Designing Evaluation of Modern Apprenticeships in Scotland

Matej Bajgar and Chiara Criscuolo

Abbreviations and Acronyms

ABS	Annual Business Survey
APS	Annual Population Survey
ASHE	Annual Survey of Hours and Earnings
CEM	Coarsened exact matching
CSS	Customer Support System
CTS	Corporate Training System
HESA	Higher Education Statistics Agency
HRMC	Her Majesty's Revenue and Customs
IDBR	Inter-Departmental Business Register
IV	Instrumental variables
LFS	Labour Force Survey
LP	Labour productivity
MA	Modern Apprenticeship
MFP	Multi-factor productivity
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary least squares
PAYE	Pay as you earn
PIAAC	Survey of Adult Skills
RCT	Randomised control trial
SDS	Skills Development Scotland

The opinions expressed and arguments employed herein are those of the authors and do not necessarily reflect the official views of the OECD or of the governments of its member countries.

M. Bajgar (✉) · C. Criscuolo
OECD Directorate for Science, Technology and Innovation, Paris, France
e-mail: matej.bajgar@oecd.org; chiara.criscuolo@oecd.org

N. Crato, P. Paruolo (eds.), *Data-Driven Policy Impact Evaluation*,
https://doi.org/10.1007/978-3-319-78461-8_18

SQA	Scottish Qualifications Authority
UK	United Kingdom
US	United States
VAT	Value-added tax
VET	Vocational education and training

1 Introduction

Apprenticeships combine paid employment and training, with the aim of developing occupational mastery and typically leading to a formal qualification. They have represented an important source of skills since medieval times, and now, in the twenty-first century, they are alive and well. Inspired by the low youth unemployment in those OECD countries where apprenticeships play the largest role—Germany, Austria and Switzerland—more countries are looking to them as a way to provide young people with the skills needed to find fulfilling and well-paid employment.

This is true for the United Kingdom, where apprenticeships are considered an important way of achieving a higher level of skills across the country and a policy worth rigorously evaluating. In England, the government has announced its intention that the number of new apprenticeships offered annually should reach 3 million by 2020 (Department for Business, Innovation and Skills 2015a) and has recently funded several microeconometric studies of the effectiveness of apprenticeship schemes (Bibby et al. 2014; Department for Business, Innovation and Skills 2015b). In Scotland, too, apprenticeships are on the rise, in the form of Modern Apprenticeships (MAs), Foundation Apprenticeships and Graduate-Level Apprenticeships. The number of new MAs available annually in Scotland has increased from 11,000 in 2008/2009 to 25,000 in recent years, and is set to reach 30,000 by 2020. As the livelihoods of more and more young people depend on skills acquired through apprenticeships, and as training subsidies and administrative costs consume an increasing proportion of public funds, Skills Development Scotland (SDS)—the public agency responsible for running MAs in Scotland—has come under mounting pressure to prove that MAs fulfil their goals: to provide young people with the right skills and to increase productivity of Scottish businesses (Audit Scotland 2014).

Up to 2015, SDS evaluated MAs through telephone surveys of employers (Skills Development Scotland 2015) and apprentices (Skills Development Scotland 2013), which asked about the impact on outcomes such as skills, career progression and productivity. Unfortunately, self-report studies are unreliable measures of impact because respondents may consciously or unconsciously adjust their answers to what they expect the evaluator hopes to hear, and simply because it is difficult for them to judge what the outcome would have been without participation. In addition, the surveys contacted apprentices within the first 6 months after leaving the training, and, as a result, they did not provide information on longer term outcomes.

In 2015, SDS decided to examine the impact of MAs through a more rigorous evaluation. SDS's aim was to develop an evaluation that would examine the causal impact of MAs in the long term, could be replicated over time and would allow the effects of MAs on different levels and in different sectors to be compared.

However, it was clear that the evaluation would not be an easy task. The first challenge was a lack of data. Apart from the telephone surveys conducted shortly after the end of training, no information was readily available on outcomes for participating individuals or firms in the longer term. Furthermore, no information had been collected on individuals who had not participated in MAs and could, therefore, serve as a control group. Finally, apprentices had not been selected randomly or through a centralised mechanism, making it hard to separate causal effects of apprenticeships from the effect of characteristics of individuals who had chosen to take up MAs.

Lack of data, absence of a control group and exogenous variation are common limitations of programmes such as apprenticeships. The aim of this chapter is to convince the reader that, despite these drawback situations, although challenging, it is still feasible and valuable to conduct an evaluation that is robust and useful. To this end, it describes the planned evaluation of MAs in Scotland. It builds on a collaboration between SDS and the Organisation for Economic Co-operation and Development (OECD), in which the OECD has prepared an 'evaluation framework' setting out recommendations for evaluating MAs (Bajgar and Criscuolo 2016a).

Designing an evaluation involves answering several central questions. On which outcomes should the evaluator focus? Which data will provide the necessary information while being reliable and accessible? How should the control group be constructed? Which estimation methods should the evaluator use to identify causal effects of the intervention?

Table 1 gives an overview of how some existing studies evaluating impacts of apprenticeships have approached these questions, and describes their findings. It focuses on studies that employ more sophisticated estimation approaches. The table reveals a large diversity of approaches applied by the studies. Many use administrative data but others rely on existing or new surveys. Analysed outcomes include wages, probability of employment, subsequent education and job characteristics. The alternative against which apprenticeships are compared is no training in some studies, but school-based vocational training, unfinished apprenticeships or general education in others. Finally, evaluation methods include the instrumental variables (IV) method, the difference in differences method and randomised control trials.

This chapter describes which evaluation choices are most suitable for evaluating MAs in Scotland, and why. It argues that an encompassing evaluation of MAs should analyse outcomes for both individuals and firms, focusing on employment, wages and several other variables for individuals, and on productivity for firms. *It should rely mainly on existing administrative data,* possibly complemented by a new survey covering information not available in administrative records. The evaluation could use individuals who started but never completed their apprenticeships (non-completers) and those who never underwent an apprenticeship experience (never-starters) as control groups, and conduct estimation over a range of time horizons and

Table 1 Selected studies evaluating impacts of apprenticeships

Authors	Data	Estimation	Findings
Adda et al. (2006)	Social security records	Compare apprenticeships with school-based vocational training; use number of apprenticeship vacancies as IV for participating in apprenticeships	Total wage return 10% after 5 years and 25% after 20 years
Fersterer et al. (2008)	Social security records	Compare completed apprenticeships with unfinished ones; use firm failures as IV for apprenticeship duration	Annual wage return 2–4%
Parey (2008)	Social security records	Compare apprenticeships with school-based vocational training; use number of apprenticeship vacancies as IV for participating in apprenticeships	Annual wage return 3% (not significant with IV); initial reduction in probability of unemployment by 15% points per year of training fades out over time
Malamud and Pop-Eleches (2010)	Census and existing survey	Compare apprenticeships with general education; use regression discontinuity design based on a large educational reform	No causal difference
Alet and Bonnal (2011)	Existing longitudinal education survey	Compare apprenticeships with school-based vocational training; use number of apprenticeship vacancies as IV for participating in apprenticeships	Increase in probability of completing high school diploma and staying in education
Reed et al. (2012)	Programme administrative data and unemployment insurance wage records	Compare apprentices with non-participants and non-achievers; control for initial earnings and use propensity score matching	Participation increases earnings by 50% and the probability of employment by 9% points compared with non-participants; completion increases earnings by 80% and the probability of employment by 15% points compared with non-completers (effects after 6 years)
Picchio and Staffolani (2013)	Administrative data on job contracts	Compare apprenticeships with other types of temporary contracts; use regression discontinuity design based on regional age cut-offs in eligibility	Increase in propensity to get a permanent contract after 2 years

(continued)

Table 1 (continued)

Authors	Data	Estimation	Findings
Bibby et al. (2014)	Linked education, benefit, employment and earning records	Compare apprentice completers to non-completers; use differences estimates with matching for other qualifications	Total wage returns of 11% for Level 2 apprenticeships and of 16% for Level 3 apprenticeships; there is an initial increase in the probability of employment of 3% points but this declines over time
Schaeffer et al. (2014)	Dedicated survey	Compare apprenticeships with attending standard public schools; randomised control trial	No effect on wages; increase in the probability of employment of 26% points; increase in enrolment in General Equivalency Diploma of 24% points; no effect on high school graduation rate
Noelke and Horn (2014)	Labour force survey	Compare apprenticeships with school-based vocational training; use difference-in-differences estimator based on county-level shifts from work-based to school-based vocational education	A 10% increase in the ratio of school- to employer-provided places corresponds to an initial increase in unemployment of 3% points, but this effect declines over time; no effect on working in non-routine occupations
Kugler et al. (2015)	Dedicated survey, and education and social security records	Compare apprentices with non-participants from among preselected candidates; randomised control trial	Total wage return of 6% if formally employed; increase in probability of formal employment of 5% points and of days in formal employment of 13%; increase in probability of completing secondary school of 1.4% points, in the probability of enrolling in college of 3.5% points and in the probability of staying in college after 5 years of 1.6% points

Note: The table is partly based on Table A.1 in Bajgar and Criscuolo (2016b)

using various econometric methodologies. The chapter illustrates how the evaluator can use narrower control groups, matching techniques, regression and, if possible, changes over time to better separate causal effects from mere correlations.

The chapter focuses on apprenticeships, but it also hopes to be informative for evaluations of other types of programmes, and, in particular, of vocational education and training (VET), active labour market policies and education.

Two things should be clarified at this point. Firstly, the chapter does not describe an evaluation that has already been undertaken but an ex ante framework strategy for an evaluation that is just starting. Secondly, this chapter focuses on counterfactual impact evaluation. Other components of programme evaluation, such as input and output monitoring, process evaluation and cost-benefit analysis, are discussed in the OECD evaluation framework for MAs (Bajgar and Criscuolo 2016a), but they are not within the scope of this chapter.

To prepare ground for the subsequent discussion, the following section explains the rationale for public intervention in apprenticeships and provides a brief background on MAs. Section 2 then sets out the scope of the impact evaluation by describing the choice of the units of analysis, the outcomes and the time horizon over which the impact is assessed. Section 3 discusses the choice of data. The next two sections provide guidance on constructing control groups and establishing causal effects for individuals (Sect. 4) and firms (Sect. 5). Section 6 gives illustrative examples of several past apprenticeship evaluations and Sect. 7 concludes.

2 Context

In Scotland, the vast majority of apprenticeships take place within the umbrella of MAs, which are co-ordinated by the government agency SDS. MAs are intended to provide employment opportunities, particularly for young people, and to support economic growth. The aim of MAs is to give young people the skills and the confidence that they need to find a good job and progress in their career. At the same time, they aim to help employers achieve higher productivity and boost staff retention and morale. MAs combine paid work with training that leads to a recognised qualification and are offered at several levels, ranging from Foundation Apprenticeships for secondary-school pupils to Graduate-Level Apprenticeships that are equivalent to a master’s degree.

A feature of MAs that is important for their evaluation is that SDS does not run them as a centralised programme but rather provides standards, oversight and financial contributions to many privately run programmes under the MA umbrella. A key role, therefore, is played by training providers (e.g. private training providers and colleges), which deliver learning and assessment. SDS allocates apprenticeship places to the training providers, and provides a contribution to the training fees, although in some cases employers are also asked to contribute to the training costs.

3 Scope of Evaluation

This section demarcates the scope of the evaluation in terms of units of analysis, examined outcomes and time horizon.

3.1 Units of Observation

The first question to ask when designing an impact evaluation is ‘impact on whom’. The answer to this question for MAs is that it should examine the impact on individuals and on firms.

Each type of unit has distinct advantages for measuring the impact of apprenticeships on economic growth. The key advantage of using individuals as the unit is that, for a young person, participating in an apprenticeship potentially represents one of the most important determinants of his or her labour market performance. This will imply a strong ‘signal-to-noise ratio’ which in turn allows for a more precise estimation of MAs’ impact on apprentices’ outcomes. The key advantage of using firms is that it is easier to measure productivity in a comparable way than it is when using individuals. However, apprentices represent only a small percentage of the workforce for most employers; thus, the effect of the apprenticeships on firm productivity may be difficult to discern even if the effect on any single apprentice is large.

3.2 Outcomes

The second key question in establishing the scope of the evaluation is ‘impact on what’: That is, which outcomes should it examine? As illustrated in Table 1, the answers to this question given by existing studies are most commonly wages and employment probability, but in some cases also job characteristics and subsequent education.

As the purpose of MAs is to develop young people’s skills and productivity, these two outcomes appear to be the most obvious options when evaluating the impact of MAs on individuals. Unfortunately, measuring either of them in a comparable way is challenging.

Skills are not readily observed in existing data and would need to be newly measured in a costly and methodologically demanding survey. Furthermore, the applied nature of apprenticeship training means that a large part of the acquired skills is industry or occupation specific. Therefore, measures of such skills cannot easily be compared across industries or occupations. This is an important limitation, because comparing performance of different MA frameworks is one of the main aims of the evaluation as seen by SDS.

Direct measures of workers’ individual productivity are similarly occupation specific and thus hard to compare across occupations, and even across tasks within a given occupation. In addition, records of such measures may exist within individual employers but are unlikely to exist for a representative set of employers.

For these reasons, evaluation of MAs should instead focus on several other individual-level outcomes:

- Employment
- Wages
- Unemployment
- Standard/non-standard employment[1]
- Career progression
- Subsequent education
- Subjective well-being

In the case of impact on employers, productivity plays a central role. MAs aim to support economic growth, and increasing firm productivity is the main way in which they can do so. Unlike individual-level productivity, firm productivity can be measured in a comparable way, most commonly as one of the following two measures:

- **Labour productivity (LP)** can be defined as output per hour worked and can be calculated as the ratio of sales or value added over the number of employees, measured as head counts or preferably, if the information is available, as full-time equivalents.
- **Multi-factor productivity (MFP)** captures the output a firm can produce with a given amount of inputs (e.g. labour, capital, total intermediate inputs, energy). It differs from LP in that it also accounts for the amount of physical capital and intermediates used by each firm. The disadvantage of MFP is that it requires more information on inputs, such as physical capital and intermediates.

3.3 Time Horizon

The third question for the evaluation is ‘impact, but when’.

This question is important because existing evaluations suggest that estimated effects vary substantially depending on the time between the end of the apprenticeships and the moment when the impact is assessed. The effect on employment is often large immediately after training but gradually declines over the next few years.[2] In contrast, some studies suggest that the effect on wages grows substantially over time.[3] Mechanisms for the effects may differ over time, too. While short-term employment effects are probably due to apprentices staying with the same employer shortly after the apprenticeship ends, the later wage increases may instead reflect acquired skills.

For individuals, it seems optimal to examine the outcomes over multiple time horizons. Such an approach will provide richer information on the impact and the mechanism behind each outcome. This approach, however, requires the use of large

[1] Standard employment here stands for full-time employment with an open-ended contract.

[2] See Bonnal et al. (2002), Parey (2008), Noelke and Horn (2014) and Bibby et al. (2014).

[3] See Cooke (2003) and Adda et al. (2006).

administrative data sources that can be accessed at different points in time and still contain information for a large number of apprentices. With survey data, a particular time horizon would need to be chosen; 2–4 years after the end of the MAs may provide a sufficient time for the effects to materialise while still allowing a reasonably precise identification of the effects.

In the case of impact on firms, the question of time horizon concerns the time between the training of the apprentices employed by a firm and the time when the outcomes are measured. While the analysis on firms, too, would ideally allow for a different effect of apprentices depending on the time since training, in reality the number of apprentices employed by most firms is small and identifying such a differential effect may not be possible.

4 Data

An essential part of any evaluation is the choice of data. In the case of individuals, the evaluation requires information on whether or not he or she took part in MAs, on the examined outcomes and on individual or firm characteristics. The first three parts of this section discuss three main ways of obtaining all this information: using a self-standing existing survey, linking multiple data (from administrative sources or from surveys) and collecting new information via a new survey specifically designed and conducted for evaluating MAs. The fourth part discusses the choice of data for evaluating the effect of apprenticeships on employers.

4.1 Data on Individuals: Existing Self-Standing Surveys

The simplest option is to rely on a single existing survey. In the Scottish context, surveys which could be considered include the Labour Force Survey (LFS) and the Annual Population Survey (APS).

The LFS gathers quarterly information on demographics, education and labour market outcomes of 100,000 individuals living in the United Kingdom. It includes a variable describing if an individual has completed an apprenticeship, but it does not specifically refer to MAs. The APS is obtained by merging waves 1 and 5 of the LFS quarters with annual boosts for England, Scotland and Wales. It samples about 340,000 individuals each year, although it contains fewer variables than the standard LFS. It also has a panel element: the households at selected addresses are interviewed annually over four waves and then leave the survey and are replaced by other households.

Using a single existing survey is a cheap and fast option, because it taps into information that already exists and is in one place. However, the number of Modern Apprentices in the surveys is rather small, the surveys cannot be linked to SDS individual records and they do not identify individuals who started but did not complete their apprenticeships. These features limit the potential precision of estimates, complicate disaggregated analysis and limit options for constructing control groups.

For these reasons, existing self-standing surveys are not a sound basis for a robust evaluation of the impact of MAs on individuals.

4.2 Data on Individuals: Linked Administrative and Survey Data

An alternative approach, one which is taken by most studies described in Table 1, is to combine information from several sources. In the case of MAs, the main potential sources of information include programme records held by SDS, Her Majesty's Revenue and Customs (HMRC) employment and earnings records, Department for Work and Pensions benefit histories, Annual Survey of Hours and Earnings (ASHE), educational records and the business register.[4]

Programme records held by SDS include the Corporate Training System (CTS) and the Customer Support System (CSS). CTS contains complete information about each apprentice's training. CSS contains information about education, training and labour market participation of 16- to 19-year-old cohorts, including young people who did not enter an apprenticeship and could thus serve as a control group.

The relevant administrative records held by HMRC include the P45 employment data, which include information on employment start and end dates, and on whether or not individuals are claiming benefits, and the P14 earnings data, which include information on employment start and end dates, earnings and hours worked. This information can be complemented with the records held by the Department for Work and Pensions, which contain detailed information on benefits claimed by each individual.

ASHE is a survey filled in by employers, and provides job and earnings information on approximately 140,000–185,000 individuals each year. It covers only a fraction of the population, but, in addition to employment and wage figures, it specifies if employees have a permanent contract and if they have a managerial role; as a result, it allows additional outcomes, such as career progression and employment stability, which are not covered by the HMRC data, to be examined.

Administrative records on education—the Pupil Census, Scottish Qualifications Authority (SQA) Attainment Data and the Higher Education Statistics Agency (HESA) Student Record—provide information on individuals' schooling and their qualifications other than MAs. The Pupil Census contains mostly demographic

[4] A similar approach, linking further education learner information with data from HMRC and the Department for Work and Pensions, has been used for evaluating outcomes of further education in England (Bibby et al. 2014).

information on all pupils in publicly funded schools in Scotland. The SQA Attainment Data contain information on entry and attainment of SQA qualifications. The HESA Student Record is an administrative dataset of all students in higher education in Scotland.

Finally, the Inter-Departmental Business Register (IDBR) can provide complementary information on employers with which apprentices undertook their training, and on their current employers. It is a comprehensive list of UK businesses based mainly on value-added tax (VAT) and pay-as-you-earn (PAYE) systems, both administered by HMRC, but also on information from Companies House, Dun & Bradstreet and business surveys. The information covered includes industry, employment, turnover and country of ownership.

Linking the datasets requires *matching records* pertaining to each person. This is preferably done based on a unique individual identifier. A more laborious and less precise method is to rely on personal information such as name, address and date of birth, and should be used if only the first option is not feasible. Since the SDS records on Modern Apprentices (CTS) include the National Insurance number, they can be linked to HMRC and ASHE data. The records on other SDS programmes (CSS) can, in turn, be linked to the educational data using the Scottish Candidate Number, which is allocated to school pupils sitting SQA exams. A link between CTS and CSS records can then connect the data linked with the National Insurance number with those linked with the Scottish Candidate Number. Finally, employer information can be matched to the individual-level data using the IDBR number and the PAYE number.

The data linking approach has some important advantages. In particular, the very large sample size of the resulting dataset allows obtaining precise estimates, decomposing results by MA types or individual characteristics, defining narrower control groups better matching people who undertook MAs, and examining outcomes at different horizons. The approach also provides information that does not appear in existing surveys, for example type and exact dates of MA or information on career progression. Moreover, once the data are set up, the marginal costs of adding new observations and re-running the estimation are relatively low.

Setting up the linked dataset requires a significant initial sunk investment in time and effort to access, link and clean the data. However, once the data infrastructure is in place, it will provide a powerful tool for evaluating not only MAs but also other public interventions in Scotland. It should, therefore, be used as the main information basis for the impact evaluation of MAs.

4.3 Data on Individuals: Dedicated Survey

The third option is to conduct a new survey specifically for the purpose of evaluating the impact of MAs. This is the choice made by studies analysing randomised control trials, such as Schaeffer et al. (2014) and Kugler et al. (2015), although the latter study combines newly collected survey data with administrative datasets to look at longer term outcomes.

The advantage of this option is that it *gives the greatest freedom in collecting the exact information* that will be useful for the evaluation. On the other hand, it is the most costly option, particularly when used for collecting all the information that is needed for evaluation purposes and in case of repeated evaluations. It also suffers from the same disadvantages as using existing surveys: limited sample size and wage information that is less reliable than administrative data, since the data are self-reported.

The advantages and disadvantages of a dedicated survey mean that it is not the best option for evaluating the impact on most core outcomes, such as employment and wages. It could, nevertheless, be beneficially used for evaluating the impact of MAs on outcomes not captured in administrative data, most notably subjective well-being and skills. To maximise value for money, the survey should be undertaken only once in several years and its results should be matched with the linked administrative and survey data.

4.4 Data on Employers

Since no existing surveys capture which firms participate in MAs, evaluating the impact of MAs on employers requires either linking employer records with information held by SDS or conducting a new survey.

The information needed to calculate productivity is available from the IDBR and from the Annual Business Survey (ABS). The IDBR covers the whole firm population but allows calculation of labour productivity only measured as turnover over employees. The ABS, in contrast, covers only a sample of firms but contains information on value added and investment.

To obtain information on participation in MAs, these datasets need to be linked to the CTS records held by the SDS. An experimental link between CTS and IDBR has already been established based on employer names and post codes. ABS data can then be added based on the IDBR number. In the future, collecting employer identifiers—IDBR, PAYE, VAT and company registration number—by SDS would allow a simpler, more comprehensive and more accurate link.

As with individual-level data, an alternative approach is to conduct a new survey. However, as apprentices represent only a small proportion of employees for most employers, accurately measuring their effect in a survey of a limited size is challenging. For this reason, the individual-level survey should be given priority or the employer survey should focus on small- and medium-sized enterprises (SMEs), where both benefits and costs of MAs may be easier to measure and which are harder to observe in the ABS.

5 Counterfactual Impact Evaluation (Individuals)

Impact evaluation aims to answer this question: ‘What would have happened to the person/firm had the intervention not taken place?’ The central challenge lies in the fact that, at a given point in time, it is not possible to observe the same person or firm both with and without the evaluated intervention (e.g. apprenticeship). The hypothetical alternative ‘counterfactual’ scenario is fundamentally unobservable.

For this reason, impact evaluation relies on ‘control groups’. These consist of units (e.g. individuals, firms) which have not been exposed to the intervention but were initially as similar as possible to units which have been exposed to the treatment, i.e. the ‘treatment group’.

As it is difficult to find a control group that is the same as the treatment group in all respects other than participation in the evaluated intervention, a range of approaches can be used for making the treatment and control groups as similar as possible, and for taking into account any remaining differences between them.

This section discusses the choice of control groups and evaluation methods for assessing the impact of MAs on individuals. The subsequent section will then focus on the impact on employers.

5.1 Control Groups

There are, in principle, two groups of people from which the control group for the evaluation of MAs could be drawn. One group are individuals who never started an MA—‘never-starters’. The other group are people who started an MA but did not complete it—‘non-completers’.

The use of each of these two groups has different advantages and challenges. The main challenge in the case of never-starters is that starting an MA is not random and may be related to a person’s characteristics, such as skills, motivation, socio-economic background and local economic conditions. These characteristics may, in turn, influence outcomes such as employment, wage and subjective well-being. Consequently, observed differences in these outcomes between never-starters and MAs may be due not only to a causal effect of taking an MA but also to differences in individual characteristics.

The problem of non-random selection into MAs does not apply to non-completers because they have themselves started an MA. Instead, one challenge with non-completers is due to a non-random selection into non-completion. Non-completers may lack skills and determination, or non-completion may be a result of more attractive outside opportunities. In either case, differences in outcomes could be due to differences in personal characteristics rather than the causal effect of completing an MA. An additional challenge with using non-completers as a control group is that apprentices are likely to learn from their MA even if they do not complete it. The difference in outcomes between MA non-completers and completers, therefore, may capture only part of the full benefit of MAs and thus underestimate the positive impact of apprentices on individuals.

The preferred solution is to use both types of control groups—non-completers as well as never-starters—for the analysis. While both control groups face problems due to non-random selection, the nature of the selection is different in each case. Finding that both approaches lead to similar results will be reassuring; if the results differ, the direction of the difference will be informative for interpretation of the estimates. Furthermore, using multiple control groups is not particularly costly when administrative data are used for evaluation. However, it may not be a suitable choice for evaluation based on newly collected survey data.

An important additional decision is how to account for other qualifications that young people in the treatment and control group achieve. Doing so is important for separating the causal effect of MAs, on outcomes such as employment and wages, from the effect of other qualifications. Restricting the treatment and control groups to individuals with no higher qualifications should be preferred in the baseline specification, and using the full sample and controlling for other qualifications would be a useful complementary analysis, providing additional information.

5.2 Evaluation Methods

The aim of the approaches discussed here is to separate the causal impact of Modern Apprenticeships from mere correlations that may be due to differences in individual characteristics between the treatment group and the control groups. While some popular approaches are not feasible in the context of Modern Apprenticeships, several others can significantly improve the reliability of the estimated impact.

A method that is widely regarded as the ‘gold standard’ in impact evaluation is *randomised control trials* (RCTs). RCTs randomly divide individuals into a treatment group and a control group, thus overcoming the non-random selection issues discussed above. Unfortunately, applying this approach to the evaluation of MAs has not proven to be possible.[5]

An alternative to conducting an experiment would be to rely on factors that make apprentices more likely to participate in MAs but do not directly affect their labour market outcomes. Such ‘instrumental variables’ could be related, for example, to the availability of MA places in a given location and year,[6] to regional differences

[5] For an evaluation of an apprenticeship programme using a RCT, see Schaeffer et al. (2014).

[6] For examples, see Adda et al. (2006), Parey (2008), Alet and Bonnal (2011) and Noelke and Horn (2014).

in relevant regulations[7] or to the intensity with which MAs are promoted in different regions or at different schools. Alternatively, the analysis could rely on factors that cut some apprenticeships short for reasons that are random from the apprentice's perspective.[8] Unfortunately, no suitable source of exogenous variation for which information is available has been identified in the context of MAs.

Several other approaches can instead be used to conduct the evaluation. To begin with, the control groups can be defined more narrowly, in order to more closely match the treatment group.

Firstly, the risk that never-starters are systematically different from apprentices in their characteristics, and that this drives the observed outcomes, can be partly overcome by restricting the control group to individuals to whom an SDS career counsellor has suggested entering MAs. This information is available in the CSS records and should be used to strengthen the estimation.

Secondly, the selection problem in the case of non-completers, which arises because the factors underlying non-completion may also affect later outcomes, can be addressed using information on the reason for non-completion, available in the CTS records held by SDS. One option is to focus only on those individuals who failed to complete their MA for a reason that is unlikely to affect later labour market outcomes (e.g. moving to a new location).[9] An alternative is to split non-completers into those whose reason for non-completion is likely to be negatively related to later outcomes (e.g. lack of motivation) and those whose reason for non-completion is likely to be positively related to them (e.g. being offered a different job). As the first control group would likely lead to overestimation of the effect of MAs and the latter control group to its underestimation, they would provide, respectively, an upper and lower bound for the estimate.

Thirdly, the issue that even incomplete training is likely to produce some benefits can be largely addressed by restricting the control group to those non-completers who left their training shortly after it began.

Matching techniques are another way of making the control group as similar to the Modern Apprentices as possible. In particular, they can be used to construct a control group (based on never-starters or non-completers) with individuals who are similar to apprentices in terms of their *observable* characteristics, such as gender, age, socio-economic background, previous education, region and previous labour market history. Based on these characteristics, apprenticeship completers can be matched with members of the control group who were initially similarly likely to start or complete an apprenticeship (using a different version of propensity

[7] See Picchio and Staffolani (2013).

[8] See Fersterer et al. (2008).

[9] A similar strategy has been used to evaluate private on-the-job training (Leuven and Oosterbeek 2008; Görlitz 2011).

score matching)[10] or whose initial characteristics were similar to those of the apprenticeship completers (coarsened exact matching (CEM)).[11]

Defining a suitable control group is essential for the evaluation and regression methods should not be seen as its substitute. But once an appropriate control group is constructed, regression analysis can be beneficially used to control for further factors that could distort the estimated relationship between participating in an MA and the examined outcomes. Such factors may include previous education and the region, size and industry of the MA employer.

Matching and regression techniques usefully account for observable individual characteristics but cannot account for the unobservable ones. A complementary approach could therefore be to analyse changes in outcomes for given individuals over time, rather than focusing on levels. This would allow controlling for individual characteristics which are constant over time and hard to directly observe in the data (e.g. motivation, talent or persistence) but which may be important in determining the evaluated outcomes. For each individual, the approach would compute changes in outcomes (e.g. wages) between the periods before and after the apprenticeships (or the time when individuals in a control group were most likely to start an apprenticeship), and then it would compare these differences between the apprentices and the control group. Importantly, this desirable approach is subject to the availability of a sufficient number of apprentices with work experience prior to their apprenticeship.[12]

6 Counterfactual Impact Evaluation (Employers)

Evaluation of impact on employers aims to compare the actual performance (e.g. productivity) of firms participating in MAs with the performance that the employer would have achieved had the firm not participated in MAs, or had it participated to a different extent. As in the case of individuals, the 'counterfactual' scenarios cannot

[10] Introduced by Rosenbaum and Rubin (1983), propensity score matching proceeds in two steps. The first estimates the probability of being treated (e.g. participating in an apprenticeship) based on observed individual characteristics. The second step then matches individuals with similar predicted probability—propensity score—from the estimation. For example, each individual in the treatment group can be matched with the person in the control group with the most similar propensity score ('nearest-neighbour matching') or can be compared with a linear combination of multiple individuals, with weights given by differences in the propensity score. For applications in the context of on-the-job training, see Almeida and Faria (2014) and Kaplan et al. (2015), and for an application to evaluating apprenticeships see Reed et al. (2012).

[11] CEM is a more robust way than PSM to construct a control group that is actually similar to the treatment group. However, CEM requires a very large sample size, and its successful application would, therefore, require comprehensive administrative data. CEM is described by Iacus et al. (2011) and has been used in the context of apprenticeship evaluation by Bibby et al. (2014).

[12] Reed et al. (2012) use changes over time to evaluate apprenticeships, and Bibby et al. (2014) do so when evaluating further education.

be directly observed and have to be approximated by other, otherwise similar, firms which differ only in terms of their participation in MAs.

This section discusses how this could be done. First, it proposes a way in which the intensity of firm participation in apprenticeships could be measured. Then it discusses evaluation approaches that can be used to strengthen the estimation.

6.1 Measuring Intensity of Participation in Modern Apprenticeships

For firms, unlike individuals, participating in MAs is not a binary decision. In addition to the question of whether or not a firm participates at all, it also matters how many apprentices it takes on relative to its size.

A recent review of literature on apprenticeship evaluations (Bajgar and Criscuolo 2016b) suggests that employing a regular worker without training is the most common alternative to training an apprentice. For this reason, the number of Modern Apprentices relative to the total number of workers employed by a firm is the most appropriate measure of firm participation in MAs.

It is also important to consider whether the number of Modern Apprentices should be based on current or past apprentices, because the effect of apprentices in training, relative to regular workers, is likely to be negative even if the long-term effects are positive and large. To measure long-term effects that operate through apprentices' individual productivity, the evaluation should focus on the effect of employing past Modern Apprentices. In contrast, a potential evaluation of the short-term effects of training Modern Apprentices (e.g. the fact that MAs spend less time than regular employees actually working, boosted staff morale) should focus on the number of apprentices currently in training.

6.2 Evaluation Methods

Evaluating the impact of MAs on employers also requires the causal effect of MAs to be separated from other factors that may be correlated with both firms' involvement in MAs and the examined outcomes. On one hand, firms that take on apprentices might be better managed. Such firms would be more productive, but not as a result of their participation in MAs. On the other hand, some firms may train apprentices in response to underinvestment in training in previous years. They would probably have lower productivity, but again not as a consequence of training apprentices.

Several methods can be employed to bring the estimated impact closer to the true value. An RCT with an 'encouragement design' could be used to induce randomly selected firms to participate (e.g. through a training voucher). Such an experiment would provide exogenous variation for identifying the true impact of MAs, and, in addition, it would provide valuable information on the extent to which financial incentives help stimulate firms' interest in MAs. However, a similar experiment would require strong policy support and would take time to allow the measurement of long-term outcomes, and as a result will not be part of the initial impact evaluation of MAs.

Instrumental variable techniques would exploit factors which increase some firms' participation in MAs without affecting firm performance other than through the apprenticeships. Such factors could be related to MA contribution rates or varying institutional arrangements for MAs across different occupations or regions.[13] Unfortunately, there do not seem to be any observed variables that could be used in this role in the Scottish context.

Instead, matching techniques can be used to construct a control group of firms that do not engage in MAs, but which are similar to the firms that do in their other characteristics. The matching could be based on, for instance, industry, size, age and availability of a training budget.

In addition, regression analysis can be used to take into account observed firm characteristics that could be related both to participation in MAs and to the examined outcomes, such as size and training intensity.

For most firms, unlike for many individuals, the examined outcomes can be observed both before and after participating in MAs. The evaluation should compare changes in examined outcomes over time between firms that start taking on apprentices and those that do not and, by doing so, account for unobservable time-invariant firm characteristics which could otherwise bias the results.

7 Examples of Past Evaluations

This chapter describes a design of an evaluation which has yet to be undertaken. In this section, the evaluation design is illustrated by briefly describing several apprenticeship evaluations which have already been conducted and have produced interesting results. The first two studies, from England and from the United States, are included because they use data and methods somewhat similar to those proposed for evaluating MAs in Scotland. Two other studies, both evaluating the same programme in Colombia, then give an example of an RCT in the context of apprenticeships.

7.1 Impacts of Apprenticeships in England

The study by Bibby et al. (2014) is among a series of studies conducted by the UK Department of Business, Innovation and Skills to evaluate effects of further

[13] Cappellari et al. (2012) examine the effect of apprenticeships on firm productivity using random staggering of a policy change roll-out across Italian regions and industries.

education in England. It also examines other types of further education, but it includes a specific section on apprenticeships. It relies on a linked administrative dataset, which this chapter also recommends for evaluating MAs in Scotland. This dataset includes individual education histories, information on benefits from the Department for Work and Pensions, and employment and earning information from HMRC.

To estimate the effects of apprenticeships, the study compares outcomes for apprenticeship completers with those for individuals who start apprenticeships but do not complete them. It estimates the effects by ordinary least squares (OLS), controlling for a number of individual characteristics such as gender, age, disability, region and prior education.

The study finds a daily wage premium associated with completing an apprenticeship of 11% and 16% for Level 2 and Level 3 apprenticeships, respectively. It also finds an employment probability premium of almost 3% points shortly after the end of the training, which nevertheless disappears by year 4 or 5. It also shows that its results are reasonably robust compared with several more sophisticated estimation approaches, although it pursues these only for further education overall and not specifically for apprenticeships. It applies difference-in-differences or difference-in-differences-in-differences methods, combined with CEM (Iacus et al. 2011), and it restricts the control group to individuals who dropped out of their training early on, finding that this leads to a larger estimated effect of the further education qualifications.

7.2 Impacts of the Registered Apprenticeships in the United States

Reed et al. (2012) estimate the impacts of the Registered Apprenticeships in ten US states using programme administrative data and unemployment insurance wage records. They estimate a regression comparing apprentices who completed different proportions of their apprenticeships, controlling for their initial earnings. Alternatively, they also compare completers and non-completers using propensity score matching, using the initial earnings as one of the variables for estimating the propensity score.

The results suggest that completing a Registered Apprenticeship is associated with annual earnings that are, on average, approximately USD 6600 higher in the sixth year after the start of the training, with the earnings premium dropping only slightly, to USD 5800, in the ninth year.

7.3 Impacts of the Youth in Action Programme in Colombia

Three studies analyse the impacts of the Youth in Action programme, which was in place in Colombia between 2002 and 2005. The programme was targeted at poor young people living in cities, and it involved a 3-month classroom-based training programme followed by a 3-month apprenticeship in a company operating in the formal sector of the economy. The studies benefit from the fact that, in 2005, the Colombian Government agreed to conduct an experiment in which it allocated the oversubscribed training places among preselected candidates based on a lottery. The evaluations can thus rely on a control group consisting of individuals who were interested in the training, and well qualified to enter it, but who were not given the opportunity to participate.

Attanasio et al. (2011) studied the short-term effects of the programme using a baseline survey carried out shortly before the start of the training and a follow-up survey conducted 13–15 months after its end. Their results suggest that the programme increased female employment by 7% points and female total earnings by 20%. The estimated increase in male employment and earnings was much smaller and not statistically significant. Both women and men were 5–7% more likely to be employed in the formal sector of the economy as a result of the programme.

Kugler et al. (2015) explored the long-term effects of the programme using administrative social security and education records matched to information on programme participants and control group. They found that even 3–8 years after the experiment the programme participants were 5% points more likely to be in formal employment; had spent, on average, 13% more days in formal employment; and, if they worked in the formal sector, their daily wage was 6% higher compared to members of the control group who also work in the formal sector. They were also 1.4% points more likely than the control group to have completed secondary school, 3.5% points more likely to have enrolled in college and 1.6% points more likely to have stayed in higher education 5 years after the training.

8 Concluding Remarks

This chapter has described the planned impact evaluation of MAs in Scotland. It has focused on the challenges the evaluation faces and on reasons for taking some approaches over others. It has aimed to encourage and inform similar evaluations elsewhere. With this aim in mind, this closing section highlights three interlinked and more general lessons learned from the evaluation strategy for MAs.

The first lesson emphasises the *potential of using administrative data* for similar evaluations. The advantages of administrative data, and especially of the large coverage they offer, have been reiterated throughout the chapter: they are instrumental for estimating effects separately for different types of MAs and individuals with different characteristics; they allow us to analyse how the effects of MAs evolve over time after the end of the training; they make it possible to use multiple control groups; and they facilitate identification of causal effects through the use of narrower control groups and matching techniques.

Second, setting up a linked dataset and conducting the evaluation can be greatly facilitated by collecting the right information in programme administrative records early on. In the case of MAs in Scotland, this would involve collecting relevant individual and company identification numbers and ensuring that the collected variables that are important for the evaluation (e.g. reason for non-completion) are coded, complete and correct.

The last lesson underlines *the importance of having an ex ante strategy* for ex post programme evaluation. Such a strategy highlights the data and variable requirements of the evaluation and, by doing so, allows early collection of the necessary information, as discussed in the second lesson. In addition, the strategy may encourage designing the programme ex ante in a way that allows robust evaluation approaches based on RCTs, regression discontinuity design or instrumental variables ex post.

Acknowledgements The authors would like to thank the staff of Skills Development Scotland for contributing important ideas and information to the analysis.

References

Adda JC, Dustmann C, Meghir C et al (2006) Career progression and formal versus on-the-job training. IZA Discussion Paper No 2260. Institute for the Study of Labor, Bonn

Alet E, Bonnal, L (2011) Vocational schooling and educational success: comparing apprenticeship to full-time vocational high-school. Toulouse School of Economics, Toulouse

Almeida RK, Faria M (2014) The wage returns to on-the-job training: evidence from matched employer-employee data. IZA Lab Dev 3:1–33

Attanasio O, Kugler A, Meghir C (2011) Subsidizing vocational training for disadvantaged youth in Colombia: evidence from a randomized trial. Am Econ J Appl Econ 3:188–220

Audit Scotland (2014) Modern apprenticeships. Audit Scotland, Edinburgh

Bajgar M, Criscuolo C (2016a) OECD evaluation framework for modern apprenticeships in Scotland. OECD Science, Technology and Industry Policy Paper 2016/35. Organisation for Economic Co-operation and Development, Paris

Bajgar M, Criscuolo C (2016b) Impact of apprenticeships on individuals and firms: lessons for evaluating modern apprenticeships in Scotland. OECD Science, Technology and Industry Working Paper 2016/06. Organisation for Economic Co-operation and Development, Paris

Bibby DF, Buscha A, Cerqua D et al (2014) Estimation of the labour market returns to qualifications gained in English further education. Research Paper 195. Department for Business, Innovation and Skills, London

Bonnal L, Mendes S, Sofer C (2002) School-to-work transition: apprenticeship versus vocational school in France. Int J Manpow 23:426–442

Cappellari L, Dell'Aringa C, Leonardi M (2012) Temporary employment, job flows and productivity: a tale of two reforms. Econ J 122:188–215

Cooke LP (2003) A comparison of initial and early life course earnings of the German secondary education and training system. Econ Educ Rev 22:79–88

Department for Business, Innovation and Skills (2015a) English apprenticeships: our 2020 vision. HM Government, London

Department for Business, Innovation and Skills (2015b) Measuring the net present value of further education in England. BIS Research Paper No 228. BIS Research, London

Fersterer J, Pischke J-S, Winter-Ebmer R (2008) Returns to apprenticeship training in Austria: evidence from failed firms. Scand J Econ 110:733–753

Görlitz K (2011) Continuous training and wages: an empirical analysis using a comparison-group approach. Econ Educ Rev 30:691–701

Iacus SM, King G, Porro G (2011) Multivariate matching methods that are monotonic imbalance bounding. J Am Stat Assoc 106:345–361

Kaplan DS, Novella R, Rucci G et al (2015) Training vouchers and labor market outcomes in Chile. Working Paper 585. Inter-American Development Bank, Washington, DC

Kugler A, Kugler M, Saavedra J et al (2015) Long-term direct and spillover effects of job training: experimental evidence from Colombia. NBER Working Paper 21607. National Bureau of Economic Research, Cambridge

Leuven E, Oosterbeek H (2008) An alternative approach to estimate the wage returns to private-sector training. J Appl Econ 23:423–434

Malamud O, Pop-Eleches C (2010) General education versus vocational training: evidence from an economy in transition. Rev Econ Stat 92:43–60

Noelke C, Horn D (2014) Social transformation and the transition from vocational education to work in Hungary: a differences-in-differences approach. Eur Sociol Rev 30:431–443

Parey M (2008) Vocational schooling versus apprenticeship training – evidence from vacancy data. Institute for Fiscal Studies. http://cep.lse.ac.uk/seminarpapers/09-03-12-MP.pdf

Picchio M, Staffolani S (2013) Does apprenticeship improve job opportunities? A regression discontinuity approach. IZA Discussion Paper No 7719. Institute for the Study of Labor, Bonn

Reed D, Liu AY-H, Kleinman R et al (2012) An effectiveness assessment and cost-benefit analysis of registered apprenticeship in 10 states: final report. Mathematica Policy Research, Princeton

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55

Schaeffer CM, Henggeler SW, Ford JD et al (2014) RCT of a promising vocational/employment program for high-risk juvenile offenders. J Subst Abus Treat 46(2):134–143

Skills Development Scotland (2013) Modern apprenticeships outcomes survey 2012. Skills Development Scotland, Glasgow

Skills Development Scotland (2015) Modern apprenticeship employer survey 2015. Report. Skills Development Scotland, Glasgow

Matej Bajgar is an economist at the OECD Productivity and Business Dynamics Division of the Science, Technology and Innovation Directorate. He focuses on using microeconomic data to evaluate impact of public support for training and R&D in the private sector and to analyse firm performance. Before joining OECD as a Young Professional, he was a doctoral student at the University of Oxford, researching performance of manufacturing exporters in the emerging economies. During his doctorate, he worked as a consultant on several projects in developing countries, including a randomised evaluation of a government social programme in Lesotho. He holds degrees from the Charles University in Prague and the University of Oxford.

Chiara Criscuolo is the head of the OECD Productivity and Business Dynamics Division of the Science, Technology and Innovation Directorate. Since joining in 2009, Chiara has also worked on climate change, innovation policies and measurement. She is co-ordinating two large cross-country microdata projects on employment dynamics and on productivity. She recently co-authored a report on the Future of Productivity. Prior to joining OECD, she was Research Fellow at the Centre for Economic Performance, London School of Economics. She has published widely in the field of productivity, innovation and international trade. She holds a doctoral degree in Economics from University College London.