Education Economics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cede20

# The joint evaluation of multiple educational policies: the case of specialist schools and Excellence in Cities policies in Britain

Steve Bradley [a] & Giuseppe Migali [a b]

[a] Department of Economics, Lancaster University Management School, Lancaster University, Lancaster, LA1 4YX, UK

[b] Dipartimento S.G.S.E.S., Universita' Magna Graecia, Catanzaro, Italy

Version of record first published: 16 Apr 2012.

To cite this article: Steve Bradley & Giuseppe Migali (2012): The joint evaluation of multiple educational policies: the case of specialist schools and Excellence in Cities policies in Britain, Education Economics, 20:3, 322-342

To link to this article: http://dx.doi.org/10.1080/09645292.2012.678715

*Education Economics*
Vol. 20, No. 3, July 2012, 322–342

# The joint evaluation of multiple educational policies: the case of specialist schools and Excellence in Cities policies in Britain

Steve Bradley[a] and Giuseppe Migali[ab]*

[a]*Department of Economics, Lancaster University Management School, Lancaster University, Lancaster LA1 4YX, UK;* [b]*Dipartimento S.G.S.E.S., Universita' Magna Graecia, Catanzaro, Italy*

(*Received 8 December 2011; final version received 16 March 2012*)

Governments frequently introduce education policy reforms to improve the educational outcomes of pupils. These often have simultaneous effects on pupils because they are implemented in the same schools and at the same time. In this paper, we evaluate the relative and multiple overlapping effects of two flagship British educational policies – the Excellence in Cities initiative and the specialist schools policy. We compare the estimates from multi-level cross-sectional and difference-in-differences (DID) matching models. The policy impacts estimated from cross-sectional models are typically positive, quite large and rise over time. The specialist schools policy had a much greater impact on test scores. However, DID matching estimates of the overlapping policies show an increase in GCSE test scores by only 0.5–1 point. We interpret this result as a small causal effect arising from complementarities between the two policies.

**Keywords:** policy evaluation; matching and multi-level models

*JEL classification* I20; I21; I28

## 1. Introduction

A perennial, and almost, universal problem facing governments and policy makers around the world is how to most effectively raise educational performance among secondary school pupils. Britain is no exception. In fact, successive British governments have introduced numerous reforms to the education system, and simultaneously pumped considerable financial resources into the educational system in an attempt to improve the educational outcomes of pupils.[1]

Many of these policies have overlapping or joint effects on pupils. They are often implemented in the same schools and at the same time. However, most analyses of the impact of education reforms typically focus on each policy separately. More generally, in those studies which do examine the effects of multiple policies (outside of the education field), they often treat the policy effects as mutually exclusive. The main contribution of this paper is that we analyse the relative and the multiple overlapping effects of two flagship educational policies that have been implemented in England – the specialist schools initiative and the Excellence in Cities (*EiC*) programme. Although all schools could, in principle, participate in the specialist schools initiative, it was

*Corresponding author. Email: g.migali@lancaster.ac.uk

ISSN 0964-5292 print/ISSN 1469-5782 online

http://dx.doi.org/10.1080/09645292.2012.678715
http://www.tandfonline.com

actually the case that better schools, as measured by their test score performance, were early adopters, which raises the possibility of a selection effect (*school selection bias*). Note, however, that by 2007, almost all schools were participating in the initiative and our analysis of test scores focuses on the 2002–2006 period. It is also possible that specialist schools cream-skimmed the best pupils so ensuring high average test scores (generating a *pupil selection bias*). Bradley, Migali, and Taylor (2012) provide some evidence for the existence of these selection effects which could generate a positive bias in the estimated effect of the specialist schools policy. Clearly, the estimated effect would not be causal if attempts are not made to mitigate these sources of bias. In contrast, the *EiC* programme focused on pupils from disadvantaged socio-economic backgrounds in deprived metropolitan areas, and the latter could generate a negative bias on the estimated effect of this policy due to unobserved *district heterogeneity*. Since some schools were part of both policy initiatives, the three sources of bias could be simultaneously present.[2]

The econometric evaluation literature has increasingly adopted propensity score matching methods to obtain suitable treated and control groups and then combined this with a difference-in-differences (DID) analysis to remove pupil and school selection bias and district-level unobserved heterogeneity. This approach requires repeated cross-sections of pupil data which we have in the form of several versions of the National Pupil Database. Another contribution of this paper is that we adapt matching methods to investigate the effects of the overlapping policies. Specifically, we estimate the joint effects of the two policies versus the effect of no policy (*multiple overlapping treatment effects*) or of one policy versus another policy (*relative treatment effects*). However, given the focus of the *EiC* policy, care has to be taken in selecting pupils for our treatment and control groups such that they come from observably and unobservably similar educational districts. To address this issue, we exploit the hierarchical nature of our data, insofar as pupils are nested in schools and schools within education districts. In the education evaluation literature, many papers have accounted for the nested nature of the data, but we are the first to combine multi-level modelling with cross-sectional matching methods in a multiple-treatment framework. This is the third contribution of our paper. Finally, a comparison of the two approaches (DID analysis and multi-level modelling) will help to reveal the extent of the bias generated by the selection effects described above.

Our work has important policy implications. The two policies were based on entirely different approaches to improving standards, yet the resource cost was similar in terms of the grant per pupil. It is therefore useful to compare their relative effectiveness in improving test score outcomes. This comparison is particularly pertinent given that expenditure on the specialist schools initiative favoured schools with above-average levels of attainment, whereas the *EiC* policy favoured schools with low levels of attainment (Bradley and Taylor 2010). Also, in view of the fact that some schools did benefit from both policies simultaneously, it is also important to understand whether the impact on test scores of having both policies in place is greater than the sum of each separate policy.

The findings of our analysis show that for those schools participating in a single initiative, it was the specialist schools initiative that had the greatest impact on pupil test scores, raising this by between 3 and 6 GCSE points over the period 2002–2006. These findings should be interpreted as correlations rather than causal effects. In terms of the effect of multiple treatment, the policy effects are always positive and statistically significant, however, the DID matching estimates are small in magnitude, raising GCSE test scores by between 0.5 and 1 point. This is interpreted as a complementarity between the *EiC* and specialist schools policies; it is also a causal effect of the multiple treatment since we are able to control for the various types of bias described above.

The remainder of this paper is structured as follows. Section 2 briefly discusses the previous literature and the two policies in more depth. Section 3 explains our econometric approach, which is followed in Section 4 by a discussion of our data. Section 5 discusses the results of our analysis, followed by our conclusions.

## 2. The policies and previous literature

Table 1 describes the key features of the two policies. The specialist schools policy began with the designation of technology schools in 1994, which now constitute approximately 20% of all schools. Significant proportions of schools also focused on Arts, Sport and Science and other specialisms, such as Business and Maths were introduced more recently in 2002.[3] It was the Government's stated aim that all secondary schools in England would ultimately have specialist status, and to date, over 80% of secondary schools have acquired specialist status. The objective of the specialist schools policy was to raise the educational performance of pupils by increasing pupil choice via two mechanisms.[4] Firstly, by allowing schools to specialise in subjects in which they had a comparative advantage, they could potentially increase allocative efficiency, insofar as pupils and teachers sort into those schools that best matches their aptitudes. We refer to this effect as a specialisation effect. Secondly, specialist schools received an increase in funding – a funding effect. Schools who were a part of the initiative received a considerable cash lump sum (£100,000) and an increase in per pupil funding of £129, which together was equivalent to a 5% increase in per pupil funding. The increase in funding was, however, limited to a 4-year period. Schools that were admitted to the specialist schools initiative also had to demonstrate that they could attract matched funding from the private sector. This is one reason why 'good' schools tended to dominate early entry to the initiative.

In contrast, the *EiC* programme focused on pupils from disadvantaged socio-economic backgrounds in deprived metropolitan areas, and extra resources, totalling £1.7b or approximately £150 per affected pupil per annum, was provided to schools. Launched in 1999, it initially targeted 471 secondary schools in 25 local education authorities in the major cities of England. This coverage was extended in 2000 (351 schools in phase 2) and again in 2001 (165 schools in phase 3). Thus, by 2001, the *EiC* policy covered approximately one-third of all secondary schools (Kendall et al. 2005). The overall objectives of the programme were to improve educational performance, by raising the motivation and expectations of pupils, improving the quality of teaching and changing the school's ethos (i.e. through local 'partnerships'). In terms of the pupils, the policy sought to diversify provision in secondary schools so that the needs of all pupils ('gifted and talented' as well as 'disadvantaged') were met.

There have been relatively few attempts to evaluate the impact of the *EiC* programme. In a detailed review, Kendall et al. (2005) conclude that the programme created a positive ethos towards learning, resulting in improved pupil motivation, behaviour and attendance. Improvements in test scores, however, were confined to maths scores for pupils aged 14 and to pupils in the most disadvantaged schools. In further work, Machin, McNally, and Meghir (2004) estimate that the short-run impact of the *EiC* programme was modest. Focusing on the effects of the *EiC* policy

Table 1. A comparison of EiC and specialist schools policies.

| Policy categories | EiC | Specialist schools |
|---|---|---|
| Target group | Schools in disadvantaged metropolitan areas | All schools, but actually ’good’ schools |
| Start date | 1999 | 1994 |
| Scale by 2006–2007 | 33% of secondary school | 85% of secondary schools |
| Funding | £140 per student per annum | £129 per student for 4 years and lump sum |
| Mechanisms | Funding effect | Funding and specialisation effects |

on exam outcomes at age 14 (i.e. Key Stage 3), they find that the policy increased the probability of a pupil attaining level 5 or better by 3.4% points. The magnitude of this effect fell to 2.4% points for Phase 3 schools. Similar results have been obtained by Bradley and Taylor (2010) using a panel of secondary schools in England; however, they do show that the impact of the *EiC* programme increased over time.

The evidence on the impact of the specialist schools policy is more extensive but is conflicting.[5] Positive and statistically significant effects of the specialist schools policy on test scores are provided by Gorard (2002), Jesson and Crossley (2004) and OFSTED (2005). These early studies typically found quite large effects – for instance, Jesson and Crossley find that the policy increased total GCSE points score by 4.2%. Schagen and Goldstein (2002) and Noden and Schagen (2006), who are especially critical of these school-level analyses, argue that multi-level modelling techniques should be used to take into account the multi-level structure of the data. Schagen et al. (2002) are one example of a study which uses pupil-level data and multi-level modelling techniques, and they suggest that estimated effects of the policy are considerably smaller at between 0.02 and 0.11 of a GCSE point with some variation by subject. Taylor (2007) also finds that the specialist schools policy has had very little impact on exam results. Benton et al. (2003), again using multi-level modelling techniques on cross-sectional pupil-level data, find that the specialist schools policy raised GCSE grades, or points, by 1.1, whereas Levacic and Jenkins (2004) found a very similar effect (1.4 GCSE points). Bradley, Migali, and Taylor (2012) use DID matching methods and find a modest causal effect on pupil test scores of between 0.4 and 0.9 of a GCSE point.

There is also a broader literature on multiple-treatment and/or multi-level matching methods – this is not extensive and is relatively recent, and in most cases the two approaches are treated separately. The seminal papers are Imbens (2000) and Lechner (2001a) which extend the binary treatment matching methodology to the multiple-treatment case, ignoring the multi-level setting. Their work is focused on the properties of the estimator, although Lechner (2002) provides a detailed empirical application with respect to Swiss labour market policies. The Lechner approach has also been applied in medical studies and in one recent paper in the field of education by Buonanno and Pozzoli (2009). They consider different university subjects as multiple treatments and study their effect on early labour market outcomes. However, their analysis ignores the effect of overlapping policies. An attempt to consider multiple overlapping treatments is that of Cuong (2009), who adopts a simulation approach and finds that when one controls for simultaneous participation in several treatments, more efficient propensity score matching in terms of the MSE is found. This approach, although useful in highlighting the problem of correlated treatments, does not provide a practical tool to implement it. Our model is closest methodologically to that of Lechner (2001a) and we study the overlapping case by defining a treatment status where both policies are active in the school at the same time.

The literature on multiple treatments in a multi-level setting is based on different approaches. For instance, Rosenbaum (1985) considers the effect of high school drop-out on academic performance, first using school dummies in the propensity score and then within-school matching, arguing that the latter also controls for between-school effects. Arpino and Mealli (2011) control for the omitted variable bias due to unobservables and individual-level covariates. They argue that matched treated and control groups are required to belong to similar, but not identical cluster groups, which in our case is the educational district in which the pupil studies. A two-stage methodology is proposed, whereby they first estimate a multi-level model for the selection process, and in the second stage they estimate the propensity score which includes the random effect obtained in the first stage (see also Hong and Raudenbush 2006; Kim and Seltzer 2007; Su and Cortina 2009). Our work is the first evaluation study to combine multiple overlapping treatments and multi-level models.

## 3. Econometric approach

The definition of the causal evaluation problem follows the standard model of Roy (1951) and Rubin (1974) which has been extended by Imbens (2000) and Lechner (2001a) to the case of multiple treatments. We consider $K+1$ mutually exclusive treatments, $T$, and in our model a pupil receives the treatment by attending a school where a policy is implemented. Two types of control groups can be constructed – one where schools have 'no policy' active and the treatment group consists of pupils in schools where one or more policies are implemented. Another control group is obtained where one policy is compared with another policy or policies.

The outcomes of the treatments (the test scores) are $K+1$ and denoted by $Y_0, Y_1, \ldots, Y_K$. For each pupil $i$ we can only observe one of them, which means that for $k=1$ we observe $Y_{i1}$ and the other $K$ outcomes are counterfactuals. The number of pupils in the population is such that $N=\sum_{k=0}^{K} N_k$, where $N_k$ is the total number of pupils in schools where policy $k$ is active.

Our parameter of interest is shown in Equation (1)

$$\theta^{k,g}=E(Y_{ik}-Y_{ig}=k)=E(Y_{ik}=k)-E(Y_{ig}=k). \tag{1}$$

$\theta^{k,g}$ denotes the expected average policy effect of policy $k$ relative to the policy $g$ for pupils in schools with policy $k$ (sample size $N_k$).[6] The problem is that the expected outcome $E(Y_{ig}=k)$ cannot be observed for the same pupil $i$. We address the fundamental evaluation problem by estimating $\theta$ using propensity score matching methodologies.

One assumption of the matching method is the *common support* or overlap condition which ensures that pupils with the same characteristics have a positive probability of attending a school of type $k$. A second and key assumption is the *conditional independence assumption* (*CIA*), which implies that selection into treatment is solely based on observable characteristics. Imbens (2000) and Lechner

(2001a) show for the case of multiple treatments that identification of $\theta^{k,g}$ comes from the CIA.[7]

If the CIA and the common support condition hold, the matching method allows us to estimate $\theta$ as an average treatment effect on the treated (ATT). The ATT estimator is the mean difference in outcomes over the common support, weighted by the propensity score distribution of participants.

$$\theta_{\mathrm{ATT}}^{k,g} = \sum_{i \in k} \left( Y_{ik} - \sum_{h \in g} W_{ih} Y_{hg} \right) w_i, \tag{2}$$

where $W_{ih}$ is the weight placed on the $h$th observation in constructing the counterfactual for the $i$th treated observation, and $w_i$ is the re-weighting that reconstructs the outcome distribution for the treated sample. A number of well-known matching estimators exist which differ in the way they construct the weights, $W_{ih}$. We use the nearest neighbour algorithm and provide analytical standard errors (Abadie and Imbens 2008). In all estimations, we only consider observations on the common support.

To allow for the fact that schools may participate in one or both policies, or they may participate in neither policy, as stated above, we let the treatment variable $T$ take one of $K$ discrete values. This multiple-treatment problem can be seen as several binary problems, and we may estimate the $K(K-1)/2$ binary conditional probabilities which will be used as propensity scores in the matching model. However, we use a more parsimonious way of parametrising the various propensity scores that are required to achieve the balancing property, which is based on the multinomial choice model, as suggested by Lechner (2002).[8] Equation (2) is estimated using cross-sectional data, which allows us to exploit a large number of observations and to estimate the overlapping treatment effects and relative treatment effects of the two educational policies.

The main problem with the estimation of Equation (2) is that it does not take account of the several forms of bias explained in the Introduction. Thus, we consider two other matching approaches. The first is a DID matching analysis which we believe is able to remove all the remaining forms of bias, and the second is multi-level propensity score cross-sectional matching, which takes into account, explicitly, the fact that the *EiC* policy was focused on schools in disadvantaged educational districts in metropolitan areas. Unfortunately, we can only compare these approaches in the case of overlapping treatment effects, that is, the case where both the *EiC* and the specialist schools policy are active versus one where only the specialist school policy is active. This restriction arises because we do not have observations on pupil test scores in schools in the pre-*EiC* period, which thus precludes a DID analysis.

### 3.1 *Multiple-treatment DID models*

Blundell et al. (2004), Smith and Todd (2005) and Machin, McNally, and Meghir (2004) combine matching methods with DID analysis. We extend this methodology to allow for multiple treatments. The DID matching estimator allows the controls to evolve from a pre-policy to a post-policy period in the same way treatments would have done had they not been treated. This approach relaxes the strong assumption of the cross-sectional matching approaches of selection based solely on observables.

The DID matching estimator for repeated cross-section data allows for temporally invariant differences in outcomes, between pupils in schools adopting one policy versus pupils in the same school who adopt an additional policy. It is obtained by re-writing Equation (2) as

$$\tau_{\mathrm{ATT}}^{\mathrm{DID}} = \sum_{i \in T_t} \left[ Y_{kti} - \sum_{h \in C_t} W_{ih} Y_{gth} \right] w_{it} - \sum_{i \in T_{t'}} \left[ Y_{gt'i} - \sum_{h \in C_{t'}} W_{ih} Y_{gt'h} \right] w_{it'}, \tag{3}$$

where $t'$ and $t$ are time periods before and after the adoption of an additional policy. If we want to compare the effect of policy $k$ versus policy $g$, $T_{t'}$ is formed by pupils in schools where policy $g$ is active and policy $k$ is not active in $t'$ – only $k$ will be active in $t$. $C_{t'}$ is formed by pupils in schools where policy $g$ is active and policy $k$ is not active in $t'$ and will remain so in $t$. $T_t$ includes pupils in schools where policy $k$ is active in $t$ and where only policy $g$ was active in $t'$. Finally, $C_t$ includes pupils in schools where policy $k$ is not active in $t$ and where only policy $g$ was active in $t'$. We use the same propensity score in both cross-sections, pre- and post-policy, and it is computed using pre-treatment variables.

## 3.2 *Multi-level propensity score matching models*

An alternative approach, which directly addresses the issue of unobserved district level heterogeneity, is to estimate multi-level propensity score models. We know that pupils are nested in schools, and schools are administered within districts, which may differ with respect to the socio-economic composition of the population and with respect to the application of the educational policy. As suggested above, it is important to take account of these differences because the *EiC* policy was explicitly targeted at schools in disadvantaged metropolitan areas. Our purpose is therefore to find more homogenous treatment and control groups to estimate our parameter of interest $\theta$ from Equation (2).

Our models differ from traditional multi-level models estimated by educational researches in a number of ways. For instance, because the policy is applied at school level, which implies that within each school pupils receive the same treatment, we cannot consider the school as a separate hierarchical level. Therefore, we have two levels – the pupil and the educational district. However, we can include observable school characteristics in the estimation of the propensity scores, and therefore potentially (indirectly) control for differences between schools in the application of policies.

We distinguish between two types of multi-level models. The first is the random intercept model, where we assume that the district unobserved heterogeneity comes only from differences in the intercept. The second is a random slope model which is less restrictive because it allows for variation in the selection of schools into each policy initiative by educational district due to differences in pupil characteristics (e.g. gender or ethnicity).

In each case, a binary logit model is estimated where the treatment variable $T = 1$ if a given school has implemented the policy $k$, and $T = 0$ if the policies $g$ are active. Defining $\pi_{ij} = E(T_{ij}ij, u_j) = P(T_{ij} = 1)$, a random intercept model is given as

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + \beta_1 X_{ij} + u_j, \tag{4}$$

where $i$ are $N$ pupils and $j$ are $M$ districts. $u_j N(0, \sigma_u^2)$ is the random effect, or level 2 residuals, and it absorbs the district-specific constants effects of unobserved district-level predictors. The coefficient $\beta_0$ is the overall intercept in the linear relationship between the log-odds and $X$, and the intercept for a given district $j$ is $\beta_0 + u_j$. $\beta_1$ is the cluster-specific effect, that is, holding $u$ constant, it shows the effect of $X$ for pupils in the same educational district.

To match pupils, we rearrange Equation (4) and the propensity score can be given as:

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 X_{ij} + u_j)}{1 + \exp(\beta_0 + \beta_1 X_{ij} + u_j)}. \tag{5}$$

A random slope logit model is an extension of Equation (4)

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + \beta_1 X_{ij} + u_{0j} + u_{1j} X_{ij}. \tag{6}$$

The district-level variation comes from the random effects, $u_{0j}$, and the interaction of the random effects and pupil-level characteristics, $u_{1j}X_{ij}$. As above, the propensity score is obtained rearranging Equation (6).

Moreover, since we are aware that cross-sectional models cannot control for all the different forms of bias outlined in the Introduction, we perform some robustness checks. In particular, we want to assess whether our ATT estimates are affected by the particular assumptions in the estimation of Equations (4) and (6). First, we relax the standard normality assumption of $u_j$, because it is possible that it produces biased estimates for the random intercept model. We therefore estimate Equation (4) adopting a non-parametric maximum likelihood (NPMLE) approach (Rabe-Hesketh, Skrondal, and Pickles 2004).[9] Second, we re-estimate Equation (6) to allow for correlation between $u_{0j}$ and $u_{1j}$.

In general, we expect that if we produce less biased estimates of the multi-level propensity score, it would marginally affect the composition of the comparison groups, and consequently the ATT should be substantially unchanged. We think that what matters in controlling for district heterogeneity is mainly allowing for random intercept variations.

To get an idea of the importance of the district-level heterogeneity, we plot the estimates of the random effect, $u_j$, with confidence intervals at 5% (i.e. a caterpillar plot) on the log-odds scale. Thus, we observe for each district the departure from the overall intercept ($u = 0$), which implies that a district $j$ whose confidence interval does not overlap with the horizontal line (at zero) differs significantly from the average log-odds of the treatment at the 5% level.

## 4. Data description

The data that we use in our analysis is the National Pupil Dataset (NPD). The NPD refers to the population of pupils attending maintained, state-funded, schools in England. The primary advantages of the NPD are that it refers to the population of pupils in secondary schooling, hence providing a large number of observations, and there are several measures of test scores. Our dependent variable is constructed from national test scores obtained by pupils at Key Stage 4, i.e. GCSE tests taken typically in around 8–10 subjects at the age of 16. We derive the total GCSE score, that is, the number of points achieved in all GCSE subjects, where grades are ranked from $A^* = 8$

points to fail = 0. Another important advantage of the NPD is that it also includes a measure of pupil attainment prior to entry into secondary schooling, that is, the Key Stage 2 tests taken at age 11.

We consider five versions of the NPD where pupils were in their final year of compulsory education in one of the years between 2002 and 2006. The original sample sizes are between 500,000 and 560,000 observations according to the year considered. We choose these cohorts of the NPD because this is the period in which many secondary schools acquired specialist school status, whereas schools joined the *EiC* programme between 1999 and 2001. Schools are clustered into one of 344 educational districts across England, and in each district, administrators monitor and liaise with local schools to ensure that the national policy is applied.[10]

To each NPD dataset, we append school-level data from the annual School Performance Tables and the annual Schools' Census, which cover the period 1994–2007. We also know when a school became specialist over the period 1994–2006 and when a school became part of the *EiC* programme over the period 1999–2001. This means that we can track each school and observe whether it has joined one or more programmes, or neither programme. To perform our multiple-treatment analysis, we create a categorical variable and schools/pupils fall into one of four treatment statuses:

- Category 0: No policy – schools that never joined the *EiC* or the specialist schools initiatives. (We also include schools that became specialist after our sample of pupils have obtained their GCSEs for reasons that are explained below.)
- Category 1: Specialist only – schools that joined only the specialist schools programme before the pupil enrolled, and never joined the *EiC* programme.
- Category 2: *EiC* only – schools that only joined the *EiC* programme before the pupils had obtained their GCSEs.
- Category 3: schools that are part of both programmes; specifically, schools that acquired *EiC* and specialist status and did so before the pupil enrolled in the school.

Category 3 is particularly relevant for our analysis because it highlights the case of overlapping policies. The same pupil receives two treatments at the same time, therefore, we have both a multiple-treatment and joint effects. In Table 2 we show the number of schools in each category by enrolment year and treatment period. Note that the size of each category varies over time as schools switch from category 0 to one of the other categories. We omit schools that acquire the specialist status while the pupils are still enrolled and hence the sample size is decreasing.

The categorical variable above becomes the dependent variable in the multinomial logit model (MNL). The conditional probabilities, used as propensity scores in each matching estimation, are computed for all combinations of the four categories.

In Table 3, we report the means and standard deviations of GCSE point scores for each cohort of pupils (2002–2006) sub-divided by category 0–3. This shows that the average GCSE score in schools which are only specialist is around 7–12 points higher than the average score in schools only in the *EiC* programme. When both policies are overlapping in the same schools the average score is between 6–9 points higher than the *EiC*-only policy and 2–4 points lower than the specialist-only policy. The beneficial effects of the multiple treatments are clear, and also increasing over time.

Table 2. Number of schools by treatment category.

| | Year of enrolment | | | | |
|---|---|---|---|---|---|
| Policy categories | 1998 | 1999 | 2000 | 2001 | 2002 |
| No policy | 1335 (64) | 1000 (58) | 653 (47) | 402 (34) | 332 (25) |
| Specialist only (from 1994) | 195 (9) | 237 (14) | 309 (22) | 394 (33) | 607 (45) |
| EiC only | 490 (23) | 407 (23) | 310 (22) | 213 (18) | 183 (14) |
| EiC−specialist | 69 (3) | 91 (5) | 124 (9) | 169 (14) | 224 (17) |
| Total | 2089 (100) | 1735 (100) | 1396 (100) | 1178 (100) | 1346 (100) |
| DID analysis | Control period | | Treatment period | | |
| | | | 2003−2004 | 2003−2005 | 2003−2006 |
| EiC only | 2002−2004 | 310 | | | |
| EiC only | 2002−2005 | 213 | | | |
| EiC only | 2002−2006 | 183 | | | |
| EiC−specialist | | | 180 | 277 | 307 |
| Total (C+T) | | | 490 | 490 | 490 |

Notes: % in parenthesis. Total sample 2645 schools. Schools that become specialist during the study period are excluded. But schools that become specialist the year after the pupil get his GCSE are included.

Table 3. Average GCSE test scores.

| | Year of GCSE | | | | |
|---|---|---|---|---|---|
| Policy categories | 2002 | 2003 | 2004 | 2005 | 2006 |
| No policy | 41.566 (18.685) | 40.944 (19.230) | 39.646 (20.163) | 39.104 (20.456) | 44.222 (17.180) |
| Specialist only | 46.557 (18.873) | 47.461 (19.960) | 46.851 (20.763) | 47.139 (21.169) | 49.971 (17.499) |
| EiC only | 36.603 (19.098) | 36.346 (19.993) | 35.330 (21.144) | 35.392 (21.548) | 42.180 (18.596) |
| EiC−specialist | 42.197 (19.094) | 43.039 (20.573) | 44.162 (22.186) | 44.884 (22.748) | 48.581 (18.284) |
| Total | 40.952 (19.046) | 41.066 (19.927) | 41.047 (21.172) | 42.555 (21.781) | 47.748 (17.936) |

Note: s.e in parenthesis.

## 5. Results

In discussing the results of the various models that are estimated, it is worth noting that in all cases we allow for possible temporal heterogeneity in the policy effects by looking at how these vary over time. The control group also varies over time and we make this clear in each section. In Figures 1 and 2, we show the density distribution of the propensity scores by treated and control groups for a selection of our models. This allows us to demonstrate that overall, for any value of the propensity scores in the treated group, we can find at least one observation in the control group with the same value.[11] In our analysis, we only consider observations on the common support.

In the next section, we discuss the effects of single and multiple treatments on test scores, which is then followed by discussion of the relative effects of educational policies on the test score outcomes of pupils.

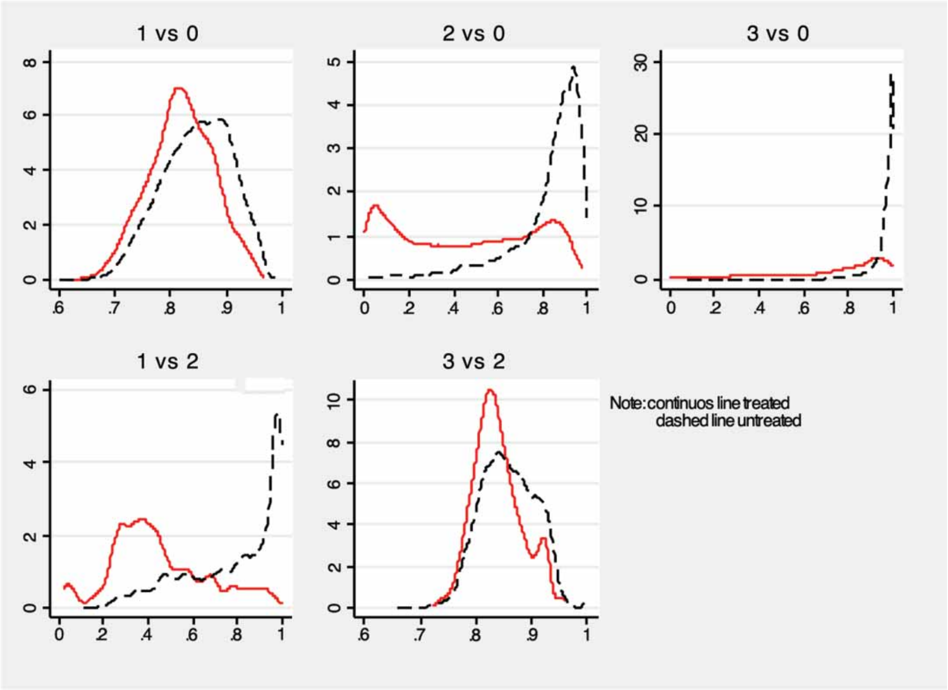

Figure 1. Selected plots of propensity scores distribution, MNL model (2002 only).

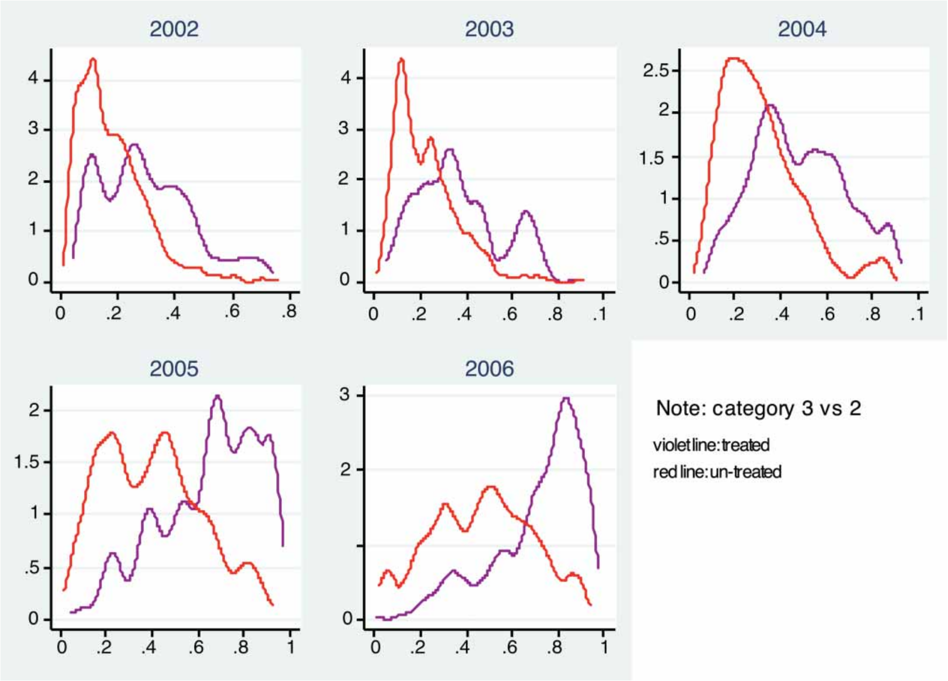

Figure 2. The propensity scores distribution, random intercept model.

### 5.1 *Single versus multiple-treatment effects of educational policies*

The first three columns of Table 4 show the effect of the specialist schools policy (category 1), the *EiC* policy (category 2) or both policies (category 3), and in each case, the control groups are those schools where no policy is implemented (category 0). Note that the control group is allowed to vary over time, insofar as category 0 includes schools that eventually become specialist after the pupil has obtained their exam results. This is because schools that will eventually become specialist schools are more like those that already are specialist. This implies that our control group (category 0) will decrease in size over time.

Focusing on the first three columns of Table 4, it is clear that in this analysis the effect of the specialist schools policy (column 1) is always positive and between 5 and 7 GCSE points, an effect that is generally increasing over time. Matching reduces this effect by between 1 and 2 GCSE points; it can be claimed therefore that the specialist schools policy increased test scores by between 4 and 6 points. These results are higher than those of previous studies discussed in the literature review and imply the presence of selection bias.

In contrast, column 2 shows that pre-matching there is a large negative, though declining, effect of the *EiC* policy on test scores whereas post-matching there is virtually no statistically significant effect of the *EiC* policy; in fact, the estimates of the ATT show that over time the effect goes from a small negative effect to a small positive effect. The exception is 2003 when the *EiC* policy has a modest effect in raising test scores by 2 GCSE points, less than half of the specialist schools policy effect. It is likely that the results for the *EiC* policy arise because of the differences between educational districts, and that matching goes part way to mitigating this problem. This is an issue to which we return below. These results are not comparable with previous studies of the *EiC* initiative. Earlier papers focused on test scores at the age of 14 in the case of Machin, McNally, and Meghir (2004), or they have focused on the proportion of pupils obtaining five or more GCSE grade A to C in the case of Bradley and Taylor (2010).

When we consider the effect of multiple overlapping treatments (column 3), the policy effects are positive, statistically significant and generally higher post-matching. The size of the effects rise over time and are quite large. Moreover, the multiple-treatment effects always exceed those of the single treatments, especially the *EiC* policy effect, which is probably unsurprising because of our findings for the specialist schools policy effect. However, of more interest is the finding that as we move beyond 2003, the effect of multiple treatment (post-matching) exceeds both single policy effects, particularly for 2006, which suggests that there are complementarities between the *EiC* and specialist school policies.[12] They are also quite large insofar as they range from 3 to 6 GCSE points. These complementarities could be due to the joint funding effect, or a combination of the joint funding effect and a specialisation effect. However, they should be interpreted with caution because no attempt is made to control for selection bias or unobserved district heterogeneity.

### 5.2 *The relative effects of educational policies on pupil test scores*

In this section, we explore the relative effects of the two policies. Here the control group consists of pupils in schools that are part of the *EiC* policy between 1999 and 2001. Two sets of treatment groups are identified. The first treatment group consists of pupils in specialist schools and the second consists of pupils in schools that have

Table 4. Estimates of multiple and relative treatment effects using the MNL model.

| | (a) Single versus multiple treatment | | | (b) Relative effects | |
|---|---|---|---|---|---|
| Policy categories | (1) versus (0) | (2) versus (0) | (3) versus (0) | (1) versus (2) | (3) versus (2) |
| 2006 | | | | | |
| Unmatched | 5.702*** (0.107) | −2.089*** (0.156) | 4.360*** (0.134) | 7.792*** (0.138) | 6.449*** (0.166) |
| ATT | 3.290*** (0.206) | 0.459 (0.397) | 6.595*** (0.404) | 5.540*** (0.654) | 5.756*** (0.279) |
| 2005 | | | | | |
| Unmatched | 7.764*** (0.112) | −3.273*** (0.143) | 5.852*** (0.145) | 11.037*** (0.139) | 9.125*** (0.171) |
| ATT | 6.258*** (0.205) | 0.675 (0.541) | 6.462*** (0.778) | 5.760*** (0.508) | 6.559*** (0.271) |
| 2004 | | | | | |
| Unmatched | 6.998*** (0.103) | −4.027*** (0.111) | 4.601*** (0.145) | 11.026*** (0.124) | 8.629*** (0.166) |
| ATT | 5.099*** (0.169) | −0.481* (0.292) | 5.370*** (0.336) | 6.480*** (1.028) | 6.981*** (0.259) |
| 2003 | | | | | |
| Unmatched | 6.444*** (0.104) | −4.615*** (0.091) | 2.094*** (0.154) | 11.059*** (0.122) | 6.709*** (0.170) |
| ATT | 5.214*** (0.174) | 2.080*** (0.702) | 4.844*** (0.315) | 4.181*** (0.294) | 5.047*** (0.276) |
| 2002 | | | | | |
| Unmatched | 4.796*** (0.107) | −4.605*** (0.081) | 0.653*** (0.171) | 9.401*** (0.122) | 5.259*** (0.183) |
| ATT | 3.719*** (0.172) | −0.260 (0.206) | 3.325*** (0.307) | 3.892*** (0.365) | 5.086*** (0.288) |

Notes: (0) No policy, (1) specialist schools only, (2) EIC schools only, (3) specialist and EIC schools. Propensity includes (a) individual characteristics: gender, non-white pupils, prior attainment at age 11; (b) pre-policy school characteristics: pupils–teacher ratio, (%) of pupils eligible for free school meals, number of pupils, (%) part-time pupils, comprehensive and modern school.

both *EiC* and specialist school status. The first analysis seeks to address the question of whether one policy is better than another in improving test scores, whereas the second analysis addresses the slightly different question of whether multiple policies have a larger impact than a single policy.

Table 4, columns 4 and 5, shows the pre-matching estimates and the ATT by year. Column 4 shows that the specialist schools policy has a larger effect than the *EiC* policy. Thus, post-matching pupils in a school that has a specialist status achieve 4–6 GCSE points more than their counterparts in schools that only have *EiC* status, and these effects rise up to 2004 and then stabilise. The effects of multiple treatment versus single treatment are reported in column 5 of Table 4. The estimated ATT of multiple overlapping treatments is always positive and statistically significant, which means that a pupil in a school that has the benefit of both policies achieves 5–6 GCSE points more than their counterpart who has the benefit from only the *EiC* policy. This effect is remarkably stable over time. Furthermore, in view of the fact that the effect of the *EiC* policy is almost always zero when compared with the 'no policy' control group (column 2), the estimates in column 5 confirm the view that there are important complementarities between these policies. What is more, since the estimates in column 5 are from models that essentially hold constant the funding effect, this strongly suggests that the differences between the ATTs in columns 2 and 5 are due to the specialisation effect. This is therefore one important way in which the two educational policies complement one another. Furthermore, it is possible to observe a similar, though less pronounced, effect by comparing the ATT estimates in columns 4 and 5, the advantage being that the control group is the same (*EiC* only). The difference between the ATT effects in these two columns is always positive but decreasing, and ranges from 1.2 GCSE points in 2002 to 0.2 GCSE points in 2006.

### 5.3 *A multiple-treatment DID analysis*

As already suggested, the problem with the estimates from the analysis in Sections 5.1 and 5.2 is that they are likely to be upwardly biased by the effect of both pupil and school selection bias. Thus, although these estimates are of interest insofar as they indicate which policies are most effective in raising test scores and are suggestive of complementarities between policies, they cannot be taken at face value. They certainly cannot be interpreted as causal effects. Therefore, in Table 5, we report estimates from Equation (3) in Section 3.1, which in principle does enable us to control for time-invariant pupil and school characteristics and, at least indirectly, the unobserved district effects. The DID ATT is obtained as a difference between two cross-sectional matching models. We employ four versions of the NPDs – cohorts 2002–2006. We assume that the year 2002 is a pre-treatment period and the other three years, separately, as post-treatment periods. Once again, data limitations force us to focus on the comparison of the multiple overlapping (*EiC* and specialist schools policies) treatment versus the single *EiC* policy treatment.

In this analysis, the control group consists of pupils that obtain their GCSE test scores in 2002 in an *EiC* school. For the treatment group, we observe pupils that obtain their GCSEs in 2002 in a school covered only by the *EiC* policy and pupils that get their results in one of the following years (2003–2006) when the same school acquired specialist status. The corresponding cross-sectional matching estimates are reported in Table 5, together with their difference which gives the DID ATT estimator. We use the same propensity score in both cross-sections, including only pre-treatment pupil characteristics and school characteristics. Furthermore, the duration of the *EiC* policy effect is fixed, which is very important because it allows us to use the DID estimation as a robustness check on our previous results.

We find several interesting results. First, the joint effects are positive and highly significant, and the longer the duration of school specialisation, the higher the effect. Moreover, since the effect of the *EiC* policy is constant, hence in the pre- and post-treatment period schools have the same amount of funding, we can interpret the increase in test scores as the specific effect of the specialist schools policy.[13] In the period 2002–2006, which includes schools that have been specialist for a maximum of 4 years, the effect is around 1 GCSE point. In general, we can conclude that, unlike our estimates above, the complementary effect of the specialist policy is not very strong. Nevertheless, there is a real, albeit small, effect due to the complementarities between the two policies.

### 5.4 *Estimates from a multi-level multiple-treatment model*

We explore further the relative effect of the specialist schools and *EiC* initiatives (category 3) versus those only involved in the latter (category 2) by exploiting the hierarchical nature of the data. Specifically, we estimate Equations (4) and (6) in Section 3.2, above. The motivation here is to find more appropriate comparison groups by selecting pupils in schools that are located in unobservably 'similar' educational districts, and by addressing the problem of unobserved district heterogeneity. Although these estimates are not directly comparable with those in Table 5, since the composition of the treatment and control groups differs, they do give an insight into the size of the selection bias, for instance.

The first data restriction we have to impose in order to perform a multi-level estimation is to exclude districts where only one treatment status is observed. Specifically, for each district, we have to find schools in policy category 3, the treatment group, and schools in policy category 2, the control group, which generates a restricted sample of districts.[14]

When we estimate the propensity score using a random-intercept multi-level model,[15] assuming normal random effects, we always find a highly significant intercept variance, $\sigma^2_{u_0}$. Recall that the caterpillar plot of the residuals (Figure 3) shows how educational districts differ with respect to unobservables. It is clear from this that

Table 5. Estimates from a multiple-treatment DID model.

| Policy categories | 2002–2004<br>(3) versus (2) | 2002–2005<br>(3) versus (2) | 2002–2006<br>(3) versus (2) |
|---|---|---|---|
| $T_{t'} - C_{t'}$ | | | |
| Unmatched | 4.642*** (0.141) | 5.311*** (0.140) | 5.124*** (0.145) |
| ATT | 2.334*** (0.213) | 2.676*** (0.220) | 1.719*** (0.230) |
| $T_t - C_t$ | | | |
| Unmatched | 5.363*** (0.150) | 5.764*** (0.152) | 4.349*** (0.157) |
| ATT | 2.780*** (0.224) | 3.310*** (0.238) | 2.697*** (0.257) |
| DID | 0.721*** (0.021) | 0.453*** (0.021) | −0.774*** (0.023) |
| DID ATT | 0.445*** (0.219) | 0.634*** (0.230) | 0.977*** (0.243) |

Notes: robust s.e. in parenthesis. 2002–2004: pre-treatment 2002 and post-treatment 2003 and 2004; 2002–2005: pre-treatment 2002 and post-treatment 2003, 2004, 2005; 2002–2006: pre-treatment 2002 and post-treatment 2003, 2004, 2005, 2006. See Note to Table 4 for variables in the propensity score.

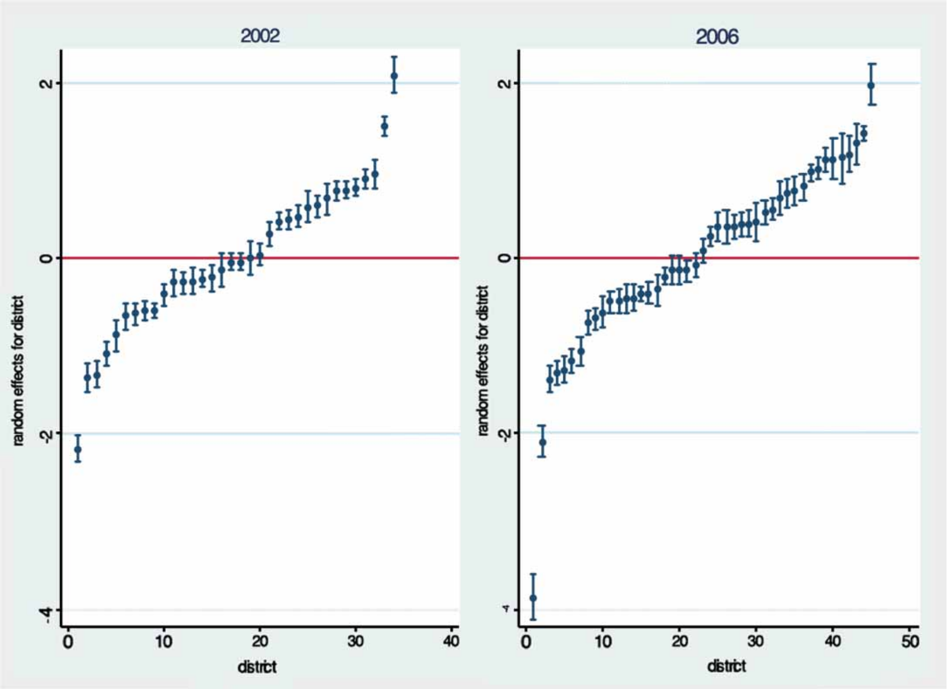

Figure 3. The caterpillar plot of the random effects distribution – random intercept models.

districts in the tails – above and below zero – are clearly different from one another. Repeating the estimation using NPMLE also results in a statistically significant intercept variance. Similarly, for the random slopes model, there is a highly statistically significant intercept variance, $\sigma^2_{u_0}$, when we assume zero correlation between the random effects. This is also true for the slope variances $\sigma^2_{u_{\text{gender}}}$ and $\sigma^2_{u_{\text{ethnicity}}}$. However, when we re-estimate the model assuming a non-zero correlation between the random effects, the intercept–slope covariances $\sigma^2_{u_{0\,\text{gender}}}$ and $\sigma^2_{u_{0\,\text{ethnicity}}}$ are still significant, whereas the slope–slope covariance $\sigma^2_{u_{\text{sex-gender}}}$ is generally insignificant. This evidence suggests that there is variation in the selection of schools (and pupils) into each policy initiative across educational districts.

The findings in Table 6 are not directly comparable with those in column 5 of Table 4, because the sample of pupils differs since we exclude all districts where both policies are not active. In this analysis, we expect the district heterogeneity to have greater effect on the *EiC* policy since it is implemented only in particularly disadvantaged areas, while the specialist school initiative is a national policy. If we look at the MLE random intercept model, the estimates of the ATT are large and increasing from 2002 to 2005. Note that the percentage of specialist schools was relatively small in this period whereas the *EiC* programme had been phased between 1999 and 2001 (Table 2). Therefore, the joint effect of the two policies on test scores is mainly driven by the *EiC* policy. In the later years, more schools are specialist for a longer period, and so it is likely that the specialisation effect prevails and the *EiC* effect fades but the ATT is higher. Looking at the variation between the unmatched and the matched effects by year, in 2002 and 2003, the drop of the ATT is around 10–12%, whereas in the subsequent years it is around 17–18%, which implies that

Table 6. Estimates of multiple treatments using multi-level models.

| Policy categories | 2006<br>(3) versus (2) | 2005<br>(3) versus (2) | 2004<br>(3) versus (2) | 2003<br>(3) versus (2) | 2002<br>(3) versus (2) |
|---|---|---|---|---|---|
| Unmatched | 6.543*** (0.174) | 9.611*** (0.182) | 8.890*** (0.173) | 6.921*** (0.177) | 4.952*** (0.191) |
| Random intercept$_{\mathrm{MLE}}$ | | | | | |
| ATT | 5.328*** (0.291) | 7.991*** (0.313) | 6.807*** (0.275) | 6.215*** (0.282) | 4.347*** (0.298) |
| Random intercept$_{\mathrm{NPMLE}}$ | | | | | |
| ATT | 5.461*** (0.299) | 7.633*** (0.296) | 6.364*** (0.278) | 5.680*** (0.281) | 3.605*** (0.300) |
| Random slopes$_{\mathrm{nocorr}}$ | | | | | |
| ATT | 5.663*** (0.530) | 6.244*** (0.542) | 5.987*** (0.397) | 6.175*** (0.362) | 12.921 (8.811) |
| Random slopes$_{\mathrm{nonzerocorr}}$ | | | | | |
| ATT | 5.306*** (0.471) | 6.229*** (0.489s) | 5.434*** (0.417) | – (–) | 7.919 (8.980) |

Notes: Random slopes are for the gender and ethnicity variables. 2003 third model non-convergence due to data sparcity. See Note to Table 4 for variables in the propensity score.

the effect of unobserved district heterogeneity is higher when the number of specialist school increases. A similar story emerges when we estimate the NPMLE random-intercept model and the random slope models. Thus, although the multi-level analysis provides a control for the effect of unobservable district effects (compare the pre- and post-matching estimates in Table 5), they do not appear to control for pupil and school selection bias since the ATT estimates are similar in magnitude to the estimates in Sections 5.1 and 5.2.

## 6. Conclusion

Successive British governments have introduced a plethora of initiatives, coupled with policies to reform the education system in an attempt to raise educational standards. Two flagship policies, introduced in the 1990s, were the specialist schools initiative and the *EiC* programme, both of which increased funding per pupil, albeit in very different ways and for different target groups. By allowing schools to specialise in particular subjects, the specialist schools initiative also allowed pupils and parents to better match their academic ability to the subject portfolio offered by a school. These initiatives, or policies, had simultaneous effects on pupils because they were implemented in the same schools and at the same time. However, in most previous analyses of the impact of education policies, the evaluation typically focuses on each policy separately. This paper therefore analyses the relative and overlapping effects of the specialist schools and *EiC* policies on the test score performance of pupils in English secondary schools. Furthermore, for the case of multiple overlapping policies, we are able to compare the findings from two different approaches to evaluation – the DID matching method and a multi-level modelling approach.

The findings of our analysis show that for those schools participating in a single initiative, it was the specialist schools initiative that had the greatest impact on pupil test scores, raising this by between 3–6 GCSE points over the period 2002–2006. In terms of the effect of multiple treatment, the policy effects are always positive, statistically significant, and the size of the effects rises over time and are quite large. The estimates of the ATT always exceed those of both single-policy effects, after 2003 and particularly for 2006, which suggests that there are complementarities between the *EiC* and specialist schools policies. However, when we re-estimate using DID matching models, the size of the effect of multiple treatment (*EiC* and specialist school policies) versus a single treatment (*EiC* only) falls substantially to around 0.5–1 GCSE point. Thus, although there are complementarities between the two policies, they are small. Furthermore, unlike the cross-sectional estimates (Table 4) and the multi-level estimates (Table 6), the estimates from the DID matching models are most likely to be causal since this technique removes the effect of pupil and school selection bias and the effect of unobserved district fixed effects. As such, they are our preferred estimates.

The main conclusion of this paper is that there does appear to be a positive, although small, multiple-treatment effect on pupil test scores, which implies that it can be important in policy evaluation to investigate the interactions between policies. The implication for policy making is that some thought does need to be given to the possible interaction effects between policies and initiatives, and it is not clear that education policy makers have actually had this in mind when launching each new initiative. Education policies can complement one another, whereas others may counteract each other. In our particular case, we find some evidence of a complementarity between the specialist schools and *EiC* policies.

## Acknowledgement

We would like to thank Petra Todd and Michael Lechner for useful advice on this paper, and the participants at the Economics seminar at Cattolica University of Milan and conference participants at IWAEE 2011.

## Notes

1. Between 1997 and 2007, for instance, expenditure per pupil increased by around 50% in real terms and total real expenditure on secondary schools increased by 60% from £9.9b in 1997/98 to £15.8b in 2006/7.
2. With the data at our disposal, it is not possible to disentangle school and pupil selection bias; therefore, in this paper we treat them as observationally equivalent.
3. See the Department for Children, Schools and Families website for more details: www.standards.dcsf.gov.uk/specialistschools.
4. It is worth noting that although specialist schools are encouraged to focus on particular subjects, all schools are also required to deliver a national curriculum. Thus, most pupils will typically study around 10 subjects in their final 2 years of compulsory schooling between the ages of 14 and 16. They then sit nationally recognised tests, the General Certificate of Secondary Education (GCSE), in each subject.

   The GCSE is a norm-based examination taken by almost all pupils, and the grades range from A* to G. Grades A* to C are considered acceptable for entry to university, together with the acquisition of advanced qualifications obtained 2 years later. Pupils of lower ability may also take General National Vocational Qualifications instead of GCSEs.
5. For a more detailed literature review, see Bradley, Migali, and Taylor (2012).
6. Note that for ease of exposition we focus on relative policy effects hereon.
7. See the Appendix for further details.
8. We assume the $T_i$ are generated by the multinomial logistic model (MNL) and from its estimation we get the marginal probabilities for each policy, and we compute the conditional probabilities (see Equation (A3) in the Appendix). The next step is to match on these propensity scores and obtain the ATT (Equation (A2) in the Appendix).
9. In practice, the random effects, $u_j$, take on a number of discrete values with a certain probability, which corresponds to assuming that the population falls into a finite number of latent classes. The estimates from this approach are considered non-parametric insofar as the data are divided into the maximum number of classes, such that the addition of further classes cannot increase the likelihood any further (Rabe-Hesketh, Skrondal, and Pickles 2004).
10. Note that educational districts can be clustered within Local Education Authorities, of which there are far fewer.
11. As noted by Caliendo and Kopeinig (2008), when estimating the ATT it is sufficient to ensure the existence of potential matches in the control group.
12. The number of schools that have both policies has also substantially increased as it is clear from Table 2.
13. This effect can be decomposed in a funding effect and specialisation effect; in this case we cannot distinguish between them but this has already been analysed in Bradley, Migali, and Taylor (2012).
14. We are therefore forced to focus on the relative effects of EiC and specialist schools versus those with EiC only because of the fact that we are dealing with a small sub-set of districts (between 34 and 45, according to the year, out of 344). All schools at least have EiC status.
15. For brevity, we omit the results which are available upon request to the Authors.
16. We omit for simplicity the index $i$.

## References

Abadie, A., and G.W. Imbens. 2008. Notes and comments on the failure of the bootstrap for matching estimators. *Econometrica* 76, no. 6: 1537–57.

Angrist, J.D. 1998. Estimating labor market impact of voluntary military service using social security data. *Econometrica* 66, no. 2: 249–88.

Arpino, B., and F. Mealli. 2011. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis* 55, no. 4: 1770–80.

Benton, T., D. Hutchinson, I. Schagen, and E. Scott. 2003. Study of the performance of maintained secondary schools in England. *Report for the National Audit Office*. Slough: NFER.

Blundell, R., M. Dias, C. Meghir, and J. van Reenen. 2004. Evaluating the employment impacts of a mandatory job search program. *Journal of European Economic Association* 2: 569–606.

Bradley, S., G. Migali, and J. Taylor. 2012. Funding, school specialisation and test scores: An evaluation of the specialist schools policy using matching models. Lancaster University Management School working paper. http://www.lums.lancs.ac.uk/publications/view/967/ (accessed February 2, 2012).

Bradley, S., and J. Taylor. 2010. Diversity, choice and the quasi-market: An empirical analysis of secondary education policy in England. *Oxford Bulletin of Economics and Statistics* 72, no. 1: 1–26, 02.

Buonanno, P., and D. Pozzoli. 2009. Early labour market returns to college subject. *LABOUR* 23, no. 4: 559–88.

Caliendo, M., and S. Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22, no. 1: 31–72.

Cuong, N.V. 2009. Impact evaluation of multiple overlapping programs under a conditional independence assumption. *Research in Economics* 63: 27–54.

Dehejia, R.H., and S. Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, no. 448: 1053–62.

Gorard, S. 2002. Let's keep it simple: the multilevel model debate. *Research Intelligence* 81: 24–5.

Heckman, J., H. Hichimura, J. Smith, and P. Todd. 1998. Characterizing selection bias using experimental data. *Econometrica* 66, no. 5: 1017–98.

Hong, G., and S.W. Raudenbush. 2006. Evaluating kindergarten retention policy: A case study of casual inference for multi-level observational data. *Journal of American Statistical Association* 101, 474, 901–10.

Imbens, G. 2000. The role of the propensity score in estimating dose–response functions. *Biometrika* 87: 706–10.

Jesson, D., and D. Crossley. 2004. Educational outcomes and value added by specialist schools, Specialist Schools Trust. http://www.specialistschoolstrust.org.uk.

Kendall, L., L. O'Donnell, S. Golden, K. Ridley, S. Machin, S. Rutt, S. McNally et al. 2005. *Excellence in cities: The national evaluation of a policy to raise standards in urban schools 2000–2003*, Research Report RR675a, Department for Education and Skills, London.

Kim, J., and M. Seltzer. 2007. Causal inference in multilevel settings in which selection process vary across schools. Working paper 708, Center for the Study of Evaluation (CSE), Los Angeles.

Lechner, M. 1999. Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business and Economic Statistics* 17, no. 1: 74–90.

Lechner, M. 2001a. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*, ed. M. Lechner and F. Pfeiffer, 43–58. Heidelberg: Physica-Verlag.

Lechner, M. 2001b. A note on the common support problem in applied evaluation studies. Discussion paper no. 2001–01, University of St Gallen, St Gallen.

Lechner, M. 2002. Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society A* 127: 59–82.

Levacic, R., and A. Jenkins. 2004. Evaluating the effectiveness of specialist schools. *Centre for the Economics of Education*. London: LSE.

Machin, S., S. McNally, and C. Meghir. 2004. Improving pupil performance in English secondary schools: Excellence in cities. *Journal of the European Economic Association* 2: 396–405.

Noden, P., and I. Schagen. 2006. The specialist schools programme: Golden goose or conjuring trick? *Oxford Review of Education* 32: 431–48.

Office for Standards in Education (OFSTED). 2005. *Specialist schools: A second evaluation*, February. Ref. HMI 2362, OFSTED, London.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2004. Generalized multilevel structural equation modeling. *Psychometrika* 69, no. 2: 167–90.

Rosenbaum, Paul R. 1985. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 387, 516–24.

Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3: 135–46.

Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.

Schagen, I., and H. Goldstein. 2002. Do specialist schools add value? Some methodological problems. *Research Intelligence* 80: 12–5.

Schagen, S., I. Davies, P. Rudd, and I. Schagen. 2002. The impact of specialist and faith schools on performance. LGA Research Report 28, National Foundation for Educational Research..

Smith, J., and P. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125, no. 1–2: 305–53.

Su, Y.-S., and J. Cortina. 2009. What do we gain? Combining propensity score methods and multilevel modeling. APSA 2009 Toronto Meeting Paper. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450058 (accessed August 21, 2009).

Taylor, J. 2007. Estimating the impact of the specialist schools programme on secondary school examination results in England. *Oxford Bulletin of Economics and Statistics* 69: 445–71.

## Appendix 1. Conditional independence assumption

Imbens (2000) and Lechner (2001a) demonstrate that the 'curse of dimensionality' is avoided by exploiting some modified versions of the balancing score properties.

$$Y_0, Y_1, \ldots, Y_K \perp T(X) = b(x) \quad \forall x \in X. \qquad \text{(A1)}$$

In particular, for the ATT, Lechner (2001b, Proposition 3) shows the following:[16]

$$\theta^{k,g} = E(Y_k = k) + E_{P^{g,g}}[E(Y_g^{g,g}(X), T = g) = k]. \qquad \text{(A2)}$$

The conditional probability of being enrolled in a school where a policy $g$ is active instead of being in a school with policy $k$ is the (one-dimensional) propensity score

$$P^{g,g} = P^{g,g}(T = g = x,\ T \in \{k, g\}) = \frac{P^g(x)}{P^k(x) + P^g(x)}. \qquad \text{(A3)}$$

and $P^g(x) = P(T = g = x)$ and $P^k(x) = P(T = k = x)$.

Thus, $\theta^{k,g}$ is identified only using information from the sub-sample of pupils in schools of type $k$ and type $g$. The validity of the CIA, Equation (A1), can now be exploited using a matching estimator (Angrist 1998, Dehejia and Wahba 1999, Heckman et al. 1998, Lechner 1999). Matching on the propensity score, $P^{g,g}$ gives a consistent estimator of the counterfactual mean $E(Y_{ig} = k)$. We have to form a comparison group of pupils treated by policy $g$ that have the same propensity score as the pupils being treated by policy $k$.