Received: 15 November 2019 | Accepted: 19 March 2021

DOI: 10.1111/rssa.12697

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Pre-apprenticeship training for young people: Estimating the marginal and average treatment effects

Richard Dorsett[1] | **Lucy Stokes**[2]

[1]University of Westminster, London, W1B 2HW, UK

[2]National Institute of Economic and Social Research, London, UK

Correspondence
Richard Dorsett, University of Westminster, 309 Regent Street, London W1B 2HW, UK.
Email: r.dorsett@westminster.ac.uk.

Abstract

This paper evaluates traineeships, a voluntary programme of work placement and preparation that aims to help young unemployed people in England compete for jobs and apprenticeships. Applying the method of local instrumental variables to administrative data, we estimate the marginal treatment effects on apprenticeship take-up and employment. The heterogeneous impacts are then aggregated to form an estimate of the average impact of treatment for all participants. The results suggest that, among younger trainees, participation increases the probability of becoming an apprentice and that this holds across the distribution of unobserved heterogeneity. For older trainees, we find no significant effect on the probability of becoming an apprenticeship on average but some evidence of a negative effect among those more resistant to participating. We find no effects on employment for either age group.

KEYWORDS

apprenticeships, employment, evaluation, impact heterogeneity, young people

JEL CLASSIFICATION

I28; J24; J48; J68

1 | INTRODUCTION

Apprenticeships are one of the key means by which the UK government aims to build skills and tackle the problem of youth unemployment. The 2015 election manifesto of the Conservative government announced a target of 3 million apprenticeship starts in England by 2020. The implied 600,000 starts a year represents a 20% hike on the level seen before the announcement of the target.

This commitment to apprenticeships is motivated by a belief that they improve skills, raise productivity and stimulate economic growth (HM Government, 2015). The most recent evaluation evidence provides some support for this, finding a positive impact on earnings, although with substantial variation by sector (Cavaglia et all., 2018). As such, it is consistent with earlier studies that likewise tend to find positive results of apprenticeships (Bibby et al., 2014; McIntosh and Morris, 2016).

However, increasing the number of new apprenticeships to the level required by the 3-million target has proved challenging. The final figures for the 2018/2019 academic year (August 2018–July 2019) show 1.9 million starts since May 2015 (Department for Education, 2019). This equates to roughly 450,000 starts a year, considerably below the target rate. Furthermore, while the manifesto presented apprenticeships as supporting young people, roughly 45% of starts have been people aged 25 or over. The relevance of the headline statistics to the youth labour market has to be seen with this in mind.

A recognised issue is that not all young people are able to secure an apprenticeship (Department for Education, 2013). To address this, a programme of traineeships was introduced in England in 2013. A traineeship provides work preparation training, work experience and, if needed, help with English and mathematics. It is designed to equip young people with the skills and experience required to secure an apprenticeship or employment.

In this paper, we provide evidence on the impact of traineeships. In line with the objectives of the programme (see, e,g. https://www.gov.uk/find-traineeship), we focus on whether they help individuals to secure an apprenticeship or employment. More specifically, we consider outcomes 12 months later.

The analysis is based on linked administrative data for the full population of trainees in the academic year 2013/2014 and a sample of non-trainees, used to form a comparison group. The data provide detailed information on background characteristics and track individuals over time, thereby allowing employment and apprenticeship outcomes to be observed.

The evaluation challenge is to adequately control for non-random selection into traineeships. We control for observed differences between participants and non-participants but also allow for selection on unobservables, using local instrumental variable (LIV) estimation to derive marginal treatment effects (MTEs), free to vary across individuals. Such an approach allows impact heterogeneity to be captured. Furthermore, the MTEs can also be aggregated across groups of individuals to give other parameters of interest (Heckman and Vytlacil, 1999). We do this for participants as a whole, thereby providing an estimate of the average effect of treatment on the treated (ATT).

The analysis distinguishes between trainees aged 16–18 and trainees aged 19–23. For younger trainees as a whole, the results show a non-significant impact on employment but a strong positive impact on apprenticeships. For older trainees as a whole, no significant effects are seen. Looking beyond the ATT, no strong pattern of impact heterogeneity is evident for younger trainees. For older trainees, the results suggest that the estimated effects on apprenticeship take-up are more negative among those more resistant to participation.

These findings suggest traineeships can be effective as a means of helping those just past school-leaving age into an apprenticeship, with the expectation that this in turn will lead to improved employment and earnings prospects. The fact that there is no strong pattern of impact heterogeneity indicates its general effectiveness rather than being particularly suited to those who are more motivated to participate, for example. The lack of an employment effect should not necessarily be viewed as a failing since it is

not clear that the optimal decision at this age is to enter full-time employment. The findings for older participants are less positive and suggest that traineeships may be less appropriate for reluctant participants.

The results contribute to the evidence on how to support young people through the school to work transition. The impact of education, training and apprenticeships has received considerable attention in the empirical literature but this paper is distinct in focusing on those young people who are not engaged in those activities and who are at risk of being left behind. As such, the findings fill an evidence gap and should be of value and interest to policymakers.

The paper also contributes by focusing directly on impact heterogeneity as captured through MTEs. We are not aware of other UK studies that do this. Furthermore, we use a semi-parametric estimator that avoids the functional form restrictiveness introduced by parametric estimators. Estimation is computationally demanding and we incorporate a pre-processing matching step as a pragmatic modification (Ho et al., 2007). This matching step provides as a by-product impact estimates based on the often-invoked conditional independence assumption that all important differences between trainee participants and non-participants can be observed. The fact that these by-product estimates differ substantially from those that allow for selection on unobservables suggests the conditional independence assumption may not be realistic.

The remainder of the paper has the following format. Section 2 sets out the key features of the traineeships programme. Section 3 describes the data and provides a summary of the key characteristics of trainees. The LIV estimation approach is described in Section 4, including an adaptation introduced in order to see impact heterogeneity in more detail and to make estimation feasible when applied to large administrative data. The results are presented in Section 5. Section 6 concludes.

2 | THE TRAINEESHIPS PROGRAMME

Traineeships offer work preparation training, work experience with an employer and English and mathematics help for those lacking basic qualifications in those subjects. They were introduced in August 2013 for 16- to 23-year-olds who are interested in securing an apprenticeship or employment but who lack the necessary skills and experience (this was extended in 2014/2015 to include 24-year-olds). They are delivered through a partnership between employers and education and training providers.

More specifically, the programme is for young people who:

a. are not working and have little work experience but who are focused on work or the prospect of it
b. are 16–24 and have qualifications below Level 3 (roughly, the equivalent of the standard age 18 academic qualification)
c. are motivated to enter training or work
d. are felt, by providers and employers, to have a reasonable chance of being ready for employment or an apprenticeship within 6 months.

Support lasts up to 6 months and, while tailored to individuals' circumstances and needs, involves three main elements:

a. high-quality work experience placement, intended to develop workplace skills
b. work preparation and job-search training (CV writing, job search skills and interview preparation)
c. English and mathematics training, to ensure trainees have the required literacy and numeracy skills.

On completing their traineeship, participants are guaranteed a job interview if a role becomes available. If a role does not become available, participants have an exit interview and written feedback to help them secure an apprenticeship or employment with another employer.

3 | DATA AND SAMPLE CHARACTERISTICS

The empirical analysis draws on four linked administrative data sets. These provide a rich source of information on pre-participation characteristics and attainment as well as post-participation outcomes. The data sets are as follows:

a. Individualised Learner Record (ILR)—a learner-level database that provides detailed information on further education (i.e. post-compulsory education outside school) and work-based learning in England. Each learning aim is separately listed and traineeships are specifically identified. We consider the first year of traineeships, the academic year 2013/14.
b. National Pupil Database (NPD)—a pupil-level database with individual characteristics and attainment for all children in England. It also provides details of other school experiences, such as attendance records and exclusions.
c. National Client Caseload Information System (NCCIS)—an individual-level database with monthly post-16 activity status for all young people in England from school-leaving age up to at least their 18th birthday.
d. Work and Pensions Longitudinal Study (WPLS)—an individual-level database with details on employment and welfare spells for all adults in Great Britain.

Although the ILR can be matched to any of the other data sets, the research was subject to legal restrictions that prevented linking the WPLS to the NPD or NCCIS. Reflecting this and the fact that the age coverage of the data sets differs, our analysis is carried out separately for 16- to 18-year-olds and 19- to 23-year-olds. This division also reflects the different eligibility criteria in respect of educational attainment; while 16- to 18-year-olds with qualifications below Level 3 were eligible, 19- to 23-year-olds were eligible if they had qualifications below Level 2 (roughly, the equivalent of the standard age-16 academic qualification). For both age groups, we focus on trainees in the 2013/2014 academic year.

3.1 | 16- to 18-year-olds

For younger trainees, the population we consider is made up of the three NPD cohorts completing Key Stage 4 (KS4) in the academic years ending in 2011, 2012 or 2013. KS4 refers to the 2 years of schooling when pupils are aged 14–16 and the end of KS4 coincides with the point at which most children can legally leave school. Children in these cohorts were aged 16–18 in the 2013/2014 academic year and so represent the appropriate population to consider for younger trainees. From the linked ILR, we can identify who participated in a traineeship in 2013/2014. We can also identify those who did not participate in a traineeship in 2013/2014. The data available for this latter group is a 10% random sample of young people within these cohorts who did not participate.

We observe a range of background characteristics and information on individuals' school experience, such as absences, special needs and exclusions. Attainment is also recorded in the form of

examination results. Outcomes are taken from linked records. The ILR provides information on apprenticeship outcomes and the NCCIS provides information on employment outcomes.

Table 1 presents summary information on the characteristics of trainees. Among those aged 16–18, trainees are less well-qualified (they are much more likely to have no good GCSEs), are more likely to have had special educational needs when at school, had more absences (including unauthorised absences) and were more likely to have been excluded from school than non-trainees.

The outcomes considered are whether an individual is employed 12 months after starting their traineeship and whether an individual is an apprentice at this time. For non-trainees, a "pseudo" start date was used in place of an actual start date; this was imputed as a random draw from the distribution of start dates observed among trainees. Among trainees, 16% are employed at this time and 33% are apprentices. These levels are substantially higher than those seen among non-trainees.

3.2 | 19- to 23-year-olds

For 19- to 23-year-olds, information on the population of school-leavers is not available. Instead, the estimation sample is drawn entirely from the ILR. As with 16- to 18-year-olds, the full population of trainees aged 19–23 in 2013/2014 is identified. However, unlike the 16–18s, the full population of similar-age non-trainees is not observed since the sample is drawn from those who participated in some sort of learning. To provide a comparison group, we use the population of similar-age individuals participating in other learning aims in the same or the previous academic year. More precisely, the estimation sample is made up of all 19- to 23-year-old trainees in the academic year 2013/2014 along with all individuals who, in that year or the previous year, were engaged in a learning activity at a level commensurate with traineeship eligibility but who were not trainees.

Outcomes are taken from the ILR and the linked WPLS. From the ILR, we can observe apprenticeship outcomes as before. From the WPLS, we observe employment outcomes. In addition, for this older group, it is possible to consider whether the individual is claiming Jobseeker's Allowance (JSA—the UK's unemployment benefit). This is not possible for the younger age group, most of whom will not have been eligible for JSA.

Since the sample definition for 19- to 23-year-olds does not include all non-trainees, it is natural to consider how this might affect subsequent impact estimates. As will be described later, the estimation approach involves first matching trainees to non-trainees with a view to identifying a subgroup among the non-trainees that is similar to trainees in respect of key characteristics. The extent to which it is possible to do this depends largely on the richness of the available data. The nature of the data available for 19- to 23-year-olds is such that all non-trainees have been FE learners at some point. Consequently, the estimation sample automatically includes only those who have demonstrated some interest in learning within the last 2 years. Since trainees have, by definition, shown such interest, limiting the non-trainees to those who have also shown this interest imposes a similarity across the groups.

The fact that some non-trainees will have participated in alternative training during 2013/2014 also influences how estimated impacts should be interpreted. We return to this later when considering the estimation results but, for now, note that this is also an issue for 16- to 18-year-olds; among this younger group, many will be in school or some form of training.

Table 1 presents summary characteristics for the 19- to 23-year-olds sample. The background information available in the ILR differs from that in the NPD. Nevertheless, we see that, as with 16- to 18-year-olds, the impression is of trainees having weak human capital relative to non-trainees: lower

levels of qualifications, less employment experience and a greater tendency to have a learning difficulty. A higher proportion of 19- to 23-year-old trainees are male; 64% compared to 49% of 16- to 18-year-old trainees.

With regard to outcomes, employment at 12 months stood at 40% among trainees, substantially higher than that among non-trainees (23%). Unemployment, as captured by JSA receipt, was 10% among trainees at 12 months, nearly double the non-trainee level. While this may appear to be inconsistent with the employment outcomes, this need not be the case; trainees may have higher levels of both employment and unemployment if there is a compensating lower level of economic inactivity. Lastly, apprenticeship levels at 12 months were 13.7% for trainees compared to 4.5% for non-trainees.

4 | ESTIMATION APPROACH

Our approach focuses on the MTE, as introduced by Björklund and Moffitt (1987), and which we estimate using the LIV estimator within the framework of a generalised Roy model (Heckman and Vytlacil, 1999). This section describes the estimation approach, loosely following the exposition of Carneiro et al. (2011).

4.1 | The econometric model

Under the Neyman–Rubin causal model (Neyman, 1923; Rubin, 1974), each individual has two potential outcomes $\left(Y_0, Y_1\right)$. Here, Y_0 is the outcome associated with not participating in a traineeship (which we denote $D = 0$) and Y_1 is the outcome associated with participating ($D = 1$). We assume potential outcomes can be modelled as a linearly separable function of observed characteristics, $\boldsymbol{X}$, and unobserved characteristics, U_j, where $j = \{0, 1\}$:

$$\begin{aligned} Y_0 &= \boldsymbol{X}\beta_0 + U_0 \\ Y_1 &= \boldsymbol{X}\beta_1 + U_1 \end{aligned} \tag{1}$$

The observed outcome, Y, depends on whether an individual participates in the treatment:

$$Y = DY_1 + (1 - D)\, Y_0. \tag{2}$$

Participation, D, is determined by its latent variable, D^*, itself a function of $\boldsymbol{Z} = (\boldsymbol{X}, \tilde{Z})$, where $\tilde{Z}$ is an instrument:

$$\begin{aligned} D^* &= \boldsymbol{Z}\beta_D - V \\ D &= 1\left(D^* \geq 0\right) \end{aligned} \tag{3}$$

Equation (3) implies that an individual will participate if $\boldsymbol{Z}\beta_D > V$. Following convention, we interpret V as resistance to participation and assume it is a continuous random variable with distribution function F_V. The probability of participation is $P(\boldsymbol{Z}) \equiv Pr\,(D = 1|\boldsymbol{Z}) = F_V\left(\boldsymbol{Z}\beta_D\right)$. The quantiles of V can be represented by $U_D = F_V\,(V)$. It then follows that $D = 1$ if $P(\boldsymbol{Z}) > U_D$. Intuitively, individuals will participate if the net benefit of doing so, D^*, is positive. With V representing the cost of participation, U_D represents the quantiles of those costs.

TABLE 1 Characteristics of trainees and non-trainees

		16–18s		19–23s	
		Non-trainees	**Trainees**	**Non-trainees**	**Trainees**
Female (%):		49.2	51.2	44	36.4
Age (%):	*16*	25.1	27.0		
	17	50.0	51.5		
	18	24.9	21.5		
	19			12.6	18.6
	20			23.6	28.4
	21			22.3	20.2
	22			21.1	19.1
	23			20.3	13.8
White(%):		79.9	83.2	69.4	72.0
Local unemp, Sep 2013 (%)		8.2	8.8	8.6	8.8
Urban		84.3	91.1	89.9	95.3
GCSEs at A*-C grade (%):	*0*	15.6	40.2		
	1	7.8	16.3		
	2	6.7	11.5		
	3	5.8	8.3		
	4	11.3	10.7		
	5+	52.8	12.9		
Special educational needs (%):		21.3	37.7		
School absences (%):	*None*	7.4	3.6		
	1	2.5	1.4		
	2-9	30.2	16.0		
	10-24	32.5	30.1		
	25-49	18.8	26.3		
	50+	8.4	22.6		
Unauthorised absences (%):	*None*	59.5	35.5		
	1	6.4	4.9		
	2-9	21.2	26.2		
	10-24	7.3	15.0		
	25-49	2.9	9.0		
	50+	2.6	9.4		
Excluded from school (%):		4.5	11.5		
Learning difficulty (%):				19.5	25.8
Months emp. 1st year before				5.1	3.0
Months emp. 2nd year before				4.3	2.7
Months emp. 3rd year before				3.6	2.1
Qualifications (%):	*None*			7.0	22.2

(Continues)

TABLE 1 (Continued)

		16–18s		19–23s	
		Non-trainees	**Trainees**	**Non-trainees**	**Trainees**
	< Level 1			9.3	13.7
	Level 1			22.3	40.6
	Level 2/3			55.7	17.4
	Unknown			5.2	4.7
Emp. 12 months later (%):		9.3	16.3	22.7	40.4
Unemp. 12 months later (%):				5.4	10.0
Appren. 12 months later (%):		5.8	33.0	4.5	13.7
Number of observations		88,724	4,282	347,155	3,163

Notes: The results for 16- to 18-year-olds are based on all trainees in the 2013/2014 academic year and a 10% sample of all similar-aged non-trainees. The results for 19- to 23-year-olds are based on all trainees in the 2013/2014 academic year and all similar-aged non-trainees who were observed to participate in a learning aim at Level 3 or lower in either the 2013/2014 academic year or the 2012/2013 academic year.

Italics denote values for categorical variables.

We write the MTE, the mean impact for individuals at a particular value of U_D, as.

$$\Delta_{MTE}\left(\boldsymbol{x}, u_D\right) = E\left(Y_1 - Y_0 | \boldsymbol{X} = \boldsymbol{x}, U_D = u_D\right). \tag{4}$$

This is identified by differentiating the conditional mean outcome with respect to the propensity score and evaluating it at u_D:

$$\Delta_{MTE}\left(\boldsymbol{x}, u_D\right) = \left.\frac{\partial E\left(Y | \boldsymbol{X} = \boldsymbol{x}, P\left(\boldsymbol{Z}\right) = p\right)}{\partial p}\right|_{p=u_D} \tag{5}$$

The conditional mean outcome can be written.

$$E\left(Y | \boldsymbol{X} = \boldsymbol{x}, P\left(\boldsymbol{Z}\right) = p\right) = \boldsymbol{X}\beta_0 + p\boldsymbol{X}\left(\beta_1 - \beta_0\right) + K\left(p\right) \tag{6}$$

where

$$K\left(p\right) = E\left(U_0 | P\left(\boldsymbol{Z}\right) = p\right) + E\left(U_1 - U_0 | P\left(\boldsymbol{Z}\right) = p\right)p. \tag{7}$$

Substituting into Equation (5) gives.

$$\Delta_{MTE}\left(\boldsymbol{x}, u_D\right) = X\left(\beta_1 - \beta_0\right) + \left.\frac{\partial K\left(p\right)}{\partial p}\right|_{p=u_D}. \tag{8}$$

We estimate this using the LIV approach of Heckman and Vytlacil (1999), implementing a semi-parametric estimator, as described in the online supplementary material accompanying Heckman et al. (2006). This estimation approach allows impacts to vary flexibly across the distribution of p. We avoid parametric models (Aakvik et al. 2005, for example) since these assume joint normal errors, which restricts the shape of the MTE curve (Cornellisen et al., 2016).

An outline of this approach is that it involves running a local linear regression of Y, $\boldsymbol{X}$ and $\boldsymbol{X}P(\boldsymbol{Z})$ on $P(\boldsymbol{Z})$ at each observed value of $\widehat{P}(\boldsymbol{Z})$ and saving the residuals as $\widehat{e}_Y$, $\widehat{\boldsymbol{e}}_{\boldsymbol{X}}$ and $\widehat{\boldsymbol{e}}_{\boldsymbol{Xp}}$, respectively. Regressing $\widehat{e}_Y$ on $\widehat{\boldsymbol{e}}_{\boldsymbol{X}}$ and $\widehat{\boldsymbol{e}}_{\boldsymbol{Xp}}$ provides estimates of β_0 and β_1. To get the remaining term of Equation (8), we first define $\tilde{Y} = Y - \boldsymbol{X}\widehat{\beta}_0 - \boldsymbol{X}\left(\widehat{\beta_1 - \beta_0}\right)P(\boldsymbol{Z})$. A local polynomial regression of $\tilde{Y}$ on the support of $P(\boldsymbol{Z})$ provides an estimate of $\partial K(p)/\partial p$, completing what is needed for the MTE estimate:

$$\widehat{\Delta}_{MTE}(\boldsymbol{x}, u_D) = \boldsymbol{x}\left(\widehat{\beta_1 - \beta_0}\right) + \left.\frac{\widehat{\partial K(p)}}{\partial p}\right|_{p=u_D} \tag{9}$$

Estimation used the Stata routine *margte* (Brave and Scott, 2014).

4.2 | Pre-processing

A drawback to LIV estimation is that it is computationally demanding. This becomes a relevant consideration when using large administrative data sets to estimate impacts, particularly since inference requires estimates to be bootstrapped. To make estimation feasible, we introduced a pre-processing step to select from the pool of non-trainees a sub-sample with observed characteristics, $\boldsymbol{X}$, similar to those of the trainees. The resulting sample—comprising the population of trainees and the subsample of selected non-trainees—was then used for the LIV estimation described above. This preliminary stage was carried out using single nearest neighbour propensity score matching without replacement, implemented using the Stata routine *psmatch2* (Leuven and Sianesi, 2003). This resulted in a group of non-trainees identical in size to the trainees group, all of whom receive a matching weight of 1 (convenient, as the Stata routine *margte* does not accept weights).

A by-product of this first stage is that a comparison of outcomes among trainees and the subsample of selected non-trainees provides a matching estimate of the ATT. Matching estimates can be interpreted as causal if the Conditional Independence Assumption (CIA) is satisfied, $Y_o \perp D \,|\, \boldsymbol{X}$, or $Y_o \perp D \,|\, P(\boldsymbol{X})$ under propensity score matching, (Rosenbaum and Rubin, 1983). Since matching estimators are non-parametric (or semi-parametric in the case of propensity score matching), participants with combinations of characteristics not represented among non-participants must be excluded from the estimation sample. This can be operationalised in various ways but a common approach is to impose the support condition, in this case requiring $0 < P(\boldsymbol{X}) < 1$.

The propensity score matching estimator compares the mean outcome of trainees with the mean outcome of the matched comparison group:

$$\widehat{\Delta}_{TT}^{PSM} = E\left(Y_1 | D = 1\right) - E_{P(\boldsymbol{x}_1)}\left\{E\left(Y_0 | D = 0, P(X)\right)\right\} \tag{10}$$

Here, the expectation of the term in braces is over the distribution of propensity scores in the treatment group, denoted $P\left(\boldsymbol{x}_1\right)$. Under the CIA, this term provides a consistent estimate of $E\left(Y_0 \,|D = 1\right)$ so $\widehat{\Delta}_{TT}^{\text{PSM}}$ in turn provides a consistent estimate of the ATT.

The CIA requires that $\boldsymbol{X}$ must capture all participation influences that also affect outcomes. If this does not hold, matching is still helpful to the extent that it eliminates two of the three bias components identified in Heckman et al. (1998). Specifically, it removes differences between participants and non-participants in the supports of those characteristics that affect outcomes and in the distribution of those characteristics in the region of common support. This suggests its usefulness as a pre-processing step, as advanced by, for instance, Ho et al. (2007).

Nevertheless, if the CIA is not satisfied, the third component—classical selection bias—remains. We might speculate that, by aligning observed characteristics across participants and non-participants, matching may indirectly result in correlated unobserved characteristics looking more similar across the two groups. However, we cannot guarantee that this has been achieved to the extent that the classical selection bias has been eliminated. Hence, matching estimates will in general be biased if the CIA is not satisfied.

In our application, the nature of the available data is such that satisfaction of the CIA may seem unlikely. While the data are rich in comparison to many administrative data sets, they lack information on motivation, attitudes and so on, which are likely to contribute to the decision to participate and might plausibly influence subsequent outcomes. Hence, despite controlling for a rich variety of background characteristics, it is still possible that an unobserved influence on participation—that is also likely to influence post-traineeship outcomes—has not been captured. This could arise through two channels. First, it may simply be that an important determinant of the decision is not recorded in the available data (motivation, for example). Having richer data is the best way of avoiding this possibility. Second, since we are unable to know whether non-trainees would have satisfied those eligibility criteria that cannot be observed in the data, ineligible comparators cannot be fully excluded from the analysis. For example, one of the eligibility criteria is that providers and employers feel there is a reasonable chance of the individual being ready for employment or an apprenticeship within 6 months. Assuming the rules have been faithfully implemented, all trainees will have been judged to have a reasonable chance of being ready within 6 months. Non-trainees, on the other hand, will not have been judged in this way so there is no way of knowing how likely they are to be ready within 6 months. Consequently, there may still be compositional differences between the treated and comparison groups identified through matching, pointing to the need to control for unobserved influences on participation (and hence motivating the LIV approach).

4.3 | Aggregating MTEs to give the ATT

MTE estimates can be aggregated to give estimates of other parameters. We concentrate on the ATT and, when reporting results, compare these with the matching ATT estimates, thus providing some insight into the role of unobservable influences. The LIV-based estimated ATT can be expressed:

$$\widehat{\Delta}_{TT}^{\text{LIV}} = \int_0^1 \widehat{\Delta}_{MTE}\left(\boldsymbol{x}, u_D\right) \omega\left(\boldsymbol{x}, u_D\right) du_D$$

where

$$\omega\left(\boldsymbol{x}, u_D\right) = \frac{Pr\left(p\left(\boldsymbol{Z}\right) > u_D \,|\, \boldsymbol{X} = \boldsymbol{x}\right)}{\int Pr\left(p\left(\boldsymbol{Z}\right) > u_D \,|\, \boldsymbol{X} = \boldsymbol{x}\right) du_D}$$

as derived in Heckman and Vytlacil (1999).

4.4 | Identifying assumptions of the LIV approach

Estimation of MTEs usually requires a continuous instrument (Although see Brinch et al. (2017) for an approach requiring only a discrete instrument). We use individuals' geographical distance from their nearest traineeship provider. The intuition here is that the greater this distance, the higher the cost

of participation, yet outcomes are unlikely to be directly affected. The use of geographical distance as an instrument has a long history (Card, 1993, provides an early example).

For distance to be a valid instrument, it must satisfy the usual IV assumptions. The first assumption is that $\tilde{\boldsymbol{Z}}$ is predictive of participation, D, after controlling for $\boldsymbol{X}$. Figure 1 provides some initial evidence to support this. It shows (using local polynomial regressions) the proportion of the sample that participated in a traineeship as a function of distance. For both 16- to 18-year-olds and 19- to 23-year-olds, participation is more common among those living closer to a provider.

As an aside, we note that the data available mean that the proportion of the sample participating cannot accurately be interpreted as a population probability of participation. In the case of 16- to 18-year-olds, the sample comprises all trainees but only a 10% sample of non-trainees, so the proportions in the chart are roughly 10 times the population probabilities. For the 19–23 group, non-trainees are selected from those participating in training in 2013/2014 or the previous year rather than the full population of non-trainees, so the inflation factor is less clear. Reflecting this, we refer to the participation proportion as the *sample* probability of participation.

Also included in Figure 1 are histograms showing the distribution of individuals' distances from the nearest traineeship provider. It is clear that the majority live quite close to a provider. Not included are those living at a distance of more than 15 km. This applies to very few people and we exclude them from the analysis since the relationship with participation becomes more erratic. This results in only a minor reduction in the number of trainees (for both age groups this is less than 1%; 37 trainees were excluded among the 16- to 18-year-olds and 17 trainees among the 19- to 23-year-olds).

While Figure 1 provides evidence of an unconditional relationship between distance and participation, identification requires that this relationship remains after conditioning on $\boldsymbol{X}$. Appendix 1

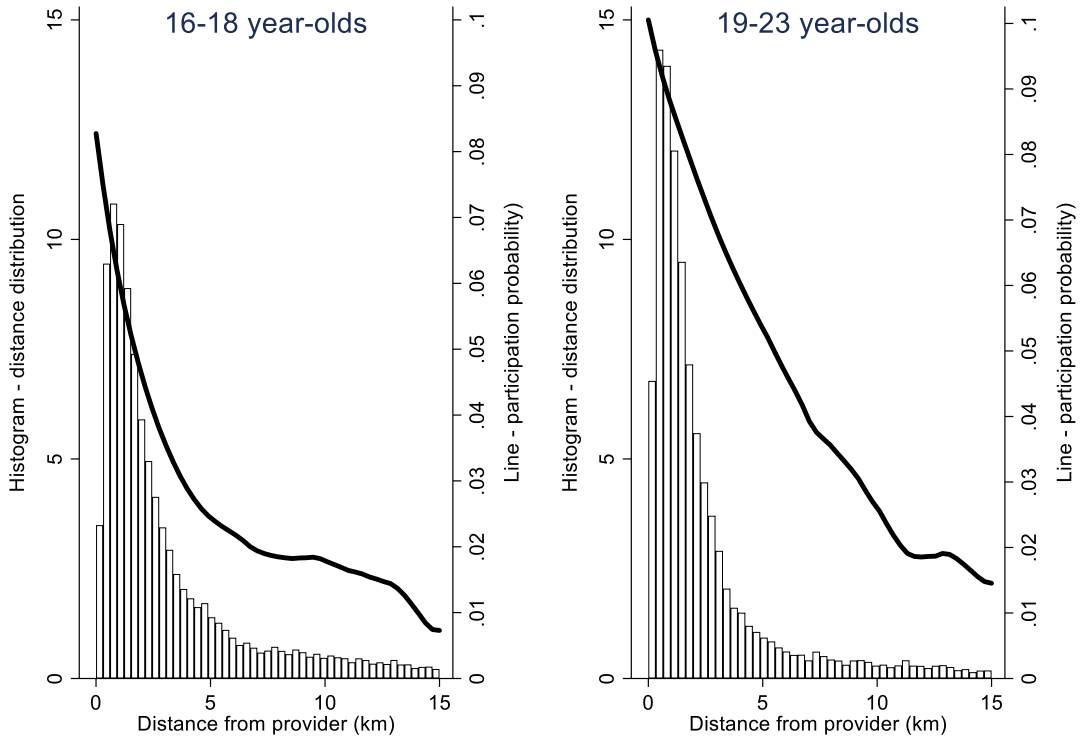

FIGURE 1 The sample probability of traineeship participation as a function of distance from provider

provides the results of estimating a probit regression of traineeship participation on the pre-processed samples for both age groups. The relationship between distance and participation is strongly significant in both cases. For 16- to 18-year-olds, the results suggest a relationship that is quadratic while for 19- to 23-year-olds, the relationship was adequately captured through a linear term alone. The key point in both cases is that the negative relationship between distance and participation is evident in the matched samples suggesting that the first condition for the instrument to be valid is satisfied.

The second IV assumption is that the instrument is conditionally independent of the unobserved components of the outcome and participation equations, $\tilde{\boldsymbol{Z}} \perp U_0, U_1, V \mid \boldsymbol{X}$. As usual, this is more difficult to justify and we may have suspicions that distance might be correlated with outcomes. This would arise if the location of training providers were influenced by local characteristics correlated with outcomes. It should be noted that traineeships were delivered by providers who were already established. Local availability did not require that providers made a new decision to locate within an area; rather, the decision is around whether they should add traineeships to their other offerings. This is unlikely to be influenced by labour demand considerations to the same degree as that of the provider's initial location decision which, from the provider's perspective, is clearly more consequential. As some evidence in support of the local availability of traineeships provision being distinct from providers' location decisions more broadly, note that the correlation between individuals' distances to traineeship provision and distances to other learning provision is only 0.33. Living further from a traineeship provider need not imply that there are no other providers located nearby.

Nevertheless, it is still possible that labour market conditions play a role. We can go some way towards addressing this concern. As already mentioned, the estimation sample is limited to individuals living within 15 km of a traineeships provider. Should providers choose to deliver traineeships only where there is employer demand, retaining in the estimation sample those individuals with no traineeships provider nearby would introduce a correlation between distance and outcomes. Restricting the sample to those living within 15 km mitigates against this. Furthermore, we include in $\boldsymbol{X}$ the unemployment rate in the local travel-to-work area (as a proxy for the strength of the local labour market) and therefore control for such variation directly. As another precaution, we include in $\boldsymbol{X}$ a measure of population density (rurality). This is prompted by the possibility that individuals living in rural areas may tend to live further from a provider of traineeships while also having fewer job opportunities locally; unaddressed, this could introduce a correlation between distance and outcomes.

It is also possible that the instrument may be correlated with individuals' unobserved characteristics. As some indication that this could be the case, regressing distance on those characteristics used to estimate the propensity score reveals some statistically significant associations. These results (available on request) underscore the importance of controlling for observed characteristics in the LIV estimation. By doing this, we hope to increase the credibility of the identifying assumption that the instrument is conditionally independent of unobserved influences on outcomes. However, it remains the case that this cannot be directly verified.

To assess empirically the extent to which this assumption is likely to be hold, a placebo test was conducted. This test is related to that described in Imbens and Wooldridge (2009, pp 48–50), which amounts to estimating the impact of treatment on a pseudo outcome uninfluenced by the treatment (such as a lagged outcome). The approach followed here, instead creates a pseudo treatment and estimates its effect on an observed outcome known to be unaffected by the treatment. The pseudo treatment (or placebo) is a function of the instrument (distance). If the effect of the placebo treatment were found to be statistically significant, it would suggest that distance was

correlated with unobserved variables influencing outcomes, thereby compromising its credibility as a valid instrument.

The placebo test was implemented as follows. First, the probability of traineeship participation was estimated using a probit model specified in the same way as in Appendix 1 but using the full (i.e. not pre-processed) data. The estimated coefficients were used to generate a linear projection, $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$, to which was added a standard normal error term, e, in order to give $\widehat{p} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + e$. The sample participation proportion, k, was calculated and then participants ($D = 1$) were dropped. Among the remaining sample of non-participants ($D = 0$), a placebo treatment, $\tilde{D}$, was assigned according to the condition $\tilde{D} = 1\left(\Phi\left(\widehat{p}\right) > 1 - k\right)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

In the resulting sample, the placebo treatment is allocated in the same proportion as seen in the full sample and the probability of participation is a function of distance. The approach described earlier was then followed to estimate the impact of this placebo: a pre-processing step, MTE estimation and lastly ATT estimation. Since this is a placebo treatment (nobody in the sample participated in a traineeship) we expect to find no statistically significant effect. This is indeed the case, as shown in Table 2. Although not presented, the estimated MTEs of the placebo treatment were also found to be non-significant at all points for both age groups and all outcomes. If significant effects had been found, it would indicate that the placebo was capturing some correlation between distance and outcomes that was not otherwise controlled for. The fact that no significant effect was found increases the credibility of the second assumption and therefore provides support for the claim in this paper that the estimated impacts presented in the next section are valid and not merely capturing an effect of distance on outcomes.

5 | RESULTS

All analysis was conducted separately for the two age groups. In both cases, 200 bootstrap samples were drawn. Ho et al. (2007) argue that it may not be necessary to allow for the matching step to influence variance estimation. However, we do allow for this, conducting the matching step followed by the LIV step for each bootstrap sample. This section describes the results across all bootstrap samples. We first present the matching estimates (from the pre-processing step) and then proceed to the LIV estimates, where we show both the estimated effect of treatment on the treated and the estimated marginal treatment effects (to capture impact heterogeneity).

TABLE 2 Estimates of the average impact of placebo treatment

Outcome	Impact	SE	95% CI
16- to 18-year olds			
Employed at 12 months	−0.167	1.289	(−2.693, 2.358)
Apprenticeship at 12 months	0.144	0.573	(−0.978, 1.267)
19 to 23-year olds			
Employed at 12 months	0.282	2.185	(−4.001, 4.565)
Unemployed at 12 months	−0.137	0.913	(−1.927, 1.653)
Apprenticeship at 12 months	0.098	2.047	(−3.915, 4.111)

Notes: Standard errors based on 200 replications. Asterisks denote statistical significance: *90%, **95%, ***99%.

Italics operate as headings within the table, distinguishing results for the younger age group from results for the older age group.

5.1 | Pre-processing step

Table 3 presents the results of estimating the sample probability of traineeship participation for 16- to 18-year-olds and 19- to 23-year-olds, respectively. The variables included as regressors were chosen on the basis that they were expected to be correlated with the outcomes considered and were also found to be significant predictors of participation. The results in Table 3 confirm the significant differences between trainees and non-trainees evident from their summary characteristics presented earlier.

The coefficients from these probit models were used to generate the propensity scores for the matching step. Matching was implemented without replacement and without imposing common support, so the resulting matched comparison group was equal in size to the number of trainees (for the respective age group).

5.2 | Estimates of the marginal treatment effects

Following this, MTEs were estimated at percentiles of U_D, the participants' subjective resistance to participation. The $\boldsymbol{X}$ variables are the same as those used to estimate the propensity scores in the pre-processing step. The first step is to estimate the propensity score. The results of doing this are presented in Appendix 1.

MTEs were estimated using the approach described in Section 4.1. Figure 2 shows the estimated impacts on the probabilities of being employed or an apprentice 12 months after traineeship start, for 16–18 year olds. As with subsequent charts, the MTE at each percentile is shown by the thick solid line, with 95% bootstrap confidence intervals indicated by the dashed lines. Note that the MTE is only shown for that range of the U_D distribution that is empirically relevant in the sense of accounting for a non-negligible number of participants (operationalised as meaning at least 20 participants). This is for presentational convenience. The estimated MTEs are extremely imprecise in regions of sparse support and including them re-scales the y-axis such that the pattern of impact heterogeneity across the rest of the distribution becomes more difficult to visualise.

For employment, it is clear that the MTE is negative across much of the relevant U_D distribution. The estimates are more precise closer to the centre of the distribution but still do not suggest a statistically significant effect. The degree of variation means it is not possible to say anything about the shape of the MTE curve. For apprenticeships, on the other hand, the MTEs are positive and significantly so towards the centre of the distribution.

Figure 3 shows the corresponding estimates for 19- to 23-year-olds. Here, the estimated impacts on employment are positive for the majority of trainees. However, the confidence intervals are wider than those seen for 16- to 18-year-olds and at no point does the estimated MTE come close to conventional statistical significance. It should be remembered that the qualification criterion for the older age group is such that it selects a group of 19- to 23-year-olds that is lower-attaining than the 16- to 18-year-olds. As such, differences between the age groups in their estimated impacts may reflect impact heterogeneity by skill level. The estimated impacts on the probability of being an apprentice are, if anything, negative. The MTEs are mostly some way short of statistical significance at the conventional level but, with this caveat in mind, the pattern of results is suggestive of an effect that is more negative at higher U_D levels. In other words, the estimated effects are more negative among those more resistant to participation. Lastly, Figure 3 shows that the unemployment MTEs are not statistically significant at any value of U_D. It is notable that, like the employment MTEs, they are positive. This suggests the possibility of an offsetting negative effect on economic inactivity (neither working nor looking for work).

TABLE 3 The results of estimating a probit model of traineeship participation

		16–18s			19–23s		
		Coeff		**SE**	**Coeff**		**SE**
Age:	*16*	*0.142*	***	*0.023*			
	17	*0.121*	***	*0.02*			
	19				0.28	***	0.025
	20				0.214	***	0.022
	21				0.11	***	0.024
	22				0.107	***	0.024
Female		*0.119*	***	*0.016*	−0.084	***	0.014
White		*0.06*	***	*0.021*	0.125	***	0.016
Special educational needs:	*Non-statemented*	*0.027*		*0.02*			
	Statemented	*−0.192*	***	*0.04*			
Excluded from school		*0.07*	**	*0.029*			
School absences:	*None*	*−0.118*	***	*0.032*			
	1	*−0.138*	***	*0.044*			
	2-9	*−0.03*		*0.03*			
	10-24	*−0.01*		*0.032*			
Unauthorised absences:	*None*	*−0.192*	***	*0.043*			
	1	*−0.033*		*0.064*			
	2-9	*−0.136*	***	*0.026*			
	10-24	*−0.002*		*0.022*			
GCSEs at A*-C grade:	*None*	*0.718*	***	*0.026*			
	1	*0.732*	***	*0.028*			
	2	*0.65*	***	*0.03*			
	3	*0.602*	***	*0.033*			
	4	*0.457*	***	*0.028*			
Qualifications:	*None*				0.08		0.067
	Below level 1				−0.238	***	0.067
	Level 1				−0.152	**	0.065
	Levels 2 or 3				−0.749	***	0.066
	Unknown				−0.384	***	0.071
Learning difficulty					0.002		0.017
Status in month before start:	*Education*	*−0.337*	***	*0.019*			
	Training	*0.741*	***	*0.036*			
	NEET seeking EET[‡]	*0.807*	***	*0.031*			
Emp. in month before start					−0.312	***	0.027
Months emp. 1st year before					0.005		0.003
Months emp. 2nd year before					0.008	***	0.003

(Continues)

TABLE 3 (Continued)

	16–18s			19–23s		
	Coeff		**SE**	**Coeff**		**SE**
Months emp. 3rd year before				−0.009	***	0.002
Local unemp, Sep 2013	*0.04*	***	*0.003*	0.01	***	0.003
Urban	*0.163*	***	*0.026*	0.282	***	0.031
Constant	*−2.565*	***	*0.057*	−2.455	***	0.079
N	93,006			350,318		

Notes: Asterisks denote statistical significance: *90%, **95%, ***99%.

†EET is Employment, Education or Training. NEET is Not EET.

Italics denote values for categorical variables.

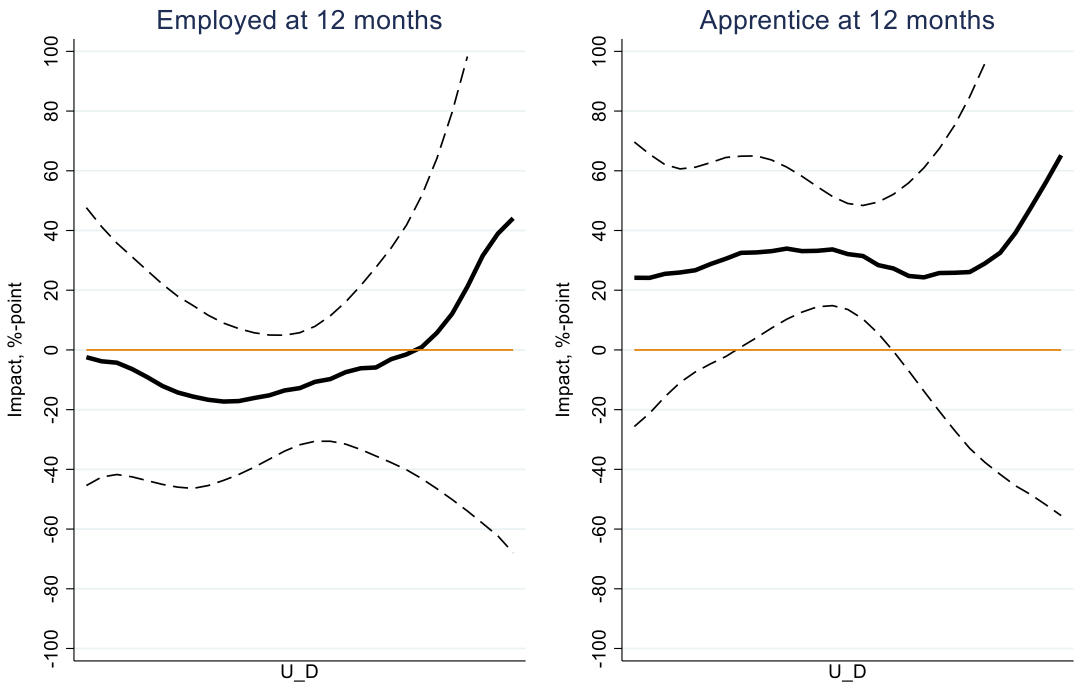

FIGURE 2 Marginal treatment effects for 16–18 s

5.3 | Estimates of the average effect of treatment on the treated

Since the pre-processing step used propensity score matching, differences in outcomes between trainees and their matched comparison group might be viewed as estimates of the ATT. Of course, a causal interpretation of this type relies on the assumption that the analysis has controlled for all relevant influences on outcomes.

By comparison, the MTEs can be aggregated to produce an estimate of the ATT that does not rely on the CIA but instead is based on the assumptions described in Section 4. A practical issue with this is that the MTE estimates are extremely imprecise at some values of U_D. This arises from the fact that, while $F_V(V)$ is uniformly distributed, this is not necessarily true of $F_V(V|D = 1)$, so

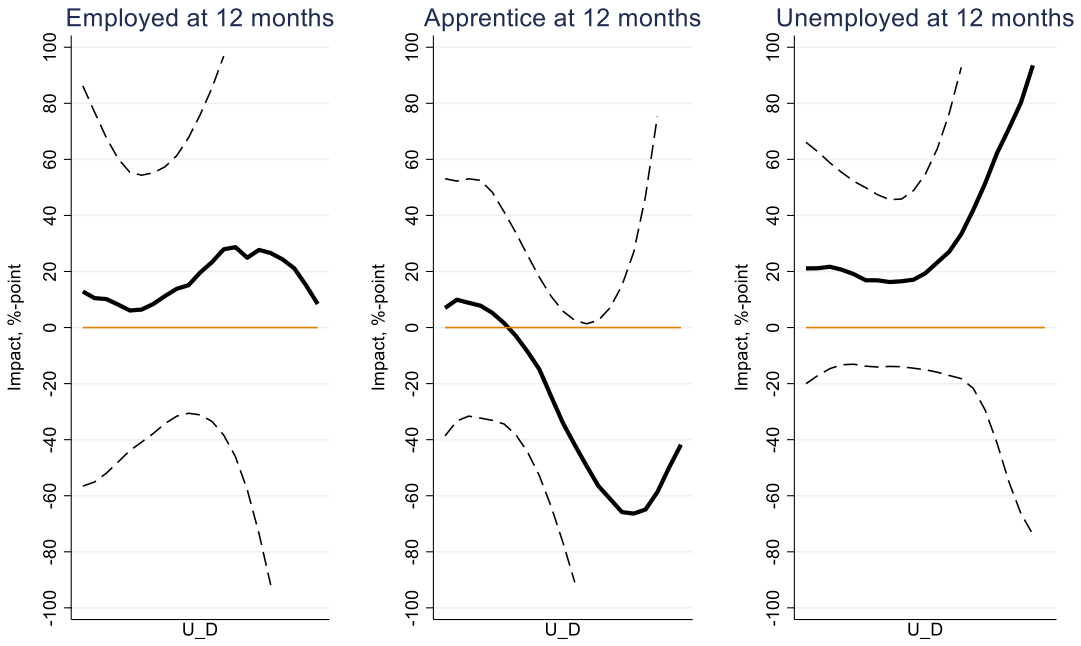

FIGURE 3 Marginal treatment effects for 19- to 23-year-olds

that at some values of U_D there are very few trainees. To address this, two MTE-based ATTs are constructed. The first is a straightforward weighted average of MTEs across the full distribution of U_D (using the weights described in Section 4.3). The second is similar but excludes from the calculation those MTE estimates at values of U_D that account for a very small number of trainees. This is operationalised by ignoring MTEs at values of U_D for which there are fewer than 20 trainees (similar to the approach taken when constructing the charts above). This 'trimmed' estimate is preferred since it reduces the noise arising from imprecise estimates in those regions of the U_D distribution that are empirically less relevant. To provide some support for this approach, we include in Appendix 2 a simulation illustrating the extent to which the reliability of estimates can decline in regions of sparse support.

Estimates of these three ATT variants are shown in Table 4. The upper panel presents results for 16- to 18-year-olds (employment and apprenticeship) while the bottom panel presents results for 19- to 23-year-olds (employment, unemployment and apprenticeship).

The PSM-based estimate of employment impact for 16- to 18-year-olds is close to zero. The LIV-based estimate is roughly −15 percentage points while the preferred estimate—the trimmed LIV—is −5 percentage points. It is notable that the PSM-based estimate is very precisely estimated compared to the LIV-based estimates. This is unsurprising in view of the wide confidence intervals around the MTEs, as discussed above. The value of trimming the LIV-based estimate is apparent since doing so achieves a substantial reduction in the standard error relative to the untrimmed LIV-based estimate, and therefore a narrower confidence interval. Nevertheless, in all three cases, no statistically significant employment effect is found across participants as a whole. By contrast, both the PSM- and LIV-based estimates suggest a positive impact on apprenticeships. Again, the trimmed LIV estimate is smaller. This suggests that participating in a traineeship increased the probability of being an apprentice 1 year later by 16 percentage points.

For 19- to 23-year-olds, there is a stark difference between the PSM- and LIV-based estimates of employment effects. The PSM-based estimates suggest significant positive impacts on employment,

TABLE 4 Estimates of the average impact of traineeship participation, PSM and LIV

Outcome		Impact		SE	95% CI
16-18 year olds					
Employed at 12 months	PSM	0.015	*	0.008	(−0.001, 0.031)
	LIV	−0.146		0.326	(−0.785, 0.494)
	LIV, trimmed	−0.052		0.052	(−0.153, 0.050)
Apprenticeship at 12 months	PSM	0.252	***	0.008	(0.236, 0.268)
	LIV	0.343		0.252	(−0.152, 0.838)
	LIV, trimmed	0.162	***	0.051	(0.062, 0.262)
19- to 23-year olds					
Employed at 12 months	PSM	0.182	***	0.012	(0.159, 0.205)
	LIV	0.285		0.382	(−0.464, 1.035)
	LIV, trimmed	0.045		0.065	(−0.081, 0.172)
Unemployed at 12 months	PSM	0.030	***	0.007	(0.017, 0.044)
	LIV	−0.030		0.267	(−0.554, 0.493)
	LIV, trimmed	0.073	*	0.037	(0.000, 0.147)
Apprenticeship at 12 months	PSM	0.107	***	0.007	(0.094, 0.121)
	LIV	−0.221		0.324	(−0.855, 0.414)
	LIV, trimmed	−0.028		0.044	(−0.114, 0.059)

Notes: Standard errors based on 200 replications. Asterisks denote statistical significance: *90%, **95%, ***99%.

Italics operate as headings within the table, distinguishing results for the younger age group from results for the older age group.

unemployment and apprenticeships while the LIV-based estimates do not provide strong evidence of an impact on any of these outcomes.

When considering impacts, it is possible that estimates are over-stated due to displacement effects. For instance, the guarantee of an interview on completing the traineeship if a role becomes available may disadvantage non-trainees; indeed, in a follow-up survey of 2014/2015 trainees, 19% reported that their current or most recent job was with the same employer as during their traineeship (Fitzpatrick et al., 2017). Evaluations that explore such externalities are few in number and offer mixed findings. Crépon et al. (2013), for example, provide experimental evidence of employment displacement among young, educated job seekers in France. Blundell et al. (2004), on the other hand, find no such effects among young unemployed people in United Kingdom.

Since our estimates suggest no significant effect on employment, it is mainly when considering the impact on apprenticeships, that the displacement effect of traineeships becomes a potential concern. We are not aware of any empirical evidence on apprenticeship displacement. There are two reasons why, intuitively, we might expect it to play little role. First, Crépon et al. (2013) find that employment displacement is more of an issue in a slack labour market. The period considered in this paper was characterised by a rapidly tightening labour market. Official statistics show that the number of unemployed per vacancy fell from roughly 4.5 in August 2013 to 2.4 in July 2015 (see https://tinyurl.com/ruevs6ww). Hence, if apprenticeship displacement is similar to employment displacement in how it varies with labour market tightness, we would expect it to have less effect during the economic conditions prevailing during the period considered. Second, this period also saw a government commitment to increasing the number of apprenticeships, even before the introduction of the 2020 target. This commitment included increased funding for apprenticeships and the use of public procurement

as a lever to increase employer engagement (Department for Business, Innovation and Skills, 2010). Expanding the capacity for apprenticeships in this way reduces the likelihood of a displacement effect, which might be more expected were the availability of apprenticeships fixed.

A further reason why any apprenticeship displacement effects are likely to be minor is that the number of trainees is small relative to the number of apprentices as a whole. In 2013/2014, there were 440,400 apprenticeship starts compared to 10,400 traineeship starts. Maximum displacement would occur under the scenario whereby all apprenticeships filled by trainees would otherwise have been taken by non-trainees. Among the full intake of trainees, 1,850 were apprentices 12 months after starting their traineeship; less than half a percent of apprenticeship starts. Consequently, while displacement may exist, it would not alter the estimated impact by more than half a percentage point and so does not qualitatively alter the interpretation of the results.

A separate concern arises from the possibility that the definition of trainee and non-trainee groups implicitly conditions on future outcomes. If individuals who are successful in the labour market during 2013/2014 are less likely to select into traineeships then including all participants in the trainees group—regardless of the point in the year at which they participate—may create a situation whereby non-participants have systematically better outcomes than trainees would have had in the absence of participation. If so, impact estimates that do not control for this will be biased downwards.

At its heart, this difficulty arises from using a static evaluation technique to analyse a dynamic participation decision. While dynamic techniques exist (Abbring and van den Berg (2003) and Heckman and Navarro (2007) provide two prominent contributions), these do not allow impact heterogeneity to be considered to the same degree as static models. Using imputed pseudo-start dates for non-participants increases the comparability between participant and non-participant groups. By construction, the pseudo-start dates have the same distribution as the start dates among trainees. This reduces the degree to which any systematic difference may exist since the information on circumstances immediately prior to (pseudo-) participation is included among the regressors of the propensity score model, thereby ensuring balance between participants and (matched) non-participants groups at this point.

As a more general comment, the results have revealed differences between ATT estimates based on a selection-on-observables assumption (the PSM results) and ATT estimates that are based on a selection-on-unobservables assumption (the LIV results). In view of this, it is natural to consider which assumption is more credible.

Because of the preliminary matching step, LIV estimation is applied to a pre-processed sample that, under the selection on observables assumption, should have already adequately controlled for selection. The fact that a significant relationship between participation and distance remains in the matched sample and that LIV estimation finds some significant impacts using this pre-processed data suggests a role for unobservables that cannot be addressed through matching.

This is perhaps not a surprising finding. The assumption that all important influences on participation can be observed in the data is a strong one. Despite the data being rich, they can only imperfectly proxy some of the attitudinal orientation towards participation. This is particularly relevant due to the fact that participation is voluntary rather than compulsory, so subjective choice plays a greater role. Furthermore, as noted already, we are not able to observe the eligibility criteria fully for non-trainees but trainees will, in principle, meet these criteria by virtue of being accepted onto the programme.

6 | CONCLUSION

This paper has presented the results of evaluating the impact of traineeships on employment and apprenticeships. It has used LIV to allow for the possibility that unobserved factors influence the

decision to participate and also to gain an insight into impact heterogeneity. These impacts are always less positive than those found under the assumption that all important influences are observed. This suggests that individuals with more favourable unobserved characteristics select into traineeships. Such positive selection is unsurprising given the eligibility criteria. Specifically, participants must be motivated and have a reasonable chance of securing an apprenticeship or a job. Both of these judgements are made by traineeships providers and are not recorded in the administrative data.

The substantive findings suggest that traineeships, a programme intended to help young people enter work or an apprenticeship, have had mixed results. Employment among 16- to 18-year-olds was, if anything, reduced and among 19- to 23-year-olds the apparent positive impact was too imprecisely estimated to be regarded as reliable. Apprenticeships, on the other hand, were significantly increased among 16- to 18-year-olds but not 19- to 23-year-olds. For the younger group, this positive effect was seen across the full distribution of resistance to participation. Among the older age group, the results suggest that, for those more resistant to participating, traineeships may actually reduce the probability of becoming an apprentice.

It is important to bear in mind that, while these results estimate the impacts of traineeship participation relative to non-participation, this should not be interpreted as the impacts of traineeship participation relative to doing nothing. Non-participants may be involved in alternative activities that might be expected to influence outcomes in their own rights. While the estimation approach aims to ensure trainees and non-trainees are similar with regard to their circumstances immediately preceding participation (or pseudo-participation), this approach is static and does not prevent non-trainees from subsequently participating in other activities. In view of this, the results should be interpreted as capturing the impact of traineeships relative to the type of activity among non-trainees which, for some, will involve education or training. This has a pragmatic relevance, indicating how traineeships affect outcomes relative to existing provision.

With regard to how we assess the overall performance of the programme, a fundamental question is whether promoting employment among 16- to 18-year-olds is an optimal aim when set against the alternatives of, for example, an apprenticeship. Longer-term, one might expect the latter to be associated with higher earnings. In this light, a negative employment effect for younger trainees might not be viewed as a negative social outcome and the positive impact on apprenticeships may be sufficient to regard the programme as a success. For the older age group, there is less evidence of programme effectiveness.

ACKNOWLEDGEMENT

We are grateful to the Department for Education for granting access to the data used in this study and for funding the research upon which this paper is based. We thank Bart Cockx, Thomas Cornelissen, participants at the IZA/UCPH workshop on the evaluation of labour market policies in Copenhagen and the 2019 Administrative Data Research conference in Cardiff, the Editor, Associate Editor and three anonymous reviewers for very helpful comments and suggestions. The usual disclaimer applies.

ORCID


Richard Dorsett https://orcid.org/0000-0002-4180-8685

REFERENCES

Aakvik, A., Heckman, J. & Vytlacil, E. (2005) Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics*, 125, 15–51.

Abbring, J. & van den Berg, G. (2003) The non-parametric identification of treatment effects in duration models. *Econometrica*, 71, 1491–1517.

Bibby, D., Buscha, F., Cerqua, A., Thomson, D. & Urwin, P. (2014) Further development in the estimation of labour market returns to qualifications gained in English Further Education using ILR-WPLS Administrative Data. Research paper 195. Department for Business, Innovation and Skills. (Available from).

Björklund, A. & Moffitt, R. (1987) The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 69, 42–49.

Blundell, R., Costa Dias, M., Meghir, C. & Van Reenen, J. (2004) Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association*, 2, 569–606.

Brave, S. & Walstrum, T. (2014) Estimating marginal treatment effects using parametric and semiparametric methods. *Stata Journal*, 14, 191–217.

Brinch, C.N., Mogstad, M. & Wiswall, M. (2017) Beyond LATE with a discrete instrument. *Journal of Political Economy*, 125, 985–1039.

Card, D. (1993) Using geographic variation in college proximity to estimate the return to schooling. Working paper w4483. National Bureau of Economic Research. (Available from https://www.nber.org/papers/w4483.pdf).

Carneiro, P., Heckman, J. & Vytlacil, E. (2011) Estimating marginal returns to education. *American Economic Review*, 101, 2754–81.

Cavaglia, C., McNally, S. & Ventura, G. (2018) *Do apprenticeships pay? Evidence for England*. Oxford: Oxford Bulletin of Economics and Statistics.

Cornelissen, T., Dustmann, C., Raute, A. & Schönberg, U. (2016) From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, 41, 47–60.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R. & Zamora, P. (2013) Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Quarterly Journal of Economics*, 128, 531–580.

Department for Business, Innovation and Skills. (2010) Skills for sustainable growth: strategy document. (Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32368/10-1274-skills-for-sustainable-growth-strategy.pdf).

Department for Education. (2013) Traineeships: Supporting young people to develop the skills for Apprenticeships and other sustained jobs. A discussion paper (Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/197592/Traineeships_discussion_paper.pdf).

Department for Education. (2019) Further Education and Skills. England: Department for Education 2018/19 academic year (Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/848534/FE_and_Skills_commentary_November_ 2019.pdf).

Fitzpatrick, A., Coleman, E., Shanahan, M., Coleman, N. & Cordes, A. (2017) Traineeships: Year Two Process Evaluation. Department for Education Research report 694. (Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/847348/Traineeships_Year_Two_Process_Evaluation.pdf).

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998) Characterizing selection bias using experimental data. *Econometrica*, 66, 1017–1098.

Heckman, J. & Navarro, S. (2007) Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, 136, 341–396.

Heckman, J., Urzua, S. & Vytlacil, E. (2006) Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88, 389–432.

Heckman, J. & Vytlacil, E. (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

HM Government. (2015) English Apprenticeships: Our 2020 Vision. (Available from https://www.gov.uk/government/publications/apprenticeships-in-england-vision-for-2020).

Ho, D., Imai, K., King, G. & Stuart, E. (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15, 199–236.

Imbens, G. & Wooldridge, J. (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47, 5–86.

Leuven, E. & Sianesi, B. (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components S432001,

Boston College Department of Economics. Revised 01 Feb 2018. (Available from http://ideas.repec.org/c/boc/bocode/s432001.html).

McIntosh, S. & Morris, D. (2016). Labour Market Returns to Vocational Qualifications in the Labour Force Survey. Discussion Paper 002. Centre for Vocational Education Research, London School of Economics. (Available from http://cver.lse.ac.uk/textonly/cver/pubs/cverdp002.pdf).

Moffitt, R. (2008) Estimating marginal treatment effects in heterogeneous populations. *Annales d'Economie et de Statistique*, 91–92, 239–261.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. section 9. *Statistical Science*. 5, 463–480. [1990] reprint. Translated by Dabrowska, D. and Speed, T.

Rosenbaum, P. & Rubin, D. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rubin, D. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66, 688–701.

How to cite this article: Dorsett R. & Stokes L. (2021) Pre-apprenticeship training for young people: Estimating the marginal and average treatment effects. *J R Stat Soc Series A*, 00:1–24. https://doi.org/10.1111/rssa.12697

APPENDIX 1

PROBIT REGRESSIONS OF TRAINEESHIP PARTICIPATION USING PRE-PROCESSED DATA

		16–18s			19–23		
		Coeff		**SE**	**Coeff**		**SE**
Age:	*16*	*0.056*		*0.040*			
	17	*0.020*	*	*0.033*			
	19				*−0.044*		*0.056*
	20				*−0.004*		*0.053*
	21				*0.014*		*0.056*
	22				*0.001*		*0.055*
Female		*0.065*		*0.029*	*−0.032*	*	*0.031*
White		*0.051*		*0.040*	*0.040*		*0.038*
Special educational needs:	*Non-statemented*	*−0.024*	*	*0.031*			
	Statemented	*0.029*		*0.066*			
Excluded from school		*0.061*		*0.047*			
School absences:	*None*	*0.110*		*0.048*			
	1	*0.167*		*0.074*			
	2-9	*0.092*		*0.047*			
	10–24	*0.049*		*0.046*			
Unauthorised absences:	*None*	*0.055*		*0.086*			
	1	*0.288*		*0.136*			
	2–9	*0.012*		*0.051*			

		16–18s			19–23		
		Coeff		SE	Coeff		SE
	10–24	*0.008*		*0.038*			
GCSEs at A*-C grade:	*None*	*−0.092*	**	*0.047*			
	1	*−0.097*	*	*0.051*			
	2	*−0.064*		*0.054*			
	3	*−0.095*		*0.062*			
	4	*−0.074*	*	*0.059*			
Qualifications:	*None*				*−0.180*		*0.142*
	Below level 1				*−0.088*		*0.139*
	Level 1				*−0.196*		*0.135*
	Levels 2 or 3				*−0.178*		*0.141*
	Unknown				*−0.264*		*0.151*
Learning difficulty					*0.064*		*0.038*
Status in month before start:	*Education*	*−0.023*	*	*0.036*			
	Training	*0.091*		*0.044*			
	NEET seeking EET†	*0.033*		*0.042*			
Emp. in month before start					*−0.106*	**	*0.060*
Months emp. 1st year before					*0.013*	**	*0.007*
Months emp. 2nd year before					*0.000*	**	*0.006*
Months emp. 3rd year before					*0.004*	**	*0.006*
Local unemp, Sep 2013		*−0.013*	***	*0.006*	*−0.015*	***	*0.007*
Urban		*−0.424*	***	*0.051*	*−0.299*	***	*0.078*
Distance to provider		*−0.141*	***	*0.016*	*−0.071*	***	*0.009*
Distance, squared		*0.006*	***	*0.001*			
Constant		*0.637*		*0.108*	*0.681*		*0.186*

† EET is Employment, Education or Training. NEET is Not EET.* 90%, **95%, ***99%.

APPENDIX 2

SIMULATION ILLUSTRATING THE RELIABILITY AND POWER OF THE ESTIMATED MTEs

The requirements of a continuous instrument are more demanding than those of a binary instrument. Moffitt (2008) points out that instruments can be strong in some ranges of the distribution of individuals' subjective costs, U_D, but weak elsewhere. Our exploration of the relationship between participation and distance shows a negative slope at most distances. There is a slight deviation from this for younger learners but only at distances that account for a relatively small proportion of trainees.

Even if the relationship is seemingly strong, there is the question of whether there are enough observations at a given point to regard the resulting estimated MTEs as statistically significant. To explore this, we conducted a simulation study. Each replication drew a sample of 10,000 individuals, whose treatment participation was determined on the basis of $d^* = 0.5x + 0.5z - \upsilon$, where x, z and υ are independent standard normal. Participation is denoted by d, where $d = 1(d^* > 0)$. An outcome, y, was constructed as $y = .2x + \beta d\Phi(\upsilon)^2 + \epsilon$, where Φ is the standard normal cumulative distribution function and ϵ is standard normal. This sets impact heterogeneity to be a quadratic in the quantiles of υ. The parameter β was chosen such that there is an 80% chance of detecting the overall impact at the 5% level of significance (i.e. the conventional power and significance levels).

Figure A1 presents estimated MTEs. The impacts from the data-generating process are shown as a red line. The LIV estimates are shown with a black line. Within each replication, standard errors were generated by bootstrapping (100 replications). Hence, the simulation was nested in the sense that each replication involved a simulation of its own. Those MTE estimates for which, across all 'outer' replications, at least 80% suggest a significant impact at the 95% level are marked on the chart by a diamond. That is, diamonds highlight those MTEs for which the standard 80% power level has been attained. In addition, the mean distributions of the propensity score for participants and non-participants are shown (blue and green respectively), with their overlap also visible

The chart shows that the MTEs are close to the true values for much of the distribution of U_D, and that they capture the curvature introduced by the quadratic specification of the outcome equation. However, the MTE estimates are erratic at low and high values of U_D, where there is little common support (that is, where the overlap in the propensity score distributions is less). The lack of diamonds indicates that statistical power is also reduced in the tails of the overlap distribution. The suggestion from the simulations is that MTEs are more reliably estimated in the region of the U_D distribution where both participants and non-participants are adequately represented. This finding provides some support for the trimming approach used in the impact analysis.

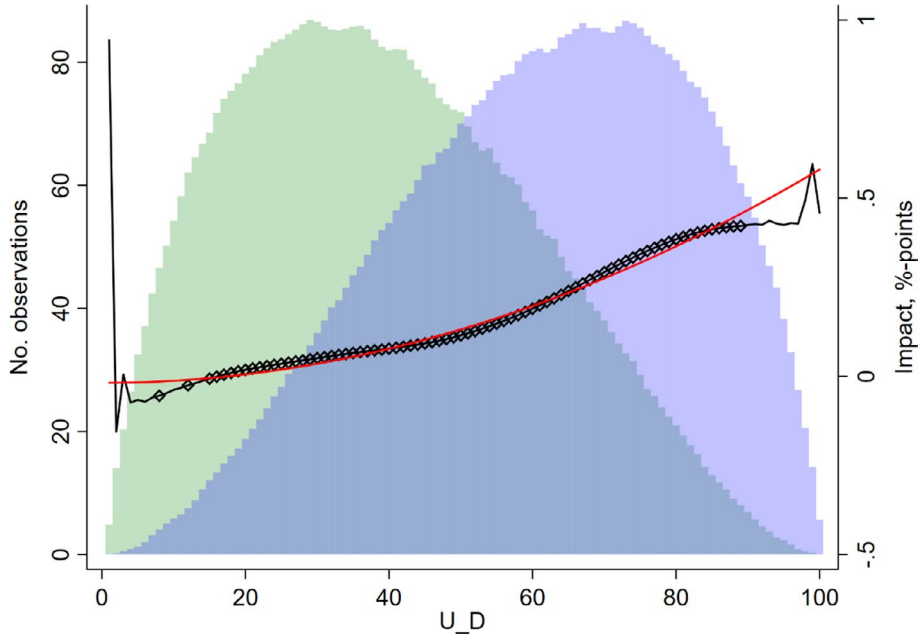

FIGURE A1 Simulated marginal treatment effects (MTEs) with quadratic impact heterogeneity (100 replications)—red line shows true MTEs, black line shows local instrumental variable estimates with markers indicating 80% power; distribution of participants in blue, non-participants in green