Funding, school specialisation and test scores: An evaluation of the specialist schools policy using matching models*

Steve Bradley[†]
Giuseppe Migali[‡]
Jim Taylor[§]

April 20, 2011

Abstract

We evaluate the effect on test scores of a UK education reform which has increased funding of schools and encouraged their specialisation in particular subject areas, enhancing pupil choice and competition between schools. Using several data sets, we apply cross-sectional and difference-in-differences matching models, to confront issues of the choice of an appropriate control group and different forms of selection bias. We demonstrate a statistically significant causal effect of the specialist schools policy on test score outcomes. The duration of specialisation matters, and we consistently find that the longer a school has been specialist the larger is the impact on test scores. We finally disentangle the funding effect from a specialisation effect, and the latter occurs yielding relatively large improvements in test scores in particular subjects.

Keywords: Matching models, Subject Specialisation, School Quality.
JEL Classification: I20, I21, I28

1 Introduction

In many countries around the world, such as the US and the UK, there has been widespread debate about the best way to improve the educational performance of school pupils. A common theme in terms of policy is a shift away from centralised funding and provision to a

[*] The authors would like to thank: the Nuffield Foundation for their financial support; the University of Pennsylvania for providing Migali excellent research facilities; and Petra Todd, Ian Walker, Colin Green and Geraint Johnes for helpful comments. We would also like to thank the participants at the European Economic Association Annual conference 2009, the Scottish Economic Association Annual conference 2009; seminar participants at the Institute of Education, and at the Departments of Economics of the University of Sheffield and Lancaster University.

[†] Department of Economics, Lancaster University Management School

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

[§] Department of Economics, Lancaster University Management School.

decentralised approach to educational provision (Hoxby, 1996). In the US the 'No Child Left Behind Act', introduced by the Bush administration, coupled school choice with accountability measures, to allow parents of children in under-performing schools the opportunity to choose higher performing schools. However, there is debate about whether this policy has had a positive effect on the educational performance of pupils, in fact some claim that it has been harmful to America's schools (The Economist, 2009). As a result, the Obama government is considering spending an extra $10 billion on education in an attempt to turn around failing schools, so bringing the issue of whether increased school resources will improve educational outcomes back on to the political agenda.

In the UK a number of educational policy reforms have been introduced, such as the 1988 Education Reform Act, which led to the creation of a quasi market in education, at the heart of which is enhanced parental choice and competition between schools for pupils. However, funding for secondary schools has also been increased substantially since 1997, rising from £9.9bn to £15.8bn in 2006/7. Over the same period real expenditure per pupil increased by over 50%, from £3206 to £4836 (in 2005/6 prices). One of the key policy initiatives that has led, in part, to this increase in funding is the specialist schools policy, which was introduced in 1994. To obtain specialist status, state maintained schools are required to raise unconditional sponsorship from the private sector of £50,000 and to have a development plan. Selected schools then received a capital grant of £100,000 from central government, and around £130 per pupil over a four year period.[1] This amounts to an approximate increase in funding per pupil of 5%.

In addition to increasing the funding of schools, the specialist schools policy also simultaneously enhanced parental choice of school, and competition between schools for pupils, because schools were encouraged to specialise in particular subjects.[2] The earliest specialist schools were Technology schools, starting in 1994, now constituting approximately 20% of all schools, with significant proportions of schools focusing on Arts, Sport and Science. Other specialisms, such as Business and Maths were introduced more recently in 2002.[3] Specialist schools are encouraged to spread good practice to non-specialist schools in the same educational district with respect, for instance, to teaching methods. Over 80% of secondary schools are now specialist, and the intention is that all schools will eventually become specialist.

The key objective of the specialist schools policy is to improve the test score performance of secondary school pupils. Evaluating whether this policy has had the desired effect provides important guidance for policy makers contemplating whether, and how, to spend increasingly scarce resources on schooling. However, there are very few studies which focus on the

[1] The capital grant has been reduced to £25,000 in recent years but was much higher for the time period covered by this study.

[2] 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 two years of compulsory schooling between the ages of 14 and 16. They then sit national 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 two years later. Pupils of lower ability may also take General National Vocational Qualifications instead of GCSEs.

[3] See the Department for Children, Schools and Families website for more details: www.standards.dcsf.gov.uk/specialistschools.

evaluation of the specialist school policy, and the evidence from this literature is mixed. Gorard (2002), Jesson and Crossley (2004) and OFSTED (2005) find a positive effect on test scores. Schagen and Goldstein (2002) raise some issues regarding the methodological approach of this work, which uses school-level data, arguing that pupil level data and multi-level modelling techniques should be used. Taylor (2007) finds that the specialist schools policy has had very little impact on average test scores, though there is evidence of more substantial impacts for specific areas of specialisation, for example, business and technology. Bradley and Taylor (2008) estimate the impact of the specialist schools policy, as well as other educational policies, using school-level panel data, and find a small positive effect of specialist schools on test scores. However, many of these papers fail to allow for the bias that often arises in programme evaluation settings, which calls into question whether they have been able to identify a causal effect of the specialist schools policy. Furthermore, many of these studies do not explicitly consider the mechanisms by which the specialist schools policy could affect the test score outcomes of pupils.

Our contributions in this paper are twofold. This is the first paper that we are aware of to evaluate the impact of the specialist school policy on a variety of test score outcomes, including test score gain, combining matching methods at the pupil and school levels with a difference-in-differences analysis.[4] This approach enables us to deal with various forms of selection bias that we identify below, as well as the effect of unobserved pupil and school heterogeneity. The second contribution of this paper is that we provide an *exploratory* analysis of the relative importance of two different mechanisms by which the policy could affect test scores - *funding* and *specialisation* effects. The increase in resources to specialist schools creates a funding effect whereby increased spending on books and equipment, for instance, improves the quality of the educational experience throughout the school and hence may improve test scores in all subjects. By allowing greater subject specialisation, parents can select those schools that 'match' the aptitudes and skills of their children, thereby increasing allocative efficiency. 'Better' subject specialist teachers may also move to schools that specialise in their subject area. Hence test scores in particular subjects may increase - a specialisation effect.

There are likely to be several sources of bias in an evaluation of the specialist schools policy, arising primarily from selection on unobservables (so-called 'hidden bias'), that must be mitigated. First, there is the non-random selection of schools into the programme (hereafter *school selection bias*). Figure 1 shows the variation in the number of schools becoming specialist between 1994-2006. The Figure also shows that there is an output trend in GCSE test scores and it is clear that specialist schools have out-performed non-specialist schools throughout the period.[5] Moreover, some schools were early adopters whereas others joined

[4] Previous studies using matching methods are mainly focused on the estimated effects of training programmes on the unemployed. For example, Blundell et al. (2004) study the effects of the New Deal for Young People in the UK. Aakvik (2001) evaluates the Norwegian vocational rehabilitation programme by comparing employment outcomes of participants and nonparticipants. Diprete and Gangl (2004) analyze the impact of unemployment insurance on several outcomes such as post unemployment wage or probability of relocation. Machin et al. (2004) adopt a similar approach to ours in evaluating the Excellence in Cities programme.

[5] 5+ A*-C grades is an important policy measure and, when combined with suitable grades at A level, permits entry to HE.

the programme much later. 'Older' specialist schools may have had a first-mover advantage insofar as they have more time to exploit the additional resources to generate better test scores. Moreover, as their reputation grows, these schools attract better pupils, particularly in their specialism, which, via a peer effect also enhances the test score performance of other pupils in the school. This process is cumulative and self-reinforcing. There is some evidence in Figure 2 to support this view, which reports the total number of specialist schools between 1994-2005, insofar as schools that were early adopters have higher growth rates in the proportion of pupils with 'good' exam results. Figure 3 disaggregates the average test score performance of Figure 1 into quintiles and plots the proportion of specialist schools in the lowest (quintile 1) and highest (quintile 5) categories. For example, in 2003 the proportion of specialist schools in the 5th quintile of average test score performance is 60%, whereas it reduces to just 26% in the 1st quintile. What is immediately clear is that specialist schools are increasingly likely to have test scores in the highest quintile, which is strongly suggestive of non-random assignment of certain types of school into the specialist schools initiative.

A closely related source of bias is the non-random selection of pupils into specialist schools (*pupil selection bias*), insofar as unobservably more able pupils are 'cream-skimmed' by 'good' (specialist) schools. Figure 4 provides some evidence of cream-skimming based on observable characteristics that are correlated with test scores performance. Panel A of Figure 4 shows that pupils from the poorest social backgrounds, as reflected by their eligibility for free school meals, are less likely to attend specialist schools. Specifically, specialist schools are more likely to have a higher percentage of pupils in the 1st quintile (pupils less eligible for free school meals), than in the 5th quintile (more eligible for free school meals). In Panel B we show the distribution of ethnic minorities, where we plot percentage differences for the first and fifth quintiles between non-specialist and specialist schools. Specialist schools are more likely to have a lower percentage of ethnic minority pupils.

Figures 1-4 suggest that school selection bias and pupil selection bias may be very real. In principle, it would be helpful to disentangle these two sources of bias, however, this is not possible with our data. In the evaluation of the specialist schools policy that follows we therefore treat these two sources of bias as observationally equivalent and try to deal with them in a 'reduced-form' cross-sectional matching approach. This could, however, be problematic for our evaluation of the specialist schools policy because pupils and schools will clearly differ in all kinds of unobservable characteristics. We investigate whether unobservables are a problem for our cross-sectional models by using two indirect tests that have been developed in the literature and involve the inclusion of a confounder variable (Ichino et al. 2008, Rosembaum, 1987). An alternative, and more robust, approach is to exploit the availability of repeated cross-sectional data. Specifically two different models are estimated - a linear difference-in-differences (DID) model, which is combined with matching at school level following Machin et al. (2004), and a DID matching model following Heckman et al (1997 and 1998).

Figures 1-4 also imply that there are likely to be heterogeneous policy effects on pupil test scores depending on the length of time a school has been exposed to the policy. Machin et al. (2004) do, in fact, find heterogeneous effects in their evaluation of the Excellence in Cities policy. We explore the possibility of heterogeneous effects throughout our analysis.

The data used to estimate the matching models were obtained from several sources: the National Pupil Database (NPD, cohorts 2002-2004), the Youth Cohort Surveys (YCS) and the Longitudinal Survey of Young People in England (LSYPE), and to each of these datasets we append school level data from the annual School Performance Tables and the annual Schools' Census. Using these data we discuss in Section 2 how we tackle the sources of bias identified earlier.

Our main finding is that the specialist schools policy has had a positive and statistically significant *causal* effect on the test score outcomes of secondary school pupils in England. In the YCS cross-sectional matching model we investigate the effect on pupils test scores when each school switches from non-specialist ('policy-off') to specialist ('policy-on') status. The estimates suggest that the policy has raised GCSE scores by between 2-3 GCSE points. A bigger effect is obtained at the upper (5+ GCSE grades A*-C) and high end (10+ GCSE grades A*-C) of the test score distribution. These results imply that the policy has had larger effects for more able students. The duration of specialisation matters and is a common thread in our results. In the LSYPE cross-sectional estimates, the peak of the policy effect (around 3 GCSE points) is reached after approximately four years, at which point the additional funding typically ceases. However, the policy effect does not decline to zero beyond that point, rather it remains positive and statistically significant. This suggests that we are not simply capturing a funding effect. These results are substantially confirmed by the DID analysis. In particular, the linear DID model shows that pupils in schools that have been specialist for longest have higher test scores. In the DID matching model, where we consider the switch from a pre-treatment status of non-specialist schools to a treatment status of schools specialised for one or two years, the impact of the policy on test score is around 0.4 GCSE points but we do not find any effect on test score gains. This finding confirms that the length of the specialization matters, and for schools that have been in the programme less than four years the effect of the policy on test scores is small.

Finally, models that attempt to disentangle the funding effect from a specialisation effect suggest that there is a specialisation effect. This amounts to between 21-50% of the total effect depending on the matching estimator used.

The remainder of this paper is structured as follows. In Section 3 we explain the econometric approach, in Section 2 we discuss the data and how we select the treatment and comparison groups. Section 4 discusses the propensity score estimates and the findings from the cross-sectional matching models and from the difference-in-differences matching models. Section 5 draws some conclusions.

2 Data and dependent variables

We use three different datasets in our analysis, two of which are cross-sectional (the YCS and the LSYPE) and one of which has a panel element (the NPD). Tables 1 and 2 show various measures of test score outcome for each dataset. The first is the total GCSE score (*GCSEscore*), taken at age 16, that is, the number of points achieved in all GCSE subjects, where grades are ranked from A*=8 points to fail=0. The second is a binary variable indicating whether a pupil obtained 5 or more GCSE grades A*-C (*GCSEbin*). The third measure is also a dummy variable indicating whether a pupil obtained 10 or more GCSE grades A*-C (*GCSEbin10*), which refers to the upper end of the ability distribution. A fourth measure refers to tests taken at age 11, that is the total score in English, maths and science in the so-called Key Stage 2 tests, taken in the final year of primary schooling. Standardising both the *GCSEscore* and the Key Stage 2 score and taking the difference we obtain test score gain, or value added, between the ages of 11 and 16.

The YCS is a major programme of longitudinal research designed to monitor the behaviour and decisions of representative samples of young people (around 14,000 per survey) aged 16 and upwards. The survey records educational outcomes and provides more socio-demographic data on the pupil and their family than in the NPD. A further advantage of the YCS is that we observe precisely when each school switches from non-specialist ('policy-off') to specialist ('policy-on') status to investigate how the test scores of different cohorts of pupils change. Specifically, we link schools in YCS11 with the same schools in YCS12 and restrict attention to those pupils in a non-specialist school in 2001/02 ('policy-off') and compare them with pupils in the same school which acquired specialist status during 2002/04 ('policy-on'). This reduces the original sample to 5,244, see Table 1. This approach allows us to go some way to controlling for school selection bias since we essentially difference out unobserved school fixed effects. Pupil selection bias should also not be a problem since all of the pupils in the analysis had chosen a non-specialist school, which then becomes specialist during their period of secondary schooling (policy on). One potential drawback of this type of analysis is the fact that we cannot distinguish year effects and cohort effects, because the inclusion of these dummies would be perfectly collinear with the treatment variable. Consequently, we assume that temporal exogenous shocks affect all schools in the same way, hence shifting the distribution of pupil attainment.

The LSYPE is a panel study of young people started in 2004, when its sample of young people (around 15,700) were aged 13 to 14. The study brings together data from a wide range of sources and reflects the variety of influences on learning and pupil progression. Annual interviews obtain information from the pupil and from parental interviews. The main advantage of these data are that we can exploit a rich set of family covariates, such as parental education and employment. Information on pupil behaviour is also included in the survey, for instance, the use of PCs to support learning and the attitude towards their school, proxied by use of a school sports facilities. The LSYPE allows us to control for the duration of specialist school status, to investigate whether the effect, if any, of the specialist schools policy declines over time for a specific cohort of pupils. We can track schools that have been specialised for two, four or five or more years. The downside of the LSYPE is that our sample is quite small, decreasing according to the year of specialisation from 3,837 to 2,933. Nevertheless, the GCSE scores are very similar for each year of specialisation (see Table 1).

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 variables are constructed from national test scores obtained by pupils at Key Stage 2 and Key Stage 4 (i.e. GCSE tests). One 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 three versions of the NPD where pupils were in their final year of compulsory education in either 2002, 2003 or 2004. The original sample sizes are 504,555, 523,658 and 560,493 observations, respectively. We choose these three cohorts of the NPD because this is the period in which many secondary schools acquired specialist school status. Also, the percentage of pupils in specialist schools in 2002 was 30%; in 2003 it was 50%, and in 2004 it was 70%. This allows us to select representative treatment and control groups. By pooling the three cohorts we can observe the same school across cohorts and identify when the specialist school policy is 'switched on' for any particular school. Since we know when a school became specialist over the period 1994-2006 we can also investigate whether there is heterogeneity in the policy effects.

The NPD data is used in three separate analysis. First, to estimate a linear difference-in-differences model of pupil test scores at Key Stage 4 (GCSEscore). We initially pool the three cohorts. Using a school level data set we identify only those schools that became specialist in the 1997-2003 period because this corresponds to the time the pupils from the three NPD cohorts attended secondary school. For example, the 2002 cohort began their secondary schooling in 1997/98, whereas the 2004 cohort began in 1999/00. Our control group is drawn from those schools that become specialist between 2004 and 2006 on the grounds that these schools are likely to be most similar to the current group of specialist schools. To find more suitable treatment and control groups we estimate a propensity score matching at school level, using only pre-policy school characteristics. We further restrict our sample of schools to those that lie on the common support, which we then merge with the pooled pupil-level data from the NPD. Panel A of Table 2 provides the total number of post-matched specialist schools (1109), sub-divided by the year in which they became specialist, as well as the number of non-specialist schools (894). Panel A, Table 2, also reports descriptive statistics on the GCSE performance and the raw DID.

The second analysis is the estimation of a DID matching model of pupil test scores, we consider the year 2002 as our pre-treatment period, while 2003 and 2004 are separately used as post-treatment periods. We first pool separately the NPD cohorts 2002-2003 and 2002-2004. For the first case, 2002-2003, in the school level data we identify schools that are non-specialist in 2002, but that will become specialist in 2005 or 2006, we exclude schools that never become specialist. Then we identify schools that are non-specialist in 2002 but will be specialist in 2003. We finally merge the school data to the NPD. We can identify now four categories: in $C_{t'}$ we have pupils in 2002 that attended non-specialist schools that will remain so in 2003. In $T_{t'}$, we have pupils in schools that were non-specialist in 2002 but that become specialist in 2003. In C_t we observe pupils in schools that are both non-specialist in 2002 and 2003. In T_t, we observe pupils in schools that became specialist in 2003, but which were non-specialist in 2002. For the second case, 2002-2004, we repeat the above steps but 2003 is replaced by 2004. The main difference between these two analyses is that in the first case schools have only been specialist for one year, whereas in the second case schools have been specialist for two years. Panel B of Table 2 reports the total number of schools in each group, as well as the raw DID for each of our analyses. Using these data we estimate a difference-in-differences matching model on test scores at KS4 and on test score gain between KS2 and KS4.

In a third analysis we estimate matching models at the pupil level by subject of specialism, in an attempt to disentangle the funding effect from the specialisation effect. In this analysis we only use the 2003 cohort because this is most 'balanced' in terms of the percentage of pupils in specialist schools, that is, approximately 50% of pupils.

3 Econometric Approach

3.1 Cross-sectional matching methods

Our approach is based on the concept of the education production function wherein test scores are a function of personal, family and school inputs, as well as specialist school status. However, to estimate the effect of the specialist schools policy on the test scores of pupils requires a solution to the counterfactual question of how pupils would have performed had they not attended a specialist school. We adopt the non-parametric matching method which does not require an exclusion restriction, or a particular specification of the model for attendance at a specialist school. Thus, the main purpose of matching is to find a group of non-treated pupils who are similar to the treated in all relevant pre-treatment characteristics, $\mathbf{x}$, the only remaining difference being that one group attended a specialist school and another group did not. In the first stage we therefore estimate the propensity score (PS) using a discrete response model of attendance at a specialist school.

One assumption of the matching method is the *common support* or overlap condition which ensures that pupils with the same $\mathbf{x}$ values have a positive probability of attending a specialist school. A second, and key assumption is the *conditional independence assumption (CIA)*, which implies that selection into treatment is solely based on observable characteristics.[6] There may, however, be a problem of hidden bias due to unobserved effects, and any positive association between a pupil's treatment status and test score outcomes may not therefore represent a causal effect. If the assumption of ignorability (i.e. no hidden bias) fails, the treatment is endogenous and the matching estimates will be biased (Heckman et al. 1998). Several tests have been developed to assess whether hidden bias is a problem in cross-sectional models and we adopt the method proposed by Ichino et al. (2008) and Rosenbaum (1987). The details of these tests are reported in Appendix A.

Given these two assumptions, the matching method allows us to estimate the 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.

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

$$\tau_{ATT} = \sum_{i \in T} (Y_{1i} - \sum_{j \in C} W_{ij} Y_{0j}) \, w_i \tag{1}$$

where T and C represent treatment and control groups, respectively. W_{ij} is the weight placed on the jth observation in constructing the counterfactual for the ith treated observation. Y_1 is the outcome of participants and Y_0 of non-participants; 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_{ij}. We use two matching algorithms, nearest neighbor (NN) and kernel with bandwidth 0.1. Analytical standard errors are provided for the first estimator (Abadie and Imbens, 2008) and bootstrapped standard errors for the second (Heckman et al. 1998).

[6] Conditional on a set of pre-treatment observable variables $\mathbf{x}$, potential outcomes are independent of assignment to treatment.

3.2 Difference-in-differences

Given the availability of the NPD data, which has a short panel element for repeated cross-sections of pupils, we report the estimates from the two DID analyses. Recall that the first is a linear DID estimator coupled with matching based on schools, whereas the second is a DID matching estimator at pupil level.

Recall that to obtain the first estimator, we initially match schools using a propensity score model which includes several pre-policy school variables. Only those schools that lie on the common support are included in the analysis to ensure that the treatment and control groups of schools are as like as possible. A DID model is then estimated at the pupil level, see equation 2, which has the advantage of eliminating unobserved time-invariant differences between treated and untreated pupils.

$$Y_{istj} = \alpha_j + \beta_{tj} Policy_{ist} * Yearspec_j + \gamma_t X_{it} + \delta_j Z_s + \lambda_j Y_{is,t-5} + \epsilon_{istj} \tag{2}$$

Y is the test score at age 16 of pupil i in year t in school s which is specialist from year j. $Policy$ is an individual level dummy indicating whether the pupil is assigned to a specialist school in year t, whereas $Yearspec$ is a dummy indicating when a school becomes specialist ('policy on') in year j in the period 1997-2002. X denotes individual characteristics, Z is a set of lagged (pre-policy) school characteristics, and the lagged dependent variable $Y_{is,t-5}$ indicates the pupil's prior (primary school) attainment measured by the total test score at age 11. The coefficient of interest is β_{tj} which represents the ATT, or the difference-in-differences estimate of the specialist schools policy. These models allow us to control for time-invariant unobserved pupil effects and school fixed effects.

The second estimator adopted is the difference-in-differences matching estimator, which has been implemented by Blundell et al (2004), Smith and Todd (2005) and Machin et al (2004). This approach relaxes the strong assumption of the cross-sectional matching approaches of selection based solely on observables and allows for temporally invariant differences in outcomes between pupils in specialist and non-specialist schools. The DID matching can be seen as an extension of simple matching, because the bias is not required to vanish for any covariates but just to be the same before and after treatment.

The DID matching estimator for repeated cross-section data can be obtained by rewriting equation 1 as

$$\tau^{DID}_{ATT} = \sum_{i \in T_t} [\, Y_{1ti} - \sum_{j \in C_t} W_{ij} Y_{0tj}] \, w_{it} - \sum_{i \in T_{t'}} [\, Y_{0t'i} - \sum_{j \in C_{t'}} W_{ij} Y_{0t'j} \,] \, w_{it'} . \tag{3}$$

where t' and t are time periods before and after the acquisition of specialist school status. Specifically, $T_{t'}$ is formed by students in schools non-specialist in t' that will be specialist in t, $C_{t'}$ is formed by students in schools non-specialist in t' that will remain non-specialist in t. T_t includes students in schools specialist in t that where non-specialist in t', C_t includes students in schools non-specialist in t which were also non specialist in t'.

The specialist schools policy may have had a disproportionate effect on test scores at age 16 (the GCSE) because these are 'high stakes' exams that influence the ranking of schools in local and national league tables. Therefore, we also estimate a value added DID matching model where we compute the difference between the standardized tests taken at age 16 and at age 11, and then estimate equation 3 except that Y is replaced by $\Delta Y = Y_{age16} - Y_{age11}$.

A further issue arises insofar as expenditure per pupil has risen for reasons other than the specialist schools policy and we must control for this. Therefore, we estimate a post-matching regression including real expenditure per pupil, measured in 2003 prices, of the form:[7]

$$Y_{is} = \alpha + \beta Spec_{is} + \gamma \widehat{p}(x_i) + \rho Spec_{is}(\widehat{p}(x_i) - \mu_p) + \theta Exp_{is} + \varepsilon_{is} \tag{4}$$

where $\widehat{p}(x_i) = P(Spec_{is} = 1|x_i)$ is the estimated propensity score, μ_p is the sample average of $\widehat{p}(x_i)$ and Exp_{is} are the expenditure per pupil in school s. Since we only have data on expenditure from 1999, we restrict our sample to schools that become specialist from 1999 to 2002, and schools that are non-specialist in the same period but which become specialist between 2003 and 2005.

4 Findings

4.1 Estimation of the propensity score models

In the estimations of the propensity score models, the choice of the covariates to be included is an issue (Heckman et al. 1997, Bryson et al. 2002). There is some discussion in the literature which emphasises that the balancing condition is satisfied, hence reducing the influence of confounding variables (e.g. Dehejia and Wahba 1999, DiPrete and Gangl, 2004). Therefore, in the estimation of the propensity score with most of our data sets we include only those variables that satisfy both the balancing test and the common support condition. However, given the extremely large sample size of the NPD, with more than 300,000 observations in our analyses, it is very difficult to pass this test. In this case it is important to include in the propensity score model all variables that are strong predictors of attendance at specialist schools and outcomes. These variables include factors that affect a pupil's choice of school, and also those variables that affect a schools' acquisition of specialist status. Table 3 reports the estimates for a selection of the estimated PS models, including those that use the NPD.

A larger number of covariates are included in the propensity score model using the YCS data, for instance, controls for family background (i.e. parental occupation) and whether the pupil is from a single parent background. A very important variable to include is the school performance lagged five years, which is likely to be an important influence on school choice for pupils at age 11 and selection of a school into the specialist schools initiative.[8] Most variables are statistically significant and we note a strong effect on the lagged school performance on school choice (the marginal effect is also very strong, 0.993 with a s.e of

[7] In the post-matching analysis we adopt a control function approach using the propensity score (Wooldridge, 2005 and Rosembaum and Rubin, 1983). We regress our dependent variable (test scores) on the treatment dummy variable (*Spec*), the estimated propensity score and its deviation from the mean interacted with *Spec*. The coefficient of the *Spec* dummy consistently estimates ATE. In our specific case, we add in the post-matching regression the expenditure per pupil variable and its deviation from the mean interacted with *Spec*. In this way we want to get the treatment effect corrected for the fact that the control group changes over time.

[8] This variable measures the proportion of pupils in the school (lagged five years) who obtained five or more A*-C grades in the GCSE examinations.

0.173); estimates on the family background variables are also quite large and statistically significant.

The analysis in the LSYPE is more complex, since the treatment group refers to pupils in specialist schools disaggregated by the duration that the school has been in receipt of specialist school funding; the control group comprises pupils in non-specialist schools. Specifically, we consider three possible treatments: schools that have been specialist for five or more years, for four years and for two years. Many more covariates are included in these models than in the previous models for the YCS. The estimates typically have the right sign and most of them are statistically significant.

In the propensity score model estimated for the school level matching (not reported) we include only pre-policy school characteristics, that is, variables measured before the start of the specialist school programme in 1994. A large number of school and pupil characteristics are included, such as school size, the pupil-teacher ratio, the proportion of pupils eligible for free school meals, the ethnic composition of the school and school type. All of these variables are highly statistically significant and they pass the balancing test.

To compute the DID matching estimator we estimate four propensity score models, corresponding to the level and value added matching models for the year 2002-2003 and 2003-2004. In Table 3 we only report the model constructed from the NPD 2002-2003, but the others have similar results. We include a measure of prior attainment (KS2 test score), which captures the cumulative effect of the history of family, pupil and school inputs that determined test scores up to age 11 (Todd and Wolpin, 2002). Its effect is highly statistically significant with a marginal effect of 0.02 (s.e.=0.002). This suggests that more able primary school pupils sort into, or are selected by, specialist schools. We also include dummies for ethnicity and gender, both of which are statistically significant. A measure of family income, that is, whether a pupil is eligible for free school meals, is also included. The marginal effect is negative and highly significant and suggests that pupils from poorer backgrounds are less likely to attend specialist schools. We also include a number of school level variables, that are measured in the pre-policy period, including school type, pupil-teacher ratio, the proportion of pupils eligible for free school meals and school size which may influence whether a school is selected into the specialist schools initiative. All of these variables are highly statistically significant.

In sum, the covariates in the propensity score models work in the expected direction and confirm some of the claims made in the Introduction regarding the sorting of schools and pupils into the specialist schools programme.

4.2 'Cross-sectional' matching estimates

In this Section we investigate the potential impact of the specialist schools policy on test scores assuming no hidden bias, that is, that there is no correlation between treatment status and unobserved variables. However, we do need to assess the effect of estimation on the reduction in bias on observables and the CIA. We therefore report the results with and without the inclusion of the confounder variable, and assess matching quality by reporting the standardized bias associated with each matching estimator (Caliendo et al, 2005).[9] In

[9] This requires that for each covariate we compute the standardised bias (SB) in the unmatched and matched sub-samples as the difference in sample means between treated and control observations, divided

most empirical studies a bias reduction of 3% to 5% is seen as a success of the matching procedure.

4.2.1 *A 'policy-on' versus 'policy-off' analysis using the YCS*

Table 4 shows that, prior to matching, pupils in a given school during a 'policy-on' period obtain around 2.8 GCSE points more than their counterparts in the *same* school in the 'policy-off' period (*Gcsescore*). After matching we observe a reduction in the effect on GCSE points score by between 15-37%, with the estimated impact falling to between 1.7-2.0 points, depending on which estimator is used. Interestingly, there is no statistically significant difference in the proportion of pupils obtaining 5 or more GCSEs graded A*-C (*Gcsebin*). However, at the very top of the attainment distribution (*Gcsebin10*) a positive and statistically significant effect is observed. In fact, the pre-match estimate of 0.09 falls to between 0.07-0.08 implying that the specialist schools policy increased the probability of obtaining 10+ GCSE grades A*-C by between 7-8 percentage points.

The inclusion of a confounder variable does not dramatically change our results. This is also confirmed by the Rosembaum test (see Table 4, lower panel) in a situation of up to 50% hidden bias. As an additional robustness check we estimate equation 4 which corresponds to a post-matching regression that includes expenditure per pupils at school level. We notice that the effect of the policy remains substantially unchanged compared to the matching estimates. In particular, for *Gcsescore* the effect is 1.86, a values that lies between the nearest neighbour and kernel estimates. The same happens for *Gcsebin10*, while the effect for *Gcsebin* is still insignificant. Note that in this analysis we mitigate the bias arising from school selection bias, this is because we remove unobserved school fixed effects and due to the short time framework of the analysis it is also unlikely to be affected by pupil selection bias.

4.2.2 *The effect of the duration of specialist school status using the LSYPE*

Insofar as specialist schools receive extra funding per pupil for 4 years after they have become specialist, and given that subject-specific 'reputation' effects take time to develop, then one would expect the positive effect on test scores that we observe to be larger the longer the school has had specialist status. Table 5 shows that this is, in fact, what we observe. Compare the results for schools that have been specialist for 4 and 5 or more years with those that have been specialist for only 2 years, where the heterogeneity of the specialist schools policy becomes apparent.

Prior to matching, pupils in schools that have been specialist for longest obtain 4 GCSE points more than their counterparts in non-specialist schools. The equivalent figure for schools that have been specialist for only 2 years is lower at 2.8 GCSE points. After matching, these effects fall to 2.3 and 1.3 GCSE points, respectively, and the latter are statistically insignificant. The duration of specialist status clearly matters, however, it is also worth comparing the estimates for schools that have been specialist for 4 years with those that

by the square root of the average of sample variances in both groups. We then average over the SB of each covariate in the two subsamples in order to obtain the absolute value of the SB before matching and after matching.

have been specialist for 5 or more years. Recall, that funding typically lasts for up to 4 years. What we observe is that both pre- and post-matching estimates for the schools that have been specialist for 4 years are larger by between 0.5-1 GCSE point, implying that once the funding begins to dry up, the effect on test scores begins to wane. Importantly, however, it does not fall to zero.

In sum, this analysis suggests that the longer the time a school has been specialist, the better the test scores of the pupils, however, this effect falls as funding declines. It is also worth noting in passing that the estimated effects for schools that have been specialist for 4 or more years are consistent with those from the previous analysis using the YCS.

The standardized bias is substantially reduced, but the estimates are less robust to the inclusion of a confounding variable. Moreover, the Rosembaum's test in Table 5, lower panel shows that our estimates are not sensitive to hidden bias up to a level of 25%.

4.3 Difference-in-differences estimates using the NPD data

4.3.1 *School Matching and linear difference-in-differences in test scores*

Table 6 reports two sets of estimates from three models, one for the comparison of the 2004 with the 2002 cohort, and the other for the comparison of the 2003 and 2002 cohorts. In each case the dependent variable is the unstandardised GCSE score. In model 1 we include only the policy effect and year dummies, and these estimates correspond to the 'raw' difference-in-differences reported in Panel A of Table 2. For instance, the average effect of the specialist schools policy is 0.266 for the comparison of the 2004 and 2002 cohorts (see the penultimate column). This suggests that pupils in specialist schools achieved an improvement of approximately one quarter of a GCSE point more than their counterparts in non-specialist schools. This estimate is retrieved in Table 6 (see model 1, column 1). To reduce the effect of unobservable school differences, model 2 adds a large number of school and pupil composition variables, along with a smaller number of pupil characteristics. In model 3 we also add school fixed effects.

Model 1 of Table 6 shows the heterogeneity in the estimates of the specialist schools policy effect; the estimates range from zero to 0.9 of a GCSE point for the 2004-2002 cohorts and zero to 1.4 GCSE points for the 2003-2002 cohorts. Adding pupil and school covariates (model 2) increases the magnitude of the average effect to around 0.3-0.4 of a GCSE point and this does not change much when we add school fixed effects (model 3). The policy effects for each year also tend to increase once covariates and school fixed effects are included. What is interesting is that the largest effects are observed for schools that have been specialist for longer - compare the estimates from model 3 for the years 1999-2000 with those of 2002. For instance, for pupils in schools that became specialist in the period 1999-2000, GCSE scores improve by between 0.8-1.2 points. These estimates are slightly lower than those obtained in the cross-sectional analysis using the YCS and about half those obtained for the LSYPE. Nevertheless, a positive and statistically significant effect of the policy is observed after the funding has been switched off.

4.3.2 *Difference-in-differences matching in test score*

In Table 7 we report the estimates of the DID matching estimator (equation 3) from two models. In model 1 the dependent variable is *Gcescore*, and we compare the post-treatment cohort in 2004 with the pre-treatment cohort in 2002, which is then repeated for the 2003 and 2002 cohorts. We also report the cross-sectional matching estimates corresponding to the first and the second difference of the model together with the DID estimates. The unmatched values are slightly lower than those reported in Panel B of Table 2 because we are now considering only observations on the common support. Looking at the results for 2002-04, the unmatched effect of the policy for the second difference $T_t - C_t$ is positive and significant and higher than the estimate for the first difference $T_{t'} - C_{t'}$, this gives a DID value of around 0.05 GCSE points. This can be interpreted as the increase in test scores that a pupil in a non-specialist school in 2002 would have obtained in 2004 by attending a specialist school, compared to a similar pupil who attended a non-specialist school in both 2002 and 2004. After matching, we notice that the estimates of the second and the first difference drops substantially compared to the unmatched case, however the drop in the first difference is much higher and consequently this gives a DID ATT effect that is positive and much larger than the unmatched estimate at around 0.37 GCSE points. This implies that, after controlling for unobservable time invariant characteristics, the policy still has a positive effect on test scores. Similar findings are observed when we compare the 2002-2003 cohorts, however the DID ATT is now statistically insignificant. This is probably because the schools that are specialist have only been so for one year. Hence, the duration of specialist status matters.

In model 2 we repeat the same analysis but replace the level of the test scores with a value-added measure, that is, test score gains between KS4 and KS2. Since we use standardized test scores, the effect is measured in standard deviation points. All estimates are highly significant before matching, post matching the results are statistically insignificant. Again we think that this is due to the fact that schools have only been specialist for one or two years of the pupils secondary school education.

4.4 The relative importance of the funding and specialisation effects

So far we have considered the total impact of the specialist schools policy by simply looking at the test score outcomes in all subjects for pupils in specialist schools compared to various control groups. In this Section we construct a test to try to disentangle the funding and specialisation effects of the specialist schools policy using the NPD.

We focus on test score differences solely for the subjects in which the schools specialised, using the NPD. We restrict our analysis to schools that had become specialist in one of the following subject areas - Languages, with and without English, and Technology, which comprise the majority of pupils.[10]

To disentangle the specialisation effect from the funding effect we compare the estimates from Panel A with those from Panel B in Table 8. Panel A compares the test score outcome

[10] We tried to include more subjects but we did not have enough observations to perform a matching analysis.

in say, Languages, of pupils in a specialist school which specialises in that particular subject (the treatment group) with the test score outcome of pupils in Languages in specialist schools that do not specialise in that subject (the control group). Since both schools are specialist they receive the same funding and so the funding effect is constant. In contrast, Panel B compares our treatment group with a different control group - pupil's test score in Languages in non-specialist schools. Since the latter do not receive extra funding, any difference in test score outcomes in Panel B must arise from both the funding and specialisation effects. The difference in the estimates from Panel A and Panel B gives the specialisation effect.

Table 8, Panels A and B show that after matching *Gcsescore* falls substantially, and they are robust to a confounder variable. However, what is of most interest is the fact that the estimates from Panel B are higher than those for Panel A and the magnitude of this difference depends on the matching estimator. For instance, compare the nearest neighbour estimates for Technology of 0.23 (Panel B) and 0.18 (Panel A) with the equivalent for the kernel matching method, that is, 0.32 and 0.16, respectively. The difference between the estimates in Panels A and B for the nearest neighbour method is roughly 0.05 for all subjects, which implies that the specialisation effect constitutes around 22% of the total effect of the specialist schools policy. For the kernel method the implied percentage contribution of the specialisation effect varies from 38% for English to 50% for Technology. Thus, although the actual magnitude of the impact of the specialist schools policy on test scores in the subjects analysed is modest, when compared to the findings in earlier sections, the contribution of the specialisation effect is quite large when measured in percentage terms.

A word of caution is necessary, however, insofar as the specialisation effect might be picking up the fact that 'good' specialist schools, typically the early adopters, simply applied for specialist status in their strongest subject. If this view is correct then we should observe that a specialist schools performance in tests in subjects other than that in which it specialises is lower when compared to a non-specialist school. To assess whether this is the case we examine the raw data and make pairwise comparisons between specialist schools in specialisms m (English, Technology or Languages) versus non-specialist schools for subjects n, where $m \neq n$.[11] For schools specialising in English the raw data suggests superior performance in all n subjects by around 0.3 of a GCSE point; for Language and Technology schools the differential falls to 0.1-0.2 of a GCSE point. This evidence is encouraging insofar as it suggests that our specialisation effect is real.

5 Conclusions

In this paper we evaluated whether there is a causal association between the specialist schools policy, which can be regarded as a structural change in UK education policy beginning in 1994, and the test score outcomes of secondary school pupils in England. Our approach has been to use matching methods, which have become popular in the context of programme evaluation, especially with respect to the effectiveness of training schemes and programmes for the unemployed. To our knowledge there has been no previous attempt to apply such

[11] The n subjects refer to maths, science, history and business studies; we also examine languages, English and technology but only where $m \neq n$.

methods to an evaluation of the specialist schools policy. By adopting this approach we explicitly confronted the twin problems of the choice of suitable control groups, to answer the counterfactual question of what would have happened in the absence of treatment, and the potential bias arising from a correlation between the treatment status and observed and unobserved covariates.

We used several datasets in our analysis, the NPD, several versions of the YCS and the LSYPE, which allow us to construct different control groups and hence test the robustness of our estimates. Three measures of test score outcome, relating to particular points on the test score distribution (*Gcsebin* and *Gcsebin10*) or a summary of the entire distribution (*Gcsescore*), were considered. We estimated both individual cross-sectional matching and difference-in-differences matching models, and also linear difference-in-differences models combined with matching at school level. We investigated the effect of the specialist schools policy on both tests scores and change in test scores between the ages of 11 and 16.

Our main findings are as follows. The estimates from the cross-sectional models were generally consistent suggesting that the specialist schools policy has increased the GCSE points score by approximately 2-3 GCSE points, and a more substantial effect has been observed at the upper (*Gcsebin*) and high end (*Gcsebin10*) of the test score distribution. These results imply that the policy has had a more beneficial effect on more able students. Furthermore, the longer a school has been a specialist school the larger the impact on test scores, however, we found some evidence that the impact begins to fall after 4 years once the additional funding associated with the policy begins to decline. Importantly, the policy effect did not fall to zero. These results were robust to tests that controlled for the presence of unobservable confounding factors. The linear DID model evaluated the effect on test scores gains of school becoming specialist while the pupils were still enrolled. The results confirmed that the effect on GCSE scores is slightly higher for those schools that have been specialist for longest, but the specialist schools policy has had a modest effect on GCSE test scores, and no effect on KS3 tests.

In the DID matching models we estimated the effect of the policy on test score levels and test score gains from a pre-treatment status to a post-treatment status. We have been able to control for unobserved individual and school fixed effects and time-invariant unobserved characteristics. We found a positive and highly statistically significant effect of the policy on test scores for schools that have been specialist for two years. The effect is statistically insignificant if schools have been specialist for only one year and when we use value-added test scores. These results confirm that duration of specialization matters and that ‘older’ schools may attract better pupils.

Finally, the impact of the specialist schools policy on test scores could arise from a funding effect, a specialisation effect or a peer effect. We attempted to disentangle these effects. Our findings for GCSE points score suggested that between 21-50% of the total effect in particular subjects arises from the specialisation effect. These findings were consistent with our evidence on the duration of the specialist school policy effect.

In conclusion, having controlled for various types of selection bias, we argue that the specialist schools policy has had a statistically significant causal effect on test score outcomes and test score gain.

A Appendix: Testing the validity of the CIA assumption

The CIA is not directly testable, because the data are uninformative about the distribution of $Y_i(0)$ for the treated and of $Y_i(1)$ for the control group. We therefore use two indirect tests from the literature. One test, developed by Imbens (2004), proposes an indirect way of assessing the CIA, based on the estimation of a 'pseudo' confounding factor that should, if the CIA holds, have zero effect. We adopt the method proposed by Ichino et al. (2008). This is based on the prediction of a confounding factor, A, by simulating its distribution for each treated and control unit. Then, estimates of the average treatment effect of the treated (ATT) are derived by including the confounding factor in the set of matching variables. Different assumptions on the distribution of A imply different possible scenarios of deviation from the CIA.

For simplicity, let A be a binary variable, its distribution is given by fixing the following parameters

$$P(A = 1|T = i, Y = j) = p_{ij} \quad i, j = 0, 1$$

Where Y is a binary test score outcome (e.g. 'high ability'$= 1$ and 'low ability'$= 0$) and T is the pupil's treatment status. In this way, we can define the probability of $A = 1$ in each of the four groups identified by the treatment and the outcome.[12] In our analysis, we assume that the confounding variable follows the same distribution as that of the pupil test scores prior to entry to secondary school (i.e. Key Stage 2 scores). Therefore A can be thought as a measure of the ability of the pupil that secondary schools 'observe' in making selection decisions. The effect of bias on the estimation of the policy varies depending on the dataset used.[13] The variable A is included in the set of variables used to estimate the propensity score and the ATT is estimated using the nearest neighbour algorithm.[14] The ATT is re-estimated 500 times, and the values presented in our Tables are an average over the distribution of A(See *NN with confounder*).

Given this set up, if the confounded estimates are still significant, but with the same sign and (similar) in magnitude when compared to the 'true' estimates, we can be fairly confident of the robustness of our results.

The second method we adopt has been proposed by Rosenbaum (1987) and involves only one parameter, representing the association of T and A, and derives bounds for significance levels and confidence intervals. Specifically, it computes the upper and lower bounds on the Mantel and Haenszel (MH, 1959) test-statistic used to test the null hypothesis of no treatment effect. In particular, e^{γ} measures the degree of departure from a situation that is free of hidden bias ($e^{\gamma} = 1$) and we use e^{γ} in the range [1,2]; γ represents the effect of an unobserved variable on the probability of attendance at a specialist school.[15] The test

[12] The simplifying assumption that the simulation of A does not depend on X does not change the interpretation of the test. For a complete explanation of the test see Ichino et al. (2008).

[13] For example, for the NPD we obtain the following parameters: $p_{11} = 0.76$, $p_{10} = 0.30$, $p_{01} = 0.73$, $p_{00} = 0.29$; p_{11} can be interpreted as the proportion of 'high ability' pupils in specialist schools who get high test scores. In contrast in the LSYPE we get $p_{11} = 0.72$, $p_{10} = 0.28$, $p_{01} = 0.75$, $p_{00} = 0.33$, where there is a slightly higher probability of 'high ability' pupils obtaining high test scores attending non-specialist schools.

[14] We omit the results for different matching methods because they are very similar.

[15] Thus $Pr(D_i = 1|x_i, u_i) = F(\beta x_i + \gamma u_i)$ is the probability of attending a specialist school and F is the

can be interpreted as the difference in the relative odds of attending a specialist school for two pupils that appear similar in terms of observable covariates, $\mathbf{x}$. If those most likely to go to specialist schools are more able, then there is positive unobserved selection and the estimated treatment effects overestimate the true treatment effect. In general, DiPrete and Gangl (2004) stress that the results of this test are worst-case scenarios, insofar as they only reveal how the hidden bias might alter inference.

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.

Aakvik, A. (2001) Bounding a Matching Estimator: The Case of a Norwegian Training Program, *Oxford Bulletin of Economics and Statistics* 63(1), 115-143.

Blundell, R. Dias M. (2002) Alternative approaches to Evaluation in Empirical Microeconomics, *Portuguese Journal of Economics* 1, 91-115.

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. (2008) Diversity, choice and the quasi-market: An empirical analysis of secondary education policy in England, *Lancaster University Management School working paper 2007/38.*

Bryson, A. Dorsett, R. and Purdon, S. (2002) The use of Propensity Score Matching in the Evaluation of Labour Market Policies, *Working paper n.4*, Department of Work and Pensions.

Caliendo M. and Kopeinig S. (2005) Some Practical Guidance for the Implementation of Propensity Score Matching, *IZA DP N.1588.*

Cochrane, W. and Chambers, S. (2003) The Planning of Observational Studies of Human Populations, *Journal of the Royal Statistical Society*, series A 128, 234-266.

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.

DiNardo, J. and Tobias, J. (2001) Nonparametric Density and Regression Estimation, *Journal of Economic Perspectives* 15(4), 11-28.

DiPrete, T. and Gangl M. (2004) Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments, *Sociological Methodology* 34, 271-310.

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

logistic distribution. The odds that pupil i attends a specialist school is given by $\frac{P_i}{(1-P_i)} = \exp(\beta x_i + \gamma u_i)$, and the odds ratio of receiving this treatment is $\frac{\frac{P_i}{(1-P_i)}}{\frac{P_j}{(1-P_j)}} = \exp(\gamma(u_i - u_j))$. For simplicity, u is assumed to be a dummy variable and the previous equation may be rewritten as $\frac{1}{e^\gamma} \leq \frac{\frac{P_i}{(1-P_i)}}{\frac{P_j}{(1-P_j)}} \leq e^\gamma$. In our work, we apply the routines *mbound* and *rbounds* available in Stata. A detailed explanation of the method can be found in Rosenbaum (1995), Aakvik (2001), DiPrete and Gangl (2004).

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

Heckman, J. Hichimura, H. and Todd, P. (1997) Matching as an Economtric Evaluation Estimator: Evidence from Evaluating a Job Training Programme, *Review of Economic Studies* 64, 1017-1098.

Hoxby, C.M. (1996) Are Efficiency and Equity in School Finance Substitutes or Complements?, *Journal of Economic Perspectives*, American Economic Association, 10(4), 51-72, Fall.

Ichino, A. Mealli, F. and Nannicini T. (2008) From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics* 23, 305-327.

Imbens, G. (2004) Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Survey, *Review of Economics and Statistics*, 86, 4-30.

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

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

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 .

Mantel, N. and Haenszel, W. (1959) Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, *Journal of the National Cancer Institute* 22, 719-748.

Rosenbaum, P. R. (1995) Observational Studies, *Springer-Verlag*, New York.

Rosenbaum, P. R. Rubin, D. (1987) The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika* 70, 41-50.

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

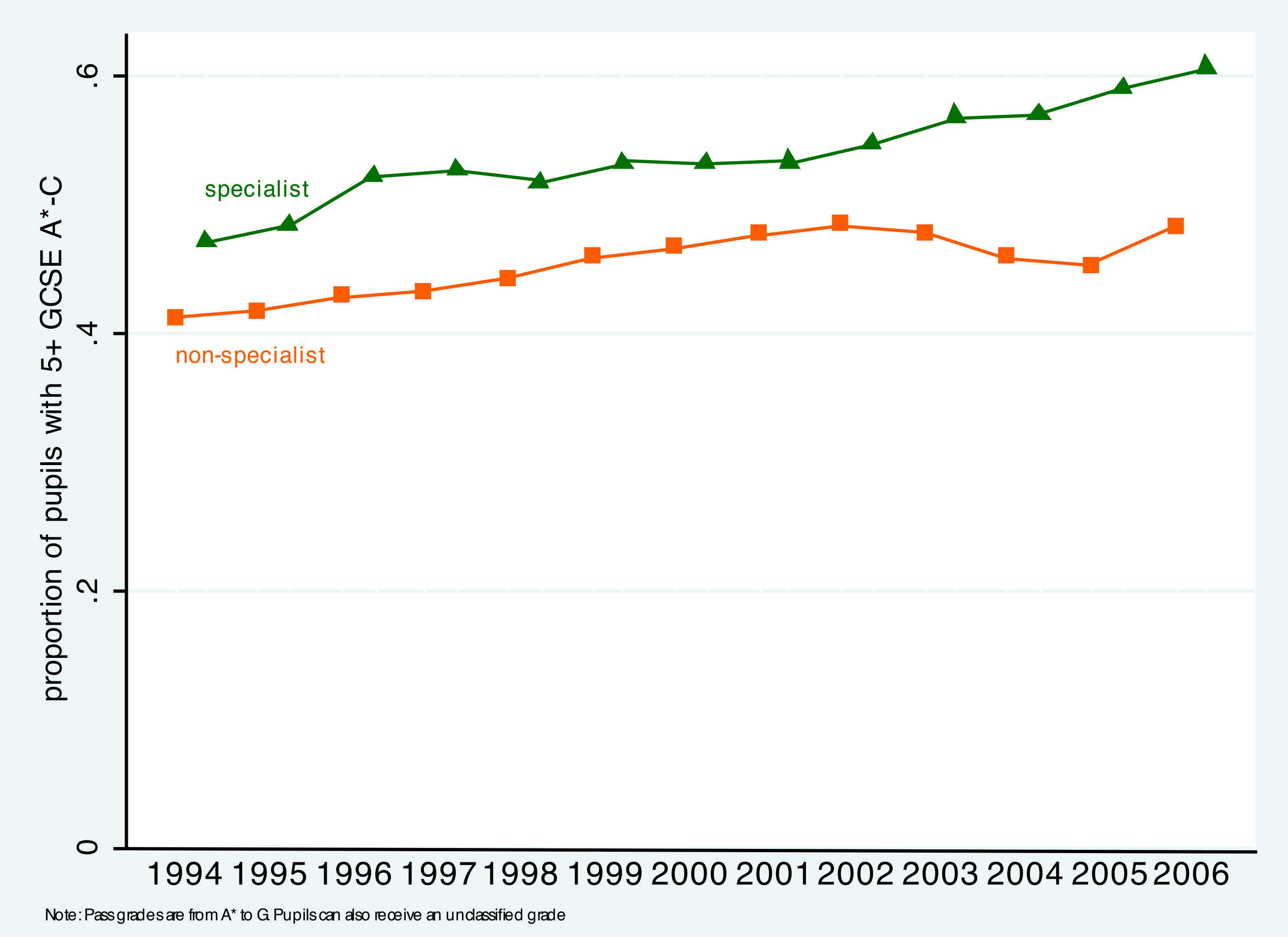
Smith, J. and Todd, P. (2005) Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?, *Journal of Econometrics* 125(1-2), 305-353.

Smith, J. (2000) A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies, *Swiss Journal of Economics and Statistics* 163(3), 1-22.

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.

The Economist (2009) *Ready, set, go*, from The Economist print edition Oct 1st 2009.

Figure 1: The test score performance of specialist and non-specialist schools over time



Note: Pass grades are from A* to G. Pupils can also receive an unclassified grade

Figure 2: School entry to the specialist schools initiative and test score performance

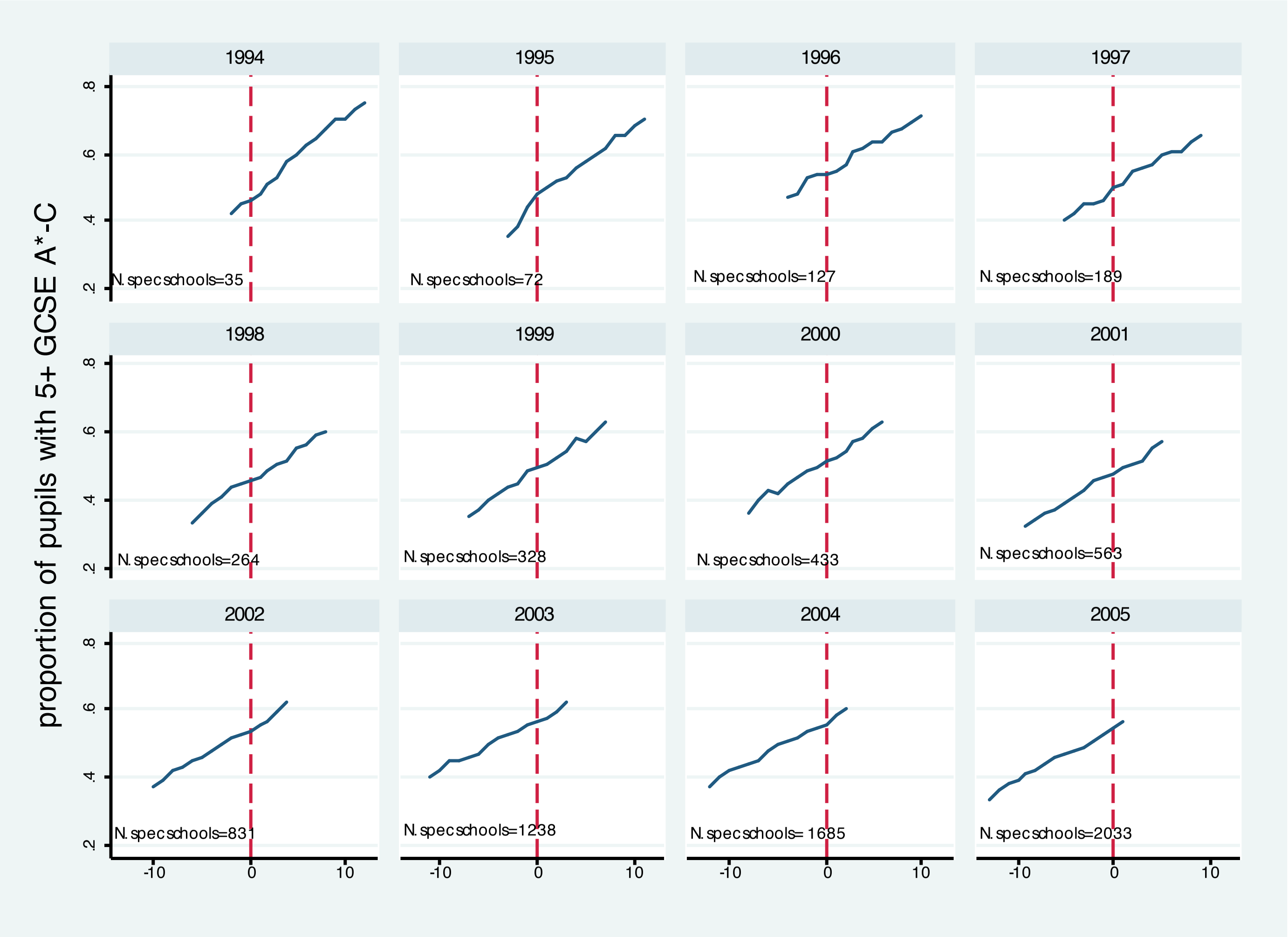


Figure 3: Specialist schools exam performance by quintile

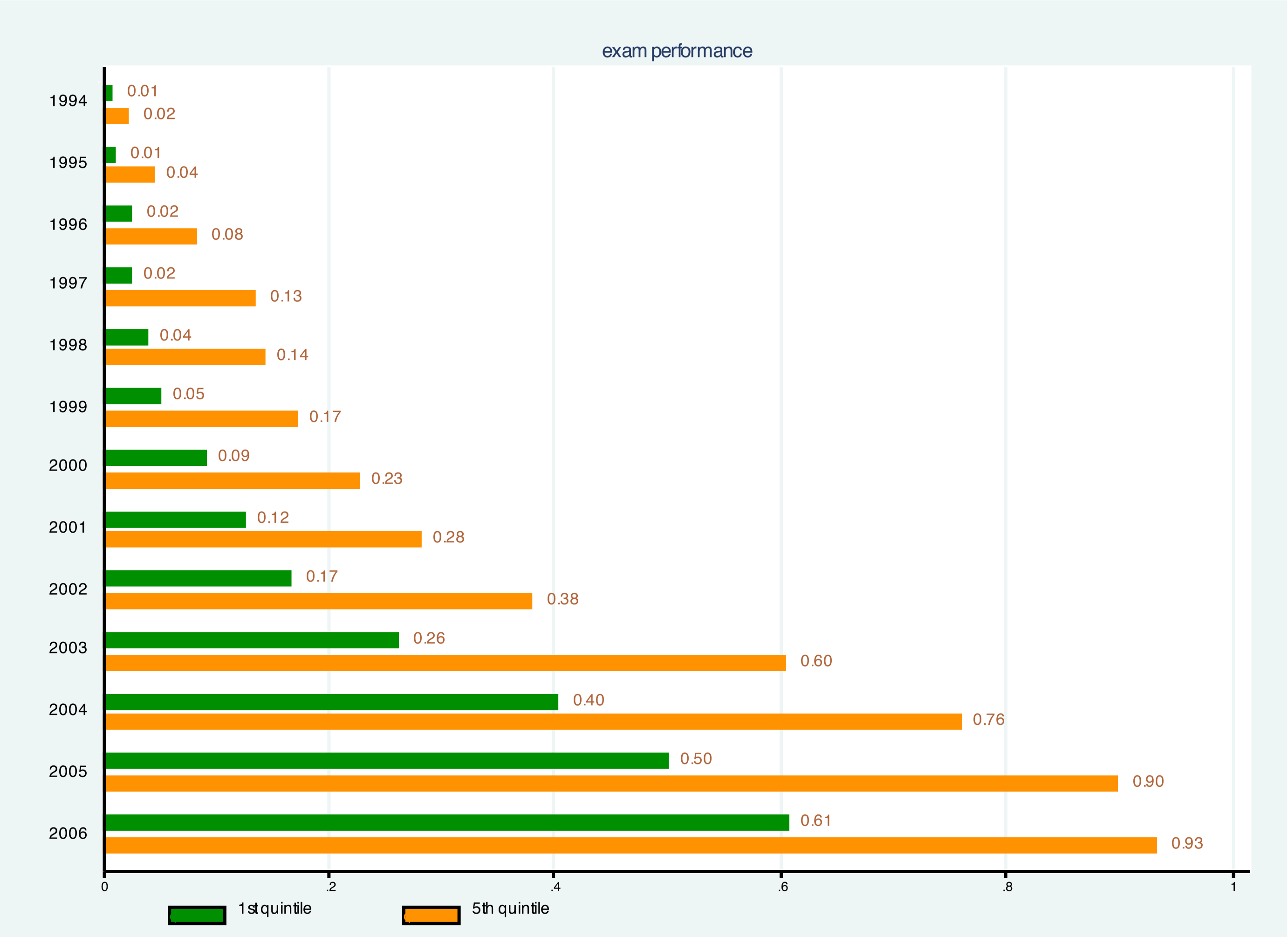


Figure 4: Specialist schools pupil composition and cream-skimming

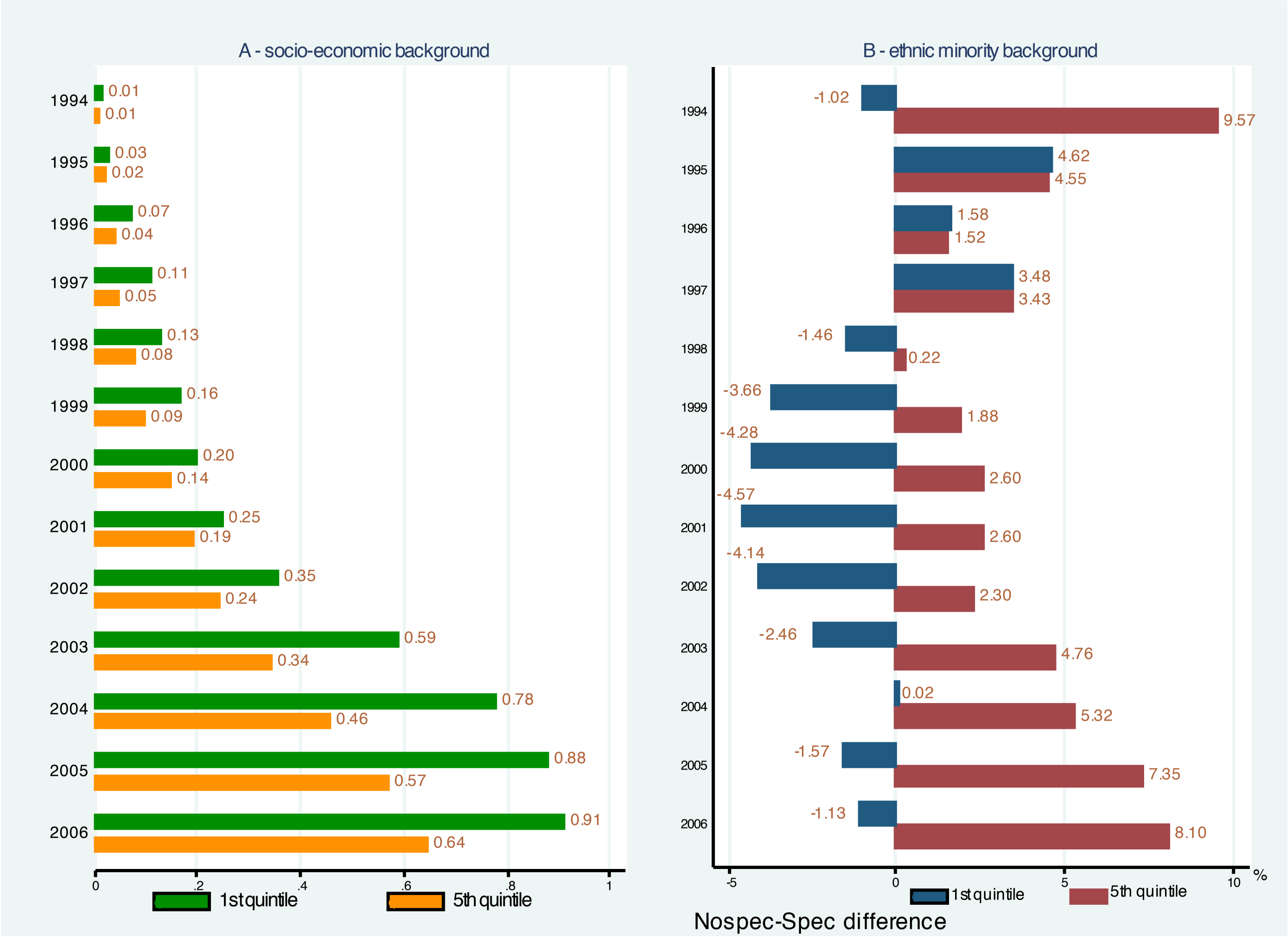


Table 1: Dependent variables

	LSYPE			**YCS11-12**
	Duration specialist			
	5+years	4 years	2 years	
Gcsescore mean				
Non-specialist	45.805	44.303	45.805	43.124
N	1,455	1,873	1,455	2,796
Specialist	49.859	49.871	48.491	45.904
N	2,382	1,150	1,476	2,448
Gcse proportions				
Gcse A*-C <5				
Non-specialist				54.19
Specialist				45.81
N				1,705
Gcse A*-C >5				
Non-specialist				52.90
Specialist				47.10
N				3,539
Gcse A*-C <10				
Non-specialist				55.73
Specialist				44.27
N				4,536
Gcse A*-C >10				
Non-specialist				37.85
Specialist				62.15
N				708

Table 2: Descriptive statistics for the DiD analysis using the NPD

Panel A: Number of schools and GCSE score for linear DID

	No. of schools	*(2002)*	*(2003)*	*(2004)*	Δ 2004-02	Δ 2003-02	DID 2004-02	DID 2003-02
all	1,109	42.874	43.613	43.902	1.028	0.739	0.266	0.278
No. of Specialist schools by								
2002	268	42.617	43.232	43.789	1.172	0.615	0.410	0.154
2001	130	41.077	41.739	41.758	0.681	0.662	-0.081	0.201
2000	105	42.915	44.747	44.446	1.531	1.832	0.769	1.371
1999	62	42.978	44.860	44.675	1.695	1.882	0.933	1.421
1998	73	43.182	43.979	44.649	1.467	0.797	0.705	0.336
1997	64	44.902	45.758	45.270	0.368	0.856	-0.394	0.395
Non-specialist	894	40.766	41.227	41.528	0.762	0.461		

Panel B: Number of schools and GCSE score for DID matching

	No. of schools	*(2002)*	*(2003)*	*(2004)*	$Y(T_{t'}) - Y(C_{t'})$	$Y(T_t) - Y(C_t)$	DID
Specialist 2003	407	43.286	43.623		3.490	3.460	-0.03
Non-specialist	448	39.796	40.163				
N obs		133,967	139,376				
Specialist 2003-2004	851	42.555		43.347	2.759	2.791	0.032
Non-specialist	448	39.796		40.556			
N obs		204,414		225,467			

Y= GCSE score, $T_{t'}$=non specialist 2002 but specialist in 2003 (2004), $C_{t'}$=Non-specialist in 2002, C_t= non-specialist in 2003 (2004), T_t= specialist in 2003 (2004).

Table 3: Probit estimates from selected propensity score models

	YCS	NPD		*LSYPE*	
			spec 5+ years	*spec 4 years*	*spec 2 years*
KS2 point score		0.021***	0.014**	0.019***	0.012*
		(0.002)	(0.005)	(0.007)	(0.006)
girl	-0.035	-0.010**	0.084***	0.196***	0.090**
	(0.035)	(0.004)	(0.024)	(0.050)	(0.044)
ethnic	0.064	0.106***	-0.039***	0.040	0.109*
	(0.056)	(0.006)	(0.007)	(0.054)	(0.049)
parents degree			-0.030	0.222**	0.047
			(0.073)	(0.088)	(0.081)
parents HE below degree			0.026	0.169*	0.115
			(0.069)	(0.087)	(0.077)
parents A-level			0.006	0.177**	0.097
			(0.066)	(0.082)	(0.073)
parents GCSE a-c			0.024	0.144**	-0.007
			(0.050)	(0.065)	(0.057)
parents no. qual.	0.193***				
	(0.072)				
parents professional	0.096**				
	(0.042)				
parents employed			-0.035	0.136**	0.026
			(0.116)	(0.063)	(0.124)
parents unemployed	0.267***		-0.165		-0.143
	(0.098)		(0.117)		(0.126)
parents on income support			-0.135*	-0.135	-0.063
			(0.064)	(0.085)	(0.077)
parent missing	0.144***			0.003	-0.110*
	(0.046)			(0.060)	(0.054)
smokes cigarettes			-0.009	0.042	-0.102
			(0.065)	(0.082)	(0.076)
uses school sports facilities			-0.019*	-0.025	-0.050**
			(0.017)	(0.022)	(0.019)
uses pc at home			0.134**	0.203**	0.023
			(0.058)	(0.078)	(0.065)
School test score$_{(t-5)}$	2.499***				
	(0.435)				
eligibility free school meals		-0.016***			
		(0.008)			
cons	-0.930***	-0.010	-0.110	-1.211***	-0.220
	(0.127)	(0.041)	(0.200)	(0.218)	(0.226)
N	5,122	264,286	3,434	3,434	
log-likelihood	-3508.106	-176314.59	-2943.747	-2341.202	-2341.2028
Chi-square	62.07	13737.03	89.93	78.01	78.01

In the NPD model we also include a number of school level variable that are measured in the pre-policy period including school type, pupil-teacher ratio, the proportion of pupils eligible for free school meals and school size. All of this variables are highly statistically significant.

Table 4: Policy-off policy-on analysis using the YCS

	Gcsescore	*Gcsebin*	*Gcsebin10*	St.Bias
unmatched	2.761***	0.012	0.085***	5.201
	(0.443)	(0.013)	(0.009)	(4.202)
NN(1)	1.732***	-0.013	0.072***	1.601
	(0.593)	(0.018)	(0.013)	(1.438)
NN(1) with counfounder	1.385**	-0.030	0.063***	
	(0.691)	(0.021)	(0.015)	
$\text{Kernel}_{(0.1)}$	1.988***	-0.012	0.078***	0.937
	(0.461)	(0.012)	(0.001)	(0.571)
post-matching regression				
	1.866***	-0.006	0.080***	
	(0.453)	(0.013)	(0.009)	

Bounds M-H statistics

	$e^{\gamma} = 1$	$e^{\gamma} = 1.25$	$e^{\gamma} = 1.50$	$e^{\gamma} = 1.75$	$e^{\gamma} = 2$
Gcsescore	2.65***	1.0-4.3***	-0.40-5.6	-1.55-6.7	-2.55-7.65
Gcsebin10	4.0***	2.60-5.45***	1.47-6.66*	0.53-7.72	0.11-8.66

Significance levels : $*$: 10% $**$: 5% $* * *$: 1%

Balancing Property and Common Support satisfied.Analytical s.e. for NN, Bootstrap (500 repetitions) for Kernel

Confounder follows same KS2 test score distribution in the NPD sample.

Standardized Bias= = $\frac{100(\overline{x}_{non-sp}-\overline{x}_{spec})}{\sqrt{(s^2_{non-sp}+s^2_{spec})/2}}$ where: $\overline{x}_{non-sp}$ = mean of the non-specialist schools group

$\overline{x}_{spec}$ = mean of the specialist school group, s^2_{non-sp} = variance of the non-specialist schools group

s^2_{spec} = variance of the specialist school group.

Significance of MH statistic bound indicates treatment effect is not sensitive to selection bias.

Note: bounds computed with the kernel method .

Table 5: The effect of the duration of specialist school status on GCSE score

	spec 5+ years		*spec 4 years*		*spec 2 years*	
	Coef.	St.Bias	Coef.	St.Bias	Coef.	St.Bias
unmatched	4.020***	7.823	5.007***	11.179	2.821***	9.748
	(0.577)	(6.240)	(0.716)	(9.423)	(0.635)	(6.017)
NN(1)	2.315***	1.178	2.821***	2.839	1.321	2.528
	(0.773)	(0.867)	(0.969)	(2.296)	(0.863)	(1.621)
NN $_{\text{with confounder}}$	2.297***		2.254***		0.776	
	(0.903)		(1.143)		(1.014)	
Kernel$_{(0.1)}$	3.257***	0.884	3.663***	1.075	1.947***	0.863
	(0.569)	(0.731)	(0.659)	(0.696)	(0.622)	(0.583)

			Bounds M-H statistics		
	$e^{\gamma} = 1$	$e^{\gamma} = 1.25$	$e^{\gamma} = 1.50$	$e^{\gamma} = 1.75$	$e^{\gamma} = 2$
Gcsescore					
5+ years	2.025***	0.460-3.591*	-0.825-4.852	-1.918-5.905	-2.864-6.804
4 years	3.00***	1.592-4.410***	0.433-5.565	-0.547-6.533	-1.405-7.372

Significance levels : $*$: 11% $**$: 5% $***$: 1%

Balancing Property and Common Support satisfied. Confounder follows KS2 test score distribution.

Analytical s.e. for NN, Bootstrap (500) for Kernel.

Significance of MH statistic bound indicates treatment effect is not sensitive to selection bias.

Note: bounds computed with the kernel method.

Table 6: The effect of the duration of specialist school status on GCSE score using DID at school level

	(1) Policy on year dummies		*(2) All controls*		*(3) Controls and school FE*	
	(2004-02)	(2003-02)	(2004-02)	(2003-02)	(2004-02)	(2003-02)
Year Policy-on						
all	0.266***	0.278***	0.358***	0.404***	0.330***	0.351***
	(0.096)	(0.095)	(0.070)	(0.075)	(0.067)	(0.071)
N	1024590		990494		990494	
F-test	755.856		23994.098		61740.469	
2002	0.410***	0.154	0.602***	0.420***	0.563***	0.346***
	(0.147)	(0.145)	(0.107)	(0.113)	(0.103)	(0.108)
2001	-0.081	0.201	-0.020	-0.059	0.060	-0.094
	(0.196)	(0.194)	(0.145)	(0.157)	(0.138)	(0.148)
2000	0.769***	1.371***	0.852***	1.143***	0.832***	1.106***
	(0.216)	(0.214)	(0.160)	(0.168)	(0.156)	(0.163)
1999	0.933***	1.421***	1.060***	1.322***	0.923***	1.253***
	(0.278)	(0.277)	(0.204)	(0.216)	(0.198)	(0.208)
1998	0.705***	0.336	0.884***	0.342*	0.984***	0.422**
	(0.255)	(0.252)	(0.195)	(0.208)	(0.188)	(0.199)
1997	-0.394	0.395	-0.072	0.588***	-0.193	0.442**
	(0.252)	(0.249)	(0.186)	(0.194)	(0.178)	(0.187)
N	1024590		990494		990494	
F-test	217.846		13776.753		20603.998	

Notes: robust s.e. in parenthesis. Model 2-3: control for gender, prior attainment at age 11, year dummies pre-policy school characteristics: number of pupils, pupils teacher ratio, (%) of pupils with special education needs, (%) of pupils eligible for free school meals, (%) non white pupils, (%) pupil from other ethnic groups, average performance of school, all boys school, all girls school, religious and grammar school.

Table 7: Difference-in-differences using pupil-level matching

	Model 1: Gcsescore		*Model 2: KS4-KS2*	
	(2002-04)	(2002-03)	(2002-04)	(2002-03)
$T_{t'} - C_{t'}$				
unmatched	2.578***	3.280***	0.037***	0.056***
	(0.088)	(0.102)	(0.003)	(0.004)
ATT	1.178***	1.279***	0.021	0.046
	(0.139)	(0.162)	(0.028)	(0.035)
$T_t - C_t$				
unmatched	2.629***	3.422***	0.039***	0.080***
	(0.091)	(0.103)	(0.003)	(0.004)
ATT	1.549***	1.452***	0.055*	0.057
	(0.143)	(0.164)	(0.030)	(0.046)
DID	0.050***	0.141***	0.001***	0.024***
	(0.008)	(0.010)	(0.000)	(0.000)
DID ATT	0.370**	0.172	0.033	0.010
	(0.141)	(0.163)	(0.029)	(0.041)

Notes: analytical s.e. in parenthesis. Model 1: control for gender, prior attainment at age 11, ethnicity, pre-policy school, characteristics. Model 2 excludes prior attainment and the coefficients are in standard deviation points.

Table 8: The relative importance of funding and school specialisation on GCSE score

	English (s.e.)	*Technology* (s.e.)	*Languages* (s.e.)
Panel A: pupils in schools specialising in subject m vs pupils in schools specialising in subject n			
unmatched	0.383***	0.237***	0.180***
	(0.045)	(0.047)	(0.045)
NN	0.181***	0.181***	0.144***
	(0.034)	(0.037)	(0.029)
NN $_{\text{with confounder}}$	0.128***	0.135***	0.083**
	(0.042)	(0.051)	(0.037)
Kernel$_{(0.1)}$	0.180***	0.161***	0.101**
	(0.0453)	(0.058)	(0.042)
Panel B: pupils in schools specialising in subject m vs pupils non-specialist schools taking subject m			
unmatched	0.432***	0.460***	0.334***
	(0.039)	(0.033)	(0.049)
NN	0.226***	0.231***	0.180***
	(0.032)	(0.034)	(0.035)
NN $_{\text{with confounder}}$	0.215***	0.214***	0.101**
	(0.037)	(0.041)	(0.045)
Kernel$_{(0.1)}$	0.288***	0.318***	0.146
	(0.050)	(0.051)	(0.042)

Significance levels : $*$: 10% $**$: 5% $***$: 1%

Balancing Property and Common Support satisfied

Analytical s.e. for NN, Bootstrap (500) for Kernel

Confounder generated using KS2 test score distribution.

Sample sizes: English Literature Panel A 48,804 Panel B 166,321;

Technology Panel A 89,062 Panel B 195,219;

Foreign Languages Panel A 55,762 Panel B 155,138.