Leading education and social research
Institute of Education
University of London

Department of Quantitative Social Science

# Matched panel data estimates of the impact of Teach First on school and departmental performance

**Rebecca Allen**
**Jay Allnutt**

DoQSS Working Paper No. 13-11
September 2013

## Disclaimer

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science, Institute of Education, University of London
20 Bedford Way, London WC1H 0AL, UK

# Matched panel data estimates of the impact of Teach First on school and departmental performance

**Rebecca Allen[1] and Jay Allnutt[2]**

**Abstract**

In this paper we evaluate whether the placement of Teach First's carefully selected, yet inexperienced new teachers into deprived secondary schools in England has altered the educational outcomes of pupils at the age of 16. Our difference-in-difference panel estimation approach matches schools participating early on in the scheme to those within the same region. We find the programme has not been damaging to schools who joined and most likely produced school-wide gains in GCSE results in the order of 5% of a pupil standard deviation or around one grade in one of the pupil's best eight subjects. We estimate pupil point-in-time fixed effect models to identify core subject departmental gains of over 5% of a subject grade resulting from placing a Teach First participant in a teaching team of six teachers.

**JEL classification:** I20, I21, C23

**Keywords:** Teacher preparation, school performance, teacher effectiveness

[1] Corresponding author: Department of Quantitative Social Science, Institute of Education, University of London, 20 Bedford Way London, WC1H 0AL (R.Allen@ioe.ac.uk)
[2] Teach First, 4 More London Riverside, London, SE1 2AU (jallnutt@teachfirst.org.uk)

Acknowledgements:
Thanks are due to Teach First and the Department of Education for providing the data that is used in this study. Thanks also to John Micklewright and others for reading drafts of this paper. Rebecca Allen's time has been funded by the Economic and Social Research Council through a Future Research Leaders award (number R00909). Part of the analysis in this paper is based upon Jay Allnutt's MSc dissertation, which was written whilst a secondary school teacher in London. He now works for Teach First.

## Introduction

Teach First, a programme which has many similarities with Teach for America and other 'Teach for All' schemes across the world, has been placing graduates into schools in challenging circumstances since 2003. These schools have traditionally struggled to recruit high quality teachers and maintain low teacher turnover (Allen et al., 2012). The Teach First participants commit to teach up to 80% of a standard teaching load for two years following six weeks of intensive basic training and are able to achieve fully qualified teacher status by the end of the programme, with in-school and partner university support throughout. After two years, over half the cohort chooses to remain in teaching in a Teach First-eligible school for at least a third year, with the rest pursuing careers in other education and non-education-related fields. Over the past decade the scheme has grown from 186 graduates in 2003/4 to 1000 graduates in 2012/13, has extended its reach from London into what will be nine English regions and Wales from September 2013, has expanded its recruitment to include later career participants and since 2008 has placed participants in primary schools.

As increasing numbers of Teach for All programmes are established across the world it is important its success is evaluated and understood in the very wide variety of contexts under which it operates. The programme appears to challenge the traditional model of college-based teacher training and asserts that it is possible for graduates with limited intensive training to thrive in often very challenging school environments. In England and Wales, Teach First is a key programme on which successive governments have placed a great deal of emphasis. It is expected to continue to grow over coming years and has inspired a wider move towards school-led teacher training across the country.

In this paper we evaluate whether the placement of these graduates has altered the educational outcomes of pupils at the age of 16 for the first three years of a school's participation in the scheme. We do not have matched teacher-pupil data in England so are forced to identify an impact across the whole school and within the key departments within which these graduates are placed. There is no random element in the sign-up of schools to the scheme so the obvious threats to validity are that (i) headteachers who choose to join this scheme are particularly dynamic and so presiding over improving schools; (ii) conversely, that schools using the scheme have particularly severe teacher recruitment and retention problems that may reflect underlying difficulties at the school; and (iii) that Teach First launched in the London region which was subject to multiple interventions to improve pupil attainment that ran concurrently with Teach First.

To identify any school impact we match early participating Teach First schools to those within

the same region which participate later in the scheme and then estimate impact within a difference-in-difference panel to control for any further time-invariant endogenous and unobservable variables which could otherwise bias estimates. To identify departmental impacts (and assuming no spillovers across the school) we (i) estimate triple difference estimators that compare changes in exam outcomes across departments within the same school and (ii) estimate pupil point-in-time fixed effect models to directly account for unobserved pupil characteristics. Whilst our approaches to estimating departmental impacts are arguably more robust than our school-level impacts, all our methods reduce risk of biased estimates compared to the matched multi-level cross sectional models used by Muijs et al. (2010) in the only existing quantitative evaluation to date.

The analysis in this paper is limited in the sense that it can only estimate the impact upon attainment of a school hiring a Teach First participant, but does not draw conclusions about the relative merit of this programme over other teacher training routes or about the value of investment in the programme compared to other programmes or interventions. We do not know what type of teachers are displaced when a school joins the scheme, but it is likely they are a combination of newly qualified teachers from other training routes, experienced teachers, and temporary/substitute teachers (we give descriptive evidence on this in the discussion section).

This is critical to interpretation of the estimates since we calculate the impact compared to the school's results in the two years prior to joining the scheme and compared to similar schools who have not yet joined the scheme. The Teach First participants are likely to have stronger academic credentials but will be less experienced than the average teacher they displace. However, there is no UK evidence on whether either of these characteristics is associated with more or less effective teaching. Overall, the international literature suggests little relationship between academic credentials and teacher effectiveness (e.g. Aaronson et al., 2007; Hanushek et al., 2005). Furthermore, it is widely accepted that teachers are at their least effective in their first year of teaching and then improve for the next four years (Clotfelter et al., 2007; Hanushek et al., 2005; Rivkin et al., 2005; Rockoff, 2004). However, even if Teach First participants are inexperienced, they still may be more effective than the alternative if (i) their replacements are also relatively inexperienced teachers; and/or (ii) on the margin they allow schools to displace or avoid recruiting teachers they know are likely to be weak.

## Background Literature

There are a number of qualitative studies of Teach First that tend to evaluate the quality of the training and support the programme offers participants, rather than the direct impact on pupil experience, *per se*. The English schools inspectorate, Ofsted, judges the Teach First training

provision as outstanding in every category and particularly praises the scheme in its encouragement of "*… the participant's relentless focus on the learning and progress of their students*" (Ofsted, 2011, p. 8). Importantly though, they also report that their visits suggest that participants may be having an impact "*on the professional development of other staff as well as on their students*" (p. 5). This possibility of spillover effects onto other teachers in the schools, and whether it occurs only within departments or encompasses wider school ethics and practice is critical to our estimation strategy, as discussed in the method section.

The qualitative studies also make it clear that Teach First participants influence students beyond their exam grades measured in this paper. An earlier report by Ofsted stated that *"[TF teachers'] professional commitment to the students in lessons and in the broader life of the school, such as clubs, was one of the major factors in the beneficial impact Teach First participants had on the schools in which they were placed"* (Ofsted 2006, p. 10). Similarly, Hutchings et al. (2006) found that *"Schools reported that Teach First teachers had a positive impact, delivering high quality lessons, undertaking extra-curricular activities and in some cases reinvigorating other staff."* (p. 75).

The only existing quantitative study of the impact of Teach First finds consistently positive and substantial impacts on attainment at age 16 of around one third of a GCSE grade per subject from two years onwards after joining the programme (Muijs et al., 2010). The estimation strategy is significantly different to the one employed in this paper and, in particular, it is not clear whether it sufficiently accounts for trends in GCSE performance at schools prior to joining the scheme. They identify the association between the number of Teach First participants in a school and average GCSE attainment using a propensity score matched sample of non-participating schools. This relationship is estimated in a series of cross-sectional multi-level models with lagged 3-year weighted average background control variables included to ensure stability and take account of cohort effects. Only schools that had participated in the programme for at least four of the following six years were included in the analysis, which introduces an obvious positive selection bias into the estimates since any schools that decided to discontinue involvement due to a poor experience are dropped from the analysis.

There is now a great deal of evidence on Teach for America, which has been operating since 1990. Teach First and Teach for America share similar missions and structures so the research findings are directly relevant, although there are some notable differences. First, the size of Teach for America results in far less centralisation in the type and level of support available to participants, which has led to differences in approach to certification and partnership between Teach for America regions. Second, a large percentage of Teach for America corp. members are

placed in elementary schools, whereas almost all Teach First participants have been placed in age 11-16/18 secondary schools. The former has a methodological advantage for evaluation of pupil outcomes because pupils largely stay with a single teacher for all or at least large parts of the curriculum, whilst high/secondary school pupils have a large number of teachers and are often grouped differently for subjects.

Evidence from Teach for America for elementary and middle schools is mixed, but typically shows a small, statistically significant positive effect in maths and science, with inconsistent results for other subjects (key studies include Boyd et al., 2006; Darling-Hammond et al., 2005; Kane et al., 2008; Raymond et al., 2001). Decker et al. (2004) stands out as the most robust of these evaluations since it avoids endogenous allocation of teachers to students through an experimental design that randomises classroom assignment of a Teach for America corp. member with another teacher. They studied grades 1-5 across six regions and found that *"average student test scores in Teach for America classrooms were higher than in control classrooms in mathematics and were about the same as control classrooms in reading. These results are found broadly across subgroups of teachers and students and are robust to a variety of tests and assumptions."* (Decker et al., 2004, p. 29). The results show a larger gain in achievement when the Teach for America teacher is compared to other novice teachers (teachers with fewer than 3 years teaching experience) than with veteran teachers (teachers with more than 3 years teaching experience). Antecol et al. (2013) reanalyses the same data to confirm Teach for America impacts are fairly uniform across the ability distribution and are particularly large for female students in maths.

The four studies which focus on High School outcomes all show that there is a positive achievement impact of Teach for America across all subjects, which concords with a perspective that a teacher's academic qualifications are more important with older students (Goldhaber and Brewer, 2000). Xu et al. (2009) use a North Carolina matched teacher-grade 10-12 student dataset with student fixed-effects to show that *"... Teach for America teachers are more effective, as measured by student exam performance, than traditional teachers ... [and] this exceeds the impact of additional years of experience, implying that Teach for America teachers are more effective than experienced secondary school teachers."* (p. 3) (see US Department of Education IoES, 2010, for criticisms of this paper). Schoeneberger et al. (2009; 2011) use fixed effects GLS and multi-level models to show Teach for America teachers perform favourably compared to other teachers teaching the same course in the same school. Henry et al. (2010) show a greater impact of Teach for America teachers in North Carolina across English, maths and science, compared to teachers with fewer than five years experience who entered via other routes. Ware et al. (2011) also find a positive and significant result for Teach for America teachers compared

to other novice teachers in Texas.

This US high school evidence suggests that positive impacts should be found in England since placement is overwhelmingly with older students. However, the relative success of the programme within a local context clearly depends not just on the efficacy of its implementation, but more importantly on local teacher labour market conditions and on the quality of the pre-existing recruits into new teaching positions in deprived schools. Furthermore, while the data quality is high in the US studies due to the availability of matched teacher-pupil data, many of these studies do make strong assumptions regarding absence of spillover impacts of Teach for America corp. members across the school and as a result may actually understate the success of the programme.

## Data

Ideally, teacher-pupil matched data would be available to allow us to directly measure the success of pupils taught by a Teach First participant compared to those taught by others. Unfortunately this is not routinely collected anywhere in the UK so instead our analysis aggregates datasets to create school and departmental level data to estimate overall direct and indirect impacts on the school overall. We combine two sources for this paper: Teach First's database of participant records from 2003/4 to 2012/3 and the National Pupil Database for England, which is collected and maintained by the Department for Education. The Teach First database provides us with one record for every participant since the programme started in 2003. From this we draw details of the participant's school allocation, their year of first participation, whether they withdrew prematurely from the scheme and the main subject they teach in the school. Table 1 shows the number of schools who take part in the Teach First programme by year of first participation. It illustrates the gradual roll-out across regions, starting with London in 2003/04 and followed by the North West in 2006/07, West Midlands in 2007/08 and the East Midlands and Yorkshire in 2009/10. (A North East region was launched in 2010/11 and South East in 2012/13, but data was not available for inclusion in analysis here.)

We create two school-level indicators of Teach First participation. The first metric indicates the year of first participation by the school and also separately for English, maths and science departments within the school. It is a binary indicator ($TF_j$) that takes the value one once the school has participated in Teach First for the first time, regardless of whether they continue to use the scheme in the future, thus excluding them from the pool of potential control schools. It does not distinguish between a school that chooses to take on dozens of Teach First participants each year from a school that takes on just one in only one year. This is because this response by schools beyond the first participation decision is endogenous to their experience of the Teach

First programme itself. That is, those schools with a positive experience of the participants they were allocated are more likely to continue with the programme, which would lead to an upward bias on estimates. We only measure whether Teach First participants were present in the school in the pupil's final year of compulsory schooling. Ideally, we would lag this variable to take account of the pupil's exposure to Teach First participants over five years with greater weight placed on the later years of schooling, but we do not yet have enough data to do so. The lack of lag does mean that the impact of Teach First could be understated in the models estimated here.

Our second treatment variable measures the dosage of Teach First participants for the first three academic years (and separately for the core English, maths and science departments) ($TFintensity_j$). We count the number of Teach First teachers present as either first or second year participants and reweight the variable to adjust for relative school size (i.e. x cohort size/average cohort size). Unfortunately we do not know whether participants remain in their placement school for a third or subsequent year, which means it is perfectly possible for this intensity variable to fall in year 3. To illustrate this problem, the mean (s.d.) of the intensity variable following the school's first participation is 3.5 (2.2), 5.5 (4.0) and 4.3 (4.3) Teach First teachers for years 1, 2 and 3, respectively. This intensity variable bears similarities to that used by Muijs et al. (2010) and is intuitively appealing in the sense that a larger number of Teach First participants is likely to have a greater impact on school culture and pupil performance. However, it is clearly endogenous with a likely upward bias on estimates (i.e. schools with the capacity to implement a successful Teach First programme are likely to take on larger numbers of participants).

We extract a pupil record from the National Pupil Database for every 16 year old student at a state-maintained school for the years 2002 through to 2012. This data is collapsed to a school-level dataset for the majority of the analysis in order to implement a panel data approach, though we also retain pupil-level records for the pupil fixed effects estimation described in the next section. These records provide us with a prior attainment score at age 11 in English, maths and science, indicators of the child's gender, age in months, ethnicity, free school meal status, deprivation of home neighbourhood (IDACI) and special educational needs.

The National Pupil Database gives us a series of key exam outcomes for students at the end of compulsory education (age 16). We use a broad measure of the pupil's exam performance across their best eight subjects in GCSE exams, standardising to a (pupil-level) mean of zero and standard deviation of one (*capped GCSE z-score*). The school-level average capped GCSE z-score has a standard deviation of about 0.5. We also report a threshold measure of the proportion of students gaining five or more GCSEs at grades A*-C, including English and maths. This threshold

measure is the main one used in league tables over the period of data used and is subject to considerable grade inflation/improvement over time.

Core subject departmental performance is measured by taking the child's best grade in the subject, scored on a scale of 0 (=U or no entry) to 8 (=A*). This is relatively uncontroversial for English and maths but may poorly reflect teaching quality for science where students take a variety of different exam types and anywhere between zero and three qualifications.

The two datasets are combined using school code indicators to create a longitudinal panel of school characteristics and outcomes from 2001/02 to 2011/12 (and with the Teach First indicators stretching to 2013/14 for schools in existing regions). For much of the analysis the panel is reshaped to indicate school characteristics and outcomes for the two years prior to Teach First participation and for three years following first Teach First participation. Table 2 summarises the 2002/03 characteristics of the schools by year of first participation in the Teach First programme. This is the year prior to the launch of Teach First. The third column of statistics demonstrates that participating schools have relatively poor capped GCSE z-scores. This is by construction since until 2010 Teach First schools had to meet one or both of the following criteria: they had to have below 25% 5 A*-C grade GCSE pass rate and/ or greater than 30% of its pupils were claiming free school meals (FSM). It is noticeable that the typical characteristics of Teach First participating schools has neither improved nor declined as the programme has expanded, which is not particularly surprising given eligibility restrictions and regional roll-out. This is helpful because our principal estimation approach matches early participating Teach First schools to those participating for the first time in later cohorts. It also aids any generalisations we might want to make from estimates using the first seven cohorts.

## Method

In this section we describe our approach to estimating the impact of Teach First participants on the schools and the departments in which they teach. The matching of participating schools to those schools who have not yet taken up the programme is central to dealing with quite serious potential endogeneity bias. The matching is combined with a school-level difference-in-difference regression, with school fixed effects soaking up unobserved time-invariant characteristics and background control variables that are intended to account for observable time-variant population changes at the school. The estimation of departmental impacts relies on less strong identification assumptions and we compare estimates via a standard difference-in-difference model with triple differences estimation and pupil point-in-time fixed effects. Each of these approaches relies on slightly different assumptions and comparison groups.

Our main estimation strategy aggregates pupil-level data to the level of the treatment (i.e. the school) in order to implement a difference-in-difference design using panel data. We choose to do this rather than use multi-level modelling of pupil-level data, as was implemented by Muijs et al. (2010), to exploit a quasi-experimental design strategy that provides a more robust estimation of outcomes which can thus be more confidently argued to capture the effect of observable factors, and not be confounded by time-invariant unobservable factors (Glewwe et al., 2011).

We introduce an education production function to illustrate the diverse means by which the introduction of Teach First participants might affect the attainment, $y$, of pupil $i$ in subject $h$ in school $j$ at time $t$. Suppose we can separate the impact of the pupil's own (time varying or otherwise) characteristics $X_{jit}$ and prior attainment $y_{hijt-1}$ from the impact of the school they attend, $S_{hjt}$.

$$y_{hijt} = f(y_{hijt-1}, X_{ijt}, S_{hjt}(T_{hjt}, D_{hjt}, R_{jt}, L_{jt}))$$

The impact of the school on the pupil's attainment in subject $h$ in turn depends on their subject teacher's effectiveness, $T_{hjt}$, departmental ethos and decision-making regarding curriculum and exam entry, $D_{hjt}$, the non-teaching resources in the school, $R_{jt}$, and the quality of school leadership, ethos and whole-school activities, $L_{jt}$. Of course, all these aspects of the school experience are inter-linked, especially in the long-run. For example, strong school leadership might influence teacher quality through attracting good teachers, investing in effective training and motivating staff to work hard.

A school's participation in Teach First might influence pupil $i$'s attainment in subject $h$ in a number of ways. First, pupil $i$ may be taught by a Teach First participant who is more or less effective than the teacher they would otherwise have been allocated had the school not joined the scheme (i.e. directly through $T_{hjt}$). Second, even if not directly taught by one, the presence of a Teach First participant may raise or lower the general standard of teaching in the department, though raising expectations or the creation of new shared resources or negatively through other teachers' need to provide mentoring time and support to the inexperienced teacher (i.e. indirectly through $D_{hjt}$). Finally, Teach First participants may engage in activities that impact on the ethos of the school more widely (i.e. indirectly through $L_{jt}$). Unfortunately, without linked teacher-pupil data it is difficult to distinguish between the first two of these three mechanisms so instead we estimate the impact of Teach First on departments and on the school more generally.

## Propensity Score Matching

We use matching to identify an untreated set of schools so that we can create a counterfactual which yields an unbiased average treatment effect on the treated (ATT), provided conditional independence between treatment and outcomes given a set of observable characteristics, holds. Our matching strategy attempts to deal with a number of potential sources of selection bias, without the imposition of function form assumptions or risk of insufficient common support. Participating in Teach First may reflect difficulty in recruiting teachers or high turnover, which in turn suggests higher levels of deprivation (cf Allen et al., 2012) or local reputational difficulties. Conversely, headteachers who are attracted to the program may be different to others (e.g. less conservative) and this might be correlated with improvements in effectiveness at the school. Either way, it suggests that Teach First participating schools would best be matched to others that choose to join the program at some point in the future.

Within the group of future participating schools, we may be concerned that schools joining at a later stage are somehow less committed or suited to the programme, in which case the best match would be to those future participating schools who were geographically blocked from taking part in the early years because the program did not yet operate in their area. However, on testing this type of match we encounter two problems: first, the match across regions is poorly balanced on ethnicity and English as an additional language characteristics; second, we know that different regions of England were subject to different policy regimes over this time period. Specifically, 60% of our treatment schools are in London, a city where exam results were rapidly improving over this period of time. Whilst part of this phenomenon could indeed have reflected Teach First's penetration in the city, funding and new support for schools under the London Challenge project almost certainly contributed to school improvement (Ofsted, 2010) and the city also experienced significant demographic change during this time.

We therefore decide to restrict our potential control schools to any future Teach First schools in the same region which join the program at least three years after the treatment cohort in question. We report the robustness of our results to a number of different matching strategies in the results section. The bottom section of Table 3 shows the number of potential control schools for each cohort of new Teach First schools. For example, 40 schools participated in Teach First for the first time in 2003/4, all in London. There are 187 schools participating for the first time in 2006/7 or later who could act as potential control schools, but only 90 of these are in London and we restrict our match to these schools.

We implement propensity score matching in Stata using psmatch2 (Leuven and Sianesi, 2003) to deal with the dimensionality of matching on multiple variables, instead matching on a single

propensity score which represents the likelihood of a school having been included in the treatment group conditional upon its being selected for treatment (Rosenbaum and Rubin, 1983). 16 matches with replacement are actually performed – one probit regression for every cohort-region set of participating schools since the group of potential control schools changes each time. We apply the nearest neighbour method with a calliper of 0.2 and the imposition of common support to avoid very poor matches contributing to the calculation of the average treatment effect.

We match on the characteristics of schools in 2003, immediately preceding the launch of Teach First (the robustness to matching on school characteristics in the year before the school's own adoption in Teach First are shown in Table 8). Conditional independence requires the propensity score to capture all variables that correlate with the outcome and programme participation. The following variables are chosen on the basis that they either (1) formally determine participation eligibility in Teach First (these were free school meals proportion greater than 30% and percentage achieving 5 good GCSEs less than 25%); or (2) statistically important in determining both participation and attainment:

- school average prior attainment of pupils at age 11 (i.e. mean Key Stage 2 score);
- average deprivation of pupil's small area neighbourhood (i.e. IDACI);
- proportion eligible for free school meals;
- proportion of white British ethnicity;
- proportion achieving 5 or more good GCSEs including English and maths in 2003; and
- change in GCSE results between 2000 and 2003.

The last matching covariate is particularly important since it aims to capture any underlying changes that are taking place at the school during the period of adoption of Teach First. For example, given that head teachers choose to participate, the treatment may be correlated with improvements in performance at the school prior to the programme. Alternatively, in the spirit of an Ashenfelter dip (Ashenfelter, 1978), adoption of the programme may reflect increasing difficulties in recruiting good teachers due (and exacerbating) to declining exam performance. Thus, the matching strategy does deal with changes based on unobservable factors, but only if these factors were already present and captured in the change in exam score variable prior to the treatment.

We report balancing tests in Table 4 on a wide variety of covariates at t-1. The match is very strong – none of the differences are statistically significant at the 5% level; the one-year change in the best-8 subjects GCSE z-score is significantly different at the 10% level. Matching within-region is critical to achieving this strong match, particularly on ethnicity characteristics;

restricting the match to only future Teach First schools seems less important in terms of balancing background characteristics, but is important for other reasons discussed earlier.

### Difference-in-Difference

We exploit our longitudinal data to combine matching methods with a difference-in-difference estimator (see Heckman et al., 1997). This estimator removes any variation in unobserved time-invariant characteristics between treatment and control observations and as such, provides a more reliable estimate of the effect of Teach First on attainment. The treatment effect is measured as the difference between the change in outcome over time for the treatment group and that for the matched control group, controlling for other time-varying variables. This is an unbiased estimate of the impact of Teach First under the following circumstances: there are common time effects between the treatment and control groups (captured by $\boldsymbol{time}_t$ in the equation below); the outcome is independent of assignment to treatment; and there are no unmeasured composition changes that occur over time in either group (measured composition changes are captured by $\boldsymbol{X}_{jt}$).

We assemble a balanced panel of five observations per treatment and matched control school with two observations prior to first take-up of Teach First and three observations following take-up. The school fixed effects regression equation for such panel analysis is given as:

$$Y_{jt} = \beta_1 TF1_j + \beta_2 TF2_j + \beta_3 TF3_j + \boldsymbol{X}_{jt}.\beta_4 + \boldsymbol{time}_t.\beta_5 + u_j + \varepsilon_{jt}$$

Thus, $\beta_1$ is the mean outcome for treated schools in the first year of TF participation; $\beta_2$ is the effect of the second year of TF participation; $\beta_3$ is the effect of third year TF participation. We do not extend the post-treatment period further because (1) data availability would severely restrict sample size; (2) the nature of the school's involvement in Teach First becomes less clear with just under half of participants leaving their placement school after two years, and more after 3 years; and (3) as the Teach for America literature points out (see below), it would be hard theoretically to justify a teacher's impact after three years being solely or mainly a result of their having been recruited through Teach First and not a product of internal school training and other professional development.

### Estimating Departmental Impacts

We can apply identical methods to those described above to estimate the impact of Teach First on maths, science and English attainment, replacing the school-wide participation measure with an indicator for the first participation of the department. As above, these estimates are only valid if there are no time-varying unobservable characteristics associated with the decision to

join the Teach First programme. If we assume that a department's Teach First participation does not spillover into improvements elsewhere in the school, we can implement a triple-difference estimation approach, using changes in other departments as an additional control, thus holding constant pupil characteristics and school-wide policies:

$$Y_{hjt} = \beta_1 TF1_{hjt} + \beta_2 TF2_{hjt} + \beta_3 TF3_{hjt} + \boldsymbol{X_{jt}}\boldsymbol{*subject_h*time_t}.\beta_4 + \boldsymbol{time_t*subject_h}.\beta_5 + \boldsymbol{subject_h}.\beta_6 + u_j + \varepsilon_{jt}$$

Here, $\beta_1$, $\beta_2$ and $\beta_3$ represent the impact on subject specific test scores in years 1, 2 and 3, respectively; $\beta_4$, $\beta_5$ and $\beta_6$ reflect subject-time trends in average performance and subject-time specific impacts of observed background characteristics. This approach effectively deals with any non-random assignment of the Teach First programme to schools, though clearly does not account for non-random assignment to departments within schools.

Alternatively, we can use cross-sections of the pupil-level data to estimate point-in-time pupil fixed effect models that associate the pupil's attainment in a subject with the department's Teach First participation, applying a pupil fixed effect ($\upsilon_{ij}$) to account for the pupil's attainment in other subjects.

$$Y_{hjj} = \beta_1 TF_{hj} + \beta_2 priorY_{hij} + \boldsymbol{X_{jt}}\boldsymbol{*subject_h}.\beta_3 + \boldsymbol{cohort_{ij}*subject_h}.\beta_4 + \boldsymbol{subject_h}.\beta_5 + \upsilon_{ij} + \varepsilon_{hij}$$

We run separate cross-sectional regressions for the year before treatment and the first three years following the school's first participation. $\beta_3$, $\beta_4$ and $\beta_5$ reflect subject-cohort differences in average performance and subject-cohort specific impacts of observed background characteristics.

Table 5 shows the variation between departments that is exploited to estimate the impact of Teach First placement within a department. We restrict our analysis here to the core departments of English, maths and science as these are the only subject areas with participant sample sizes large enough to offer meaningful evaluation. The data shows that if a school chooses to participate in Teach First then each of the three core departments usually do so at some point in the future. However, it is unusual for all three departments to participate together in the first year that a school takes Teach First participants (less than half do) and there are instances of core departments not yet having participated. In our analysis we exploit both variation in the first use of Teach First participants in the department and differences in the intensity of Teach First use in the first three years of the school's participation. The latter variation is shown in the last three rows of Table 5.

We have no reason to necessarily expect the estimates of impact from these three estimation

approaches to be the same. Any differences in coefficient sizes might tell us something about the relative importance of (1) sorting into schools and into departments on unobservables; (2) the size of spillovers from departments across schools; and (3) variation in effect sizes across different subjects.

## Results

We first present the results from the estimation of the impact of Teach First participation on whole school achievement, before moving onto the departmental estimates. The impact of whole school achievement is likely to reflect a combination of the relative effectiveness of the Teach First participant themselves, the impact of the participant on the teaching quality of others in their department and elsewhere in the school and any wider contribution the Teach First participants make to the life of the school.

In Table 6 we present results from eight regression equations. These are all difference-in-difference regressions estimated in a balanced panel of five observations per school (two prior and three following first Teach First participation). The sample is restricted to treatment schools matched to a control group of future Teach First schools located in the same region.

The top half of the table presents the coefficients for the impact of Teach First participation in years 1, 2 and 3. The regression results are presented here with and without time-varying control variables. These control variables do not change the substantive estimates, which is correct since we have no reason to believe that the time-varying controls are correlated with Teach First participation. However, they may marginally improve precision on estimates and so we do include them for all other tables in this paper.

The results show that Teach First participation has no impact on a school's exam performance in year 1, as measured by the pupil's best 8 subject grades (i.e. capped GCSE z-score). The impact in years 2 and 3 are positive and statistically significant at around 5 and 8 percent of a standard deviation. This pattern of no effect in year 1 and positive and increasing effects in years 2 and 3 is found throughout our results section.

An effect size of 5% of a pupil standard deviation is equivalent to the school moving up 10% of a standard deviation across distribution of school average capped GCSE z-scores, or a little less than one grade in one of the child's best eight subjects. This is seen as a relatively small effect size in research on school effectiveness (Hattie, 2003), but it is not surprising it is small since Teach First simply places a small number of inexperienced teachers in a large secondary school. Since there are multiple and complex possible pathways to impact for Teach First participants, we will reserve more detail on the magnitude of likely mechanisms to our estimation of

departmental impacts.

The finding of no effect in year 1 may occur for several reasons. The year 11 pupils would have had little direct exposure to the Teach First participants who would have only been in school for one of the pupil's five years. They would have lower chances of having been allocated a new Teach First teacher at the start of year 11, given standard practice in England of retaining the same teacher across years 10 and 11 where possible. Also, clearly as novice teachers in the first year, the Teach First participants will be less effective than in their second year. It is not possible to distinguish between these potential mechanisms in the data.

The impact of Teach First is less consistent on the threshold outcome measure of the proportion of pupils attaining five or more good GCSEs, including English and maths. In years two and three where point estimates are positive, they are in the region of just two percentage points (on a metric with an average of around 30%). This is a far smaller effect size than that reported in Muijs et al. (2010). The pattern of effects on our intensity variable roughly mirrors those of the individual year dummies. A positive impact of 0.005 with around 5 Teach First participants in the school translates to an effect of 0.025. This is a little smaller than the impacts estimating using the year dummies, but this intensity variable doesn't allow the impact to vary between years one and two. The estimated impact using this intensity variable is actually negative on the threshold measure. Overall, given the endogeneity of the variation in intensity and the interpretation of the coefficient in year three where continuing former participants are not included, it is not worth placing too much weight on this variable.

In Table 7 we explore whether the impact of Teach First varies across time and across regions. We are relieved to find that the impact of the programme is not bigger within London than it is outside London because this was a period of considerable change within the capital city and so there was a serious risk that estimates were inflated by Teach First schools receiving other interventions at the same time. The impact of the programme for schools first joining in the later years of 2008/09 to 2010/11 has not shrunk compared to earlier years; indeed estimates are slightly larger. We had expected to find that the impact of the programme shrinks as it expands because the quality of the marginal participant should decline and the programme itself may experience scale diseconomies. The finding that this is not the case does not mean it will not face expansion difficulties in the future since the largest cohort we estimate impacts on in this paper is just over 500, compared to about 1000 in 2012/13. The finding of smaller effect sizes in earlier years (the same years Muijs et al. find large effects) is not due to lower intensity of participation; nor can it be explained by particularly different school characteristics.

Table 8 reports the robustness of our main estimate to alternative matching and estimation

approaches. The top row reports a variety of different matching approaches. In column 3 we match Teach First schools to those who will participate in the future but whose participation is geographically constrained by the regional roll-out. This approach is most appealing from a quasi-experimental perspective, but may pick up the impact of other London initiatives. The estimates are all positive, even in year one, in this specification where we are concerned about the contamination of other London initiatives during this time period. But otherwise, estimates are relatively consistent across alternative matching strategies.

The bottom half of Table 8 reports a variety of other changes to the matching strategy. In the first column we match on data two years before the first participation in Teach First, rather than on 2003 data regardless of year of first participation. This might arguably lead to a slightly closer match for the start of the difference-in-difference panel, but we find no differences in estimates overall. In the second column we allow the number of control schools to be greater than one, which could potentially increase precision of estimates, but the number of control schools does not hugely rise due to limits in the number of potential match schools within our sample. In column 3 we drop the common support requirement in the propensity score match. This does marginally increase the number of schools for which we can achieve a successful match, but does not change our estimates. Finally, we perform a falsification test by creating a false treatment two years prior to the school's adoption of Teach First. The impact of the false treatment is not significant in all years, which lends support to our assumption of no time-invariant unobservables confounding estimates.

In Table 9 we turn to estimates of the departmental impact of Teach First participants. There are three estimation strategies here, so the findings are rather complex and make different assumptions regarding likely pathways to impact. The first three columns of estimates are from difference-in-difference regressions of changes in the effectiveness of English, maths and science departments in a school separately, without holding constant any changes taking place in other departments in the school. The advantage of these regressions is that they allow for cross-departmental spillovers in impact, with the corresponding disadvantage that any unobserved changes in overall school processes cannot be accounted for. The estimates here are not consistent across subjects: the impact is strongest and most precisely estimated in English, a subject where Teach First themselves will claim they find it easiest to recruit high quality participants (by contrast, some Teach First participants in maths and science do not have a degree in these subjects). No impact is ever found in maths, though the point estimates are positive.

The triple difference estimates in column four measure changes in departmental effectiveness,

holding constant changes taking place in core departments who have not yet taken on a Teach First participant. It imposes an assumption of equal potential impact across subjects and will be seriously biased downwards if, for example, Teach First participants in English departments are able to positively influence a child's maths GCSE grade. The impact estimates are zero in year one and positive in years two and three in the order of about 8% and 11% of a grade, respectively.

The pupil point-in-time fixed effects models take a cross-section of data separately for one year prior and one, two and three years following the school's first Teach First participation and estimate the impact of departmental participation in pupil-by-subject data. The pre-treatment estimate shows relative effectiveness in the year before participation for departments who take on Teach First participants in the first year that the school first participates, holding constant the effectiveness of those departments who do not take on a participant in the first year. It shows that the departments who participate early on are significantly less effective before the arrival of the Teach First participants than those who do not immediately participate. They may have staff recruitment difficulties or higher teacher turnover which presents vacancy opportunities or they might be viewed as struggling by the headteacher who therefore encourages them to try Teach First. These early participating departments are also significantly less effective in year 1 but more effective in years two and three to the tune of 15% of a grade. Clearly, the year three estimates only reflect differences between the early participating departments and others in schools where one or more core department does not participate by year three; in 22% of treatment schools all three core departments have participated at least once by year three.

The results using the intensity variable are roughly consistent with the Teach First year dummies throughout Table 9. Within a core subject department in a participating school, about one in six teachers will be a Teach First participant in each of the first three years. Our estimate of impact of the order of at least 5% of a subject grade could be as high as 30% of a grade if we assume no spillovers of participation to other teachers in the same department. This implies that Teach First participants are highly effective, on average, compared to those they have displaced. Using estimates from Slater et al. (2012), 30% of a subject grade is equivalent to one standard deviation higher teacher effectiveness. It seems unlikely that Teach First selection and training processes are this effective, so we believe there must be some spillovers to other teachers in the department.

## Discussion

In this paper we provide convincing evidence that placing carefully selected, yet inexperienced,

graduates into English secondary schools has not been damaging to pupils and most likely produced school-wide gains in GCSE results of the order of 5% of a pupil standard deviation or around one grade in one of the pupil's best eight subjects. This is a consistently estimated positive effect, though is clearly not as large as the impact of other interventions to improve teaching standards such as training to improve the quality of pupil feedback (Hattie, 2003). It is also substantially smaller than the estimate of one third of a grade per subject in Muijs et al. (2010).

Within core departments our estimates suggest a gain of over 5% of a subject grade, which could translate to as high as 30% of a grade in the Teach First participant's classroom if we assume no spillover to other teachers. If there were no spillovers of Teach First participation to other teachers this would suggest that the Teach First selection process succeeds in attracting and selecting outstanding teachers who are, on average, one standard deviation more effective than those who they displace. This figure seems rather high so, more likely, Teach First presence also raises the teaching standards of those who teach alongside them in the same department, echoing the findings of Jackson and Bruegmann (2009) who identify the importance of teacher peers.

The research design used here cannot make claims about the relative effectiveness of Teach First participants versus those trained via other routes; nor can it assess the quality of the short Teach First summer training programme which is undertaken alongside ongoing university and in-school support for participants. That said, our findings are best interpreted alongside some understanding of who participating schools might have recruited in the absence of Teach First. We can use the new School Workforce Census for 2010/11 and 2011/12 to see how Teach First participation changes the composition of teachers within the school. We explore this in several different ways: (1) comparing the staff composition before and after joining the programme for the 2011/12 cohort; (2) comparing staff composition between treatment and the matched future Teach First schools within region for the 2009/10 and 2009/10 cohorts; (3) comparing all our Teach First cohorts to a matched set of non-Teach First schools within the same region. Through this series of comparisons we can make the following generalisations about the programme.

First, participation in Teach First results in a greater number of teachers under age 30 (from around 22 percent of all staff to around 27 percent of all staff), which suggests that schools do not use the scheme to solely replace other newly qualified graduates. Second, schools participating in Teach First have greater numbers of teachers who have taught at the school for less than 5 years (as many as 60 percent with tenure less than 5 years, versus 50 percent for

comparison schools), though not larger numbers of new arrivals once the scheme is established in their school. Finally, there are no ethnic or sex differences in the staff composition at Teach First participating schools.

This School Workforce Census data suggests that schools are displacing slightly older and more experienced teachers who might have spent longer working at the school, had they been recruited. So, it is not surprising that many believed the Teach First programme could be damaging. In general, it is true that teachers are less effective at the start of their career (e.g. Rockoff, 2004 estimate teachers improve by around 10% of a standard deviation in the first two years in maths). However, the possible damage of inexperience of the Teach First participants appears to be more than outweighed by the gains to careful selection of individuals into the profession. Furthermore, if the more effective Teach First participants can be persuaded to stay in the profession, US evidence suggests they will make large improvements in their teaching standards, compared to those who are less effective first year teachers (Atteberry et al., 2013; TNTP, 2013). That said, we must recognise the disruption and recruitment costs to schools of dealing with the higher teacher turnover that the Teach First programme necessarily produces. This turnover also places limits on how large the scheme should become for any individual school.

Our estimates relate to a period of time when the Teach First programme was a fraction of the size it is now, or aspires to be in the future. We can say nothing about whether its effectiveness will fall as it expands in the type of graduates it recruits and the type of schools within which it places. With as many as 1000 participants a year from a much more diverse range of undergraduate universities, it is likely they are now recruiting many participants who would otherwise have joined the teaching profession through the traditional Post-Graduate Certificate in Education route. This reduces the value of Teach First if we believe its impact arises more from the recruitment of talented graduates rather than the efficacy of their very short intensive training programme and subsequent in-school and partner university support. Obviously, the overall benefits of the scheme are highly contingent on the proportion of participants who choose to stay in teaching in the long term. As the scheme has expanded, retention rates beyond the standard two years into year three do appear to have risen, but it is not known whether this reflects the changing composition of the intake or other reasons, such as high levels of graduate unemployment due to the recession.

Overall, this study lends strong support to studies from the US regarding the effectiveness of these types of teacher recruitment programmes, particularly where graduates are placed in classrooms with older children. The growth of similar programmes in a number of other

countries affiliated to the international 'Teach for All' umbrella organisation, created in cooperation between Teach First and Teach for America in 2007, means that our conclusions are relevant beyond Teach for America and Teach First. However, the lack of matched teacher-pupil data means we can say little about individual participant effectiveness or about how participants influence the teaching experiences of others in the school. Understanding precise mechanisms of impact is important because Teach First itself cannot expand indefinitely. This is because Teach First's greatest success has been to "detoxify teaching" for high attaining graduates (Wigdortz 2012, p. 230) and maintenance of its position as a premium brand is somewhat contingent on retaining exclusivity. However, through understanding exactly what Teach First does that makes it an effective programme, it may enable us to replicate small parts of the behaviour of the participants across the education system.

## References


Aaronson, D., Barrow, L. and Sander, W. (2007) Teachers and student achievement in the Chicago public high schools, *Journal of Labour Economics,* 25 (1)*,* pp. 95-135.

Allen, R., Burgess, S. and Mayo, J. (2012) *The teacher labour market, teacher turnover and disadvantaged schools: new evidence for England*, CMPO working paper No. 12/294 and DoQSS working paper No. 12/09.

Antecol, H., Eren, O. and Ozbeklik, S. (2013) *The effect of Teach for America on the distribution of student achievement in primary school: Evidence from a randomized experiment,* IZA discussion paper No. 7296.

Ashenfelter, O. (1978) Estimating the Effects of Training Programmes on Earnings, *The Review of Economics and Statistics* 60 (1) pp. 47-57.

Atteberry, A., Loeb, S., & Wyckoff, J. (2013). *Do first impressions matter? Improvement in early career teacher effectiveness*, Washington, DC: National Center for Analysis of Longitudinal Data in Education Research (CALDER WP 90).

Boyd, D.J., Grossman, P., Lankford, H., Loeb, S., Michelli, N.M. and Wyckoff, J. (2006) How changes in entry requirements alter the teacher workforce and affect student achievement, *Education Finance and Policy,* 1(2) pp. 176-216.

Clotfelter, C.T., Ladd, H.F. and Vigdor, J.L. (2007) Teacher credentials and student achievement: Longitudinal analysis with student fixed effects, *Economics of Education Review,* 26*,* pp. 673 - 682.

Darling-Hammond, L., Holtzman, D.J., Gatlin, S.J. and Heilig, J.V. (2005) Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness, *Education Policy Analysis Archives*, 13(42) pp. 1068-2341.

Decker, P.T., Mayer, D.P. and Glazerman, S. (2004) The Effects of Teach For America on Students: Findings from a National Evaluation, *Journal of Policy Analysis and Management*, 25(1) pp. 75–96.

Glewwe, P., Heckman, E., Humpage, S., Ravina, R. (2011) *School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010,* Cambridge, Mass.: NBER Working Paper No. 17554.

Goldhaber, D. and Brewer, D. (2000) Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement, *Educational Evaluation and Policy Analysis*, 22(2) pp. 129-145.

Hanushek, E. A., Kain, J.F., O'Brien, D.M., Rivkin, S.G. (2005) *The market for teacher quality*, Cambridge, Mass.: NBER Working Paper No. 11154.

Hattie, J. (2003) *Teachers make a difference: what is the research evidence?* Melbourne: Australian Council for Educational Research.

Heckman, J.J., Ichimura, H. and Todd, P. (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme*, The Review of Economic Studies,* 64(4) pp. 605-654.

Henry, G.T., Thompson, T.L., Fortner, C.K., Zulli, R.A. and Kershaw, D.C. (2010) *The Impact of Teacher Preparation on Student Learning in North Carolina Public School,* Carolina Institute of Public Policy, University of North Carolina. Available at: www.publicpolicy.unc.edu (accessed

24th June 2013).

Hutchings, M., Maylor, U., Mendick, H., Menter, I. and Smart, S. (2006) *An evaluation of innovative approaches to teacher training on the Teach First programme: Final report to the Training and Development Agency for Schools*, London: Teaching and Development Agency.

Jackson, C. K. & Bruegmann, E. (2009) Teaching students and teaching each other: The importance of peer learning for teachers, *American Economic Journal: Applied Economics,* 1(4) pp. 85-108.

Kane, T.J., Rockoff, J.E. and Staiger, D.O. (2008) What does certification tell us about teacher effectiveness? Evidence from New York City, *Economics of Education Review*, 27(6) pp. 615-631.

Leuven, E. and Sianesi, B. (2003) PSMATCH2: *Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing*. http://ideas.repec.org/c/boc/bocode/s432001.html. This version 3.1.5 2 may 2009.

Mujis D, Chapman C, Collins A and Armstrong P (2010) *Maximum Impact Evaluation. The impact of Teach First teachers in schools: An evaluation,* London: Teach First. Available from www.teachforall.org/articles/max_impact.pdf (accessed 24th June 2013).

Ofsted (2006) *Rising to the challenge: a review of the Teach First initial teacher training programme*, London: Ofsted.

Ofsted (2010) *London Challenge*, London: Ofsted.

Ofsted (2011) *Teach First: Initial Teacher Education Inspection Report*, London: Ofsted

Raymond, M., Fletcher, S.H. and Luque, J. (2001) *Teach For America: An Evaluation of Teacher Differences and Student Outcomes in Houston*, Texas. Palo Alto: Center for Research on Education Outcomes.

Rivkin, S.G., Hanushek, E.A. and Kain, J.F. (2005) Teachers, Schools and Academic Achievement, *Econometrica,* 73(2), pp. 417–458.

Rockoff, J.E. (2004) The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data, *American Economic Review*, 94(2) pp. 247-252.

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects*, Biometrika,* 70(1), pp. 41-55.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin.

Schoeneberger, J.A., Dever, K.A. and Tingle, L. (2009) *Evaluation of Teach for America in Charlotte-Mecklenburg Schools,* Center for Research and Evaluation, Office of Accountability, Charlotte-Mecklenburg Schools. Available at http://www.cms.k12.nc.us/cmsdepartments/accountability/cfre/Documents/TFA_Evaluation_Report.pdf (accessed 24th June 2013).

Schoeneberger, J.A. (2011) *Evaluation of Teach for America in Charlotte-Mecklenburg Schools,* Center for Research and Evaluation, Office of Accountability, Charlotte-Mecklenburg Schools. Available at: http://www.cms.k12.nc.us/cmsdepartments/accountability/cfre/Documents/TeachForAmerica_Evaluation_Report_2011.pdf (accessed 24th June 2013).

Slater, H., Davies, N.M. and Burgess, S. (2012) Do Teachers Matter? Measuring the Variation in

Teacher Effectiveness in England*, Oxford Bulletin of Economics and Statistics*, 74(5) pp. 629–645.

TNTP (2013). *Leap year: Assessing and supporting effective first-year teachers*, Brooklyn, NY: TNTP. Available at: http://tntp.org/assets/documents/TNTP_LeapYear_2013.pdf (accessed 24th June 2013).

Ware, A., LaTurner, R.J., Parsons, J., Okulicz-Kozaryn, A., Garland, M. and Klopfenstein, K. (2011) *Teacher Preparation Programs and Teach for America Research Study,* The University of Texas at Dallas, Education Research Center. Available from https://www.teachforamerica.org/our-organization/research (accessed 24th June 2013).

Wigdortz, B. (2012) *Success against the odds. Five lessons in how to achieve the impossible: The Story of Teach First,* Short Books.

United States Department of Education Institute of Education Sciences 'What Works Clearinghouse' (2008) *WWC Quick Review of the Report "Making a Difference? The Effects of Teach for America in High School"*, Washington, DC: What Works Clearinghouse, U.S. Dept. of Education. Available at: http://ies.ed.gov/ncee/wwc/quickreviewsum.aspx?sid=53 (accessed 24th June 2013).

Xu, Z., Hannaway, J. and Taylor, C. (2009) *Making a Difference? The Effects of Teach for America in High School*, Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

# Tables and figures

### Table 1: Number of schools in Teach First programme (by year of first participation)

| | London | North West | West Midlands | East Midlands | Yorkshire | East of England | South East | North East | South West | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Cohort 2003/04 | **43** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 43 |
| Cohort 2004/05 | **19** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 |
| Cohort 2005/06 | **14** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 |
| Cohort 2006/07 | **10** | **18** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 28 |
| Cohort 2007/08 | **10** | **6** | **17** | 0 | 0 | 0 | 0 | 0 | 0 | 33 |
| Cohort 2008/09 | **13** | **10** | **11** | 0 | 0 | 0 | 0 | 0 | 0 | 34 |
| Cohort 2009/10 | **8** | **11** | **11** | **12** | **20** | 0 | 0 | 0 | 0 | 62 |
| Cohort 2010/11 | 5 | 10 | 13 | 9 | 7 | 3 | 3 | 0 | 0 | 51 |
| Cohort 2011/12 | 21 | 13 | 14 | 5 | 16 | 1 | 4 | 20 | 0 | 94 |
| Cohort 2012/13 | 23 | 10 | 14 | 8 | 12 | 2 | 8 | 9 | 0 | 86 |
| Never joined TF | 267 | 421 | 323 | 249 | 270 | 349 | 499 | 141 | 318 | 2,837 |

## Table 2: 2003 characteristics of schools (by year of first participation)

| | Number of schools | Average school cohort size | Average capped GCSE z-score | % 5+A*-C at GCSE, incl Eng and maths | 3-year pp change in GCSE threshold | Average KS2 z-score | Average IDACI deprivation score | % free school meals | % white British ethnicity |
|---|---|---|---|---|---|---|---|---|---|
| Cohort 2003/04 | 43 | 170 | -0.20 | 31 | 8 | -0.29 | 0.39 | 36 | 32 |
| Cohort 2004/05 | 17 | 174 | -0.54 | 18 | 3 | -0.44 | 0.41 | 35 | 41 |
| Cohort 2005/06 | 14 | 160 | -0.34 | 25 | 5 | -0.28 | 0.39 | 37 | 39 |
| Cohort 2006/07 | 26 | 166 | -0.45 | 22 | 7 | -0.31 | 0.41 | 34 | 58 |
| Cohort 2007/08 | 32 | 176 | -0.44 | 20 | 6 | -0.40 | 0.40 | 37 | 55 |
| Cohort 2008/09 | 32 | 199 | -0.39 | 23 | 5 | -0.33 | 0.39 | 34 | 58 |
| Cohort 2009/10 | 60 | 190 | -0.51 | 19 | 4 | -0.41 | 0.39 | 32 | 69 |
| Cohort 2010/11 | 46 | 180 | -0.35 | 25 | 7 | -0.31 | 0.33 | 25 | 67 |
| Cohort 2011/12 | 93 | 190 | -0.39 | 23 | 5 | -0.30 | 0.35 | 29 | 69 |
| Cohort 2012/13 | 68 | 182 | -0.36 | 24 | 4 | -0.31 | 0.36 | 28 | 66 |
| All TF schools | 431 | 182 | -0.39 | 23 | 5 | -0.33 | 0.37 | 31 | 60 |
| Never joined TF | 2,662 | 181 | 0.09 | 44 | 4 | 0.07 | 0.19 | 12 | 82 |

Note: The Income Deprivation Affecting Children Index (IDACI) measures in a local area the proportion of children under the age of 16 that live in low income households. The codes range from 0.00 (least deprived) to 0.99 (most deprived).

### Table 3: Potential and actual matched control schools (by year of first participation)

| | | Treatment cohort year | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **2004** | **2005** | **2006** | **2007** | **2008** | **2009** | **2010** |
| Year of first participation | Cohort 2003/04 | **40** | 0 | 0 | 0 | 0 | 0 | 0 |
| | Cohort 2004/05 | 0 | **11** | 0 | 0 | 0 | 0 | 0 |
| | Cohort 2005/06 | 0 | 0 | **11** | 0 | 0 | 0 | 0 |
| | Cohort 2006/07 | 4 | 0 | 0 | **25** | 0 | 0 | 0 |
| | Cohort 2007/08 | 2 | 3 | 0 | 0 | **26** | 0 | 0 |
| | Cohort 2008/09 | 8 | 2 | 1 | 0 | 0 | **31** | 0 |
| | Cohort 2009/10 | 5 | 1 | 1 | 7 | 0 | 0 | **24** |
| | Cohort 2010/11 | 1 | 0 | 0 | 4 | 7 | 0 | 0 |
| | Cohort 2011/12 | 7 | 2 | 5 | 9 | 8 | 21 | 0 |
| | Cohort 2012/13 | 13 | 3 | 4 | 5 | 11 | 10 | 24 |
| N treatment schools with no successful match | | 3 | 8 | 3 | 3 | 7 | 3 | 38 |
| N of potential controls available | | 387 | 359 | 326 | 292 | 230 | 180 | 86 |
| *of which:* | | | | | | | | |
| London | | 90 | 80 | 70 | 57 | 49 | 44 | 23 |
| North West | | - | - | - | 44 | 33 | 23 | 10 |
| West Midlands | | - | - | - | - | 41 | 28 | 14 |
| East Midlands and Yorkshire | | - | - | - | - | - | - | 20 |

## Table 4: Balancing tests (year before treatment)

| | Number of schools | Average capped GCSE z-score | 1 yr prior change in GCSE score | Average KS2 score | 1 year prior change in KS2 score | Average IDACI deprivation score | % free school meals | % white British ethnicity |
|---|---|---|---|---|---|---|---|---|
| Treatment group | 168 | -0.297 | 0.048 | -0.319 | 0.015 | 0.398 | 0.329 | 0.472 |
| Control (future TF schools) | 168 | -0.260 | 0.024 | -0.309 | -0.008 | 0.400 | 0.316 | 0.495 |
| Difference | | -0.037 | 0.023 | -0.009 | 0.024 | -0.001 | 0.013 | -0.023 |
| (Standard error) | | (0.032) | (0.016) | (0.028) | (0.017) | (0.011) | (0.016) | (0.035) |

Note: The 168 control schools includes multiple counts of schools drawn more than once in the propensity score matching (16 drawn twice; 9 drawn 3 times; 1 drawn four times; 2 drawn six times).

## Table 5: Timing of first participation across departments within schools

| | English | | | | Maths | | | | Science | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Dept participates with:* | | | | | | | | | | | | |
| 0 year lag | 81 | | | | 76 | | | | 76 | | | |
| 1 year lag | 31 | | | | 27 | | | | 32 | | | |
| 2 year lag | 13 | | | | 20 | | | | 15 | | | |
| 3 year lag | 9 | | | | 5 | | | | 10 | | | |
| 4 year lag | 1 | | | | 3 | | | | 0 | | | |
| 5 year lag | 3 | | | | 1 | | | | 2 | | | |
| 6 year lag | 2 | | | | 2 | | | | 0 | | | |
| 7 year lag | 0 | | | | 0 | | | | 2 | | | |
| 8 year lag | 1 | | | | 2 | | | | 0 | | | |
| 9 year lag | 1 | | | | 0 | | | | 0 | | | |
| Not TF yet | 26 | | | | 32 | | | | 31 | | | |
| | **Mean** | **S.D.** | **Min** | **Max** | **Mean** | **S.D.** | **Min** | **Max** | **Mean** | **S.D.** | **Min** | **Max** |
| Intensity year 1 | 0.7 | (0.9) | 0.0 | 3.5 | 0.7 | (0.9) | 0.0 | 6.8 | 0.7 | (0.9) | 0.0 | 4.9 |
| Intensity year 2 | 1.2 | (1.3) | 0.0 | 7.0 | 0.9 | (1.1) | 0.0 | 6.6 | 1.2 | (1.2) | 0.0 | 5.0 |
| Intensity year 3 | 1.1 | (1.4) | 0.0 | 6.9 | 0.8 | (1.0) | 0.0 | 5.4 | 0.9 | (1.1) | 0.0 | 5.3 |
| *Average intensity compared to English dept in:* | | | | | | | | | | | | |
| Year 1 | | | | | 0.0 | (0.9) | -3.1 | 6.8 | 0.0 | (1.0) | -3.5 | 4.9 |
| Year 2 | | | | | -0.1 | (1.0) | -5.1 | 4.4 | 0.0 | (1.3) | -7.0 | 5.0 |
| Year 3 | | | | | -0.1 | (0.9) | -4.8 | 4.4 | -0.1 | (1.0) | -4.7 | 4.5 |

## Table 6: Matched difference-in-difference regression results

| | Capped GCSE z-score | | | Capped GCSE z-score | | | 5+A*-C, including Eng and maths | | | 5+A*-C, including Eng and maths | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** |
| Treatment year 1 | 0.019 | (0.019) | n.s. | 0.027 | (0.019) | n.s. | -0.004 | (0.008) | n.s. | -0.001 | (0.008) | n.s. |
| Treatment year 2 | 0.048 | (0.019) | *** | 0.058 | (0.019) | *** | 0.015 | (0.008) | ** | 0.019 | (0.008) | ** |
| Treatment year 3 | 0.081 | (0.019) | *** | 0.082 | (0.019) | *** | 0.020 | (0.008) | *** | 0.021 | (0.008) | *** |
| Rho (Fraction var due to $u_j$) | | 0.706 | | | 0.758 | | | 0.746 | | | 0.788 | |
| R-sq (overall) | | 0.294 | | | 0.018 | | | 0.341 | | | 0.129 | |
| Treatment intensity | 0.005 | (0.002) | *** | 0.004 | (0.002) | ** | -0.001 | (0.001) | n.s. | -0.001 | (0.001) | * |
| Rho (Fraction var due to $u_j$) | | 0.704 | | | 0.755 | | | 0.744 | | | 0.786 | |
| R-sq (overall) | | 0.294 | | | 0.018 | | | 0.345 | | | 0.132 | |
| Controls | | Yes | | | No | | | Yes | | | No | |
| N (overall) | | 1680 | | | 1680 | | | 1680 | | | 1680 | |
| N (schools) | | 336 | | | 336 | | | 336 | | | 336 | |

Notes: (a) n.s.=not statistically significant, ***=sig. at 1% level, **=sig.at 5% level, *=sig. at 10% level
(b) School and time dummy variables included, time variant controls are mean KS2 prior attainment and proportions of free school meals, English as an additional language, white British ethnicity pupils

## Table 7: Testing for heterogeneity across regions and over time

| | Main specification | | | London only | | | 2007 and earlier | | | 2008 onwards | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** |
| Treatment year 1 | 0.019 | (0.019) | n.s. | 0.025 | (0.023) | n.s. | 0.005 | (0.021) | n.s. | 0.032 | (0.032) | n.s. |
| Treatment year 2 | 0.048 | (0.019) | *** | 0.050 | (0.023) | ** | 0.005 | (0.021) | n.s. | 0.095 | (0.032) | *** |
| Treatment year 3 | 0.081 | (0.019) | *** | 0.087 | (0.023) | *** | 0.049 | (0.021) | ** | 0.115 | (0.032) | *** |
| Rho (% var due to $u_i$) | | 0.706 | | | 0.739 | | | 0.792 | | | 0.629 | |
| R-sq (overall) | | 0.294 | | | 0.441 | | | 0.516 | | | 0.091 | |
| Controls | | Yes | | | Yes | | | Yes | | | Yes | |
| N (overall) | | 1680 | | | 950 | | | 870 | | | 810 | |
| N (schools) | | 336 | | | 190 | | | 174 | | | 162 | |
| *Average characteristics of sample in year prior to treatment:* | | | | | | | | | | | | |
| Capped GCSE score | | -0.279 | | | -0.261 | | | -0.288 | | | -0.269 | |
| IDACI depr score | | 0.399 | | | 0.398 | | | 0.395 | | | 0.404 | |
| % white British ethnicity | | 0.484 | | | 0.323 | | | 0.424 | | | 0.547 | |

Notes: (a) n.s.=not statistically significant, ***=sig. at 1% level, **=sig.at 5% level, *=sig. at 10% level;
(b) School and time dummy variables included, time variant controls are mean KS2 prior attainment, free school meal proportion, English as an additional language proportion, white British ethnicity proportion

## Table 8: Robustness of results to changes in matching strategy

| | Main specification (Match to future TF school within region) | | | Match to never-TF schools within region | | | Match to future TF schools in future participating areas | | | Match to future TF school in any region | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** |
| Treatment year 1 | 0.019 | (0.019) | n.s. | 0.000 | (0.017) | n.s. | 0.030 | (0.018) | * | 0.022 | (0.018) | n.s. |
| Treatment year 2 | 0.048 | (0.019) | *** | 0.025 | (0.017) | n.s. | 0.055 | (0.018) | *** | 0.041 | (0.018) | ** |
| Treatment year 3 | 0.081 | (0.019) | *** | 0.052 | (0.017) | *** | 0.064 | (0.018) | *** | 0.034 | (0.018) | * |
| Rho (% var due to $u_i$) | | 0.706 | | | 0.722 | | | 0.670 | | | 0.676 | |
| R-sq (overall) | | 0.294 | | | 0.403 | | | 0.270 | | | 0.338 | |
| Controls | | Yes | | | Yes | | | Yes | | | Yes | |
| N (overall) | | 1680 | | | 1891 | | | 2029 | | | 2078 | |
| N (schools) | | 336 | | | 389 | | | 406 | | | 416 | |

| | Drop common support requirement in PSM | | | Match using 2-yr prior to treatment not 2003 data | | | Change to 2-nearest neighbour PSM | | | Fake policy implemented two years prior | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** |
| Treatment year 1 | 0.022 | (0.018) | n.s. | 0.026 | (0.018) | n.s. | 0.028 | (0.017) | n.s. | -0.016 | (0.018) | n.s. |
| Treatment year 2 | 0.050 | (0.018) | *** | 0.051 | (0.018) | *** | 0.041 | (0.017) | ** | -0.002 | (0.018) | n.s. |
| Treatment year 3 | 0.076 | (0.018) | *** | 0.051 | (0.018) | *** | 0.068 | (0.017) | *** | 0.008 | (0.018) | n.s. |
| Rho (% var due to $u_i$) | | 0.716 | | | 0.684 | | | 0.683 | | | 0.739 | |
| R-sq (overall) | | 0.252 | | | 0.409 | | | 0.353 | | | 0.394 | |
| Controls | | Yes | | | Yes | | | Yes | | | Yes | |
| N (overall) | | 1920 | | | 1688 | | | 2010 | | | 1498 | |
| N (schools) | | 384 | | | 338 | | | 402 | | | 336 | |

Notes: n.s.=not statistically significant, ***=sig. at 1% level, **=sig.at 5% level, *=sig. at 10% level;
School and time dummy variables included, time variant controls are mean KS2 prior attainment, free school meal proportion, English as an additional language proportion, white British ethnicity proportion

## Table 9: Departmental impact of Teach First participants

| | Difference-in-difference [a] | | | | | | | | | Triple difference [b] | | | Pupil fixed effects [c] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | English grade | | | Maths grade | | | Science grade | | | GCSE grade | | | GCSE score | | | |
| | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Beta** | **SE** | **Sig** | **Rho** |
| Pre-treatment | | | | | | | | | | | | | -0.086 | (0.007) | *** | 0.765 |
| Year 1 | -0.014 | (0.410) | n.s. | -0.003 | (0.038) | n.s. | -0.013 | (0.085) | n.s. | -0.008 | (0.034) | n.s. | -0.073 | (0.008) | *** | 0.724 |
| Year 2 | 0.057 | (0.038) | n.s. | 0.047 | (0.035) | n.s. | 0.093 | (0.077) | n.s. | 0.077 | (0.031) | ** | 0.159 | (0.009) | *** | 0.698 |
| Year 3 | 0.128 | (0.037) | *** | 0.045 | (0.033) | n.s. | 0.138 | (0.075) | * | 0.110 | (0.030) | *** | 0.146 | (0.011) | *** | 0.682 |
| Rho | | 0.736 | | | 0.757 | | | 0.596 | | | 0.540 | | | | | |
| R-sq | | 0.295 | | | 0.336 | | | 0.274 | | | 0.410 | | | | | |
| Intensity | 0.031 | (0.011) | *** | 0.011 | (0.013) | n.s. | 0.052 | (0.025) | ** | 0.041 | (0.010) | *** | 0.068 | (0.004) | *** | [d] |
| Rho | | 0.746 | | | 0.757 | | | 0.597 | | | 0.541 | | | 0.698 | | |
| R-sq | | 0.300 | | | 0.363 | | | 0.269 | | | 0.408 | | | 0.250 | | |
| Controls | | Yes | | | Yes | | | Yes | | | Yes | | | Yes | | |
| N (overall) | | 1680 | | | 1680 | | | 1680 | | | 5040 | | | 186,670 | | |
| N (schools) | | 336 | | | 336 | | | 336 | | | 336 | | | 336 | | |

Notes: n.s.=not statistically significant, ***=sig. at 1% level, **=sig. at 5% level, *=sig. at 10% level;
(a) School and time dummy variables included, time variant controls are mean KS2 prior attainment, % free school meal, % English as an additional language, % white British ethnicity;
(b) School, time, subject and time-subject dummy variables included, within-school subject differences at t-2 included, time variant controls are mean KS2 prior attainment, % free school meal, % English as an additional language, % white British ethnicity;
(c) Pupil, subject and subject-year dummy variables included, subject variant controls are pupil prior attainment in subject, sex and English as an additional language status
(d) Here we only report the year two impact of intensity in the pupil fixed effect model