Taking Teaching Further programme.
Impact evaluation	An impact evaluation seeks to draw causal inferences regarding the effects of an intervention or evaluand on an outcome or outcomes of interest.
Instrumental variable	A variable used to identify the causal effect of a programme or intervention (an evaluand) (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016). In simplest terms, an instrumental variable must be correlated with programme or intervention participation and must only affect the outcomes of interest through the intervention or programme. This latter condition is known as the ‘exclusion restriction’.
Intervention group	A group of units (in this case, colleges or teachers) formed at random, in the context of a randomised controlled trial (see below), that are subsequently exposed to the intervention or programme being evaluated (the TTF). In the context of a non-randomised study the formation of an intervention group is the result of choice or selection rather than an outcome of randomisation.
Matching	The creation of a comparison group through assembling units that are matched to

	intervention group units on the basis of observable or measured characteristics and behaviours. For example, providers in the intervention and comparison groups could be matched on characteristics such as number of employees, region, and/ or the proportion of vacant teaching posts.
Process evaluation	“An evaluation that tries to establish the level of quality or success of the processes of a program (sic)” (Gertler et al., 2016, page 233). A process evaluation assesses how well a programme is being or has been implemented, to inform efficiency and quality improvement. The findings of a process evaluation can shed light on why a programme has or has not delivered its desired impacts.
Propensity score	An estimated probability of programme or intervention participation or exposure based on pre-intervention observable characteristics or behaviours. For example, propensity to take part in TTF may be influenced by a range of factors such as type of provider, size of provider, and the extent of recruitment difficulties.
Qualitative comparative analysis (QCA)	An approach to causal attribution based on between-case comparisons, drawing on set theory and Boolean logic. QCA starts with mapping out all the different configurations of conditions associated with each case of an observed outcome. These are then subject to a minimisation procedure that identifies the simplest set of conditions that can account for all the observed outcomes, as well as their absence. The results are typically expressed in statements or as Boolean algebra. For example: a combination of Condition A and condition B, or a combination of condition C and condition D, will lead to outcome E.
Randomisation	The process of dividing a sample of either teachers or colleges into two or more groups on

	a chance basis (complete randomisation), or some form of constrained chance basis.
Randomised controlled trial (RCT)	A research design in which units from a population are assigned at random to two or more groups. Subsequently the groups are exposed to different levels of a programme or intervention. Comparison of outcomes across the groups permits, in many situations, valid estimates of uncertainty for detecting causal effects.
Realist evaluation	An approach to evaluation that draws on the philosophy of critical or scientific realism and which focuses on the context of an intervention, the causal mechanisms which the intervention brings into play, and how these contexts and mechanisms interact to produce outcomes. Realist evaluations draw heavily on mixed methods with an emphasis on qualitative research in many cases.
Regression discontinuity	Designs that exploit situations where the target population are exposed to a programme or intervention on the basis of a continuous score and where units fall in relation to a threshold or cut point on that score. For example, students might qualify for additional support if their test score falls below a certain threshold. In such cases intervention and comparison groups are defined in relation to the threshold. So, to extend the example, students whose score falls just above the threshold form a comparison group, whilst those whose score falls just below form the intervention group and receive additional support.
Statistical significance	An observed effect is considered to be statistically significant if the probability of an effect at least as large as that observed, under the null hypothesis (the hypothesis that the effect of the intervention is in fact zero), falls below a certain level, given a specified statistical

	model. Typically, arbitrary levels below which such a probability or p-value must fall are ten, five or one per cent before the null hypothesis is rejected.
Type I statistical error	Put crudely, the probability of rejecting a null hypothesis (see statistical significance) that is in fact true.
Type II statistical error	Put crudely, the probability of failing to reject the null hypothesis (see statistical significance) when it is not true.
Theory of change	A representation of the underlying logic, rationale or theory for the design of a programme or intervention that typically sets out a programme's operating assumptions, its resource requirements or inputs, activities that the intervention comprises, the outputs that are produced as well as the programme outcomes and longer term impacts. Quite often theories of change are presented diagrammatically in the form of a logic model.

Introduction

The Further Education (FE) sector has experienced considerable difficulty in recruiting and retaining suitably qualified teaching staff, particularly those with prior commercial or industrial experience. The Taking Teaching Further (TTF) programme aims to address this problem as part of a suite of policy measures. Overseen by the Education Training Foundation (ETF) on behalf of the Department for Education (DfE), the programme has two Strands. The focus of this report is the feasibility of conducting a rigorous and credible impact evaluation of future rounds of Strand 1 of TTF.

Transforming the FE sector is at the heart of the government plans to raise productivity and increase economic growth. The Productivity Plan (HM Treasury, 2015), the Post-16 Skills Plan (Department for Business Innovation and Skills & Department for Education, 2016) and the Industrial Strategy (HM Government, 2017) highlight the importance of improving investment in technical skills to strengthen the nation's industrial base and performance.

In line with this vision, the FE sector is facing major reforms, including:

- structural and system-led changes following the area review programme, which has included several rationalisations through college mergers and an overall decrease in full time equivalent (FTE) teaching staff;
- preparing for the introduction of T Levels (in 2020), including the T Level Professional Development Fund and the need to establish more links with employers to deliver industry placements;
- responding to the potential impact of Brexit on industry skills needs and the FE workforce; and
- accommodating a growth in the number of students resulting from underlying demographic trends.

These changes have considerable implications for the FE workforce, which has faced long-standing supply difficulties (Greatbatch & Tate, 2018). The TTF programme has the following long-term aims:

- to raise the profile and prestige of FE teaching, particularly among industry professionals;
- to increase the overall number of skilled FE teachers in the T level technical routes that will be taught first (Childcare and Education, Digital, Construction, Engineering and Manufacturing and other Science Technology Engineering and Manufacturing (STEM) technical routes);
- to increase the opportunity for industry-related Continuing Professional Development (CPD) for current teachers;

- to demonstrate the value of, and possibilities for, industry / FE collaboration; and
- to stimulate and support local programmes to build capacity in FE teaching and improve industry collaboration.

The TTF programme is divided into two Strands:

- **Strand 1** provides financial support for up to 150 industry experts to become FE teachers, covering the course costs of teacher training as well as support and mentoring.
- **Strand 2** supports up to 40 projects that help develop local partnerships and collaborations between FE and industry.

Although TTF is being evaluated as one programme, each Strand has its own specific aims and objectives.

To date, Strands 1 and 2 of TTF have run as pilots across two Rounds. Both Strands have been subject to a process evaluation[3]. The process evaluation, together with this feasibility study, forms a programme of activities designed to understand how TTF has operated in practice from the perspectives of both providers and newly recruited teachers (IFF Research, 2019). This report looks forward to potential future Rounds of funding under the TTF programme that essentially replicate or looks very similar to Strand 1 TTF and asks to what extent will it be possible to evaluate the impact of future Rounds.

The findings of this feasibility study are based on desk-based research into programme documentation and web-content on the nature of TTF in its Strand 1 form, and the extent and nature of take-up of Strand 1 to date among providers and trainee teachers. The findings also take into account the UK government's guidance on evaluation known as the Magenta Book (HM Treasury, 2011), and the Maryland Scale of Scientific Methods (Sherman et al., 1997) which is a way of classifying different approaches to impact evaluation. Based on this literature and previous discussions with the department, this feasibility study considers the potential for impact evaluation designs at Levels 3 and 5 of the Maryland Scale, consistent with what the Magenta Book describes as 'empirical impact evaluation'. This report uses the term 'counterfactual impact evaluation' to distinguish these from other forms of impact evaluation such as programme theory approaches, case studies or realist methods.

This report is organised as follows. Section 1 elaborates what the Magenta Book and other literature means by 'counterfactual impact evaluation' and sets out a provisional

[3] DfE commissioned IFF Research (IFF Research, 2019) to conduct a process evaluation of both Strands of TTF to understand how they have operated and what providers, teachers, employers and learners have gained from participation. As well as providing an understanding specifically of TTF that could lead to improvements for later rounds, the evaluation sought to identify good practice and scalable policies that could be rolled out as part of a wider programme, generating lessons for DfE to share with the FE sector.

high-level discussion of relevant counterfactual approaches. The average causal effect of TTF on providers/teachers or learners that take part, or in the language of counterfactuals, are 'treated', is to be the estimand[4] that is the focus of this evaluation[5]. In this situation, the average causal effect of TTF would be the average outcome for those participating in TTF, minus the same average if those that participated had not been exposed to TTF[6]. Having discussed the estimand of interest - in this case, the average effect of Strand 1 TTF on outcomes chosen to be the focus of the intervention

The existence of an estimand in the case of an evaluation assumes the existence of an evaluand, which means the object or target of the evaluation to which the estimand is related. In the case of this study, the evaluand is Strand 1 TTF or some future, closely related version of it. The nature of the evaluand has important implications for the design of an evaluation, discussed in Section 2. To understand what a future Strand 1 TTF might look like, Section 3 of this report sets out the important features of Strand 1 TTF at Rounds 1 and 2 and describes their current patterns of participation.

Section 4 elaborates further on the nature of TTF from the perspective of an evaluation and addresses the target population eligible to participate in it, relating this to further discussion of the estimand. It also considers the level or fundamental unit of analysis that might be chosen. This is done by discussing the possibilities of estimating the causal effects of TTF at the levels of the provider, teacher and learner. This discussion is supplemented by further considerations around what these various levels might mean for the definition of outcomes and indicators.

A counterfactual approach to impact evaluation is reliant on the availability and quality of data:

- from which outcome indicators can be derived;
- that captures the target populations eligible to participate;
- that enables the analyst to distinguish between that portion of the target population that participates or is exposed to TTF and that which is not exposed; and
- which enables 'control variables' to be identified that permit statistical adjustments in analysis, particularly in relation to Level 3 approaches to evaluation.

[4] The estimand is a term used when referring to the statistical quantity or parameter that we are seeking to estimate on the basis of the sample data. We can distinguish between the estimand, the target of our analysis, the estimator, the statistical model used to obtain the estimate, and the estimate itself.
[5] We assume that the relevant estimand is the average causal effect of treatment on the treated, rather than the average causal effect, the average causal effect of intention to treat or the local average causal effect. These are parameters that the evaluation might have as its target but which address causal questions that have a different substantive emphasis.
[6] Technically this the average effect of treatment on the treated.

Section 5 reviews a range of existing data sources and assesses their usefulness from this perspective.

Section 6 examines the prospects for implementing either a Level 5 impact evaluation, namely a randomised controlled trial design, or a Level 3 design based on a non-random comparison group. It looks at what it might take to implement such designs practically. This section assesses how effective TTF would have to be in order for the likely samples available to be large enough to detect any effect. Section 7 provides a summary and suggested way forward.

Section 1: A counterfactual approach to programme evaluation and the objectives of this study

The purpose of this report is to set out possible approaches to evaluating the impact of future rounds of TTF Strand 1. It is important at the outset of any feasibility work to define what is meant by impact evaluation so that what we are seeking to achieve is conceptually clear.

An impact evaluation addresses causal attribution. Impact evaluation seeks to determine whether we can infer the presence or otherwise of a causal link between the evaluand and outcomes of interest. Beyond this, there is considerable disagreement among social scientists concerning causal attribution and evaluation. This debate has spawned a range of different approaches to the evaluation of impact. The approach to impact evaluation adopted for this feasibility study is what is referred to as 'empirical impact evaluation' in the UK government's Magenta Book (HM Treasury, 2011).

The details of what is meant by 'empirical impact evaluation' are important because they frame what is discussed in the rest of this report. In essence 'empirical impact evaluation' is statistical. It involves not just demonstrating the existence or otherwise of a causal link between the evaluand and an outcome, but a quantitative estimate of its magnitude and a measure of uncertainty about this estimate. As a shorthand throughout this report, the approach to impact evaluation adopted here is referred to as 'counterfactual' impact evaluation. Counterfactual approaches to impact evaluation are widely considered to provide important and, where executed according to 'best methods', reliable evidence of effectiveness that is useful to policymakers (Gertler et al., 2016).

Applying a counterfactual based approach to the evaluation of Strand 1 of TTF assumes that the causal effect being estimated is the average effect of Strand 1 TTF on those that participate or are exposed to it[7] (known as the average effect of treatment on those treated)[8]. What this means is that, for a given outcome of interest, there is an estimate of the average for that outcome for those providers/teachers/learners that take part or are exposed to Strand 1 TTF, as well as a counterfactual outcome for this group – that is the average outcome that would have prevailed for them had they not participated or not been exposed (all else being equal). So, in the case of TTF, if the outcome of interest was, for example, unfilled vacancies held by providers, the evaluation would need to produce an average estimate of the number of unfilled vacancies for those providers that participated in TTF. In addition, the evaluation would need an estimate of the average unfilled vacancies that would have been observed, all else equal, for this group of

[7] This is in contrast to wishing to estimate the average effect of Strand 1 of TTF for the entire population of providers/teachers or learners (the average treatment effect).
[8] The focus on average treatment effects implies that effects for individual participants will vary about the average.

providers had they instead not participated in TTF. This latter quantity is the counterfactual. The estimated average causal effect for those that participate or are exposed is the difference between these two averages. Obtaining an estimate of the average outcome for those exposed, that would have prevailed had they not been exposed, is challenging and requires a control or comparison group to understand more about what would have happened had they not taken part in the programme. The control or comparison group should be as similar as possible, ideally identical, to the group exposed. For example, they should have similar average characteristics: gender, age, prior qualifications and employment histories, and be exposed to similar contextual conditions. Under the counterfactual approach, there are different approaches to obtaining both estimates of average outcomes and identifying and selecting control groups. These approaches range in the extent to which various assumptions are required for any results to be interpreted as unbiased. Approaches that require fewer and more plausible assumptions are preferred over those that require more and less plausible assumptions[9]. The precise meaning of this and the necessary assumptions will be made clear later in this report. For now, it should be noted that, of the likely available evaluation options, some require more assumptions than others for results to have a reliable causal interpretation.

One way of classifying the relative desirability of different approaches is known as the Maryland Scale of Scientific Methods (Sherman et al., 1997). The table below presents a version of the Maryland Scale developed by the UK What Works Centre for Local Economic Growth.

[9] For an example of the type of assumptions referred to here consider the following example. In the case of TTF, a strategy of comparing unfilled vacancies among providers exposed to TTF before and after TTF might be chosen as the approach to estimating the average effect of the intervention. Here the pre-TTF sample estimate of unfilled vacancies represents the average outcome in the absence of exposure (crudely the counterfactual) whilst the average turnover rate post-TTF the average outcome in the exposure condition. The assumption required for this estimate to be reliable is that in the absence of TTF the average number of unfilled vacancies would have remained unchanged.

Table 1: Adapted Maryland Scale of Scientific Methods

Maryland Scientific Methods scale	Description
Level 5 (Generally considered most reliable)	Reserved for research designs that involve explicit randomisation into treatment and control groups, with randomised control Trials (RCTs) providing the definitive example.
Level 4	Quasi-randomness in treatment is exploited, so that it can be credibly held that treatment and control groups differ only in their exposure to the random allocation of treatment.
Level 3	Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. difference-in- difference). Techniques such as regression and propensity score matching may be used to adjust for difference between treated and untreated groups, but there are likely to be important unobserved differences remaining.
Level 2	Use of adequate control variables and either: a) a cross-sectional comparison of treated groups with untreated groups; or b) a before-and-after comparison of treated group, without an untreated comparison group.
Level 1 (Generally considered least reliable)	Either: a) a cross-sectional comparison of treated groups with untreated groups; or b) a before-and-after comparison of treated group, without an untreated comparison group. No use of control variables to adjust for differences between treated and untreated groups or periods.

Source: What Works Centre for Local Economic Growth
https://whatworksgrowth.org/resources/the-scientific-maryland-scale/

Here different approaches are arranged along a scale from Level 1 to Level 5. Those methods at Level 5 are preferred to those at Levels 4, 3, 2 and 1. Those methods at Level 1 seldom provide evidence of causal effects that are plausible, whereas those at Level 5 are more likely to yield results that have a reliable causal interpretation. As can be seen, it is Randomised Controlled Trials (RCTs) that are broadly preferred. This does not mean that RCTs do not require evaluators to make assumptions before interpreting their results, or that they are always preferred. There will also usually be some challenges in generalising results from an RCT.

Following preliminary discussions held with the department, this feasibility study focuses on the prospects for a counterfactual impact evaluation of Strand 1 TTF at Level 3 or above. It appears that there is little prospect for an evaluation at Level 4. This would require either an instrumental variables[10] or regression discontinuity[11] design. An initial assessment suggested neither of these were realistic. It was felt exposure to TTF would not conform to the requirements of regression discontinuity. This is because it would require access to the scheme being determined by a strict quantifying eligibility criterion / variable and a distinct threshold. A regression discontinuity design is relevant where access to an intervention is determined by a score on some continuous variable or some other similar metric. For example, in this case circumstances might be imagined where only providers whose rate of unfilled vacancies in a given year exceeds a pre-determined threshold might be able to participate in TTF. The evaluation exploits the fact that providers just either side of such a threshold are similar to one another. Thus, those just above the threshold form an intervention group those just below a control group and their subsequent outcomes are compared. In the case of instrumental variables, an initial assessment suggested that the types of circumstances and data required for this approach were unlikely to be forthcoming. This led to consideration of the prospects for an RCT (Level 5) or a comparison group design (Level 3).

For an RCT, estimates of average outcomes under participation and non-participation are obtained from intervention and control groups created at random, where those units (provider/teacher or learner) allocated to the intervention group are exposed to TTF Strand 1. In the case of TTF, an RCT might take the form of allocating providers to the intervention or control group with only those allocated to the intervention group able to take up TTF. At some subsequent point, outcomes in the two groups are measured and compared and this comparison provides an estimate of the effectiveness of TTF. To achieve such a design, a wide range of factors need to be considered. These are discussed further below but include:

- the stage in the programme in-take process where randomisation should occur;
- how baseline measures should be collected, and;

[10] An instrumental variable is technically a variable that causes some providers/teachers or learners to be exposed to TTF on a random basis. Further, the instrument can only affect any outcome (for example the provider vacancy rate) through TTF, not through any alternative route. This randomness is not as a result of the actions of the researcher or policy maker but is in a crude sense accidental or “naturally” occurring. Instrumental variables are a highly technical approach to evaluation and require an advanced understanding of causal analysis and statistics. Therefore, a detailed discussion is beyond the scope of this report. The interested reader is referred to Chapter 13 of Paul Rosenbaum's 2017 book "Observation and Experiment: An Introduction to Causal Inference " (Rosenbaum, 2017) for a relatively non-technical introduction.

[11] A full discussion of regression discontinuity as an approach to evaluation is beyond the scope of this report. For a relatively non-technical discussion the reader is referred to chapter 3 of Thad Dunning’s book “Natural experiments in the social sciences” (Dunning, 2012).

- whether there is likely to be a sample of eligible providers/teachers/learners of sufficient size to obtain an acceptable level of statistical power to be able to detect impact.

The benefits of an RCT design are that, if implemented correctly, results can be interpreted as a good estimate of the effectiveness of TTF Strand 1 and thus indicate whether TTF has been a success (Gerber, Alan & Green, Donald, 2012; Glennerster & Takavarasha, 2013). Moreover, calculation of a margin of error under such designs is relatively straightforward compared to alternatives (Bloom, 2006).

RCTs generally tend to generate a data set explicitly through the process of implementing the study design. In comparison, study designs consistent with Level 3 tend, generally, to place a much heavier reliance on existing data sources as well as in some cases primary data sources[12]. Level 3 designs require rich data. They include a comparison group, from which counterfactual outcome estimates are obtained. Comparison groups are drawn from a part of the target population for the intervention that did not participate in it. Outcome data are required for these units too. So, in the case of TTF, the relevant outcomes for providers that participated in TTF as well as for a sample of, or all, providers that did not participate in it would need to be collected. The non-participating providers (or a subset of them) would form the comparison group. For example, if unfilled vacancies were the outcome of interest (and provider the unit of analysis) any evaluation would need a measure of this both for the providers that participated and for all or a sample of those that did not, as they would form a comparison group. But providers drawn from non-participating portions of a target population will tend, prior to exposure to TTF, to differ systematically to providers that participate. Therefore, they will potentially not provide a good estimate of average outcomes for those that participate under non-exposure. Due to these prior differences, comparisons of outcomes between exposed groups and comparison groups needs to be adjusted statistically for such analysis to have a reliable causal interpretation. This is a very demanding requirement and significant weakness of Level 3 designs. It is generally not one shared with Level 5 RCT designs.

To make the required statistical adjustments, a range of control variables are required. These are variables that are correlated with participation/exposure and outcomes, and that are used to make the statistical adjustments described above. So in the case of TTF, these will be variables that are associated with the outcome (for example unfilled vacancies) and the decision to take part (that is they also help distinguish between

[12] By primary data we mean data collected specifically for the evaluation. This is distinct from data that's existence pre-dates the evaluation and that were collected for other reasons, usually for the purpose of management, audit and/or monitoring.

providers that do and do not take up Strand 1 TTF). Values on these variables must be observed for both intervention and comparison groups (or for all potential comparators).

Implications for the evaluation

The issues discussed so far create a series of implications for any credible impact evaluation design, particularly at Level 3, where the data requirements are more demanding:

- this would require individual unit level data (provider/teacher/learner – depending on the unit of analysis);
- these data must record individual unit level outcomes for both exposed and unexposed cases (a fuller discussion of outcomes is presented below);
- these data must indicate which cases are exposed and which cases at the point in time outcomes were measured remain unexposed; furthermore the data must also record any previous or prior participation in TTF Strand 1 (that is participation in Rounds 1 and 2); and
- these data must also contain 'control' variables – these are variables that are unaffected by TTF (usually measured pre-exposure) that capture important differences between exposed and unexposed units, but are also correlated with outcomes.

The purpose of the rest of this feasibility study is to determine if it is possible:

- to produce a credible estimate of the causal effect of exposure to TTF for those exposed;
- to determine whether data sets exist or need to be created that enable us to measure outcomes in ways consistent with a Level 3 or Level 5 evaluation design;
- to determine whether there are data sets that exist or need to be created that enable us to identify cases or units in the target population that have been exposed to TTF and that remain unexposed; and
- to determine whether there are data sources that also contain a plausible set of control variables, or a source that could be created.

Section 2: The nature of the evaluand

The nature of the evaluand will determine what is possible from the perspective of evaluation design. This section describes key features of how TTF Strand 1 has operated in the past, to get a sense of how it might perform in the future. This provides important information that will be considered in designing an evaluation for future rounds of Strand 1 TTF.

Strand 1 of Taking Teaching Further

Strand 1 of TTF has run for two years as a pilot and at the time of writing is ongoing. Strand 1 has had two rounds so far: Rounds 1 and 2. This feasibility study is looking at the potential for an impact evaluation of future Rounds of Strand 1. The total budget for TTF is £5million split across Strands 1 and 2.

Strand 1 of TTF provides funding and a range of support that aims to encourage skilled and experienced staff working in industry to enter teaching in the FE sector. FE providers are encouraged to apply to the ETF for funding that covers:

- costs of Initial Teacher Education (ITE) of up to £4,000 per trainee, with each provider able to apply for up to five trainee places. Trainees undertake a Level 5 diploma in Education and Training[13];
- cost of providing intensive support to teacher trainees during the first eight weeks of their appointment; and
- costs to cover a reduced workload for the teacher trainee for the remainder of the first year.

The total value of each award to the provider is around £18,200 per trainee for the Rounds of TTF conducted to date. The programme makes funding available to providers who, along with trainee teachers, are direct beneficiaries. The TTF theory of change (see Annex A) suggests that learners are also potential beneficiaries of the programme. It anticipates learners will receive an improved quality of teaching by being taught by teachers with recent industry experience. Beneficiaries and effects are multi-layered and clustered due to this. Learners are nested within teachers, and teachers within providers. This also raises the question of the appropriate outcome measures for different beneficiaries and the time periods over which it is realistic to expect such outcomes to be observed. Moreover, the range of beneficiaries considered in any evaluation also raises

[13] Note this qualification can be completed through a range of different modes of delivery, including online and at a pace determined by the student. Typically, the qualification takes up to 2 years to complete and involves around 1,200 hours of study. Students must have amassed 100 hours of teaching experience prior to study

the question of how to identify a valid comparison or control group for them. Finally, particularly in the case of a randomised design, there is a related question as to whether declared vacancies could even constitute a unit of analysis (nested within provider), an issue that is discussed in detail later in this report (see Section 6).

Eligibility criteria for TTF Strand 1 funding are as follows[14]:

Trainees must:

- be an industry professional;
- be from one of four prioritised sectors: childcare and education, digital, construction or STEM;
- not already hold a level 5 teaching qualification or equivalent; and
- "be of suitable quality/calibre as determined by the providers' recruitment processes".

Providers must:

- be an FE provider: general or specialist Further Education college, National College, Independent Training Provider, employer-led provider, third sector training provider, Local Authority provider and/or an Adult/Community Learning provider:
    - show that the post(s) subject to recruitment are at least 0.5 FTE;
    - confirm that trainees will undertake a level 5 qualification in education and training;
    - confirm that trainees will commence study by September 2019 (Round 2 – September 2018 Round 1) and finish by July 2021 (Round 2 – July 2021 Round 1); and
    - confirm that the vacancy is hard to fill (e.g. vacant for at least three months).

These eligibility criteria are important from an evaluation perspective because they define the populations (providers and teacher trainees) at which TTF is currently targeted, and from which unexposed comparison groups might be selected, as well as contribute to an understanding of the evaluand (i.e. the TTF Strand 1 programme) and estimand (i.e. the average causal effect of the intention to treat).

[14] These criteria can be found in the document "Taking teaching further: Round 1 application guidance" published in June 2018 and available at: https://www.et-foundation.co.uk/supporting/support-teacher-recruitment/taking-teaching-further/Strand-1-financial-support-initial-teacher-education-ite/

Providers had to complete an application form for TTF[15]. The information provided in this form represents the main source of information regarding which providers applied for TTF funding under Strand 1. It therefore acts as a potential sampling frame or source of information about applicants for a future evaluation, should a similar form be used in future Rounds. The form identifies the lead organisation, the lead contact, their contact details and contains URN/UKPRN fields. Other information required from applicants in completing this form (and therefore potentially available to evaluators as background/classificatory information) is:

- incorporation and ownership;
- size of the organisation;
- details of whether the application is made in collaboration with other partners thereby forming a consortium;
- financial information (audited accounts for the last two years);
- strand applied for (organisations can apply for both Strands);
- confirmation that the applicant meets the provider and trainee eligibility criteria set out above;
- the sector that trainees will work in is childcare and education, digital, construction, or STEM; and
- the provider's objectives, activities, performance indicators, target dates and roles/responsibilities in the form of free text fields.

[15] https://www.et-foundation.co.uk/supporting/support-teacher-recruitment/taking-teaching-further/Strand-1-financial-support-initial-teacher-education-ite/

Section 3: Participation in Taking Teaching Further Strand 1 Rounds 1 and 2

The population of providers that could bid for support under Strand 1 of TTF is those in receipt of Education and Skills Funding Agency (ESFA) funding. This represents the population from which participating providers are drawn. This population also forms that from which future cohorts of providers might be recruited and, should a Level 3 approach to impact evaluation be chosen, the population from which comparison samples might be drawn. If a Level 5 RCT approach is pursued, it is this population from which a sample would be recruited. In 2018/19 this population of providers in receipt of ESFA funding amounted to 1,194 providers across England[16].

Table 2: Population by provider type (2018/19) & participating providers, Rounds 1 and 2

Provider type	Total number	Participation Round 1 only	Participation Round 2 only	Participation Rounds 1 & 2	Total
Employer Providers	37				
General FE colleges	186	12	34	5	51
HE institutions	61				
Independent Training Providers	690	1	3	1	5
Local Authorities	134	1			1
Other publicly funded (e.g. British Army)	10				
Sixth form colleges	49				
Specialist colleges (e.g. agricultural colleges)	24	1	2		3
Post-16 institution	3				
Total	**1,194**	**15**	**39**	**6**	**60**

[16] These population estimates are based on ESFA funding allocations for 2018/19, sourced in September 2019 at: https://www.gov.uk/government/publications/funding-allocations-to-training-providers-2018-to-2019

Source: Education and Skills Funding Agency and programme participation records

Table 2 reveals the vast bulk of the population of providers are Independent Training Providers, but these make up only a small proportion of providers that have participated in TTF over the two rounds (8%). Generally, however, Independent Training Providers tend to be quite small in terms of the overall FE student body (IFF Research, 2019). The bulk of participating providers are General FE colleges (85%). Of the 186 colleges in receipt of ESFA funding in 2018/19 nearly a third have participated in TTF. Tables 3 and 4 below provide a full list of participating providers at Rounds 1 and 2.

In total, 21 providers were successful in applying to receive funding via Strand 1 of TTF in Round 1 (note of these 21, six also took part in Round 2). From the 21 providers, according to programme records two dropped out. Recent evidence suggests that for the 19 providers that participated in Round 1, and for whom data is available, 50 teachers in total were recruited to posts through TTF (IFF Research, 2019). This means that on average 2.6 trainee teachers were recruited per provider out of a possible five. Further, our best estimate is that across Rounds 1 and 2 we can expect roughly 170 teachers to have been exposed to Strand 1 support. These figures are based on data provided by the ETF as of March 2020.

These data provide important information about the likely sample sizes available to any future evaluation, as well as the potential make-up of that sample in terms of the mix of provider type. They represent the best source of data available to assess the likely size of samples that might be available for the evaluation of future rounds of TTF. The mix of providers participating to date suggests that it is General FE colleges that are likely to dominate recruitment in the future[17], thus any evaluation design will need to take into account the plausibility of extending the evaluation population and sample beyond the General FE college population. At present is seems unlikely that any evaluation could focus on providers other than General FE colleges; simply not enough other types of provider have taken part thus far. In the future we assume similar participation and therefore exposure to TTF will be concentrated among General FE colleges.

[17] That is unless the scheme is changed in some way to make it more attractive to the non-college provider base or some form of successful enhanced marketing campaign is conducted targeting private providers.

Table 3: Participating providers Strand 1 Round 1 (correct as of March 2020)

Provider name	Provider type	Drop out	TTF Round 1 Strand 2?	TTF Round 2 Strand 1?	TTF Round 2 Strand 2?
BLACKPOOL AND THE FYLDE COLLEGE	General FE college incl tertiary				
BRIDGWATER AND TAUNTON COLLEGE	General FE college incl tertiary			Y	
BUCKINGHAMSHIRE COLLEGE GROUP	General FE college incl tertiary				
CALDERDALE COLLEGE	General FE college incl tertiary		Y	Y	
CITY COLLEGE NORWICH	General FE college incl tertiary				Y
EKC GROUP	General FE college incl tertiary		Y		Y
GRIMSBY INSTITUTE OF FURTHER AND HIGHER EDUCATION	General FE college incl tertiary				
KIRKLEES COLLEGE	General FE college incl tertiary				
LAKES COLLEGE WEST CUMBRIA	General FE college incl tertiary			Y	Y
LEARNING SKILLS PARTNERSHIP LTD	Independent Training Provider		Y		
LEICESTER COLLEGE	General FE college incl tertiary	Y			
NORTHAMPTON COLLEGE	General FE college incl tertiary		Y		
PETROC	General FE college incl tertiary	Y	Y		
REASEHEATH COLLEGE	Special colleges				
ST HELENS CHAMBER LIMITED	Independent Training Provider			Y	
TAMESIDE COLLEGE	General FE college incl tertiary				

Provider name	Provider type	Drop out	TTF Round 1 Strand 2?	TTF Round 2 Strand 1?	TTF Round 2 Strand 2?
THE NORTHUMBERLAND COUNCIL	Local Authority				
THE OLDHAM COLLEGE	General FE college incl tertiary			Y	
THE WKCIC GROUP (Capital City College Group)	General FE college incl tertiary				
WAKEFIELD COLLEGE	General FE college incl tertiary			Y	
WALSALL COLLEGE	General FE college incl tertiary		Y		

Table 4: Participating Providers Strand 1 Round 2 (correct as of March 2020)

Provider name	Provider type	Drop out of Round 2 Strand 1	TTF Round 1 Strand 2?	TTF Round 2 Strand 1?	TTF Round 2 Strand 2?
ACTIVATE LEARNING	General FE college incl tertiary			Y	Y
ADA NATIONAL COLLEGE FOR DIGITAL SKILLS	General FE college incl tertiary	Y		Y	
BOSTON COLLEGE	General FE college incl tertiary			Y	
BRIDGWATER AND TAUNTON COLLEGE	General FE college incl tertiary			Y	
BROMLEY COLLEGE OF FURTHER AND HIGHER EDUCATION (LONDON SOUTH EAST COLLEGES)	General FE college incl tertiary			Y	
BURY COLLEGE	General FE college incl tertiary			Y	
CALDERDALE COLLEGE	General FE college incl tertiary		Y	Y	
CAMBRIDGE REGIONAL COLLEGE	General FE college incl tertiary			Y	
CHELMSFORD COLLEGE	General FE college incl tertiary			Y	
CHICHESTER COLLEGE GROUP	General FE college incl tertiary			Y	Y
CITY OF WOLVERHAMPTON COLLEGE	General FE college incl tertiary			Y	
CORNWALL COLLEGE GROUP	General FE college incl tertiary	Y		Y	Y
DERBY COLLEGE	General FE college incl tertiary	Y		Y	Y

Provider name	Provider type	Drop out of Round 2 Strand 1	TTF Round 1 Strand 2?	TTF Round 2 Strand 1?	TTF Round 2 Strand 2?
EALING, HAMMERSMITH & WEST LONDON COLLEGE	General FE college incl tertiary			Y	
EAST SUSSEX COLLEGE GROUP	General FE college incl tertiary			Y	
FURNESS COLLEGE	General FE college incl tertiary			Y	
GRANTHAM COLLEGE	General FE college incl tertiary	Y		Y	
HARLOW COLLEGE	General FE college incl tertiary		Y	Y	
HAVANT AND SOUTH DOWNS COLLEGE	General FE college incl tertiary			Y	Y
HUDDERSFIELD TEXTILE TRAINING LIMITED	Independent Training Provider			Y	
ISLE OF WIGHT COLLEGE	General FE college incl tertiary		Y	Y	Y
LAKES COLLEGE WEST CUMBRIA	General FE college incl tertiary			Y	Y
LEEDS CITY COLLEGE	General FE college incl tertiary			Y	Y
LINCOLN COLLEGE	General FE college incl tertiary			Y	Y
MYERSCOUGH COLLEGE	Specialist College			Y	Y
NATIONAL COLLEGE FOR HIGH SPEED RAIL	Independent Training Provider			Y	
NEW CITY COLLEGE	General FE college incl tertiary			Y	
NEWHAM COLLEGE OF FURTHER EDUCATION	General FE college incl tertiary			Y	
NORTH HERTFORDSHIRE COLLEGE	General FE college incl tertiary	Y		Y	Y

Provider name	Provider type	Drop out of Round 2 Strand 1	TTF Round 1 Strand 2?	TTF Round 2 Strand 1?	TTF Round 2 Strand 2?
OLDHAM COLLEGE	General FE college incl tertiary	Y		Y	
PTP	Independent Training Provider			Y	
RNN GROUP	General FE college incl tertiary			Y	
SOUTH DEVON COLLEGE	General FE college incl tertiary			Y	Y
SOUTH ESSEX COLLEGE OF FURTHER AND HIGHER EDUCATION	General FE college incl tertiary			Y	
SPARSHOLT COLLEGE	Specialist College			Y	
ST HELENS CHAMBER LTD	Independent Training Provider			Y	
STOCKTON RIVERSIDE COLLEGE	General FE college incl tertiary			Y	
SUNDERLAND COLLEGE	General FE college incl tertiary			Y	
THE COLLEGE OF WEST ANGLIA	General FE college incl tertiary	Y		Y	
TRAFFORD COLLEGE GROUP	General FE college incl tertiary	Y		Y	
WAKEFIELD COLLEGE	General FE college incl tertiary			Y	
WALTHAM FOREST COLLEGE	General FE college incl tertiary			Y	
WESTON COLLEGE	General FE college incl tertiary		Y	Y	
WIGAN AND LEIGH COLLEGE	General FE college incl tertiary			Y	

Provider name	Provider type	Drop out of Round 2 Strand 1	TTF Round 1 Strand 2?	TTF Round 2 Strand 1?	TTF Round 2 Strand 2?
WINDSOR FOREST COLLEGE GROUP	General FE college incl tertiary			Y	Y

Implications for the evaluation

There have already been two Rounds of Strand 1 funding. The impact of these funding activities has not been evaluated due to the small sample sizes at the first Round and the need to explore the feasibility of conducting an impact evaluation, before any decision about commissioning one. It must be noted, however, that any future impact evaluation of Strand 1 of TTF will need to consider the existence of the first two rounds of funding and their consequences. For example, should providers that have participated in previous rounds of TTF be considered as potential control group members in future evaluations? At the time of writing (i.e. March 2020) this report it is unknown how many of the colleges that have taken part in TTF thus far will do so again in the future.

As noted above, participation in the programme is dominated by General FE colleges. This leaves open the question as to how far other types of provider will engage in future Rounds of TTF. Moreover, if the programme continues to attract overwhelming interest from General FE colleges, there is a question as to whether any impact evaluation should restrict itself to focusing on the effects of TTF for General FE colleges only. In other words, this would mean restricting any future evaluation to the population of General FE colleges (n=186) rather than the wider population of providers. To date, participation among non-General FE college providers has been quite low. Any evaluation would probably need to treat different provider types as distinct subgroups in any analysis[18]. This calls into question whether it would ever be possible to look at the effects of TTF on non-General FE college providers. There would simply not be enough of them, based on existing volumes of participation in the programme and the programme remaining at its current scale.

At Round 1 of TTF Strand 1 , 50 teachers were exposed to the intervention (IFF Research, 2019). It is worth noting that the total capacity of the programme at Round 1 would have been 95 teachers – implying a take up rate of 53%. A total of 106 teachers were recruited through Round 2 of the programme (79 of these recruits were still in post as of March 2020).

These estimates are important because they provide a sense of the reach of TTF across the population of providers thus far and enable us to make informed judgements about what might occur in future.

[18] It is likely that the response to the intervention could vary considerably by provider-type. Ideally this variation in response would be explored in the impact analysis requiring samples for each provider-type of a sufficient size.

Section 4: Target population, level of analysis and outcome indicators

This report has considered the most appropriate framework for evaluation (an approach based on counterfactuals) and looked at the nature of the evaluand as it operated in Rounds 1 and 2, to give a sense of how it might operate in future. In this section, attention turns to a suitable definition of the target population for the study, and the level at which the analysis will be undertaken (e.g. learner/teacher/provider).

Target population and estimand

The question of the appropriate population for the study is important because it determines several crucial features of an evaluation.

As is true for any evaluation, we can never observe the true impact of TTF – we can only provide an estimate of it and an associated margin of error. However, to make progress we need to define what it is that we are attempting to estimate, or the estimand. We have already touched on this question earlier in the report (see Section 2). Here we consider it in greater detail. Essentially, we have two choices, and these relate to the target population for the study. First, we could estimate the average effect of TTF for all General FE colleges – that is, what would the effect on outcomes be if all General FE colleges took part - known as the **average causal effect**. Alternatively, we could seek to estimate the average effect of TTF for colleges that elect to take part (the **average causal effect on the treated**). These are different estimands and relate to different populations. Given that participation in TTF is never likely to be obligatory, the second of these two estimands appears most relevant and therefore is the chosen estimand for the proposed evaluation.

These estimands are defined in relation to General FE colleges. Given the take-up of TTF, it appears that it would be very difficult to estimate the effectiveness of TTF for other forms of provider (e.g. adult and community learning providers, Independent Training Providers, Sixth Form Colleges, Specialist Colleges, private providers, specialist providers, Local Authority providers) due to low sample numbers; though this decision should be kept under review. If the number of providers coming forward to participate from these other groups did increase in future, as will be shown, both forms of impact evaluations considered at Levels 3 and 5 would be more likely to yield conclusive findings. For now, discussion proceeds on the basis that the target population is General FE colleges in England that opt to participate in TTF, or some lower level units within them (e.g. teachers or learners within these colleges).

Proposed approaches at Levels 3 and 5 are discussed later in this report. For now it is worth noting that if an RCT were conducted (Level 5), the estimated average causal effect obtained through such a study could be interpreted as the average causal effect on

those that elect to take part, as long as General FE colleges recruited to the trial did not drop out of the trial after being randomised.

If an evaluation using a non-randomised Level 3 design were conducted, which involves selecting a comparison group, this would yield results directly interpretable as the average causal effect on the treated; assuming it is possible to make the necessary statistical adjustments for unbiasedness.

In conclusion, the estimand of interest and the target population are the average causal effect of Strand 1 of TTF on General FE colleges that take part in the programme. This does not mean we are not interested in the portion of General FE colleges that do not participate. It is from these colleges that we will obtain a statistically adjusted estimate of counterfactual (unexposed) outcomes. As discussed below, an initial assessment suggests that a sample of 350 providers would be required given standard statistical assumptions.

Level of analysis

The theory of change that has been developed for TTF suggests that its effects might be felt at least three different levels (Annex A): the provider, teacher and learner (potential effects among employers are discounted as they are too distant from the intervention). In the General FE college population, learners can be thought of as being nested or clustered within teachers and teachers within these General FE colleges. This section considers each of these levels in turn. It discusses the extent to which an impact evaluation might take the respective level into account in its design and analysis.

Looking first at the provider level, it is providers that are the initial beneficiaries of funding. They secure resources to fill vacancies, which while unfilled are costly to them and potentially undermine the quality of teaching, and the learner experience. It is therefore reasonable to expect that an intervention such as TTF would impact on outcomes for providers. As can be seen in the theory of change (Annex A), it is anticipated that participation in TTF Strand 1 by providers might:

- increase applicants for teaching posts (specifically from industry);
- reduce unfilled vacancies (discussed further below)[19];
- reduce costs of recruitment in absolute terms; and

[19] It will be challenging to assess the effectiveness of TTF on outcomes such as unfilled vacancies due to the multiplicity of factors that are likely to influence such measures. Nonetheless, the research designs discussed in this report attempt to take into account these factors either explicitly or through the creation of a randomised control group and through doing so provide for the possibly of identifying the effect of TTF from among these other factors.

- reduce costs associated with vacancies remaining unfilled.

Whilst in theory TTF might affect these outcomes this does not mean that we would be able to demonstrate any such effects quantitatively. A sample of sufficient size (see Section 6 for a discussion of sample size) is required to ensure that the chances of missing an effect are reduced to an acceptable level.

During Rounds 1 and 2, in the case of General FE colleges, however, sample sizes are quite limited. There are about 186 General FE colleges in England and of these 50 have so far taken part in TTF. A Level 3 impact evaluation design would require a comparison group of non-participating colleges, that would be drawn from those that have not taken part in TTF. A Level 5 RCT needs to be able to recruit enough colleges to form both intervention and control groups in sufficient numbers. Both designs also need to consider participation of the sample in previous Rounds of Strand 1 TTF (the issue of sample size and what can be realistically achieve is discussed in Section 6).

Turning attention to teacher trainees, again this is a clear group of beneficiaries among whom it would be reasonable to anticipate TTF effects. The TTF theory of change reflects this (Annex A). The anticipated outcomes of TTF for teachers include:

- increased completion of Initial Teacher Education;
- higher levels of skill derived from an improved training experience for teacher trainees;
- fewer debts resulting from the reduced costs of ITE among trainees;
- greater levels of job satisfaction; and
- increased teacher retention.

Some of these outcomes are difficult to measure – for example the acquisition of skills, job satisfaction and teacher quality. Although improving teacher quality is an important objective of TTF, it would be a very complex outcome and controversial to measure from a quantitative perspective. It may therefore be better addressed in qualitative research. Moreover, in any analysis, account will need to be taken of the clustering[20] of teachers by General FE college. This means that the effective sample of teachers may be lower than it appears due to this clustering. Nonetheless, in theory it is desirable for an impact evaluation to attempt to measure and explore the effectiveness of TTF on these teacher-level outcomes. Unfortunately for Rounds 1 and 2, the delivery partner responsible for overseeing the pilot did not keep records of the identities of teachers that benefited from

[20] Sampling theory generally assumes that each member of a sample are statistically independent. In the case of TTF any sample of teachers will be drawn from colleges such that teachers will be grouped or clustered within colleges. This grouping or clustering means that teachers from the same college are likely to share certain factors in common such that teachers will not be statistically independent. This potential feature of the data is referred to as clustering throughout this report.

TTF. At Round 2 there was some attempt to obtain such information from providers that met with limited success. If individual teacher trainees that have received funding and support through TTF cannot be identified, then it is impossible to conduct an evaluation that estimates the impact of the programme on outcomes for these teachers. Clearly any future Round of funding for Strand 1 should require that each teacher trainee funded through the programme can be identified and their progress followed over time. An impact evaluation cannot address the effectiveness of TTF on exposed teachers within participating General FE colleges at present - unless data identifying individual teachers is forthcoming.

A related issue, discussed in Section 6 below, is whether declared vacancies could form a unit of analysis. As is the case with a teacher level unit of analysis, vacancies will be clustered or nested within General FE colleges. This is a particularly pertinent issue in the case of a Level 5 RCT design, where it appears most practical to randomise General FE colleges at the point they have declared vacancies, but before they have attempted to recruit to them. This further means that outcomes, such as whether a vacancy has been filled by a permanent employee with an industrial background at say six months subsequent to randomisation, can be considered as an outcome. There may, however, be further issues with accepting vacancies as a unit of analysis connected with, for example, re-organisation of staffing resources within colleges, mergers etc., that render particular vacancies no longer relevant for administrative reasons. Despite this, the prospect for adopting vacancy as the unit of analysis in the case of an RCT is discussed below. One further point is also worth bearing in mind. If teacher was adopted as the unit of analysis in the case of a RCT, analysis of outcomes at the teacher level (as distinct from vacancy) would most likely have to be undertaken non-experimentally; with the attendant disadvantage that results would have a less convincing causal interpretation. The reason for this is that randomisation produces control and intervention samples that are statistically equivalent (in expectations) at the point of randomisation. It is unlikely that an RCT in which teachers are recruited and then randomised would be ethically acceptable from the perspective of teachers. It would mean they would be recruited and then subsequently informed that they have been assigned to a control group and therefore cannot benefit from TTF. This means that randomisation would have to take place before teachers are recruited, at the point at which vacancies are declared. Teachers would only be recruited after randomisation and thus we cannot assume that randomisation would lead to a balanced sample of teachers appointed to declared vacancies. For this reason, a teacher-level analysis would be non-experimental.

Finally, the TTF theory of change (Annex A) suggests that the programme's impacts will be felt by learners. The causal pathway (the mechanism through which this is achieved) is that TTF, through attracting teachers with industry experience, improves the quality of teaching received both in terms of the relevance of what is taught and the consistency with which teaching is delivered (achieved through the avoidance of relying on supply teachers). It is hypothesised that both these factors will lead to improvements in

knowledge as well as in attainment among learners, and finally to better employment prospects. If a sample of learners exposed to teachers who have been receiving funding and support through TTF can be identified, and outcomes such as those listed below obtained, and an appropriate comparison sample can be identified, the following effects on learners could in theory be calculated:

- improvement in attainment (test scores); and
- improvement in employment outcomes (wages, non-wage benefits, and progression).

Again, any analysis would have to take account of clustering.

So far this report has discussed the challenge of whether the effects of TTF could be statistically observable among a sample of learners (see discussion in Section 6) and the complexity of the sample itself (due to multiple layers of clustering), Notwithstanding these challenges, it is not possible to identify exposed teachers at present (due to reasons discussed). Therefore, it is not possible to identify the relevant learners either.

Even if teachers could be identified, it would still be very challenging to obtain data on learners that have been exposed to their teaching. Test scores are recorded for learners in the Individualised Learner Records data set. Whilst the ILR does record the learning aims of learners it does not reveal the identity of their teachers, though in some cases it may be possible to identify teachers through a manual process bearing in mind that this would be resource intensive. SIR data could potentially be linked to the ILR but, as we will see, the coverage of SIR is poor and the challenge of not knowing the identities of teachers and thus learners exposed to TTF remains an issue. Although it is in theory possible and advantageous to link these data sets together, in practice it is unlikely this could be achieved.

Outcome indicators

The issues discussed in this report so far show that the current way TTF is arranged (in terms of the application process and the recording of beneficiaries) and its scale means that conducting an impact evaluation at the level of the General FE college is the only viable strategy (with the exception of extending this to the level of the vacancy in the case of an RCT). This situation could change given different arrangements and these changes are discussed further below. For now, based on the TTF theory of change, the report discusses outcome measures and plausible indicators of these measures, that might be considered as primary or co-primary outcomes in an impact evaluation conducted at General FE college level. In Section 5, we examine some of the data sources that might be consulted and that might yield relevant indicators and other data items required.

It might be reasonable to select the number or proportion of unfilled vacancies, at a given point, as the primary outcome for General FE colleges, particularly due to the fact that declared vacancies are the most plausible unit of an analysis in the case of an RCT (as discussed above). TTF would be expected to reduce the number of such vacancies, but this does not mean that it would do so at a rate that would result in an effect on vacancies that would reach statistical significance at the 95% confidence level[21] (there would be a high chance the results were consistent with there being no effect). In other words, even where a reduction in unfilled vacancies was observed findings may still be inconclusive due to a high degree of uncertainty in the statistical estimates.

Secondary outcomes might be staff turnover and posts filled by teachers employed on non-permanent contracts (refer to Annex A for the Theory of Change). Positive effects on both these variables can be converted into measures of provider cost-savings thus facilitating the reallocation of resources within colleges to more productive ends and raising efficiency. To construct measures on these outcomes several indicator variables will be required. Specifying these provides a sense of the types of data that will be required from both participating, and in the case of a Level 3 evaluation design, non-participating colleges. (Please note at this point in the discussion the possibility of either a Level 3 or 5 design is left open).

The evaluation would need to first identify an academic year that determines pre- and post-exposure periods of time, and for which the measures would be constructed. This would be a complete teaching year from September to July. In the case of an RCT, it would be desirable to collect indicators that relate to the pre- as well as post-exposure teaching years. Given the likely sample size for an RCT, gathering indicators in a pre-exposure teaching year would be necessary as these could be used to increase the statistical power of any analysis, to some extent compensating for the otherwise low sample numbers. In the case of a Level 3 design, it would be essential to collect indicators relating to pre-exposure years, so that variables derived from them can be used in the necessary statistical adjustment.

[21] Although a 95% level of statistical significance is discussed here, as this is typically the level generally accepted for hypothesis testing, this is an arbitrary level. Alternatively, a 90% level of statistical significance might be chosen, or indeed some other level. The choice of significance level is related to the level of Type I statistical error that is acceptable. At the 95% level we accept a five per cent Type I error rate. This means that we accept that in five per cent of hypothesis tests we will declare results to have reached statistical significance where in fact the null hypothesis of no effect is true – that is we make an error. In evaluation, it is generally accepted that the Type I error rate that is acceptable for a given study is a policy decision. Essentially what level of risk is acceptable to policy makers, bearing in mind that declaring an effect to be statistically significant in error might lead to a policy being introduced, or continued, which is ineffective but which requires ongoing resourcing.

The following indicators for both pre- and post-exposure periods of time are required, per provider (these outcomes have been discussed and agreed with the Department and other stakeholders):

- number of FTE teaching positions at 1st September;
- number of FTE teachers permanently employed and in post at 1st September;
- number of FTE teachers temporarily employed and in post at 1st September;
- number of FTE teaching positions as at 31 July in the following calendar year;
- number of FTE teachers permanently employed and in post at 31 July in the following calendar year; and
- number of FTE teachers temporarily employed and in post at 31 July in the following calendar year.

These indicators would need to be sought for each subject area of interest.

Asking for colleges to provide data on the indicators above and for the subject areas listed would provide finer indicator measures, more sensitive to change, because it is posts in these areas that would be directly affected.

Implications for the evaluation

The implications for the evaluation arising from the discussion in this section are as follows:

- the estimand of interest is the average causal effect on outcomes among General FE colleges that participate in TTF (the average causal effect of treatment on the treated);
- at present data are not collected that permit teachers that receive support and funding through TTF to be identified. This rules out the possibility of looking at the effects of TTF on teachers and by extension learners. Thus, the primary unit of analysis will be the General FE college; and
- there are a range of indicators that, if observed, would enable outcomes of interest to be calculated for General FE colleges exposed to TTF. Ideally these would be collected by subject area to allow for finer measures which are more sensitive to change.

Section 5: Sources of data

Thus far the report has defined the relevant estimand and evaluand, considered the unit of analysis that might form the basis for an evaluation, and examined the relevant outcomes and indicators of interest. It has also noted that two types of design are likely to be relevant to any evaluation of TTF that conform to a 'counterfactual' approach to impact evaluation and that meet aspirations around the quality of evidence required; these are designs at Levels 3 and 5 on the Maryland Scale. Finally it is noted that a Level 3 design will require variables for the population of General FE colleges that capture outcome indicators/measures, record the identities of providers that participate or are exposed to the intervention and a range of potential control variables that enable estimates of impact to be statistically adjusted for potential biases.

This section examines a number of the most notable sources of existing data that could potentially be used in an impact evaluation, covering:

- data from administrative or management systems;
- programme data; and
- existing survey data.

These data were collected for reasons other than the evaluation of TTF. The fact that these sources were created without the evaluation of TTF in mind yields an advantage but also presents challenges. The advantage is that the costs of generating them are already incurred and do not fall to the evaluation, although they would still incur costs from any required data checking, cleaning and restructuring.

The challenges posed by using these various forms of data for evaluation can be considered under the following headings:

- Coverage – do they include observations on the population of interest, or some subset of this population, e.g. all General FE colleges?
- Content – do they contain data from which outcome indicators might be derived, details on participation or exposure, and/or do they contain data from which control variables might be constructed?
- Timeliness – are the data available for time periods relevant to the requirements of the evaluation? For example, are data from which outcome measures are derived collected for periods of time both prior to exposure as well as post-exposure?
- Linking – it is unlikely we will find a single source of data that fulfils all the data-requirements of an evaluation design. Therefore, consideration should be given to whether it is possible to link various relevant sources together to form the types of data sets required. To link across sources requires

common identifying information to be held on each source as well as appropriate permissions.

These issues give rise to further considerations around the nature of legal access to data, data protection and privacy issues, as well as compliance with GDPR. These are highly complex considerations with a range of technical challenges associated with them. These issues will need to be fully addressed in future work but have already been raised in discussions with the data owners of the various sources considered.

Data sources considered and an assessment of their usefulness

Table 5 below provides a description of each of the data sources consulted. For each source a series of advantages and disadvantages are set out. In this section each source is discussed in turn, with an overall assessment of its usefulness from the perspective of an impact evaluation of TTF. The table also includes observations on how sources might be tweaked or adjusted to make them more useful for evaluation in future, and how the intake/eligibility process might be adjusted so that an evaluation of TTF might make better use of these sources.

Staff Individualised Record (SIR) data holds much promise when it comes to a single data set that might be of use in an impact evaluation of TTF. It would be worth exploring whether programme records held by ETF from Rounds 1 and 2 could be linked to General FE college records held on SIR, using the URN or UKPRN fields. During the course of this study, it was announced that SIR data as currently constituted was being discontinued. However, we proceed with a discussion of SIR as the exact form of the data that will replace it is unknown, and the observations and issues raised in relation to SIR could be used to inform the development of the data that will replace it. As Table 5 indicates, however, the main problem is that only around half of General FE colleges make a return to the SIR. From the perspective of evaluation, it would be extremely advantageous if completion of SIR could be made mandatory at least for General FE colleges; however, it is understood from discussions with officials that mandating is or would not have been possible. Making SIR returns mandatory or any replacement data set, however, would mean, as a source, it would provide full coverage of the population of interest. This is particularly important for a Level 3 evaluation design. Once data from TTF participation records were linked to extracts of the SIR or its replacement it would be possible to identify both those General FE colleges exposed to TTF and those that remained unexposed and from whom control General FE colleges might be sampled. Moreover, if TTF participation records also captured details of the teacher trainees funded through the programme (see below) these teachers' records on the SIR could also potentially be identified, and likewise a control group of teachers. Though it should be noted that records on SIR are arranged by contract rather than teacher, so this would

add a complication to any attempt to link records at the teacher level. It is understood that the proposed data set that will replace SIR will be organised by teacher. It is recommended that the department explores the possibility of linking participation records at the General FE college level from Rounds 1 and 2 of TTF to existing SIR data for the academic year 2017/18, to examine more precisely the coverage of SIR data with respect to TTF participation.

TTF programme participation records capture the identities of providers exposed to or participating in TTF. These records are generated by application forms submitted to ETF by providers wishing to participate in the programme. For this reason, they are the single most important source of data for any impact evaluation. The data are organised by URN/UKPRN which means they can be linked to other sources such as the SIR (as described previously), the Individualised Learner Record (ILR) and college accounts data. These potential linkages are at present of limited importance because of limitations with each of these data sources discussed in this section. Important and relevant background information about some providers in the case of the SIR can be linked to participation records, and some useful information about learners (a provider's student body) could in theory be linked to participation records from the ILR. However, due to coverage issues with the SIR and the fact that teachers and learners cannot be linked on the ILR, such linking can only be of limited use.

The main advantage of the ILR is that it contains a complete census of all learners in the FE sector, their learning aims and outcomes. It is an individual learner-level data set and contains details of which provider(s) the learner is attached to. The data are organised by unique learner number. Unfortunately, the ILR does not identify which teachers have taught learners and at present, the identities of teachers funded through TTF are unknown. However, if in future TTF captured the identities of funded teacher trainees and these could be linked to learners in the ILR, the ILR could be an important source of outcome measurement for evaluations of future programmes that target student attainment through delivering programmes to teachers, such as TTF.

There are two forms of survey data that collect information about the General FE college workforce. It is natural to ask therefore whether such sources could be used in any impact evaluation going forward. The two surveys are the College Staff Survey[22] commissioned by DfE (with fieldwork that took place between April and June 2018), and the College Workforce Survey[23] overseen by the Association of Colleges (with data related to an academic year but undertaken on a regular basis). If we consider an evaluation of TTF at the level of the provider, focusing on General FE colleges, then the main challenge facing us in utilising these sources is their partial coverage. The total

[22] See https://www.gov.uk/government/publications/college-staff-survey-2018
[23] See https://d4hfzltwt4wv7.cloudfront.net/uploads/files/AoC-College-Workforce-Survey-2019-20.pdf

number of General FE colleges in England is relatively small (186). This means that any Level 5 or Level 3 impact study would face a relatively small population and a rather small sample of exposed colleges. If the evaluation were reliant on either or both of these two survey sources, this sample would be likely diminished further as it could not be guaranteed that participating providers would have responded in enough numbers to either survey. For example, in the College Staff Survey the response rate to the Principal's Questionnaire was 70% and for the staff returns 59%. Likewise, for the AoC's College Workforce Survey the response rate for the main questionnaire for 2017/18 academic year stood at 49%. At the present time we do not know the proportions of General FE colleges within these samples that also participated in TTF.

Table 5: Potentially relevant sources of provider, teacher trainee and learner data

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
Staff Individualised Record (SIR)	Education and Training Foundation (ETF) More information: ETF SIR webpage	The staff individualised record (SIR) data set records basic information about the Further Education Workforce both teaching and non-teaching staff. The latest available data set contains information on the FE workforce over the academic year 2018/19 in England and contains over 91,000 individual level records (the entire FE workforce is thought to comprise roughly 217,000 individuals). All Providers funded by the Education and Skills Funding Agency are 'requested' to submit data via a web portal.	If provider level records from TTF programme participation data (see next entry in this Table) could be linked to SIR this would enable participant providers in SIR to be identified and a potential group of control providers. Moreover, providers could be tracked over time given that SIR is an annual exercise thus providing a crucial longitudinal element. Furthermore, if TTF programme participation records captured the identities of teachers funded through the programme, these teacher trainees could also be found on the SIR, and potentially also a suitable teacher level control group. SIR data appear also to enable some of the	The main disadvantage of SIR is its coverage. Providers are not mandated to provide a return. For evaluation purposes, this is a serious deficiency. At 2018/19 for example, of the 186 providers in the data set, 93 were General FE colleges meaning that just over half of the target population completed a return. As a result, it is likely that a significant proportion of General FE colleges participating in TTF do not appear in the SIR data. It should also be noted, that at the time of writing SIR data was in the process of being replaced by a new data source.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
			provider-level outcome indicators described above and also contain some useful control variables.	
TTF programme participation records	Education and Training Foundation More information: ETF Taking Teaching Further webpage	These are records capturing the details of which provider applies for funding through Strands 1 and 2 of TTF. In the section of this report that describes the evaluand we discuss the nature of these data.	These data enable us to identify providers that applied for funding from TTF and also obtain some useful background information about them. These records can be linked to other sources such as SIR and the college accounts using UKPRN. It should also be possible to identify which applicants were successful in acquiring funding and therefore which providers have been exposed to the programme. The records are complete and held by the ETF.	These records do not capture the identities of teacher trainees funded within each provider. There is no other record kept of the individual identities of beneficiary teacher trainees.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
Individualised Learner Record	Education and Skills Funding Agency (ESFA) More information: ILR webpage	ILR is a continuous data collection exercise which seeks to collect a range of important data on learners in FE colleges, former External Institutions, Sixth Form colleges, Training Organisations, Local Authorities, Academies, and Voluntary and Community Organisations. ILR returns are required from providers who receive funding directly from the ESFA, or through Advanced Learner Loans. FE colleges must send data for all learners, including those that are not funded by the ESFA. This source contains a wide range of data items on Leaners but importantly these include: learning	ILR is an important source that would enable evaluations to determine the effects of interventions on learner attainment/outcomes. Records are organised by individual learner. They can be linked to the college or provider, so in theory learners in providers that are participating in an intervention can be contrasted to those in non-participating providers for interventions delivered at the provider level such as TTF. In theory records can be linked across waves of ILR to create longitudinal data sets that could be used in Level 3-type programme evaluations. As can be seen ILR could be an important	Given that TTF currently does not collect information on teachers receiving funding through the programme and that there is currently no way of linking teachers to learners on the ILR, at present, ILR cannot be used as part of an evaluation of interventions such as TTF.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		agreements/aims and objectives, learning start and planned/actual end dates and funding details; learner details including name, sex, date of birth, home postcode (and other contact details), ethnicity, campus/college ID, learning difficulties, employment and destinations, progression and learning outcomes. Records are held on individual learner and organised by the Unique Learner Number	source of control variables at the provider level through aggregating learner level records by provider.	
College Staff Survey	Department for Education More information: CSS main report	Commissioned by DfE the College Staff Survey collected data in three forms (fieldwork of mainstage took place April to June 2018): a Principals	The College Survey contains data from which very useful outcome measures and control variables could be derived. For example, at the provider	In theory it should be possible to link programme participation records with the Principal survey returns, thus potentially using the survey returns as a baseline, and in future it will be possible to

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		survey, a teachers and leaders questionnaire and a summary staff return. The research was designed to provide insights into the experience, qualifications and perceptions held by teachers and leaders in general and specialist FE colleges. The summary staff return captures the number of leadership, teaching and agency staff employed by the college the nature of their employment, subject as well as number of vacancies. Principals were asked how easy it was to fill vacancies, by subject and	level it contains snapshot information about vacancies, recruitment difficulties and a lot of information about teachers including their previous experience of industry and future work intentions. The timing of the survey, running between April and June 2018 leaves it well placed to act as a baseline for Round 1 of TTF. Future related data collection, should a survey of a similar nature be considered, could provide follow-up observations on post-exposure outcome measures for future rounds of TTF. So, coverage in terms of question subject matter is certainly of interest to any evaluation, and the	identify participating providers and even teachers in the data. However, as the survey was administered permissions do not exist for linking data although any future surveys of a similar nature might be designed to make this legally possible. Clearly the 2018 survey could not have been used to identify teacher trainees that had benefited from TTF due to its timing. Generally, however the coverage of this survey will not necessarily overlap with TTF programme participants and funded teacher trainees to the extent that participants will be present in the survey in sufficient numbers to permit the types of analysis envisaged.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		how this was changing over time. The teachers and leaders survey contains detailed information about the background of staff in the form of individual level data, including experience outside of FE and experience in FE, qualifications and experience in industry. Respondents were also asked about how likely they were to leave FE in the future. Although described as a survey it is in effect an attempt at a Census in that efforts are made to collect data from all college principals and all teaching and leadership staff. The teachers and leaders	timing of data collection could also be an advantage.	

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		questionnaire is distributed to college staff by participating principals. The survey targeted 199 colleges and college groups (22 were college groups, 177 single colleges). Response rates were 70% for the principal's questionnaire (n=140), 14% for the teachers and leaders questionnaire (n=9,603) and 117 staff returns were received (59 percent). A follow-up survey of teachers and leaders was conducted in 2019.		
College accounts	Education and Skills Funding Agency More information: ESFA college	The ESFA maintain a data base of college accounts. For the academic year 2017/18 the file contains financial returns for 258 of	The college accounts data base has good coverage and is likely to capture data for participating college providers. The data are	Whilst these data are of interest, the data has good coverage and they are linkable to other records through the UKPRN identifier, they are realistically

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
	accounts webpage	266 institutions[24]. The data are organised at the college level and by the UKPRN identifier. This means that variables from this file can easily be match to other sources organised by UKPRN. The accounts contain detailed information on the size of the teaching and administrative workforce at each college and their costs of employment. The data contain measures of the total premises size, IT resources, college income, staff costs and a range of other financial indicators including balance sheet	organised by academic year and so could be used to construct some longitudinal measures on variables that might be considered good control variables. It should be possible to link data from the college accounts to the TTF participation records using the UKPRN identifier. In future years colleges maybe asked directly whether they have received TTF funding.	only a source of control variables to be use in either as baseline measures in a Level 5 RCT or as control variables in a Level 3 design. Other data would be required in order to derive outcome measures.

[24] The figure quoted here in relation to the total returns to the college accounts is different to that quoted elsewhere, in for example Table 2, relating to General FE colleges. This is because the collect accounts include returns from agricultural colleges, specialist, sixth form and arts colleges that are not included in the definition of General FE colleges used in this report.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		items and financial health measures. The 2017/18 data obviously did not record TTF income but future years data may capture this.		
College Workforce Survey	Association of Colleges (AoC) More information: College workforce survey 2017/18 main report	This data source is a survey (in effect an attempted census) of FE sector, Sixth form and Specialist Colleges in England. It collects data on the college workforce, staff turnover, recruitment challenges, working conditions, workforce development and workforce sickness and absence. Data are collected in the Spring and relate to the previous completed academic year. So the 2019 data collection exercise related to the	The AoC college workforce survey is a very important data source. It contains many variables that could be used as outcomes. In theory it should be possible to link TTF programme participation records to these data in order to determine which sample members participated in TTF and which did not and could therefore at as a control group member. The data set is also a potential source of important control variables and the timing of data collections could	The main disadvantage is that this source does not allow us to identify participating TTF providers and would have to be linked to TTF participation records. The fact that this source is owned by the AoC may present legal barriers to doing so. Furthermore, the survey suffers from a less than 50% response rate and therefore we cannot assume that all TTF participating colleges will provide data to this survey.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		2017/18 academic year. The target population for the 2017/18 survey consisted of 251 institutions. 124 returns were received consistent with a 49% response rate. Crucially the survey collects measures on workforce composition, staff turnover, type of employment contract, reasons for staff turnover, redundancies, vacancies and hard to fill posts.	potentially be used as baseline and follow-up sources.	
UK Register of Apprenticeship Training Providers	Education and Skills Funding Agency More information: Register of Apprenticeship Training Providers webpage	This source of data, held and maintained by the ESFA, lists organisations that are eligible to receive government funding to train apprenticeships. Applicants provide a range of details regarding how they deliver	This list clearly extends beyond FE colleges to all providers of apprenticeships training. It is a useful source as a sampling framework for drawing samples of apprenticeship providers. It contains some limited data that could be used as	The source has good coverage of apprenticeship providers but has a limit range of variables that might prove important in any impact analysis and certainly no relevant outcome measures.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		apprenticeships, evaluating the quality of delivery, experience of their employees as well as their financial status. Institutions that wish to join the register do so through completing their application via an online portal.	control variables in a Level 3 impact evaluation design, or as baseline variables in a Level 5 RCT.	
UK Register of Learning Providers	Education and Skills Funding Agency More information: UK Register of Learning Providers online portal https://www.ukrlp.co.uk/	A register of providers across the UK used by government and agencies to share information about them. There are approximately 30,000 providers registered on the system. Each provider that has been validated against some external source to the register is issued with a UKPRN identifier. This is the unique ID used to share information across	This register is again another important source for drawing samples of providers, particularly those providers beyond the scope of other surveys and SIR. It is also useful in defining the extent of the population for any evaluation, given that certain potential populations are likely to be subsets of the records held on this data base. If it becomes possible to extend the scope of an	At present the scope of an evaluation is unlikely to extend beyond FE colleges this source is only likely to be of use in defining the population. In and of itself, it does not contain many data items likely to be used in impact analysis other than some possible control variables.

Data set title	Owner	Description	Assessment: Advantages	Assessment: Disadvantages
		multiple systems such as ILR and SIR. The online portal (link to the left) enables users to search the data base and view information such as the providers contact details, legal address, website, links to Ofsted inspection reports, general details of the provider[25], regulatory data from the Office for Students and a link to attainment data (though this could not be made to work for many providers).	evaluation beyond considering the effects of TTF on General FE colleges and their student body, such as source may well prove useful.	

[25] For example, the record for Burnley College

The ESFA maintain a data set of colleges' financial standing. The records are collected and organised by academic year and UKPRN and so could be linked to existing TTF programme participation records. Although this source is unlikely to provide useful outcome measures, it could be used to extract possible control variables under a Level 3 design, or baseline measures for an RCT Level 5 design (for example variables that capture the financial stability of the college and their expenditure on staff costs).

Finally, the two sets of register data considered here (UK Register of Learning Providers and UK Register of Apprenticeship Training Providers) could potentially perform an important function in a Level 3-type impact evaluation. They would enable identification of the population of General FE colleges and other providers at which TTF is targeted and from which participating providers were drawn. Thus, they can act as sampling frames for an evaluation. Beyond this, however, it is difficult to see what further contribution they might make.

Implications for the evaluation

This section has examined a number of data sources and assessed them for their potential contribution to an impact evaluation of TTF. There are a number of sources that appear to contain data in rich enough form to be used for a Level 3-type impact evaluation if they can be linked to TTF participation records. However, each of these sources – the SIR, the College Staff Survey and the College Workforce Survey - suffer from coverage problems. Those sources that are less affected, at least in theory, by coverage problems have other deficiencies. TTF programme participation records do not indicate the identities of teachers funded through the programme. The ILR does not permit learner records to be linked to the teacher they were taught by within each provider. The college accounts appear to have good coverage but do not contain variables that might act as outcome measures, despite possessing data from which some useful control variables might be derived.

For these reasons it appears likely that any impact evaluation of TTF would need to rely on at least some primary data collection.

In terms of useful exploratory work, this feasibility report recommends that:

- TTF participation records are linked to the 2017/18 SIR and/or the College Staff Survey[26], after consideration is given to the practical constraints and GDPR-related barriers (bearing in mind that these may prevent linking), to examine the extent to which survey records overlap with participation records. Both these sources provide a range of useful indicators (see Table 5) that can be used to understand the characteristics of colleges that have and have not participated to date in TTF and therefore provide information for planning future evaluations; and

[26] We are not recommending attempting to link TTF participation records to the College Workforce Survey as this is owned by AoC and there may not be a legal gateway making such linking lawful.

- TTF participation records are linked to the college accounts so that an analysis of the financial standing of colleges that participated in TTF relative to those colleges that did not do so, might be undertaken.

Section 6: Proposed approaches

This section discusses two approaches to the evaluation of potential future rounds of TTF. Before commencing this discussion, some reflections are offered on the main challenge facing an evaluation of TTF. This is the challenge of its scale. Linked to this, there are challenges with sample sizes that any statistical approach to impact evaluation will encounter. This is likely to pose a challenge unless participation among General FE colleges and other providers rises appreciably in future Rounds.

The funding available to support teacher trainees for each declared vacancy under TTF is £18,200. At a maximum, participating providers can apply for five places each. This means that participation in TTF could provide up to £91,000 per provider per year. This assumes that each provider can recruit five qualifying trainees and, as we have seen based on the experience of Rounds 1 and 2, this is not always possible.

To put this figure in context, the college accounts reveal that the 174 General FE and technical colleges responding and providing the necessary data had total teaching staff costs of £1,819m in 2017/18, or an average of approximately £10m per college. The financial contribution to a college of £91k would represent just under one per cent of the annual staff budget of the average General FE college. From an evaluation perspective, what this means is that in terms outcomes at the provider level (such as staff recruitment and turnover) the effects of TTF are likely to be small. It is therefore harder (but not impossible) to detect in statistical terms given the number of colleges exposed. This is because when viewed in total resource terms the contribution financially of TTF is rather modest. We cannot therefore expect its effects to be large. A more generous financial contribution would be expected to result in a larger impact (i.e. if a greater incentive is created for teacher trainees). Indeed, this was a view expressed by industrial professionals that had not applied to TTF who took part in the process evaluation. Although they were positive about TTF as a concept, they felt it would need to offer more financial support to encourage more industry professionals to apply. They felt a potential reduction in income was the biggest barrier preventing them from leaving industry to become a FE teacher.

Causal effects may be easier to detect at a teacher level, but we do not know the identities of the teachers funded through the programme. This means we cannot follow them over time to examine job retention and advancement (though it may be possible to track teachers in future). In the case of a randomised design, where there is interest in estimating effects at a teacher level, teachers would have to be recruited first prior to randomisation, but only those in providers allocated to the intervention receiving funding. This would be difficult as it would mean denying funding to teachers that had already accepted posts. We discount this as an option due to this. As mentioned previously, any RCT may have to focus attention on the effects of TTF at the level of the vacancy rather than teacher, where vacancies are declared by participating providers at the point of randomisation.

With this in mind, the report now examines the prospects for both an RCT (Level 5) and a Level 3 comparison group design, paying particular attention to the practical requirements for each approach and the likely sample size constraints that will be encountered.

Randomised controlled trial design (RCT)

This section considers an RCT design that might be integrated into the in-take/application process for TTF in future Rounds of the programme. The designs discussed are all based on the assumption that the majority of any applicants to future Rounds of TTF Strand 1 will be General FE colleges, and thus any design will attempt to estimate the effects of the programme for General FE colleges only.

Figure 1 below provides a visual depiction of a proposed randomised design that involves the allocation at random to intervention and control groups of applicant General FE colleges. Much like the pilots which ran at Rounds 1 and 2, applicant providers would be asked to complete an application form. On that form they will be required to provide the same level of detail as was the case in Rounds 1 and 2. In addition, details of the following information for the last complete academic year will be required:

- number of FTE teaching positions at 1st September;
- number of FTE teachers permanently employed and in post at 1st September;
- number of FTE teachers temporarily employed and in post at 1st September;
- number of FTE teaching positions as at 31 July in the following calendar year;
- number of FTE teachers permanently employed and in post at 31 July in the following calendar year; and
- number of FTE teachers temporarily employed and in post at 31 July in the following calendar year.

For the college as a whole and broken down also by the subject areas.

These variables will be used to test the extent of balance between the intervention and control sample, as well as potentially used in analysis of the results.

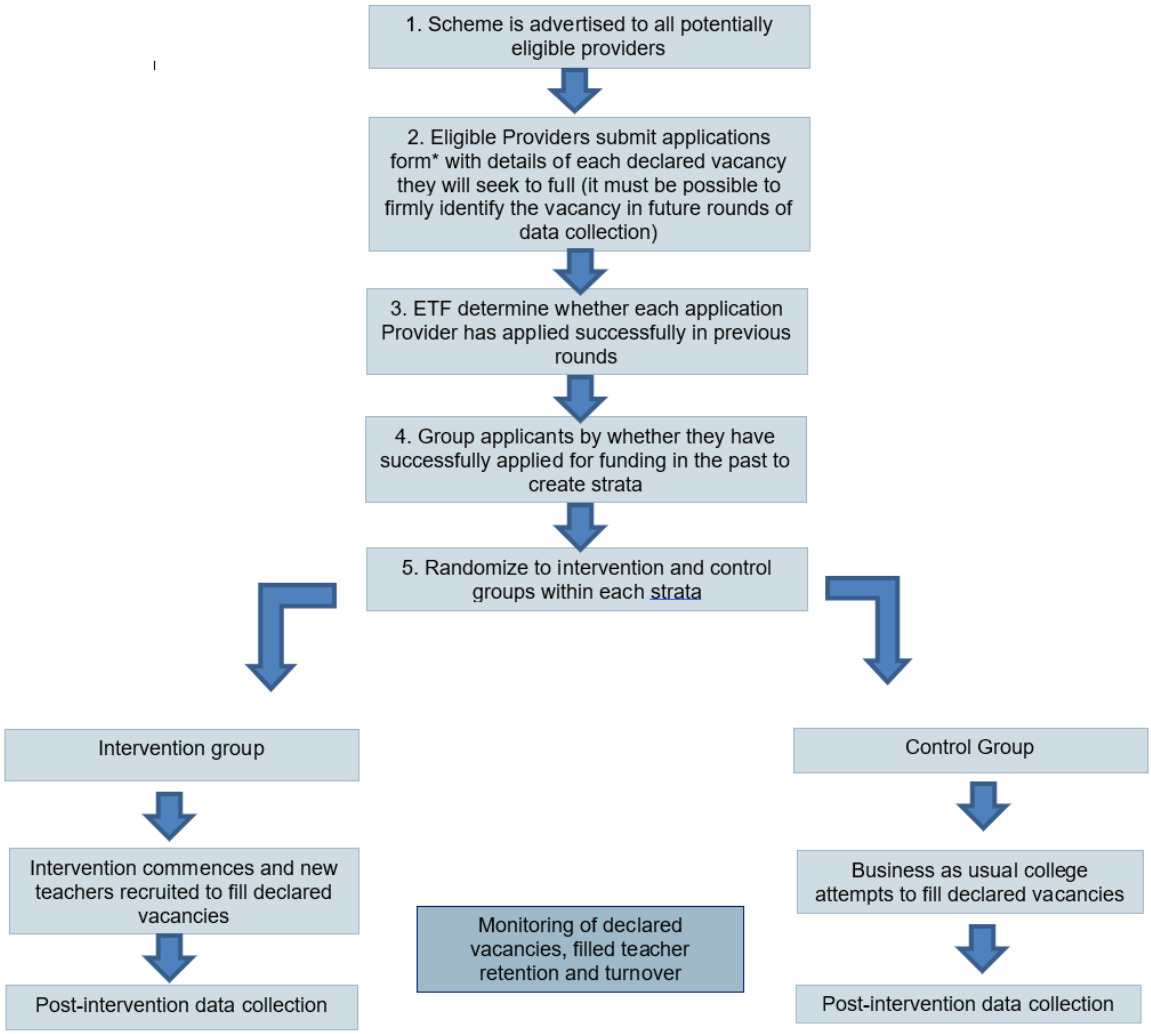
As was the case in Rounds 1 and 2, it assumes that providers could indicate up to five vacancies into which they intended to recruit a teacher trainee previously employed in industry. These must be firm vacancies (that is, likely to exist over a sustained period and not due to illness or temporary reorganisation, etc.) for which it would be possible to track their status over time (though note previous comments relating to doubts about this); that is whether the vacancy was filled by a particular date by a permanently employed teacher trainee, by what type of trainee (by level of prior experience) and for how long the vacancy remained filled, etc.

Once forms are received and checked for completeness, applying colleges could be grouped by whether they had previously received funding through TTF. Randomisation might then be performed separately within each group at the college level, with vacancies clustered or nested within the college. This ensures that the resulting intervention and control groups contain equal proportions of General FE colleges that had previously taken part in TTF.

Table 6 provides analysis of how sensitive various randomised samples are to differences in outcomes between intervention and control groups, based on different assumptions regarding sample size. Generally, a larger sample size is preferred because it will enable a small difference between the two groups to be detected statistically.

The analysis assumes that there are two future Rounds of TTF, Rounds 3 and 4, that at each Round randomisation occurs as depicted in Figure 1, and the samples across both Rounds are pooled together for the purpose of analysis in order to boost sample size. This means that results will only be forthcoming at the end of Round 4 and relate to the aggregate effects of the programme over both Rounds.

Figure 1: Preliminary proposed design for an impact evaluation of future Rounds Taking Teaching Further Strand 1



The key conclusion is that TTF would have to have quite a large effect on outcomes if we are to avoid a trial with inconclusive findings. For example, if teacher turnover were the main outcome of interest, TTF would be required to reduce turnover by a substantial amount relative to the control group position. As previously discussed, TTF Strand 1 is currently quite a modest intervention. Therefore, unless the number of providers recruited to the programme, and thereby to the trial, is much higher than previously seen, and greater in number than that assumed in Table 6, there is a high chance that any trial would fail to find an effect at the 95% level of statistical significance. This does not mean that TTF does not produce an effect, but it is likely the effect would not reach statistical significance.

In more detail, Table 6 presents a range of sample sizes and what are known as effect sizes[27]. The upper panel of the Table contains three assumed sample sizes of General FE colleges that could be assigned to intervention and control. The lower panel shows a range of assumed sample sizes based on declared vacancies. The effect size is a metric that shows the scale of the difference in outcomes between intervention and control groups that a given sample size could detect (that is, would yield a difference that is statistically significant given certain assumptions). The Table seeks to show how sensitive a particular sample size is to differences between average outcomes in intervention and control groups, first if analysis were conducted at the General FE college level and second at the level of declared vacancies.

[27] These are standardised mean differences in outcomes – the mean of the outcome in the intervention group minus the mean in the control group divided by the pooled standard deviation.

Table 6: Effect sizes for different experimental analysis conducted at the General FE college and vacancy level at various levels of statistical power

Total approved applications per round (total sample of approved General FE college applicants Rounds 3 & 4):	General FE college level (standardised mean differences)
30 (56)	0.68
40 (74)	0.59
50 (93)	0.53

Number of observed vacancies associated with General FE college level samples - Rounds 3 & 4	Vacancy level estimates (standardised mean differences)
280	0.44
370	0.38
465	0.34

Notes: For the General FE college level analysis (upper panel)

- Calculations performed using PowerUp (Dong, Kelcey, Maynard, & Spybrook, 2015)
- Given that data are pooled over two Rounds of TTF intake in the second Round (Round 4) we exclude from the randomised sample those providers already randomised in Round 3 – though these providers could still have participated in Rounds 1 or 2. From Rounds 1 and 2 we estimate that for a given round approximately 15% of colleges will submit repeat application. This adjustment also affects the samples of vacancies available for analysis reported in the lower panel of this table.
- Effects reported are minimum detectable effects sizes
- Some allowance for statistical adjustment of effects is made with the proportion of the variance explained by the inclusion of a single covariate assumed to be 0.20.
- It is assumed there is no attrition

For the vacancy level analysis (lower panel)

- Calculations performed using PowerUp (Dong et al., 2015) and assume a continuous outcome or dependent variable for ease of calculation whereas in reality outcomes are likely to be binary or even survival rates (though formulae are very similar for the continuous and binary case (Hox, Moerbeek, & Van de Schoot, 2018))
- A single covariate assumed at the Provider level that accounts for 0.10 of the outcome variance
- Intra class correlation coefficient is set at 0.20, a value typically used in the planning of school experiments

To understand Table 6, consider the first row. Here it is assumed that 30 General FE colleges are recruited to TTF Strand 1 at Round 3 and the same number again at Round 4. Of these 60 colleges, four applied at both rounds and therefore they are not randomised again at Round 4 giving a total sample of 56 colleges over two rounds (these estimates on based on provider applicant behaviour at Round 1). The effect size associated with this sample is 0.68. An effect size of 0.68 is considered large for

educational and social interventions. This means that TTF Strand 1 at Rounds 3 and 4 would together have to have a large effect for any evaluation to have a good chance of finding a statistically significant difference between intervention and control groups. To illustrate further, we can convert the 0.68 effect size into a percentage point difference as a way of making the effect size metric more meaningful. If we assume that an outcome is binary, for example whether a college has filled all its declared vacancies by six months after the commencement of the programme, and that 30% of control group colleges had done so, a 0.68 effect size would be equivalent to a difference of 31 percentage points between intervention and control groups. This is another way of saying that to have a good chance of a trial with a sample size of 54 General FE colleges declaring a result that reaches statistical significance, the underlying effectiveness of TTF would have to be equivalent to a result in which 61%[28] of intervention group (16 colleges of the 27 in the intervention group) colleges have filled all their vacancies at six months relative to 30% of control colleges having done so (8 colleges of the 27 in the control group). It is important to bear in mind that these figures, although not unreasonable estimates, are primarily illustrative.

Looking at the third row of the upper panel, here a pooled sample size over Rounds 3 and 4 of 93 General FE colleges is assumed. Based on participation rates seen at Rounds 1 and 2, we would judge this as the best-case scenario in terms of recruitment of General FE colleges to the programme. With a sample of this size the associated effect size is 0.53. For social and educational intervention this is considered a medium effect size. In percentage point terms, where a rate of 30% is observed in the control group, it is equivalent to an improvement of 24 percentage points at the 95% level of statistical significance, which is sizeable. (Please refer to the Glossary for a definition of statistical significance and the footnote to page 44). Put differently, a future round(s) of TTF that attracted 93 colleges to participate, would need to increase the rate at which colleges filled their declared vacancies by 24 percentage points over the control group. Thus, the rate of filled vacancies would have to be 54% among those colleges participating compared to 30% among those that have not participated.

The broader point to take from this Table is that if the true impact of TTF Strand 1 at Rounds 3 and 4 is lower than the effect sizes and percentage point differences we have discussed, then the chance of a result from any trial, with the associated sample sizes, reaching statistical significance falls. This is not to say that a result which reaches statistical significance will not be found, just that the chances of an inconclusive finding is greater. Therefore policy makers and analysts have to decide whether it is reasonable to expect the ‘medium’ and ‘large’ effect sizes discussed, and if it seems that these are the

[28] If the percentage is 30% in the control group and the percentage point difference between control and intervention equal to 31 percentage points then the percentage in the intervention is 61%.

upper end of what might be expected then accept that inconclusive findings are more likely than usually accepted.

The Table shows that a design based on analysis at the level of declared vacancies is more sensitive – the associated effects sizes are smaller. But as mentioned previously, this assumes that it is possible to identify declared vacancies at the point of randomisation and track them over time. We also assume that providers will not pull out of the programme once they complete their application and are randomised. This might not be an unreasonable assumption as the number of dropouts at Round 1 and Round 2 has been low. Any monitoring system put in place to receive updates as to progress in filling vacancies would also need to capture other data at the end of the trial relating to the academic year subsequent to the commencement of funding (such as the indicators discussed previously).

To conclude, it has been shown that implementing a randomised impact evaluation of TTF is practically achievable. The decision on whether to proceed with such a design, however, is not a straightforward one. The department needs to consider the magnitude of effects that TTF is likely to yield. The department also needs to take a view as to how far it can accept the risk that TTF does not lead to effects consistent with these expectations and the associated risk of inconclusive findings that may emerge from the trial as a result of this.

Comparison group designs at Level 3 on the Maryland Scale

In this section the potential for an impact evaluation of future Rounds of TTF consistent with the requirements of Level 3 on the Maryland Scale is examined. The benefits and challenges of this approach are highlighted through reference to a running example, which, based on the previous discussion, is most plausible given the conditions that have prevailed at Rounds 1 and 2 of TTF. This running example seeks to examine the impact of TTF on General FE colleges, rather than on teachers.

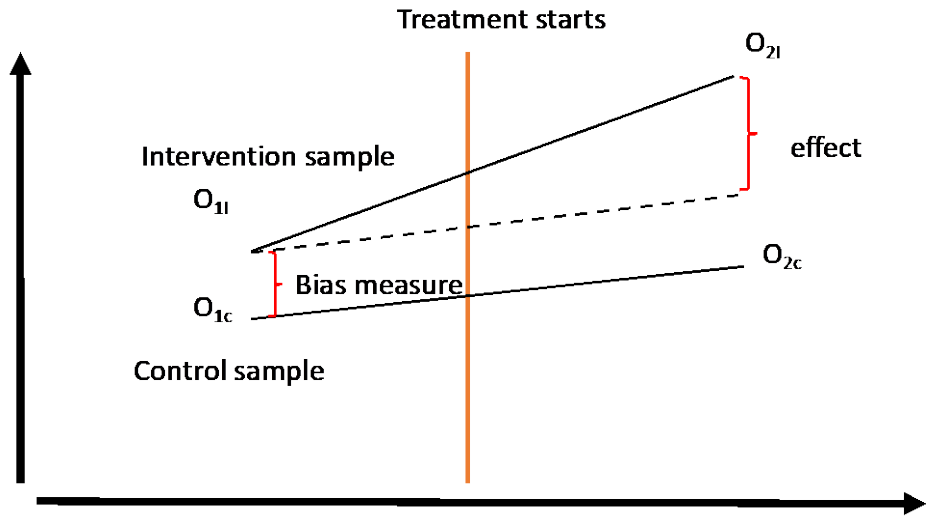
With Level 3 comparison group designs, counterfactual outcomes, for those exposed to the programme are estimated from a comparison group of similar providers. The need to select such a comparison group requires sources of data that possess a wide range of relevant control variables to statistically adjust estimates of impact.

Level 3 designs often involve the use of either or both matching and difference-in-differences. The matching approach typically uses a metric such as the propensity score to ensure that average outcomes for participating providers are compared to average outcomes among non-participating or comparison providers that are similar to the providers (those with similar propensity scores). Difference-in-differences attempts to achieve a fair comparison in average outcomes across participating and comparison group providers, by statistically adjusting the comparison for differences in outcomes between the two groups pre-intervention. Both rely on important assumptions. The

propensity score approach relies on the assumption of conditional independence. Put very briefly, this assumes that the correct propensity scores are calculated, which in turn relies on assumption that the propensity scores have been modelled correctly using all the necessary variables. Difference-in-differences relies on the common trends assumption. Again, briefly, this means that the change over time observed in the comparison group is that which would have been observed in the intervention group in the absence of TTF. A complete account of these assumptions is beyond the scope of this report but nonetheless they are important and many sources provide accessible accounts of them (for example, Angrist & Pischke, 2008; Morgan & Winship, 2015).

Box 1: What is difference in differences?

Differences-in-differences is a form of Level 3 analysis which is applied to evaluations where outcome measures have been collected for a sample, in this case, of intervention colleges that are exposed to TTF and colleges that act as a control group and that remain unexposed, both before and after intervention commences. The Figure below seeks to illustrate the approach further:



This chart plots on the x-axis time and the y-axis the value for a particular outcome. The vertical orange line which divides time into pre-intervention period to the left and post-intervention periods to the right. Average outcomes among intervention colleges start at O_{1I} pre-intervention and rise to O_{2I} post-intervention. Likewise, among control colleges they rise from O_{1C} to O_{2C}. Difference-in-differences works by first subtracting the post-intervention average outcomes among control colleges from that for intervention colleges O_{2I} - O_{2C}. The approach allows for the fact that the intervention and control group may have different average outcomes to start with captured by the pre-intervention difference in average outcomes between the two groups, or O_{1I} minus O_{1C}. Therefore the pre-intervention difference O_{1I} minus O_{1C} is subtracted from the post-intervention difference O_{2I} - O_{2C} to get the effect. Hence difference-in-differences

A difference-in-differences (DiD) approach

For the purposes of considering the viability of a comparison group design in the case of TTF the running example that is referred to will assume that a difference-in-differences (DiD) approach is chosen and that the level or unit of analysis will be the provider. In practical terms this would involve collecting pre-intervention data from all eligible General FE colleges (eligible General FE colleges can be identified through consulting the register data discussed above). An assessment of existing data sources has found that they either suffer from coverage problems or a lack of relevant data items.[29] This is notwithstanding the fact that future developments around data collection will require mandatory completion of data equivalent to those sources considered here. Therefore, the collection of pre-intervention data would involve a limited primary data collection exercise (with the exception of data from the college accounts).

What data items would such an exercise need to collect? The answer to this question is the outcome variables discussed previously (see Outcome Indicators section on page 43) and a limited range of further classificatory variables – college size, number of sites, etc. If TTF commences operation in a given academic year, the data collection exercise would relate to the previous full academic year. The pre-intervention data might most effectively be collected via telephone interview. For those colleges that **do participate**, a short telephone interview[30] with an appropriate respondent during the application process would be required (completion of this telephone interview could be a mandatory component in the application process). Immediately after the period of time for applications ends, telephone interviews would be conducted with providers that have not applied (though providers in this group may have participated in previous Rounds of TTF and a record of previous participation will need to be created for participant and non-participant providers alike). To supplement the pre-intervention data, we suggest matching extracts from the college accounts for the same academic year to telephone interview records, subject to the necessary legal gateways being in place. The coverage of college accounts data is good, and it appears to provide additional important classificatory measures that can supplement the new primary data.

Pre-intervention data will attempt to measure outcomes in the academic year immediately prior to the commencement of a given Round of TTF. There is also great merit in attempting to collect data on outcome measures not just relating to the year immediately preceding the commencement of the Round of TTF under consideration but relating to a number of preceding years. This is so that the situation among the eligible

[29] For example the data sources: SIR, ILR, College Staff Survey, College Accounts and College Workforce Survey.
[30] An alternative would be to collect the required data items in the scheme application form as suggested previously. However, because data would be collected from non-participating providers through a telephone interview important mode effects could emerge as an important source of bias.

population of General FE colleges can be assessed over time more fully (for example, trends in unfilled vacancies across different subject areas can be examined).

It is assumed that TTF will commence at the start of a given academic year and that outcomes will also therefore need to be collected for this year, and possibly the following academic years, in order to capture post-intervention/exposure outcomes. Through extending data collection to cover multiple years both pre and post the Round of TTF under consideration, the data at the disposal of the impact evaluation become more fully longitudinal in nature yielding several advantages. The disadvantages are that such data become more complex to collect, more burdensome on respondents, and potentially more prone to non-response. Some of these limitations would not be relevant where administrative data that met the requirements of this evaluation design be available. To simplify the discussion hereon, we focus instead on a design incorporating single pre and single-post intervention data collection exercises, whilst noting the benefits of more extensive longitudinal data.

Difference-in-differences - sample sizes

Essentially the sample size challenges noted in our discussion of an RCT design apply also in the case of a DiD approach. Unlike the RCT design, however, the entire sample of General FE colleges applying to TTF forms the intervention group. Moreover, the entire non-participating sample of General FE colleges can be considered potential or actual comparisons group members. As was the case with the RCT, it needs to be ensured that the data collected capture participation in previous Rounds of TTF. So, if it is assumed that this evaluation considers the impact of a future Round 3 of TTF, data will need to capture the extent to which colleges across the population (whether they choose to participate in Round 3 or not) participated in Rounds 1 and 2.

It is assumed that a future Round 3 of TTF would attract approximately 35 valid applications from General FE colleges (based on estimates of take-up at Rounds 1 and 2). It is also assumed that the total population of eligible colleges would stand at around 190. Thus, the potential comparison group would come from approximately 150 colleges not participating in Round 3. The design relies on the collection of primary data directly from participating and non-participating colleges. Given the response rates seen in both the College Workforce Survey and the College Staff Survey, it seems reasonable to assume a response rate of around 50% for any initial data collection exercise among non-participating colleges, and that roughly 75% of these respondents would respond to a follow-up post-intervention survey. For participating General FE colleges, we assume full response at pre-intervention and 75% response at post-intervention.

Table 7: Projected sample sizes for a difference-in-differences Level 3 design based on the collection of primary survey data via telephone

Participation/ non-participation	Pre-intervention achieved survey sample size	Post intervention achieved survey sample size	Total
Participating General FE colleges	35	26	61
Non-participating General FE colleges	75	56	131
Total observations	110	82	192

Notes:

We assume that:

- All general FE colleges applying to TTF respond to the pre-intervention survey interview
- That half non-participating General FE colleges respond at the pre-intervention interview stage

At post-intervention, that 75% of participating General FE colleges respond and that 75% of non-participating General FE colleges that have already provided data at the pre-intervention stages provide post-intervention data.

If an attempt to estimate the effect of TTF on unfilled vacancies in, for example, construction, at the end of the appropriate academic year is considered[31]; under certain assumptions[32], it is estimated that an effect size of 0.46 (which is medium size) would be forthcoming. This assumes that there is no residual bias (please refer to the Glossary for a definition of bias) in the estimate once we account for pre-intervention differences in outcomes, which will be open to challenge. This effect size is smaller than those discussed above in the case of an RCT because the sample sizes are larger, but the Level 3 DiD design is more likely to suffer from bias than the Level 5 RCT. Again, if the true impact of TTF is lower than effect size 0.46 the chances of findings being inconclusive rise. This means that TTF must produce an impact equivalent to an effect size of this magnitude to us to statistically detect an effect in our sample at the 95% level of statistical significance. Put more simply, TTF will have to produce quite a big effect to avoid results that are inconclusive.

[31] We set aside the likelihood that some General FE colleges in the sample will not offer courses in construction

[32] We assume there is no residual bias once covariates are included within the analysis; 80% power, 95% statistical significance, a final analytical sample of 164 colleges and that pre-post correlation in the outcome measure of 20%.

Implications for evaluation

The discussion in this section highlights the following key issues:

- Generally, TTF Strand 1 is an intervention that provides a modest increase in resources available to providers even though such resources are likely to be more significant for teacher trainees.
- Counterfactual impact evaluation designs at Level 5 and 3 can practically be implemented and the necessary data obtained for an evaluation of Rounds 3 onward of TTF (depending though on precisely when Round 3 commences and bearing in mind that the designs discussed above required pooling data from a future Round 3 and 4), though may require new primary data collection due to deficiencies in the coverage of existing data sources particularly for a Level 3 DiD design. Clearly the requirement for primary data collection will increase the costs of any evaluation.
- An RCT could provide estimates of effects at both provider and vacancy levels (effects on teacher trainees that fill posts would have to be conducted non-experimentally). If the true impact of TTF is small or modest, results at the provider level are likely to be inconclusive.
- A comparison group design at Level 3 of the Maryland Scale is also considered where estimates are obtained from a difference-in-differences analysis, and a limited range of data are collected pre- and post-intervention, with the resulting sample linked to records held on the college accounts. On this basis, it is reasonable to expect a sample of some 160 observations to be obtained (what is meant here is that we could expect to obtain the necessary data pre- and post-intervention from around 160 General FE colleges). However, given such a sample, the risks of inconclusive findings are still quite high if the impacts of TTF are modest, and the risk of bias is higher with a Level 3 design relative to a Level 5 design.
- Generally, given the assumptions we have made based on the operation of TTF at Rounds 1 and 2, there is a high chance that were the department to commission an impact evaluation, based on a counterfactual approach, at Level 3 or higher, the results that emerge from such as study are likely to be inconclusive. This is unless the rate at which providers participate can be raised and greater number of providers other than General FE colleges can be attracted to take part, thereby raising the numbers participating and the sample sizes available to any evaluation. If this were achieved then the number of participating General FE colleges would be greater, as would the number of other providers. Following on from this sample sizes available for analysis would be larger. This would mean that smaller differences in outcomes between intervention and control groups could be detected.

Section 7: Summary, conclusions and recommendations

The FE sector has experienced considerable difficulty in recruiting and retaining suitably qualified teaching staff, particularly those with prior commercial or industrial experience. The Taking Teaching Further (TTF) programme aims to address this problem as part of a suite of policy measures. The programme has two Strands overseen by the Education Training Foundation (ETF) on behalf of the Department for Education (DfE). This report discusses the feasibility of conducting a rigorous and credible impact evaluation of future rounds of Strand 1 of Taking Teaching Further. Strand 1 TTF ran in the form of a pilot at two previous rounds, referred to as Rounds 1 and 2.

The findings of this feasibility study draw on desk research that aimed to shed light on the nature of TTF in its Strand 1 form, and the extent of take-up of Strand 1 among providers and teacher trainees at Rounds 1 and 2.

A 'counterfactual impact evaluation' approach is adopted as a framework through which the impact of TTF might be evaluated; a framework consistent with the government's guidance on impact evaluation set out in the Magenta Book. This report looks at the types of approaches/designs, based on counterfactuals, that might be adopted. Specifically, the report examines the prospects for either a Level 5 impact evaluation, in the form of a randomised controlled trial, or a Level 3 design based on a non-random comparison group design.

Conclusions

The choice of either a Level 3 or 5 impact evaluation design on the Maryland Scale has implications for the types of data that are required. A Level 3 design with a non-random control group is particularly reliant on obtaining rich data, namely:

- individual unit level data – depending on the chosen unit of analysis - this will be data comprising individual cases at, most realistically the General FE college level;
- data that record individual unit level outcomes for both exposed and unexposed cases;
- data that indicate which cases are exposed and which cases, at the point in time outcomes were measured, remain unexposed; and
- data that also contain 'control' variables – these are variables that are unaffected by TTF (usually measured pre-exposure) that capture important differences between exposed and unexposed units but that are also correlated with outcomes.

Any future impact evaluation of Strand 1 of TTF will need to consider the existence of the first two rounds of TTF Strand 1 funding and their consequences. To date, participation in the programme has been dominated by General FE colleges. This leaves open the question as to how far we can expect other types of providers to engage in future Rounds of TTF. Future engagement is very difficult to predict as it depends on a very wide range of factors. For example, whether the nature of the scheme and its resourcing are changed, the alternative sources of funding available to providers and wider economic circumstances with their consequences for the jobs market at a time of some considerable uncertainty. Moreover, if we anticipate that the programme will continue to attract overwhelming interest from General FE colleges, an impact evaluation will probably be restricted to focusing on the effects of TTF on colleges only, with the attendant risk of inconclusive findings (findings that fail to reach statistical significance).

This report suggests the most relevant statistical quantity to estimate is the average causal effect on outcomes among General FE colleges that participate in TTF. This choice is partly driven by practical considerations. One of these considerations is that data are not collected that permit teachers that receive support and funding through TTF to be identified. This rules out the possibility of looking at the effects of TTF on teachers and by extension learners. However, this limitation could be addressed in future. Thus, the primary unit of analysis will be the General FE college. The report suggests a range of indicators that, if they could be observed, would enable outcomes of interest to be calculated for General FE colleges exposed to TTF.

There are a number of data sources that appear to contain data in rich enough form to be used for a Level 3-type impact evaluation if they could be linked to TTF participation records. Unfortunately, each of these sources - the SIR, the College Staff Survey and the College Workforce Survey do not have full coverage of our population of interest. The TTF programme participation records are a vital source that enables identification of providers who are exposed and participate in TTF. However, given GDPR and other legal constraints, they do not currently indicate the identities of teachers funded through the programme. The college accounts appear to have good coverage but do not contain variables that might act as outcome measures. For these reasons it appears likely that any impact evaluation of TTF would need to rely on primary data collection.

This report shows that impact evaluation designs at Levels 3 and 5 of the Maryland Scale are practically obtainable given the collection of primary data and that enough time is available before Round 3 commences for the necessary data collection processes to be put in place. However, the effects of TTF Strand 1 are likely to be modest and therefore difficult to identify statistically, particularly given the likely size of the samples available to the evaluation. An RCT is practically viable and could be designed to consider outcomes by provider and/or by declared vacancy. However, due to the likely modest scale of any impact and samples that are relatively small, results are likely to be, at least in statistical terms, inconclusive. This does not mean results will be inconclusive, just that the risk that

they will be so is higher than would be generally acceptable at the commencement of most evaluations. Such results could not be interpreted as revealing that TTF did not work, but instead that the data were not consistent with a strength of effect that reached statistical significance. A Level 3 design is discussed as an alternative. Broadly, the challenges face a Level 3 design are similar and relate to the limited samples available leading to a high chance of inconclusive findings if TTF produces modest impacts.

Recommendations

This report anticipates that the department would not accept the greater risk of inconclusive findings likely if either a Level 3 or 5 impact evaluation was attempted, given levels of participation in TTF similar to that seen at Rounds 1 and 2. Therefore, this report recommends proceeding with a counterfactual impact evaluation only if substantial numbers of independent, non-FE college providers could be attracted to the programme. Very roughly, around 350 providers would need to take part in total in TTF, over two future Rounds, before the risk of inconclusive findings would stand at levels generally accepted at the planning stage of most evaluations. This number of participants exceeds the total number of General FE colleges (or consortia of colleges) and therefore implies attracting substantial interest among other providers.

If participation of this order of magnitude is not felt feasible, then the recommendation would be to consider other forms of evaluation that attempt to shed light on impact. These approaches are non-statistical, but instead aim to provide a convincing narrative around the effects of interventions derived from evidence that comes from mixed-methods evaluation designs. Such methods include realist evaluation (Pawson & Tilley, 1997), theories of change (Funnell & Rogers, 2011), contribution analysis (Mayne, 2012) and possibly sophisticated case study approaches such as qualitative comparative analysis (Schneider & Wagemann, 2013). The Department may wish to consider consulting experts in these approaches to assess their potential usefulness.

Bibliography

Angrist, J. D., & Pischke, J.-S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Bloom, H. S. (2006). The Core Analytics of Randomized Experiments for Social Research, (August).

Department for Business Innovation and Skills, & Department for Education. (2016). Post-16 Skills Plan. London: HMSO. Retrieved from https://www.gov.uk/government/publications/post-16-skills-plan-and-independent-report-on-technical-education

Dong, N., Kelcey, B., Maynard, R., & Spybrook, J. (2015). PowerUp! Tool for power analysis. Retrieved from www.causalevaluation.org.

Dunning, T. (2012). Natural experiments in the social sciences: A design-based approach. (C. Elman, J. Gerring, & J. Mahoney, Eds.), Strategies for Social Inquiry. Cambridge, UK: Cambridge University Press.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. Journal of Pediatric Psychology, 34(9), 917–928.

Funnell, S. C., & Rogers, P. J. (2011). Purposeful program theory: Effective use of theories of change and logic models. San Francisco, CA: Jossey-Bass.

Gerber, Alan, S., & Green, Donald, P. (2012). Field experiments: Design, analysis, and interpretation. New York, NY: W. W. Norton & Company.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). Impact evaluation in practice. The World Bank.

Glennerster, R., & Takavarasha, K. (2013). Running randomized evaluations: A practical guide. Princeton, NJ: Princeton University Press.

Greatbatch, D., & Tate, S. (2018). Teaching, leadership and governance in Further Education: Research report. London: Department for Education.

HM Government. (2017). Industrial Strategy Building a Britain fit for the future. London: HMSO. Retrieved from https://www.gov.uk/government/publications/industrial-strategy-building-a-britain-fit-for-the-future

HM Treasury. (2011). The Magenta Book Guidance for evaluation. London: HM Treasury.

HM Treasury. (2015). Fixing the foundations: Creating a more prosperous nation. London: HMSO. Retrieved from https://www.gov.uk/government/publications/fixing-the-foundations-creating-a-more-prosperous-nation

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). Multilevel Analysis: Techniques and Applications (3rd ed.). NewYork: Routledge.

IFF Research. (2019). TTF Evaluation: Interim findings Research report. London, Department for Education.

Mayne, J. (2012). Contribution analysis: Coming of age? Evaluation, 18(3), 270–280. https://doi.org/10.1177/1356389012451663

Morgan, S. L., & Winship, C. (2015). Counterfactuals and causal inference: Methods and principles for social research (2nd ed.). New York, NY: Cambridge University Press.

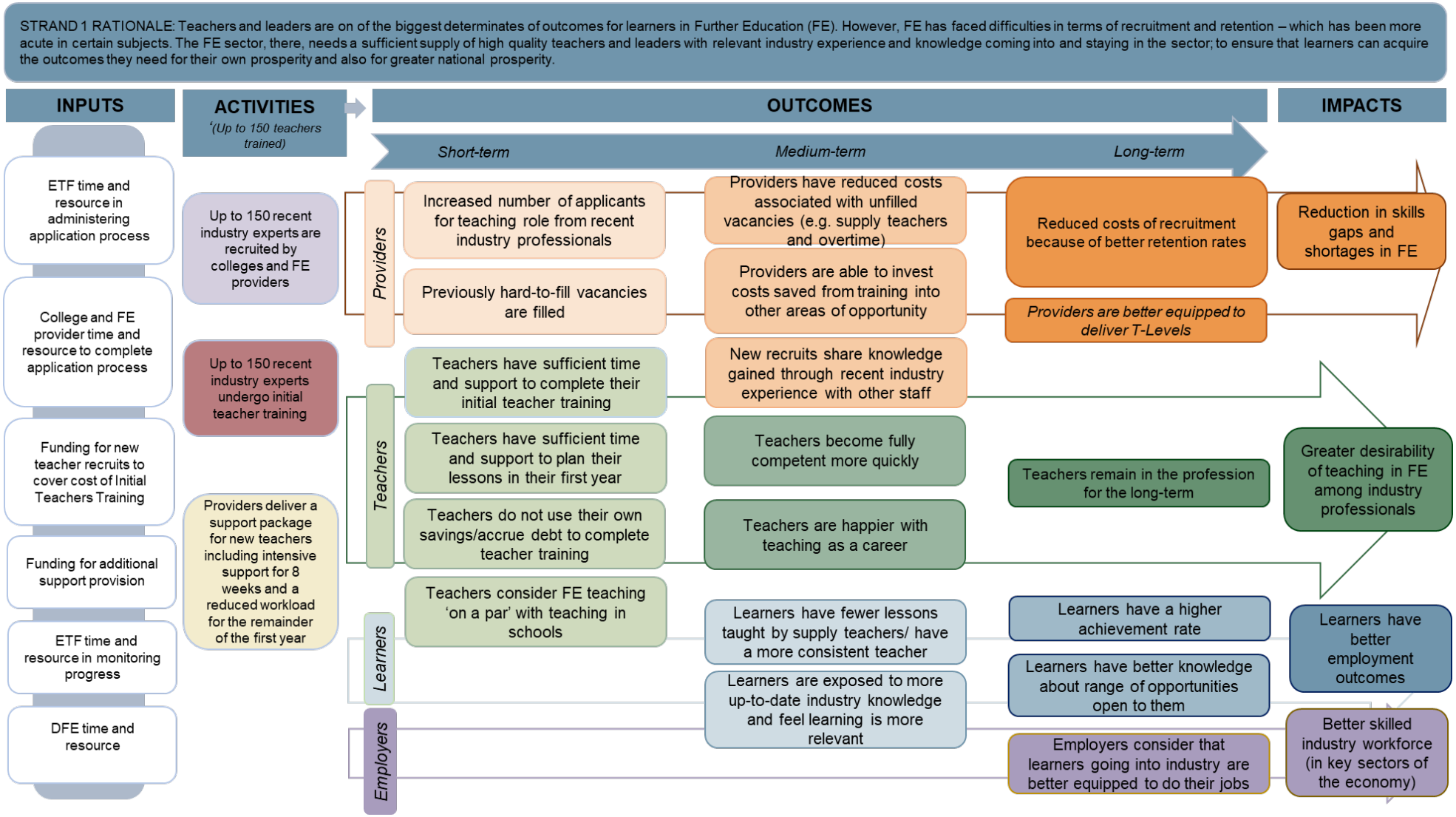
Pawson, R., & Tilley, N. (1997). Realistic evaluation. London: Sage Publications.

Rosenbaum, P. R. (2017). Observation and experiment: an introduction to causal inference. Cambridge, Massachusetts: Harvard University Press.

Schneider, C. Q., & Wagemann, C. (2013). Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis. Cambridge, UK: Cambridge University Press.

Sherman, L. W., MacKenzie, D. L., Gottfredson, D. C., Eck, J., Reuter, P., & Bushway, S. D. (1997). Preventing crime: What works, what doesn’t, what’s promising: A report to the United States Congress. US Dept. of Justice, Office of Justice Programs.

Annex A – Theory of Change



Department for Education

Reference: DFE- RR1249

ISBN: 978-1-83870-382-0

For any enquiries regarding this publication, contact us at www.education.gov.uk/contactus

This document is available for download at www.gov.uk/government/publications