An Evaluation of Multiple Overlapping Policies in a Multi-Level Framework: The Case of Secondary Education Policies in the UK*

Steve Bradley†
Giuseppe Migali‡

July 4, 2011

Abstract

Successive British governments have introduced numerous reforms to the education system, which have increased school choice and raised real expenditure per pupil, in an attempt to improve the educational outcomes of pupils. Many of these policy reforms have simultaneous effects on pupils because they are often implemented in the same schools and at the same time. In this paper we evaluate the multiple overlapping treatment effects on pupil test scores of two flagship policies - the Excellence in Cities initiative and the specialist schools policy. Using the National Pupil Database (2002-2006) combined with school level data, we show that there are overlapping multiple treatment effects, which we interpret as a complementarity between the two policies. This effect is modest when evaluated for the average pupil, at around 0.2-1.2 GCSE points. However, when we allow for the hierarchical nature of our data and distinguish between 'good' and 'bad' educational districts, the effects are much larger and increase in magnitude over time (1-7 GCSE points). We conclude that it is important to evaluate the overlapping effects of policies and to simultaneously allow for the hierarchical nature of the educational process.

Keywords: School Quality, Subject Specialisation, Matching models, Multilevel 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 amongst secondary school

*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.

†Department of Economics, Lancaster University Management School, Lancaster University, UK, LA1 4YX.

‡g.migali@lancaster.ac.uk, Department of Economics, Lancaster University Management School, Bailrigg Lancaster LA1 4YX, UK. DOPES, Universita' Magna Graecia, Catanzaro, Italy.

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 reforms, or policies, have simultaneous effects on pupils because they are often implemented in the same schools and at the same time. Moreover, since schools are nested in educational districts it is likely that there are interaction effects between the policies and the socio-economic characteristics of those districts. Indeed, administrators in educational districts may take these characteristics into account when implementing national education policies. However, in most analyses of the impact of education policies the evaluation typically focuses on each policy separately. Moreover, in studies which examine the effects of multiple policies they are often treated as mutually exclusive effects, and they also often ignore the hierarchical nature of their data.

In this paper we analyse the relative and 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. A second important contribution of the paper is that we combine the two overlapping policies in a multi-level setting, reflecting the nesting of pupils in schools and schools in educational districts. In both cases we adopt a matching methodology. As far as we are aware this is the first paper to bring together these strands of the evaluation literature in an application to British educational policy.

The specialist schools initiative aimed to raise educational attainment by increasing pupil choice by allowing schools to specialise in subjects in which they had a comparative advantage. Schools who were a part of the initiative also 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 which was limited for a 4 year period.[2] All schools were eligible to apply for specialist school status. In contrast, the EiC programme focused on pupils from disadvantaged socio-economic backgrounds in deprived metropolitan areas, and extra resources, totaling £1.7b or approximately £150 per pupil per annum, was provided to schools to support pupils, including those deemed gifted. These schools typically also had a high proportion of pupils from ethnic minorities. Table 1 describes the other key features of the two policies and Bradley and Taylor (2010) provide a more detailed discussion.

[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]Schools that were admitted to the selective 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.

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 2006-07	33% of secondary school	85% of secondary schools
Funding	£140 per student per annum	£129 per student for 4 years & lump sum
Mechanisms	Funding effect	Funding & specialisation effects

Although all schools could, in principle, participate in the specialist schools initiative it was actually the case that better schools, as measured by their test score performance, that were early adopters, which raises the possibility of a selection effect. 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. As noted above, both policies increased funding per pupil, however, schools were allowed to specialise in particular subjects, such as technology, languages or business, which may have improved the allocative efficiency of schools and hence test scores. This arises because pupils and teachers sort in those schools that best matches their aptitudes. We refer to this effect as a specialisation effect.

Schools that were part of the EiC programme could also attract specialist schools status and hence pupils potentially benefited from both policies. It is therefore important to investigate the impact of these overlapping policies on pupils exam performance. However, in doing so 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. Multi-level modelling does allow us to control for both sources of district level heterogeneity. To control for pupil level unobserved heterogeneity we also adopt a difference-in-differences analysis, the findings from which we compare with the multi-level modelling approach.

Our work has important policy implications. First, it is clear that these two policies are based on entirely different approaches to improving standards, yet the resource cost was similar in terms of the grant per pupil, therefore it is useful for policy purposes to compare their relative effectiveness in improving test score outcomes. This comparison is particularly pertinent given that expenditure on the specialist schools initiative favored schools with above average levels of attainment, whereas the EiC policy favored schools with low levels of attainment (Bradley and Taylor, 2010). Second, in view of the fact that some schools did benefit from both policies simultaneously, it is also important from a policy perspective 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 3-6 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 rise over time and are quite large. The estimates of the multiple treatment 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

school policies. The combined effect of the funding is one possible reason, however, we show that the specialisation effect is also likely to be a major contributory factor. However, the magnitude of any complementarity is modest at 0.2-1.2 GCSE points.

Furthermore, controlling for observed and unobservable district effects we show that there are large variations in the multiple treatment effects over the distribution of district level random effects. Pupils in schools in unobservably 'good' districts have witnessed a dramatic improvement in their GCSE scores because of the joint effects of the policies, rising from less than a 1 GCSE point advantage, compared to the single EiC treatment, in 2002 to over 7 GCSE points by 2006. It is also the case that pupils in schools in unobservably 'bad' districts have benefited much more from the joint effect of the policies, however this advantage more or less disappears by 2006. Finally, the estimates from DID matching estimation, although smaller in magnitude than the cross-sectional estimates, show that the longer a school has been specialised the larger the effect. This confirms that two policies taken together have some degree of complementarity.

The remainder of this paper is structured as follows. Section 2 briefly discusses the previous literature on the evaluation of the specialist schools and EiC policies, and reviews the smaller literature on the multiple treatment and multi-level matching. 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, which is followed by our conclusions.

2 Literature Review

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 has been modest. Similar results have been obtained by Bradley and Taylor (2008) 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.[3] Positive and statistically significant effects of the specialist schools policy on test scores is provided by Gorard (2002), Jesson and Crossley (2004) and OFSTED (2005). Schagen and Goldstein (2002) and Noden and Schagen (2006), who are especially critical of the school-level analyses, argue that multi-level modelling techniques should be used to take into account the multi-level structure of the data. Taylor (2007) finds that the specialist schools policy has had very little impact on exam results on average. These results are substantially confirmed by Bradley, Migali and Taylor (2011) which using several data sets and matching methodologies, finds a modest 'causal' effect on pupil test scores.

The literature on multiple treatment and/or multi-level matching methods is not extensive and is relatively recent, and in most cases the two approaches are treated separately.

[3]For a more detailed literature review see Bradley, Migali and Taylor (2011).

The seminal papers are Imbens (2000) and Lechner (2001) 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 on Swiss labor market policies, where several sub-programs act as multiple treatments. 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 the 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 (2001) 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 quite mixed. For instance, Rosembaum (1986) 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 (2008) control for the omitted variable bias due to unobserved cluster level and individual level covariates. They argue that matched treated and controls are required to belong to a similar, but not identical cluster. A two-stage methodology is proposed, whereby they first estimate a multilevel 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 and Su and Cortina, 2009.) Our work is the first evaluation study to combine multiple overlapping treatments and multi-level models.

There are therefore two major differences between our approach and that of previous work in this field. The first difference is that we are allowing for overlapping treatment status (i.e. specialist school and EiC overlapping in the same school in the same year) and we study their joint effects on pupil test score outcomes. The second difference is the combination of the overlapping policies in a multi-level setting, that is, we take into account the fact that pupils attend schools where both policies are active and those schools are nested in educational districts.

3 Econometric Approach

Our econometric approach focuses on propensity score matching methodologies in a multi-site analysis. As Table 1 shows these policies differ in several important respects, such as the target group of pupils, the focus of the policy, the duration of the treatment, selection rules and finally the mechanisms through which they seek to raise pupil test scores. It is important to take into account these possible sources of heterogeneity. Equation 1 illustrates the evaluation problem in terms of multiple treatments, multi-level effects and time-varying effects.

$$Y_{ijkqt} = \alpha_{jkt} + \theta_{kqt} T_{jkt} + \gamma \mathbf{X} + \epsilon_{ijkqt} \tag{1}$$

where $i = 1 \ldots N$, $j = 1 \ldots J$, $k = 0 \ldots K$, $q \in [0, 1]$. Y are the test scores of the N pupils in the sample. T is treatment status, which denotes one of the K types of policy active in a given school. The vector X includes pupil and school characteristics that may define the probability of each treatment. The index j accounts for the district where the type k school is located, and index q denotes the quantile in which the distribution of random district effects has been divided. The index t is the calendar year in which the pupil obtains his or her test score.[4]

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 (2001) to the case of multiple treatments. We consider $K + 1$ mutually exclusive treatments, and in our model a pupil receives the treatment by attending a school where a policy is implemented. Two types of control group can be constructed - one where 'no policy' is active and the treatment group is either pupils in schools where one or more policies are implemented, and another where one policy is compared with another policy or policies. For simplicity the remainder of this section our discussion focuses on the relative effects of two 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.

The parameter of interest, θ, from Equation 1 is usually estimated in the evaluation literature as the average treatment on the treated (ATT), whose multiple treatment version is

$$\theta^{k,g} = E(Y_{ik} - Y_{ig}|T = k) = E(Y_{ik}|T = k) - E(Y_{ig}|T = k). \tag{2}$$

$\theta^{k,g}$ in Equation 2 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). The problem is that the expected outcome $E(Y_{ig}|T = k)$ cannot be observed for the same pupil i. Imbens (2000) and Lechner (2001) show for the case of multiple treatment that identification of $\theta^{k,g}$ comes from the Conditional Independence Assumptiom (CIA). They also show for the same setting that the "curse of dimensionality" is avoided by exploiting some modified versions of the balancing score properties.

$$Y_0, Y_1, ..., Y_K \perp T|b(X) = b(x) \quad \forall x \in X. \tag{3}$$

In particular, for the ATT Lechner (2001a, proposition 3) shows the following:[5]

$$\theta^{k,g} = E(Y_k|T = k) + E_{P^{g|k,g}}[E(Y_g|P^{g|k,g}(X), T = g)|T = k]. \tag{4}$$

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

$$P^{g|k,g} = P^{g|k,g}(T = g|X = x, T \in \{k, g\}) = \frac{P^g(x)}{P^k(x) + P^g(x)}. \tag{5}$$

[4]In the following equation we omit this index for simplicity, but we allow for time variation.

[5]We omit for simplicity the index i.

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

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 3, can now be exploited using a matching estimator (see Angrist (1998), Dehejia and Wahba (1999), Heckman, Ichimura, and Todd (1998), and Lechner (1999)). Matching on the propensity score $P^{g|k,g}$ gives a consistent estimator of the counterfactual mean $E(Y_{ig}|T = k)$. 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.

All matching estimators are weighted estimators, derived from the following general formula:

$$\theta^{k,g}_{ATT} = \sum_{i \in k} (Y_{ik} - \sum_{h \in g} W_{ih} Y_{hg}) w_i \tag{6}$$

W_{ih} is the weight placed on the hth observation in constructing the counterfactual for the ith 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 neighbor (NN) algorithm and provide analytical standard errors (Abadie and Imbens, 2008). In all estimations we only consider observations on the common support.

Since we also seek to control for several sources of heterogeneity we use different estimation methods of the propensity score (Equation 5) depending on the case of interest.[6]

3.1 Evaluating the effect of multiple treatment using the multinomial logit model

Let the random variable T take one of K discrete values that, as stated above, is an index, $k = 0, \ldots, K$. 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, a more parsimonious way of parameterizing the various propensity scores that are required to achieve the balancing property is based on the multinomial choice model, as suggested by Lechner (2002).

We assume the T_i are generated by the multinomial logistic model (MNL)

$$P(T_i = k|x_i) = \frac{\exp(X'\beta_k)}{\sum_{h=0}^{K} \exp(X'\beta_h)} \tag{7}$$

From the estimation of the MNL model we get the marginal probabilities in Equation 7 for each policy, and we compute the conditional probabilities in Equation 5. The next step is to match on these propensity scores and obtain the ATT as in Equation 4. This analysis does not take into account the fact that the EiC policy was focused on schools in disadvantaged educational districts in metropolitan areas, therefore matching is performed between pupils in schools that belong to districts that can be very different both observably and unobservably. The advantage of this approach is that because we have more data we can consider several combinations of each type of policy treatment - specifically the joint effects

[6] Other matching methods are discussed by Heckman, Ichimura, and Todd (1998).

of the two policies versus the effect of no policy (*multiple overlapping treatment effects*) or of one policy versus another policy (*relative treatment effects*), as suggested above.

3.2 A multi-level modelling approach

This model differs from the previous one because it takes into account an important source of heterogeneity related to the multi-level structure of our data.Pupils are nested in schools, and schools are administered within districts, which may differ between each other not only with respect to their socio-economic make-up but also with respect to the application of educational policy. Our multi-level model differs from the traditional multi-level modeling because the policy is at school level, which implies that within each school pupils receive the same treatment. Therefore we cannot consider the school as a separate hierarchical level and practically we only have two levels - the pupil and the educational district. However, the fact that we have a school-based policy allows us to include school characteristics in the estimation of the propensity scores, and therefore indirectly control for school differences.

We distinguish between two multi-level models - the random intercept and the random slopes models - to allow for variation in the selection of schools into each policy initiative by educational district. We estimate in each case a binary response model, that is the treatment variable $T = 1$ if a given school has implemented the policy k, and $T = 0$ if the policy g is active.

3.2.1 A random intercept model

A generalized linear random intercept model for binary responses is represented by

$$F^{-1}(\pi_{ij}) = \beta_0 + \beta_1 X_{ij} + u_j \tag{8}$$

where $\pi_{ij} = E(T_{ij}|x_{ij}, u_j) = P(T_{ij} = 1)$, and F^{-1} is a link function corresponding to the inverse cumulative function of a known distribution. u_j is the random effects, or level 2 residuals, and it absorbs the district-specific constants effects of unobserved district-level predictors. In our analysis we consider a logit distribution, the link function is the log-odds that $T = 1$ and equation 8 becomes

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_0 + \beta_1 X_{ij} + u_j \tag{9}$$

where $u_j \sim N(0, \sigma_u^2)$. The variance of the intercepts across districts is $var(u_j) = \sigma_u^2$ and is also known as the between-group variance adjusted for X. Equation 9 is a random effects model that captures the true propensity score for a random intercept model. The left-hand-side is a non-linear transformation of π_{ij}, whereas the right-hand-side is linear in terms of the coefficients and the residuals. The coefficient β_0 is the overall intercept in the linear relationship between the log-odds and X, and the intercept for a given district j is $\beta_0 + u_j$. β_1 is the cluster-specific effect, that is, holding u constant it shows the effect of X for pupils in the same district. To get an idea of the importance of the district 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, which we discuss below. We therefore observe for each district the

departure from the overall intercept ($u = 0$), which implies that a district j whose confidence interval does not overlap the line at zero differs significantly from the average log-odds of the treatment at the 5% level.

The conditional probability from Equation 5, which we use to match pupils, corresponds in this model to the predicted response probability for pupil i in district j. Re-arranging Equation 9, the propensity score can be given as:

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 X_{ij} + u_j)}{1 + \exp(\beta_0 + \beta_1 X_{ij} + u_j)} \tag{10}$$

However, it is possible that the standard normality assumption for the random intercept model produces biased estimates. We therefore relax the normality assumption on the distribution of u_j by adopting a non-parametric maximum likelihood (NPMLE) approach (Rabe-Hesketh *et al.*, 2003, 2004). 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 is divided into the maximum number of classes, such that the addition of further classes cannot increase the likelihood any further (Rabe-Hesketh *et al*, 2004).

3.2.2 A random slope model

In this model the contribution of pupil-level covariates to vary across districts. The random slope logit model is an extension of Equation 9

$$\log(\frac{\pi_{ij}}{1 - \pi_{ij}}) = \beta_0 + \beta_1 X_{ij} + u_{0j} + u_{1j} X_{ij} \tag{11}$$

The district-level variation comes from the random effects, u_{0j}, and the interaction of the random effects and pupil-level characteristics, $u_{1j}X_{ij}$. The covariance matrix of the random effects is a symmetric matrix formed by the intercept variance, $\sigma^2_{u_0}$, the slope variance, $\sigma^2_{u_1}$, and the intercept-slope covariance, $\sigma^2_{u_{01}}$. In our analysis we estimate the random slope model distinguishing between two cases; in the first u_{0j} and u_{1j} are not correlated and their covariance $\sigma^2_{u_{01}} = 0$. In the second case u_{0j} and u_{1j} are allowed to be correlated and their covariance $\sigma^2_{u_{01}} \neq 0$. As in the random intercept model we compute the predicted response probabilities and we use them to match pupils from the treated and control groups.

3.2.3 A multi-level matching approach by quantiles of the random effects

In this model we further control for the heterogeneity due to differences between districts. We estimate a matching model where the treatment is a function of the quantile distribution of the district random effects. In terms of Equation 2 the ATT can be rewritten as

$$\theta^{k,g}_q = E(Y_{ikq} - Y_{igq} | T = k, u_{jq}). \tag{12}$$

where u_{jq} is the q^{th} quantile of the random variable u_j.

We assume that the propensity score follows a random intercept model and we estimate Equation 9 and obtain the random effects, $\hat{u}_j$. We divide them into quantiles with an equal

number of districts. Using Equation 10 we obtain the predicted response probabilities and we divide them according to the quantile distribution of u. We finally match pupils in schools located in districts in the same quantile of the random effects distribution. The estimation of the ATT in each quantile should show the differences between clusters of educational districts. Variation of the ATT across quantiles in the same year, and across years for the same quantile, should also highlight the effect of district-level unobserved heterogeneity.

3.3 Multiple treatment difference-in-differences (DiD) models

An alternative approach (see Blundell et al (2004), Smith and Todd (2005) and Machin et al (2004)) is based on matching with difference-in-differences but with allowance made for multiple treatments. This approach relaxes the strong assumption of the cross-sectional matching approaches of selection based solely on observables. 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. Our primary analysis uses pupil level data where we exploit repeated cross-sections, however we do also use a panel of schools in a further analysis.

The DiD matching estimator for repeated cross-section data, allows for temporally invariant differences in outcomes between pupils in schools adopting (or being selected into) one policy and pupils in the same school who adopt an additional policy. It is obtained by rewriting Equation 6 as

$$\tau_{ATT}^{DID} = \sum_{i \in T_t} [\, Y_{kti} - \sum_{h \in C_t} W_{ih} Y_{gth}] \, w_{it} - \sum_{i \in T_{t'}} [\, Y_{gt'i} - \sum_{h \in C_{t'}} W_{ih} Y_{gt'h} \,] \, w_{it'}. \tag{13}$$

where t' and t are time periods before and after the adoption of a school 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 students 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. The rationale behind the DiD analysis is to remove unobserved pupil fixed effects, in an attempt to control for pupil selection effects. It should also take into account of district heterogeneity and therefore the DiD does not require a multi-level estimation.[7]

[7] As an additional robustness check , we have also tried a DiD matching estimator using only school panel data to allow for temporally invariant differences in outcomes between schools adopting one policy and the same school adopting an additional policy. The rationale behind this analysis is that we wish to minimise the effect of unobserved school heterogeneity and hence control for the non-random selection of schools into each policy regime.

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 score. 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 six versions of the NPD where pupils were in their final year of compulsory education in one of the years between 2002 to 2006. The original sample sizes are between 500,000 and 560,000 observations according to the year considered. We choose these six 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 national policy is applied appropriately.

To each NPD dataset we append school level data from the annual School Performance Tables and the annual Schools' Census. We have a school panel formed by 2645 schools observed from 1992 to 2006. 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 4 treatment statuses:

1. Category zero: No policy - schools that never joined the EiC or the specialist schools initiative. (We also include schools that became specialist after our sample of pupils have obtained their GCSEs. We explain why this approach is adopted below.)

2. Category one: Specialist only - schools that joined only the specialist schools programme before the pupil enrolled, and never joined the EiC programme.

3. Category two: EiC only - schools that joined only the EiC programme and never acquired specialist school status before our samples of pupils have obtained their GCSEs.

4. Category three: schools that are part of both programmes; specifically, schools that acquired specialist status did so before the pupil enrolled in the school.

Category three 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 multiple treatment and joint effects. We also generated this categorical variable in the school panel and use all pre-treatment school characteristic (i.e. pre-1994). In Table 2

we show the number of schools in each category by enrollment year and treatment period. Note that the size of each categories varies over time as schools switch from category zero to one of the other categories. We omit schools that acquire the specialist status while the pupils are still enrolled, therefore in Table 2 the sample size is decreasing.

This categorical variable becomes the dependent variable in the multinomial logit model. The conditional probabilities, used as propensity scores in each matching estimation, are computed for all combinations of the four categories. Precisely, we consider as a treatment group each policy separately (schools with only EiC or only specialist school status) and overlapping policies (EiC and specialist school status). However, due to data limitations, the control group is either the no policy category or the EiC only policy.

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 untreated groups, for a selection of our models. This allows us to demonstrate that overall any combination of characteristics in the treated group can be observed in the control group.[8] 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.

5.1 Single versus multiple treatment effects of educational policies

The first three columns of Table 3 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 group 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 shrink in size over time.

Focussing on the first three columns of Table 3 it is clear that in this analysis the effect of the specialist schools policy (column 1) is always positive and the ATT suggests an effect of between 3-6 GCSE points, an effect that is generally increasing over time. It is worth noting that matching reduces this effect by between 1 and 2 GCSE points. These results are slightly higher than those found in previous studies, for example in Bradley, Migali and Taylor (2011) where this policy was shown to have increased the GCSE points score by approximately 2-3 GCSE points.

In contrast, column 2 shows that pre-matching there is large negative, though declining, effect of the EiC policy on test scores and that post-matching there is virtually no statistically

[8] As noted by Caliendo (2008), when estimating the ATT it is sufficient to ensure the existence of potential matches in the control group.

significant effect of the EiC policy on test scores; 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 because these earlier papers focussed on test scores at the age of 14 in the case of Machin *et al* (2004), or they have focussed on the proportion of pupils obtaining 5 or more GCSE grade A to C in the case 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, what is interesting is the finding that as we move beyond 2003 the effect of multiple treatment (post-matching) exceed both single policy effects, particularly for 2006, which suggests that there are complementarities between the EiC and specialist school policies.[9] These complementarities could be due to the joint funding effect, or a combination of the joint funding effect and a specialisation effect, a point we come back to below.

5.2 The relative effects of educational policies on pupil test scores

In this section we explore the relative effects of the two policies and the control group are 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 are pupils in specialist schools and the second are pupils in schools that have 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 3, columns 4 and 5 show 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 specialist status achieve 4-6 GCSE points more than their counterparts in schools that only have EiC status, and these effects rise over time suggesting a cumulative gain from the specialist schools policy. Note, however, that the ATT estimates are substantially lower than the pre-matching estimates and this analysis does not allow for district-level unobserved effects.

The effects of multiple treatment versus single treatment are reported in column 5 of Table 3. The estimated ATT of multiple overlapping treatments is always positive and statistically significant, suggesting 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

[9]The number of schools that have both policies has also substantially increased as it is clear from Table 2.

'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 difference 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 Estimates from a multi-level multiple treatment model

We further explore the relative effect of schools engaged in both the specialist schools and EiC initiatives (category three) versus those only involved in the latter (category two) by exploiting the hierarchical nature of the data. Specifically, we estimate Equations 9 and 11 in sections 3.2.1 and 3.2.2, above. The motivation here is to find more appropriate comparison groups by selecting pupils in schools that are located in unobservably 'similar' educational districts.

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 the policy category three, the treatment group, and schools in policy category two, the control group.[10] The advantages of this analysis are the combination of multiple overlapping treatments and the simultaneous exploitation of the multi-level structure of the data.

When we estimate the propensity score using a random intercept multi-level model, assuming normal random effects, we always find a highly significant intercept variance, $\sigma^2_{u_0}$. The caterpillar plot of the residuals (see Figure 3) shows how educational districts differ with respect to unobservables. It is clear from this that districts in the tails - above and below zero - are clearly different to one another. Repeating the estimation using NPMLE also results in a statistically significant intercept variance. In contrast, 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 $\sigma^2_{u_{gender}}$ and $\sigma^2_{u_{ethnicity}}$. However, when we re-estimate the model assuming a non-zero correlation between the random effects, the intercept-slope covariances $\sigma^2_{u_{0gender}}$ and $\sigma^2_{u_{0ethnicity}}$ are still significant, whereas the slope-slope covariance $\sigma^2_{u_{sex-gender}}$ 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 4 are not directly comparable with those in column 5 of Table 3, 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

[10] 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.

the EiC policy since it is implemented only in particularly disadvantaged areas, whilst the specialist school initiative is a national policy. If we look at the MLE random intercept model, the ATT is 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 1999-2001. 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%. This implies that the effect of unobservable heterogeneity is higher when the number of specialist school increases.

Looking at the NPMLE random intercept estimator, we consistently find lower estimates than the MLE estimator between 2002 and 2005. However, the difference between the MLE ATT and the NPMLE ATT is not large and actually decreasing from 17% in 2002 to 4% in 2005, and then becoming negative in 2006. In general, our conclusions are unchanged therefore we retain the normaility assumption for the estimates that follow.

From 2002 to 2005 we observe the random slopes estimates are lower than those from the random intercept model, except for 2002 where the ATT is not significant. The difference between the coefficient of the random intercept model and the coefficient of the non-correlated random slopes increases from 0.6% in 2003 to 22% in 2005. The correlated random slopes are still not significant in 2002, and we cannot find any convergence of the estimator in 2003 due to data thinning. For the remaining years the effect is even smaller than the uncorrelated case.

Thus, allowing for ethnicity and gender differences by districts the results are substantially unchanged, although a bit lower than the random intercept case. In general, our multi-level analsysis provides a better control for the effect of unobservable district effects after matching.

5.4 Estimates from a multi-level model by quantiles of random effects

We present here a further test of the robustness of the previous effects, in an attempt to overcome the problem that EiC schools are clustered in disadvantaged educational districts and the fact that we do not have information on the characteristics of these districts. We examine how the multiple overlapping effect of the two education policies (versus the single effect of the EiC policy) varies over the quantiles of the random effects distribution, as explained in Section 3.2.3. We select comparison groups that are drawn from equivalent quartiles of the distribution of residuals reported in Figure 3. Dividing the propensity scores obtained from the estimation of a random intercept model into those quartiles we perform a matching analysis of pupils within each quartile.

Since there are a large number of estimates, Figure 4 plots the estimated ATT effects for each year, which can be compared with the estimates reported in Table 4. There is considerable variation in the multiple treatment effect, relative to the single treatment of the EiC policy, over these quantiles, suggesting that the joint effect of EiC and specialist

school status interacts in some way with (unobserved) district characteristics. Thus pupils in schools in unobservably 'good' districts (quartile 4) have witnessed a dramatic improvement in their GCSE scores because of the joint effects of the policies, rising from less than a 1 GCSE point advantage in 2002 to over 7 GCSE points by 2006. In contrast, pupils in schools in unobservably 'bad' districts (quartile 1) have benefited much more from the joint effect of the policies, when compared with quantile 4, although this advantage more or less disappears by 2006.

5.5 A multiple treatment difference-in-differences analysis

In Table 5 we report estimates from Equation 13 in section 3.3, which can be regarded as the more traditional method of controlling for time-invariant pupil characteristics and, at least indirectly, the unobserved district effects. The DiD ATT is obtained as a difference between two cross-sectional matching models. The data available for test scores at age 16 start in 2002, therefore we cannot consider a pre-treatment period before that year. We employ 4 NPDs, 2002, 2004, 2005 and 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 multiple overlapping (EiC and specialist schools policies) treatment versus the single EiC policy treatment.[11]

In this analysis the control group are pupils that get a GCSE in 2002 in an EiC only school. For the treatment group we observe pupils that get a GCSE in 2002 in an EiC only school and pupils that get a GCSE in the following years in the same school that has also become specialist during the study period. 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.

These estimates are not directly comparable with those in Table 4, because although we consider the same categories the composition of the treatment and control groups differ, and also the duration of the specialist school policy is much lower. However, both analyses have in common the fact that 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 of our previous results.

We find several interesting results. First the joint effects are positive and highly significant, and the longer the duration of school specialization the higher the effect. Moreover, since the effect of the EiC policy is constant, because in the pre- and post-treatment period schools exploit the same amount of funding, we can interpret the increase in test scores as the specific effect of the specialist school policy.[12] In the period 2002-2006 we include schools that have been specialist for a maximum of 4 years and the effect is around 1 GCSE

[11] For the Specialist school policy, we have data on schools that specialize from 2003 and we can track them when they were non specialist in 2002. However, schools joins the Eic programme only between 1999 and 2001, therefore the first possible post-treatment period would be 1999 and the last 2001. This would require in any case pre-treatment information on GCSE test scores before 2002. Therefore, in the pre-treatment period we consider schools that are only in the EiC programme and then we track the same schools that join the Specialist programme in the post-treatment period.

[12] 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 analyzed in Bradley, Migali and Taylor (2011).

point. In general, we can conclude that the effect of the specialist policy is not very strong, and this is consistent with the findings in Bradley, Migali and Taylor (2011). Therefore we can be confident, that in our previous analysis we are picking a real effect due to the complementarities between the two policies.[13]

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, and by allowing schools to specialise in particular subjects the specialist schools initiative allowed pupils and parents to better match their academic ability to the subject portfolio offered by a school. These initiatives, or policies, had in some cases 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 the each policy separately. This paper therefore analyses the relative and multiple overlapping effects, using matching methods, of the specialist schools and EiC policies on the test score performance of pupils in English secondary schools. We also combine the evaluation of these two policies with a multi-level modelling approach to reflect the fact that pupils are grouped in schools and schools in educational districts.

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 rise 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 school policies. The combined effect of the funding is one possible reason, however, we show that the specialisation effect is also likely to be a major contributory factor.

In further analysis we compare the performance of pupils in specialist schools with their counterparts in EiC schools and we show that pupils in the former achieve between 4-6 GCSE points more than pupils in EiC schools, which confirms our earlier finding. Moreover, the effects of multiple treatment versus single (EiC) are very similar insofar as a pupil in a school that has the benefit of both policies achieves 5-6 GCSE points more than their counterpart who has benefit from the EiC policy. This effect is remarkably stable over time. Furfthermore, when we compare these findings with those from single policy analyses, above, a small but positive difference in the two sets of effects emerges (0.2-1.2 GCSE points), which adds some confirmation to the view that there are complementarities between the two policies. The specialisation effect is one plausible mechanism.

Since these findings may not adequately control for observed and unobservable district

[13]Our estimates from the school level difference-in-differences analysis produced broadly similar estimates, however, because of the relatively small number of observations these were statistically insignificant. Therefore we do not report them here.

effects we estimate a multi-level model with two levels - pupils and districts. These findings essentially confirm our previous findings suggesting that district level heterogeneity plays only a modest role in pupil test score performance. However, what is more revealing are the large variations in the multiple treatment effects over the distribution of district level random effects. Pupils in schools in unobservably 'good' districts have witnessed a dramatic improvement in their GCSE scores because of the joint effects of the policies, rising from less than a 1 GCSE point advantage, compared to the single EiC treatment, in 2002 to over 7 GCSE points by 2006. It is the case that pupils in schools in unobservably 'bad' districts have benefited much more from the joint effect of the policies, however this advantage more or less disappears by 2006. The DID matching estimation, by keeping constant the EiC policy, finds that the effect of the specialist school policy alone is larger the longer the school has been specialist. However this effect is not very strong and this confirms that there are some complementarities between the two policies.

The main conclusion of this paper is that does appear to be a positive, although modest, 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. A further conclusion of this paper is that allowing for the hierarchical nature of the data can be important, and in our case this is manifested in the variation in the estimates of the ATT over the distribution of district level unobservables.

A References

Abadie A. and Imbens, G.W. (2008) Notes and Comments on the Failure of the Bootstrap for Matching Estimators *Econometrica* 76, No.6, 1537-1557.

Angrist, J. D., Estimating Labor Market Impact of Voluntary Military Service Using Social Security Data, Econometrica 66:2 (1998), 249-288.

Arpino, B. and Mealli, F. (2011), The specification of the propensity score in multilevel observational studies, Computational Statistics and Data Analysis Volume 55, Issue 4, 1 April 2011, Pages 1770-1780.

Blundell, R. Dias M., Meghir C. and Van Reenen (2004) Evaluating the Employment Impacts of a Mandatory Job Search Program, *Journal of European Economic Association* 2, 569-606.

Bradley, S. and Taylor, J. (2010) Diversity, Choice and the Quasi-market: An Empirical Analysis of Secondary Education Policy in England, Oxford Bulletin of Economics and Statistics, vol. 72(1), pages 1-26, 02.

Bradley, Migali and Taylor (2011) Funding, school specialisation and test scores: An evaluation of the specialist schools policy using matching models, Lancaster University Management School Working Paper

M. Caliendo and S. Kopeinig. Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22(1):31-72, 2008.

Cuong, N.V. (2009), Impact evaluation of multiple overlapping programs under a conditional independence assumption, Research in Economics 63 (2009) 27-54

Dehejia, R. H. Wahba S. (1999) Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, *Journal of the American Statistical Association* 94(448), 1053-1062.

Kendall, L., OÕDonnell, L., Golden, S., Ridley, K., Machin, S., Rutt, S., McNally, S., Schagen, I., Meghir, C., Stoney, S., Morris, M., West, A. and Noden, P. (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.

Gorard, S. (2002) Let's Keep It Simple: the Multilevel Model Debate, *Research Intelligence* 81.

Heckman, J. Hichimura, H. Smith, J. and Todd, P. (1998) Characterizing Selection Bias Using Experimental Data, *Econometrica* 66(5), 1017-1098.

Hong, G. and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of casual inference for multi-level observational data. Journal of American Statistical Association. 101:474,901-910.

Imbens G. (2000), The Role of the Propensity Score in Estimating Dose-Response Functions, Biometrika 87: 706-710.

Jesson, D. and Crossley, D. (2004) Educational Outcomes and Value Added by Specialist Schools, Specialist Schools Trust. (http://www.specialistschoolstrust.org.uk).

Lechner, M., Earnings and Employment Effects of Continuous Off-the- Job Training in East Germany After Unification, Journal of Business & Economic Statistics 17:1 (1999), 74-90.

Lechner M. (2001), Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in Lechner M. and Pfeiffer F. (eds.) Econometric Evaluation of Labour Market Policies, Heidelberg: Physica-Verlag: 43-58.

Lechner, M. (2001a) A note on the common support problem in applied evaluation studies, Discussion paper n. 2001-01, University of St. Gallen.

Lechner M. (2002) ÔSome Practical Issues in the Evaluation of Heterogenous Labour Market Programmes by Matching MethodsÕ, Journal of the Royal Statistical Society A 127: 59-82.

Machin, S. McNally, S. and Meghir, C. (2004) Improving pupil performance in English secondary schools: Excellence in Cities, Journal of the European Economic Association 2, 396-405 .

Office for Standards in Education (OFSTED) (2005) Specialist Schools: A Second Evaluation, February, Ref. HMI 2362, OFSTED, London.

Plesca M. and Smith, J. (2007) Evaluating multi-treatment programs: theory and evidence from the U.S. Job Training Partnership Act experiment, Empirical Economics, Springer, vol. 32(2), pages 491-528, May.

Buonanno, P. and Pozzoli, D. (2009) Early Labour Market Returns to College Subject, LABOUR, Volume 23, Issue 4, pages 559-588, December.

Hong, G. and Raudenbush, S. W. (2006) Evaluating kindergarten retention policy: A case study of casual inference for multi-level observational data. Journal of American Statistical

Association, 101:474,901-910.

Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2003a) Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. Statistical Modelling 3 (3), 215-232

Schagen, I. and Goldstein, H. (2002) Do Specialist Schools Add Value? Some Methodological Problems, *Research Intelligence* 80, 12-15.

Kim, J., Seltzer, M. (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.

Roy, A. D. (1951) Some Thoughts on the Distribution of Earnings, Oxford Economic Papers 3 (135-1460.

Rosenbaum, Paul R., and Donald B. Rubin. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70 (1):41-55.

Rosenbaum, Paul R., (1985) Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, Journal of the American Statistical Association 79:387 , 516-524.

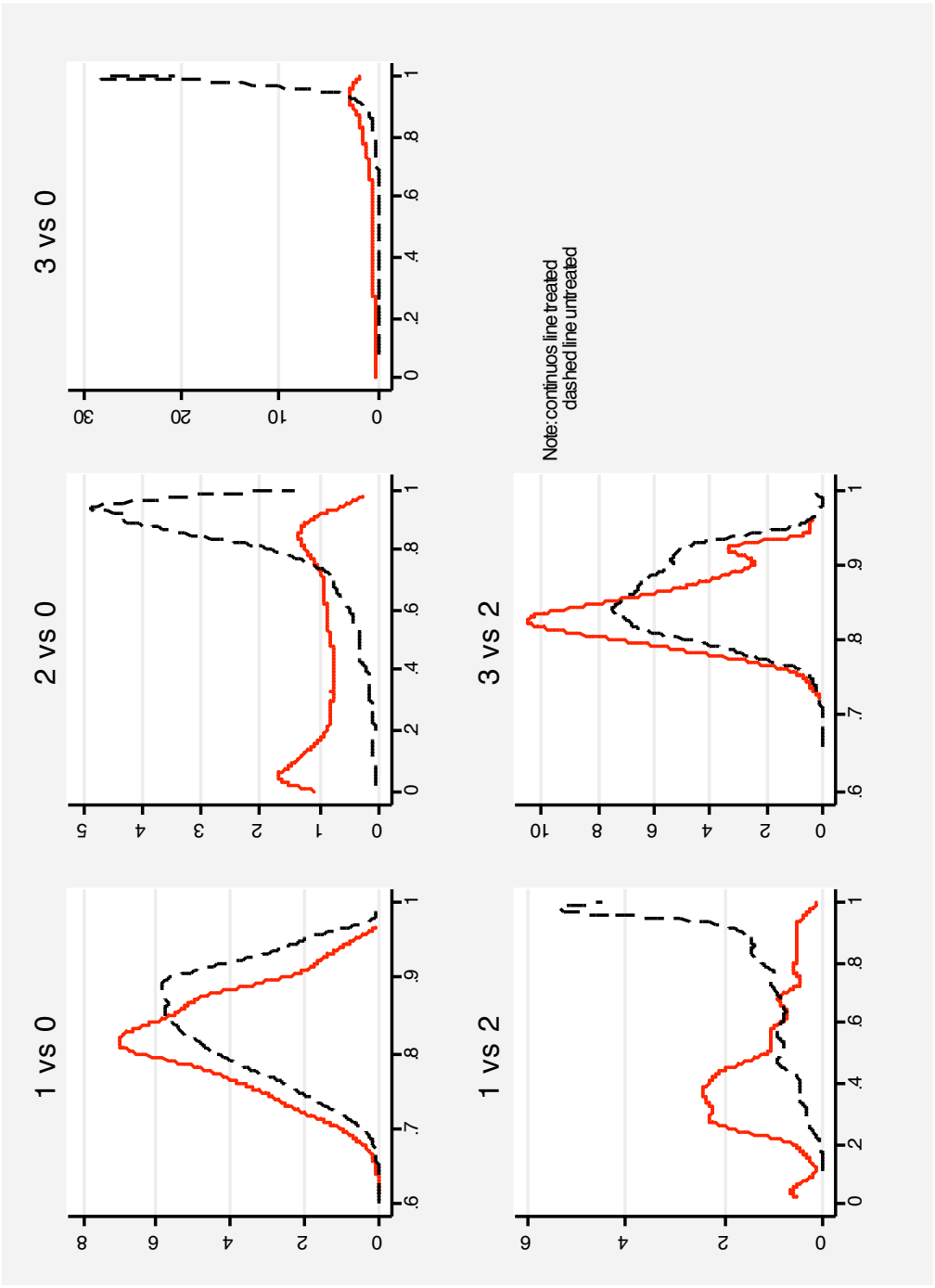
Rubin, D. B., Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Journal of Educational Psychology 66 (1974), 688-701.

Smith, J. and Todd, P. (2005) Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?, *Journal of Econometrics* 125(1-2), 305-353.

Su, Yu-Sung and Cortina, Jeronimo, What do We Gain? Combining Propensity Score Methods and Multilevel Modeling (2009). APSA 2009 Toronto Meeting Paper.

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, 445-471.

Figure 1: Selected Plots of Propensity Scores Distribution, MNL model 2002



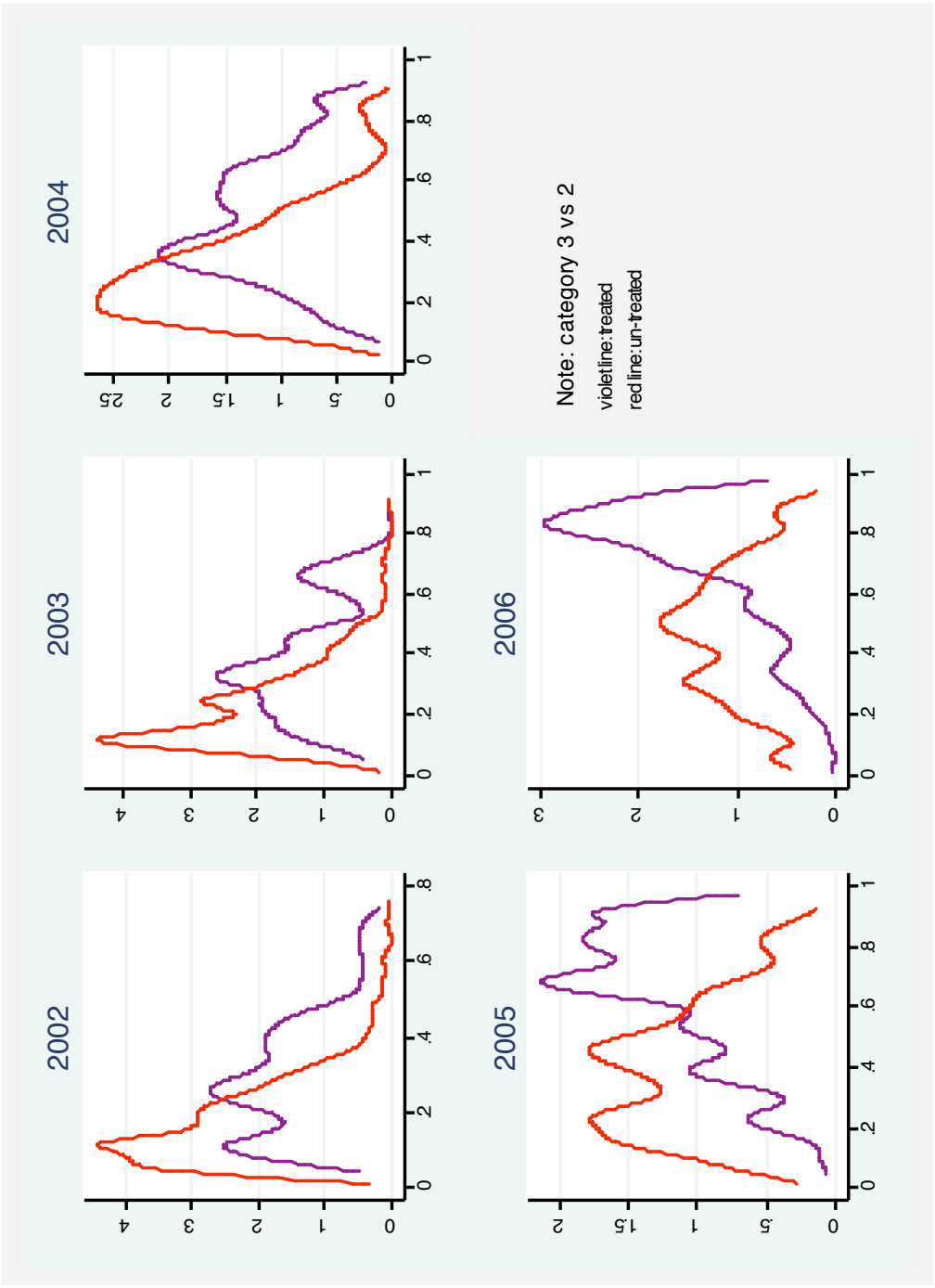


Figure 2: The Propensity Scores Distribution, RI model

Figure 3: Caterpillar Plot of the Random Effects Distribution - RI models

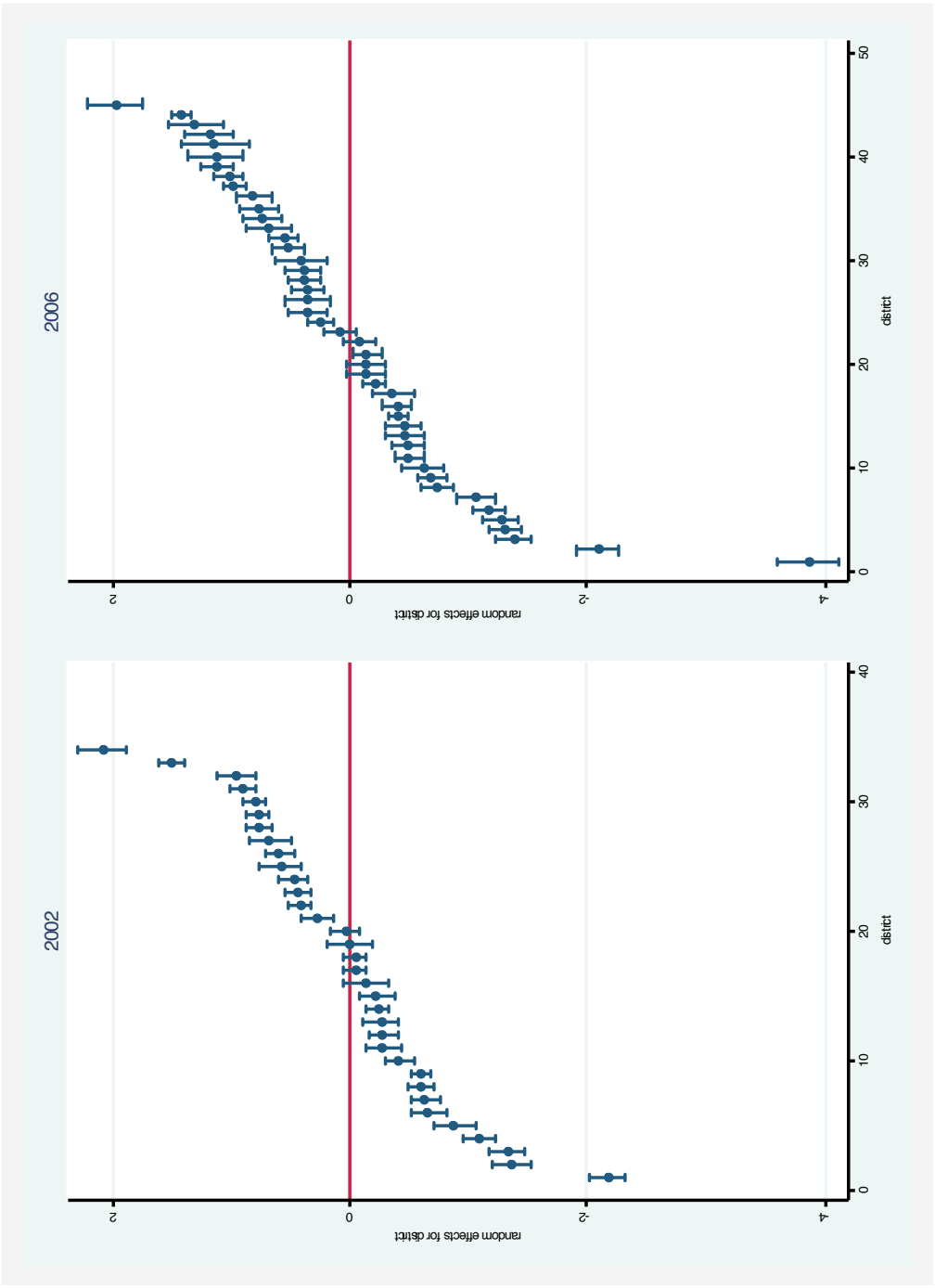
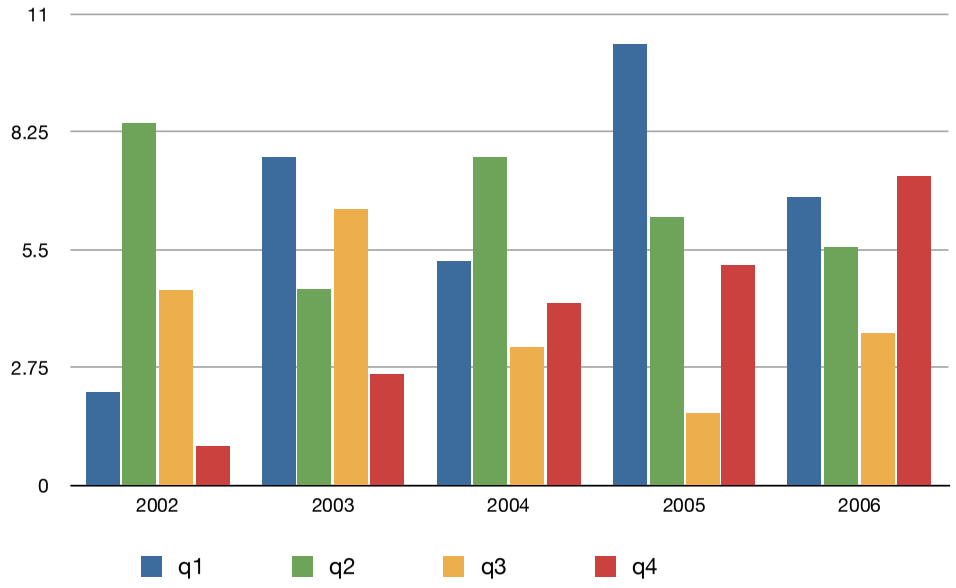


Figure 4: Estimates from a Quantile Multilevel Model



Note: Comparison of categories 3 vs 2.
All estimates are statistical signifcant at 1% level, except quarte 3 2006 (n.s),
quarter 3 2005 (10%) and quarter 4 2002 (10%).

Table 2: Number of Schools by Treatment Category

	Year of enrollment				
Policy categories	*1998*	*1999*	*2000*	*2001*	*2002*
No policy	1,335	1,000	653	402	332
	(64)	(58)	(47)	(34)	(25)
Specialist only (from 1994)	195	237	309	394	607
	(9)	(14)	(22)	(33)	(45)
EiC only	490	407	310	213	183
	(23)	(23)	(22)	(18)	(14)
EiC-Specialist	69	91	124	169	224
	(3)	(5)	(9)	(14)	(17)
Total	2,089	1,735	1,396	1,178	1,346
	(100)	(100)	(100)	(100)	(100)

DID analysis

			Treatment period		
	Control period		*2003-04*	*2003-05*	*2003-06*
EiC only	2002-04	310			
EiC only	2002-05	213			
EiC only	2002-06	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: Estimates of Multiple and Relative Treatment Effects using MNL

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

Table 4: Estimates of Multiple Treatments using Multi-level models

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

Notes: Random slopes are for the the gender and ethnicity varaibles.

2003 3rd model non-convergence due to data sparcity.

See Note to Table 3 for variables in the Propensity Score.

Table 5: Estimates from a Multiple Treatment DID model

	2002-2004	*2002-2005*	*2002-2006*
Policy categories	*(3) vs (2)*	*(3) vs (2)*	*(3) vs (2)*
$T_{t'} - C_{t'}$			
unmatched	4.642***	5.311***	5.124***
	(0.141)	(0.140)	(0.145)
ATT	2.334***	2.676***	1.719***
	(0.213)	(0.220)	(0.230)
$T_t - C_t$			
unmatched	5.363***	5.764***	4.349***
	(0.150)	(0.152)	(0.157)
ATT	2.780***	3.310***	2.697***
	(0.224)	(0.238)	(0.257)
DID	0.721***	0.453***	-0.774***
	(0.021)	(0.021)	(0.023)
DID ATT	0.445***	0.634***	0.977***
	(0.219)	(0.230)	(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-treament 2003, 2004, 2005.

2002-2006: pre-treatment 2002 and post-treament 2003, 2004, 2005, 2006.

See Note to Table 3 for variables in the Propensity Score.