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

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

February 16, 2010

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 matching methods 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 and test score gain. The effect peaks after 4 years, at which point the additional funding ceases. A specialisation effect occurs yielding relatively large improvements in test scores in particular subjects.

Keywords: School Quality, Subject Specialisation, Matching models.
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 decentralised approach to educational provision (Hoxby, 1996). In the US the 'No

[*] The authors would like to thank Nuffield Foundation for the financial support to this work; Ian Walker, Colin Green and Geraint Johnes for helpful feedbacks. We would also thank the participants to the European Economic Association Annual conference 2009, to the Scottish Economic Association Annual conference 2009, to the Institute of Education Economics Seminar and to the Lancaster University Economics seminar for useful comments.

[†] 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.

Child Left Behind Act', introduced by the previous 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 new 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 receive 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 funding to 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 approximtely 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 (see the Department for Children, Schools and Families website for more details: www.standards.dcsf.gov.uk/specialistschools/). 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 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

[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. They then sit national recognised tests, the General Certificate of Secondary Education (GCSE), in each subject.

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 multiple sources of bias that often arise 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 therefore 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 using matching methods and post-matching regression models.[3] This approach enables us to deal with various forms of selection bias as well as possible contagion effects between specialist and non-specialist schools. 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. In addition, 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.

However, there are likely to be several sources of bias, 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 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.[4] Moreover, the differential in GCSE test scores appears to be greatest from around 2002 onwards, which coincides with the period over which our analysis is conducted. Figure 2 digs a little deeper into the difference in test score performance of specialist schools. Panel A of Figure 2 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. 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

[3] 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.

[4] 5+ A*-C grades is an important policy measure and, when combined with suitable grades at A level, permits entry to HE.

school into the specialist schools initiative.

Second, and closely related to the first source of bias, there is likely to be 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 2 provides some evidence of cream-skimming based on observable characteristics that are correlated with exam performance. Panel B of Figure 1 shows that pupils from the poorest social backgrounds are less likely to attend specialist schools.[5] In Panel C we show the distribution of ethnic minorities, insofar as 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.

Third, there is what we call *dynamic selection bias*, which arises from the first source of bias but is cumulative in its effect. ‘Older’ specialist schools have 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. Figure 3 plots the mean proportion of pupils obtaining 5 or more GCSEs grades A*-C by the year in which schools acquired specialist status (i.e. year 0). Schools that acquired specialist status earlier tend to have better test score performance on entry to the initiative and superior growth (compare the slopes), providing evidence of a dynamic selection effect.

Failure to allow for these three sources of bias would lead to an upward bias on the estimated effect of the specialist schools policy on test score outcomes.

A fourth source of bias arises from a possible *contagion effect* insofar as specialist schools are required to help non-specialist schools in the same educational district. This may consequently raise test scores in non-specialist schools hence biasing down the estimated effect of the specialist schools policy.

The data used to estimate the matching models were obtained from several sources: the National Pupil Database (NPD), 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 3 how we tackle the four 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. The effect is not large insofar as it has raised GCSE scores by between 2-2.5 GCSE points, and is approximately 50% lower than the ‘naive’ pre-matching estimates. The policy has also had the effect of increasing the probability of obtaining 5+ GCSE grades A*-C by about 2-3 percentage points; the effect on the probability of obtaining 10+ GCSE grades A*-C is larger at around 7-8 percentage points. These results imply that the policy has had larger effects for more able students. Moreover, the difference-in-differences model, which allows us to control for individual unobserved characteristics as well as some of the biases outlined above, suggests that the policy led to an improvement in test scores between the ages of 14 and 16 of around 0.07 of a standard deviation.

[5] Specifically, specialist schools have the lowest number of pupils in the 5th quintile of the proportion of pupils eligibility for free school meals.

The duration of specialisation also matters, in that the peak of the policy effect is reached after 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 simple funding effect. 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.

These findings are robust insofar as they ‘pass’ several tests for the presence of hidden bias - a substantial reduction in the bias due to selection on observables is found, and our estimates are robust to a test for unconfoundedness.

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

2 Econometric Approach

2.1 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 estimate the propensity score (PS) using a discrete response model of attendance at a specialist school. This approach solves the so-called ‘dimensionality problem’, insofar as we can match the treated with the control group on the basis of a mono-dimensional variable instead of the multi-dimensional vector, $\mathbf{x}$.

One assumption of the matching method is the *common support* or overlap condition. Intuitively, to estimate the counterfactual for a given pupil we need to have someone similar to that pupil in the counterfactual state. If we do not, we have a failure of the common support condition because the density in one sample is zero whereas there is positive density in the other. This condition ensures that pupils with the same $\mathbf{x}$ values have a positive probability of attending a specialist school. The choice of the covariates to be included in this first stage is an issue. Heckman et al. (1997) show that omitting important variables can increase the bias in the resulting estimation. But, in general, only variables that simultaneously influence the decision to attend a specialist school and the test score outcome, which in turn are unaffected by attendance, should be included in the model. Bryson et al. (2002) also recommend against over-parameterized models because including extraneous variables in the attendance model will reduce the likelihood of finding a common support. Others, such as Rosenbaum and Rubin (1987), Dehejia and Wahba (1999) and DiPrete and Gangl (2004) emphasize that the crucial issue is to ensure that the balancing condition is satisfied, because it reduces the influence of confounding variables. Thus, we match the potentially confounding covariates of the pupils assigned to specialist schools with pupils that attended non-specialist schools. Practically, the sample is stratified in several blocks and we carry out a series of two-sample t-tests of the equality of the means on the propensity scores and equality of the means on all covariates, between treated and untreated pupils.

We include in our selection model only those variables that satisfy both the balancing property and the common support condition. Our approach is to create reliable comparison groups and reduce the bias on observables, which makes the matching estimator more efficient.

A second, and key, assumption in the matching method is the *conditional independence assumption*, which implies that selection into treatment is solely based on observable characteristics.[6] As suggested above, there may 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. As noted by Heckman et al. (1998), matching only eliminates bias averaged over specific intervals of the propensity score. If the assumption of ignorability (i.e. no hidden bias) fails, the treatment is endogenous and the matching estimates will be biased. Several tests have been developed to assess whether hidden bias is a problem in cross-sectional models.

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. The first was developed by Imbens (2004) who suggested that there are indirect ways of assessing the CIA, based on the estimation of a 'pseudo' confounding factor that should, if the CIA holds, have zero effect. The second test was proposed by Rosenbaum (1987) and involves computing one 'sensitivity' parameter, representing the association between treatment and a confounding factor, and derives bounds for significance levels and confidence intervals.

For the first test we adopt the method proposed by Ichino et al. (2008). It is based on the prediction of a confounding factor, A, by simulating its distribution for each treated and control unit. Then, estimates of the 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 ('high scores'$= 1$ and 'low scores'$= 0$) and T is the pupil's treatment status. In this way, we can define the probability of $A = 1$ in

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

each of the four groups identified by the treatment and the outcome.[7] In our analysis, we consider the following parameters: $p_{11} = 0.65$, $p_{10} = 0.55$, $p_{01} = 0.45$, $p_{00} = 0.35$. Thus, we assume that the potential confounder variable has both a positive effect on the test scores of pupils in non-specialist schools ($p_{01} - p_{00} > 0$), and on selection into specialist schools ($p_{1.} = 0.62 - p_{0.} = 0.40 > 0$). For example, if we interpret A as unobserved pupil ability, $p_{11} = 0.65$ indicates the proportion of high ability pupils among those in specialist schools who get high test scores.

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.[8] The ATT is re-estimated 1000 times, and the value presented in our Tables is an average over the distribution of A.

Our assumptions on the probabilities p_{ij} imply that pupils in specialist schools are more able than those in non-specialist schools. This is illustrated by the estimation of two further effects provided by the test. One effect is the 'outcome effect', which measures the effect of unobserved ability on the probability of having high test scores for pupils in non-specialist schools, controlling for observables x.[9] This effect in our estimates is always positive (around $1.541 > 1$). The second is the 'selection effect', which measures the effect of unobserved ability on the probability of attending a specialist school, controlling for observables x.[10] This effect in our estimates is also positive and much higher than the previous one ($2.390 > 1$). The effect of the confounder on the probability of attending a specialist school for 'good' pupils is higher than the effect on the probability of getting high test scores in a non-specialist school. Given these assumptions, if the confounded estimates are still significant, but with the same sign and (similar) magnitude when compared to the 'true' estimates, we can be confident of the robustness of our results.

The second method proposed by Rosenbaum (1987) 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.[11] The test can be interpreted as the difference in the relative odds of attending a specialist school for two pupils that appear

[7] 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).

[8] We omit the results with different methods because they are very similar.

[9] Formally, it is the average of the estimated odds ratio of A, from the logit model $P(Y = 1|T = 0, A, X)$ in every iteration.

[10] Formally, it is the average of the estimated odds ratio of A, from the logit model $P(T = 1|A, X)$ in every iteration.

[11] 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 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).

similar in terms of observable covariates, **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.

2.2 Matching estimators

Given these two assumptions, the matching method allows us to estimate the average treatment effect of 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:

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

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 is the outcome and w_i is the re-weighting that reconstructs the outcome distribution for the treated sample. A number of well-known matching estimators exist but they differ in how they construct the weights, W_{ij}.

In this paper we present the estimates from two matching algorithms. The first is the nearest neighbor matching (NN) estimator. Here a pupil from the control group is chosen as a matched partner for a treated pupil who is closest in terms of the propensity score. In Equation 1 a unity weight is placed on the nearest observation; zero for all other observations, $w_{ij} \in [1, 0]$. A limitation of all NN estimators is that fewer observations from the control group are used to construct the counterfactual for each treated pupil. We therefore also use kernel matching, where every treated pupil is matched with a weighted average of all control pupils with weights that are inversely proportional to the distance between treated and control pupils. Thus, the variance is lower because more information is used, but a drawback is the possible use of observations which are 'poor' matches. The choice of the kernel function is relatively unimportant (see DiNardo and Tobias, 2001), and in our analysis we use the Epanechnikov function. Caliendo and Kopeinig (2005) argue that the choice of 'bandwidth' is more important - high values of the bandwidth yield a smoother estimated density function, with a better fit between the estimated and true underlying density function. However, this can smooth away underlying features and bias the estimates, therefore we use an intermediate value where the bandwidth is 0.1.[12] In general, Smith (2000) argues that, asymptotically, all matching estimators should give the same results because for increasing sample size they all get closer to comparing only exact matches.

Finally, there is an issue in the literature as to whether standard errors should be bootstrapped. Abadie and Imbens (2008) show that bootstrapping is not a valid method to make inference when employing the nearest neighbour matching estimator with a fixed number

[12] The bandwidth is defined in terms of distance of each individual from the control group, and values of 0.02 and 0.2 are usually considered low and high, respectively.

of matches. We therefore provide analytical standard errors for estimates from the nearest neighbour method. However, with kernel-based matching methods, like those used by Heckman et al. (1998), the number of matches increases with the sample size and these estimators are asymptotically linear. The standard bootstrap in this case provides valid inference, and so for these models we report bootstrapped standard errors.

Where panel data are available an alternative method can be adopted, that is, the difference-in-differences (DID) matching estimator (Blundell et al., 2004, Smith and Todd, 2005 and Machin et al., 2004). This approach requires longitudinal data and 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- to a post-attendance period in the same way treatments would have done had they not been treated (Blundell and Costa Dias, 2002). DID 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 (Heckman et al., 1998). Thus the DID matching estimator has the advantage of eliminating unobserved time-invariant differences between treated and untreated observations.

The DID matching estimator for the ATT can be obtained by rewriting Equation 1 as

$$\tau_{ATT}^{DID} = \sum_{i \in T} [\,(Y_{it_1} - Y_{it_0}) - \sum_{j \in C} W_{ij}(Y_{jt_1} - Y_{jt_0}\,)\,]\,w_i. \tag{2}$$

3 Selection bias, contagion and descriptive statistics

We use three different datasets in our analysis, each of which have different strengths with respect to the mitigation of the biases outlined in the Introduction. Table 1 shows various measures of test score outcome for each dataset. The first is the total GCSE score (*GCSEscore*), the number of points achieved in all GCSE subjects where a grade A*=8 and a fail=0.[13] 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.

The NPD refers to the population of pupils attending maintained, state funded, schools in England who were in their final year of compulsory education in 2003. 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 3 (for 13/14 year olds) and Key Stage 4 (for 15/16 year olds). 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. In one analysis we use this data to investigate the effect of the specialist schools policy on Key Stage 4 outcomes.[14]

[13] Pass grades are from A* to G. Pupils can also receive an unclassified grade which is treated as equal to a fail.

[14] The Key Stage 3 result for each pupil is the total test score for English, Maths and Science, whereas the Key Stage 4 result for each pupil is the average points scored across all subjects in the GCSE examinations taken at the end of compulsory education.

We restrict the control group to pupils in those schools that will become specialist between 2003 and 2005. This is because these schools are most like those that have already acquired specialist status, many of which did so from 2000 onwards (see Figure 3). Also, schools that acquire specialist status after 2005 can be thought of as being at the end of a specialist school 'queue', and are in some senses 'poorer' schools. Since there is a large number of observations, we also stratify the data on the basis of the percentage of specialist schools in an educational district. As this percentage rises we argue that the level of pupil selection bias will fall because pupils have less choice regarding the type of school to attend. Moreover, the bias generated by the contagion effect must also fall because there are fewer non-specialist schools that could be affected by the specialist schools in the educational district. Table 1 shows how these restrictions reduce our NPD sample from approximately half a million pupils to around 130,000 observations when we restrict the analysis to pupils in districts with 60% of specialist schools, for instance.[15] Table 1 shows how the sample size falls as we increase the proportion of specialist schools and non-specialist schools in the educational district.

The problem of dynamic selection bias may nevertheless remain and to deal with this we estimate post-matching regression models. Specifically, we compute the real growth rate in the proportion of pupils in each school obtaining 5 or more GCSEs grades A*-C from 1992 onwards. Thus, if a school is non-specialist in 2003, for instance, we will have its growth rate from 1992 up to 2003. More generally, and in terms of Figure 3, we restrict our measure of the growth rate for specialist schools to the right of the vertical dashed line in each year. For non-specialist schools the growth rate is computed for the left side of the vertical dashed line.[16] 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. However, 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.

Another way of tackling the problem of pupil selection bias is to estimate a difference-in-differences matching model. We impose the same restrictions on the data as we do for the cross-sectional matching models and identify the same individuals in two different time periods according to their test score at Key Stage 3 and 4. We only consider pupils with the same pre-treatment characteristics (we also control for prior attainment using Key Stage 2 test scores). Then we evaluate the effect of specialist schools on test score gain between Key

Pupils prepare for the General Certificate of Secondary Education (GCSE) examinations in typically no more than 10 subjects in their final two years of compulsory schooling between the ages of 14 and 16. 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.

[15] In England, there are 366 Local Education Authorities.

[16] 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 (D), the estimated propensity score and its deviation from the mean interacted with D. The coefficient of D consistently estimates ATE. In our specific case, we add in the post-matching regression the growth rate variable and its deviation from the mean interacted with D. In this way we want to get the treatment effect corrected for the fact that the control group changes over time.

Stage 3 and Key Stage 4.[17]

The YCS is a major programme of longitudinal research designed to monitor the behaviour and decisions of representative samples of young people 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. However, the primary advantage of the YCS is that we can link schools together to investigate how the test scores of different cohorts of pupils change as a school moves from non-specialist to specialist status. This allows us to construct a 'policy-off' versus 'policy-on' test of the effect of the specialist schools policy on test scores. Specifically, we link schools in YCS11 and 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 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 (policy off), which then becomes specialist during their period of secondary schooling (policy on). Dynamic selection bias may nevertheless remain.

The LSYPE is a panel study of young people started in 2004, when its sample of young people 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 the LSYPE is that we can control for the duration of specialist school status, to investigate whether the effect, if any, of the specialist schools policy declines over time. We can track schools that have been specialised for two, four or more than four years. An additional advantage of these data are that we can exploit a rich set of family covariates, such as parental education and employment, and pupil behaviour, for instance, bullying, misbehavior in school and smoking, which we include in the propensity score models in an attempt to reduce the impact of hidden bias. The real value of this analysis is in identifying the duration of specialist school effect; the relatively small number of observations does not allow us to address the various sources of bias. In the LSYPE the sample is smaller than the NPD and YCS, and 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.

4 Findings

4.1 *Estimation of the propensity score models*

The variables included in the propensity score models are those factors that affect a pupil's choice of school, and also in some models the schools' decision to apply for specialist status. However, recall that the variables must pass the balancing test and overlap condition. We do not report tables of estimates from the propensity score models, however we briefly describe

[17] Note that we only report the results of this analysis for those pupils where all schools in the district are specialist (the treatment group) which are compared to pupils in those districts that have no specialist schools (the control group). This ensures that the bias caused by contagion is minimised. We also experiment by varying the proportions of specialist and non-specialist schools in a district.

the key findings.[18]

The model constructed from the NPD includes prior attainment, that is, the standardized test score at Key Stage 2, taken at age 11, 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). This variable is highly statistically significant with a marginal effect of 0.047 (s.e.=0.004) and suggests that more able primary school pupils sort into, or are selected by, selective schools. We also include dummies for ethnicity, which has a large positive effect on attendance at a specialist school (the marginal effect is 0.180 and the s.e. is 0.017), and gender, which is statistically insignificant. The finding on the ethnicity variable may reflect the heterogeneity of ethnic groups in the UK, where Indians and Chinese pupils have relatively high test scores when compared to whites, whereas Pakistani and Bangladeshi groups have relatively lower scores.

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. This variable measures the proportion of pupils in the school five years earlier who obtained five or more A*-C grades in the GCSE examinations. Most variables are statistically significant and we note the very strong marginal effect of lagged school performance on school choice.

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 NPD and YCS. The estimates work in the expected direction, but only a subset are statistically significant. For isntance, pupils who truant and those whose parents are on income support are less likely to attend a specialist school. In contrast, those pupils with higher prior attainment or who use a personal computer at home are more likely to attend a specialist school.

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 in terms of 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).[19] In

[18] These results are available on request from the corresponding author.

[19] 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 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. Note that for the stratification method we only report the number of blocks.

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

4.2.1 *The impact of the specialist schools policy using the NPD*

Looking at Table 2, in Panel A we report the estimates for *Gcsescore* only, using the full NPD which is compared to the estimates from models with different samples of this population of pupils. Recall that the population is restricted in two ways. First, we remove pupils in schools that have not acquired specialist school status before 2005 since they are likely to be very different from specialist schools that have been in the initiative for some time. Second, we restrict the sample by looking at pupils in districts where the percentage of specialist schools is 60%, 80% and then 100% of schools, which is compared to a control comprising pupils from districts with identical percentages of non-specialist schools.

With the full sample the policy effect on test scores is around 3.2 points, which sharply decreases to 1.8 after matching. This estimate is likely to be biased by pupil and school selection bias and by the bias caused by contagion between specialist and non-specialist schools. We observe, however, that the policy effect increases as we restrict the sample - the largest impact of the policy is observed in districts which contain only one type of schools. Prior to matching the effect is around 6 GCSE points which after matching falls to 3.2 points. We argue that this model not only reduces the bias arising from a contagion effect but also minimises pupil selection bias. These estimates do not, however, control for dynamic selection bias. The final row of Panel A therefore reports the post-matching regression which includes the growth rate in test scores. These estimates are very similar to the NN estimates and lower than the kernel estimates. Looking at the last column of Panel A, where the matching estimates are not confounded by pupil self-selection and school contagion bias, the post-matching analysis can better isolate the effect of the dynamic selection bias. The latter clearly inflates the true effect; the post-matching regression estimates suggest that the specialist schools policy increases test scores by 2-2.5 GCSE points.

The impact of the specialist schools policy is stronger at the upper end of the test score distribution, increasing the probability of a pupil achieving 5 or more GCSE grades A*-C (*Gcsebin*) by between 6-10 percentage points, depending on the estimator. However, this effect is still substantially lower than the naive, pre-matching, estimates (see the last column of Panel B). Controlling for dynamic selection bias, leads to the estimated impact decreasing by 40% with NN and by around 65% with kernel - the specialist schools policy increased the probability of obtaining 5+ GCSE grades A*-C by around 3-4 percentage points.

In a further sensitivity analysis we estimate a post-matching regression with real expenditure per pupil as a covariate (see Table 8 in the appendix), and it is clear that there is little effect on our estimates from the matching model.

Furthermore, the standardized biased associated with the matching estimators (see Panel C) clearly drops and our estimates are also largely unaffected by the inclusion of a confounding variable (row three in Panel A and B), which as expected reduces their magnitude. This means that unobserved factors, such as ability, upward biases our matching estimates, but its effect is small. This is further confirmed looking at Table 7 (Panel A) in the appendix, where we report the results of the Rosembaum test. Our estimates are robust for all outcomes up to a degree of departure from the situation of no bias to one equal to 1.75. This means that allowing for an (unobserved) confounding factor which makes pupils with the same x differ in their odds of attending a specialist school by 75%, thus we are confident that we are not overestimating the true treatment effect.

4.2.2 *A 'policy-on' versus 'policy-off' analysis*

Table 3, 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. 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. Note that in this analysis we mitigate the bias arising from school selection bias, and also note that these estimates are broadly in line with those from the estimates obtained for the NPD. Moreover, these estimates are also very similar to those obtained for the NPD.

Interestingly, there is no statistically significant difference in the proportion of pupils obtaining 5 or more GCSEs graded A*-C (*Gcsebin*), which is clearly inconsistent with the results obtained above. 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 gives results consistent with those in the NPD, since it is lowering the magnitude of the estimates but has little effect on their pattern. This is also confirmed by the Rosembaum test (see Table 7, Panel B, in the appendix) in the cases up to 50% hidden bias.

4.2.3 *The effect of the duration of specialist school status*

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 4 shows that this is, in fact, what we observe. Compare the results for schools that have been specialist for 4 and more than 4 years with those that have been specialist for only 2 years.

Prior to matching, pupils in schools that have been specialist for longest obtain 4.1 GCSE points more than their counterparts in non-specialist schools. The equivalent figure for schools that have been specialist for only 2 years is only 2.7 GCSE points. After matching, these effects fall to 1.8 and 0.5 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 have been specialist for 5 or more years. Recall, that funding lasts for up to 4 years. What we observe is both pre- and post-matching the estimates for the schools that been specialist for 4 years are larger by almost 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 and more than 4 years are consistent with those from the previous analysis.

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 7 (Panel C) in the appendix shows that our estimates are not sensitive to hidden bias but only up to a level of 25%. Nevertheless, the estimates from this analysis are broadly consistent with those from the NPD and the YCS.

4.2.4 *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 consider a restricted sample which includes pupils in schools within educational districts with only specialist schools, and pupils in schools within educational districts with only non-specialist schools.[20] We also 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.[21] We attempt to reduce the bias due to pupils self-selection into schools and the contagion effect, however, we cannot control for dynamic selection bias given the small size of the resulting sample but we assume that it impacts the funding and specialisation effects equally.

To disentangle the specialisation effect from the funding effect we compare the estimates from Panel A with those from Panel B in Table 5. Panel A compares the test score outcome 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 5, 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

[20] Compared to the previous analysis, we include all the non-specialist schools, i.e. also those that will become specialist from 2006 onward.

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

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.

4.3 The selective schools policy and test score gain

To assess the effect of the selective schools policy on test score gain we implement a difference-differences matching estimator. We use the most restricted sample - the treatment group are pupils in schools within educational districts with only specialist schools whereas the control group are pupils in schools in districts with only non-specialist schools. From this sample we generate a short panel using as time varying variables the standardized test scores at age 14 in 2001 (Key Stage 3) and at age 16 in 2003 (Key Stage 4). We argue that this analysis allows us to control for individual unobserved characteristics, pupil selection bias and the bias caused by any contagion between specialist and non-specialist schools.

We match pupils only on the basis of pre-treatment characteristics, because using post-treatment data could in principle affect our ability to identify the correct counterfactual since the matching variables may themselves be affected by the attendance status of the pupil. In order to generate a reliable comparison group we use the same observable covariates used in the cross-sectional analysis with the NPD (see Section 4.1), but we also add a variable to capture whether the pupil changed school between KS3 and KS4. We include only those variables that pass the balancing test. Once we have the matching weights in the pre-treatment status, we use them to compute the difference between treated and controls before and after treatment. In the second difference the value added from the standardized test score is used as dependent variable. We impose the condition that the overlapping support is the same in the two periods.

In Table 6 we present our results based on three matching methods; the standard errors are analytical for nearest neighbour and bootstrapped for Kernel and stratification. When we smooth the counterfactual outcomes with a kernel based method or when we use a stratification method, the estimates are of similar magnitude, positive and statistically significant. However, the results based on the nearest neighbour weighting scheme turn out to be much less precise. Looking at the unmatched estimates there is a small increase in test scores between age 14 and 16 of around 0.09 standard deviation points. After matching, the estimates fall by around 13-26% and remain statistically significant. We can conclude that the effect of specialist schools in terms of improvement in test scores from KS3 to KS4, for the same pupil, is around 0.07 standard deviation points relative to a pupil that attended a non-specialist school at KS4.

5 Conclusions

In this paper we evaluate 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 methods to an evaluation of the specialist schools policy. By adopting this approach we can explicitly confront 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 use 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 cross-sectional 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*), are analysed. In addition, we use difference-in-differences combined with matching methods to investigate the effect of the specialist schools policy on the change in test scores between the ages of 14 and 16.

Our main findings are as follows.

First, there is a positive and statistically significant impact of the specialist schools policy on test score outcomes, which is approximately 50% lower than our 'naive' estimates that do not allow for matching. Nevertheless, all three cross-sectional models suggest that the specialist school policy has increased the GCSE points score by approximately 2-2.5 GCSE points. The specialist schools policy has, however, had a more substantial effect at the upper (*Gcsebin*) and high end (*Gcsebin10*) of the test score distribution. Our estimates suggest that the specialist schools policy increases the probability of obtaining 5+ GCSE grades A*-C by 3-4 percentages points and the probability of obtaining 10+ GCSE grades A*-C by 7-8 percentage points. These results imply that the policy has had a more beneficial effect on more able students.

Second, the longer a school has been a specialist school the larger the impact on test scores, however, there is some evidence that the impact begins to fall after 4 years once the additional funding associated with the policy begins to decline. Importantly, however, the policy effect does not fall to zero.

Third, the impact of the specialist schools policy on test scores could arise from a funding effect, a specialisation effect or a peer effect. We attempt to disentangle these effects. Our findings for Gcsescore suggest that between 21-50% of the total effect in particular subjects arises from the specialisation effect. This finding is consistent with our evidence on the duration of the specialist school policy effect.

Finally, we estimated a difference-in-differences matching model to control for time-invariant unobserved differences between treated and untreated pupils. The specialist schools policy improves test scores between the ages 14 and 16 by about 0.07 of a standard deviation.

In conclusion, having controlled for various types of selection bias, and having observed that our estimates are robust to so-called 'hidden bias', we argue that the specialist schools policy has had a statistically significant causal effect on test score outcomes and test score gain.

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.

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

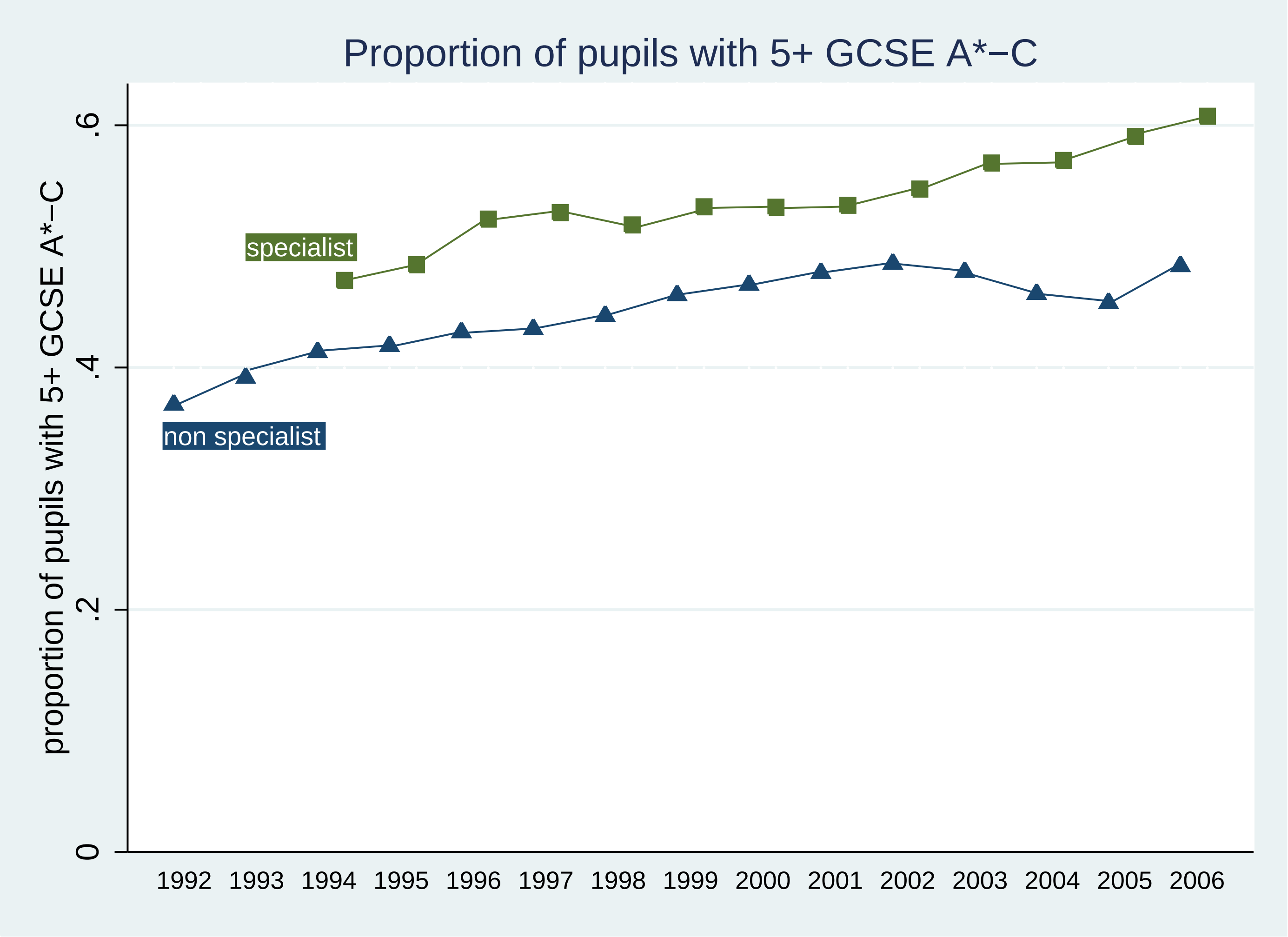
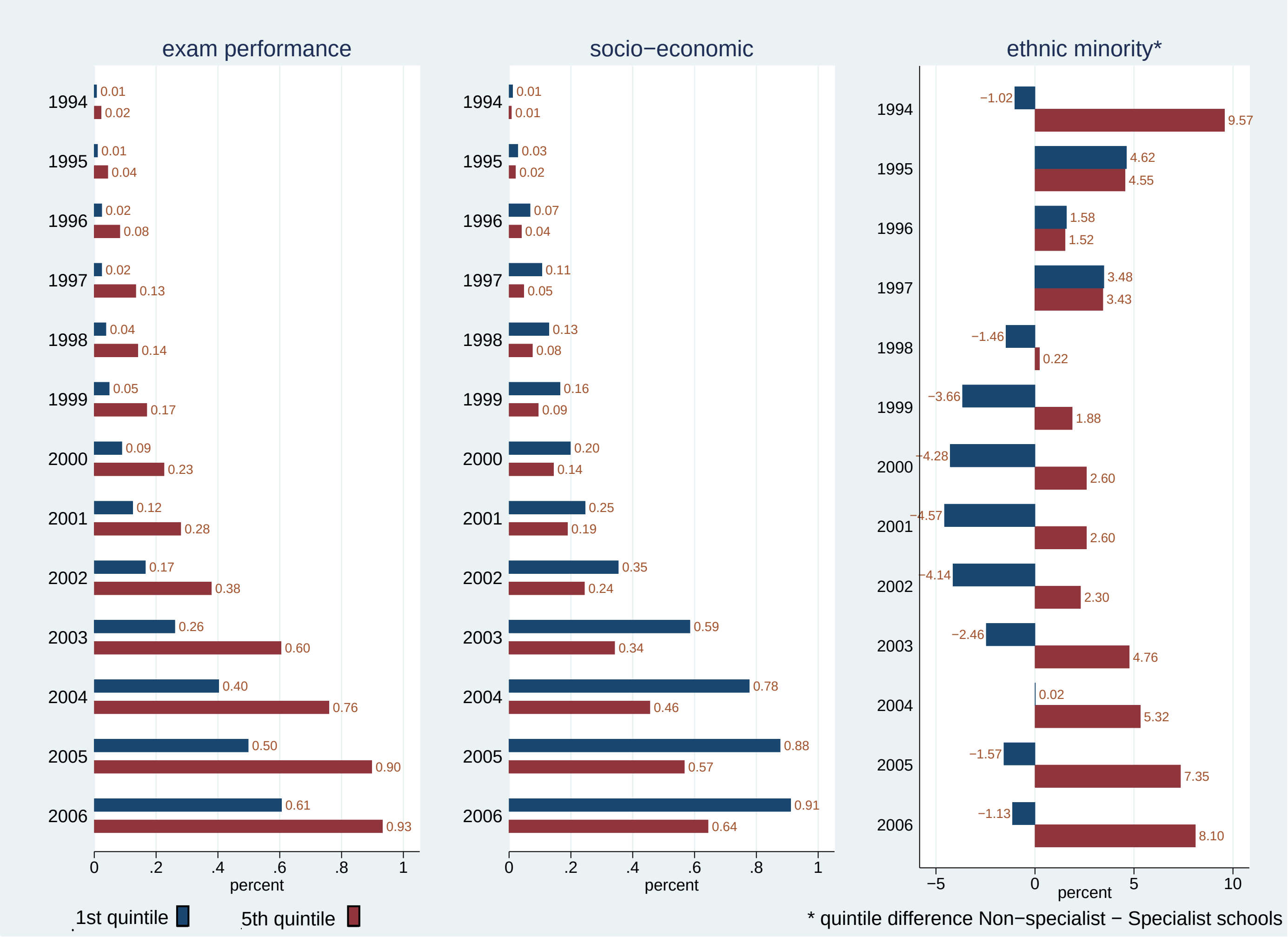


Figure 2: Specialist schools test score performance and pupil composition



* quintile difference Non−specialist − Specialist schools

Figure 3: School entry to the specialist schools initiative and test score performance

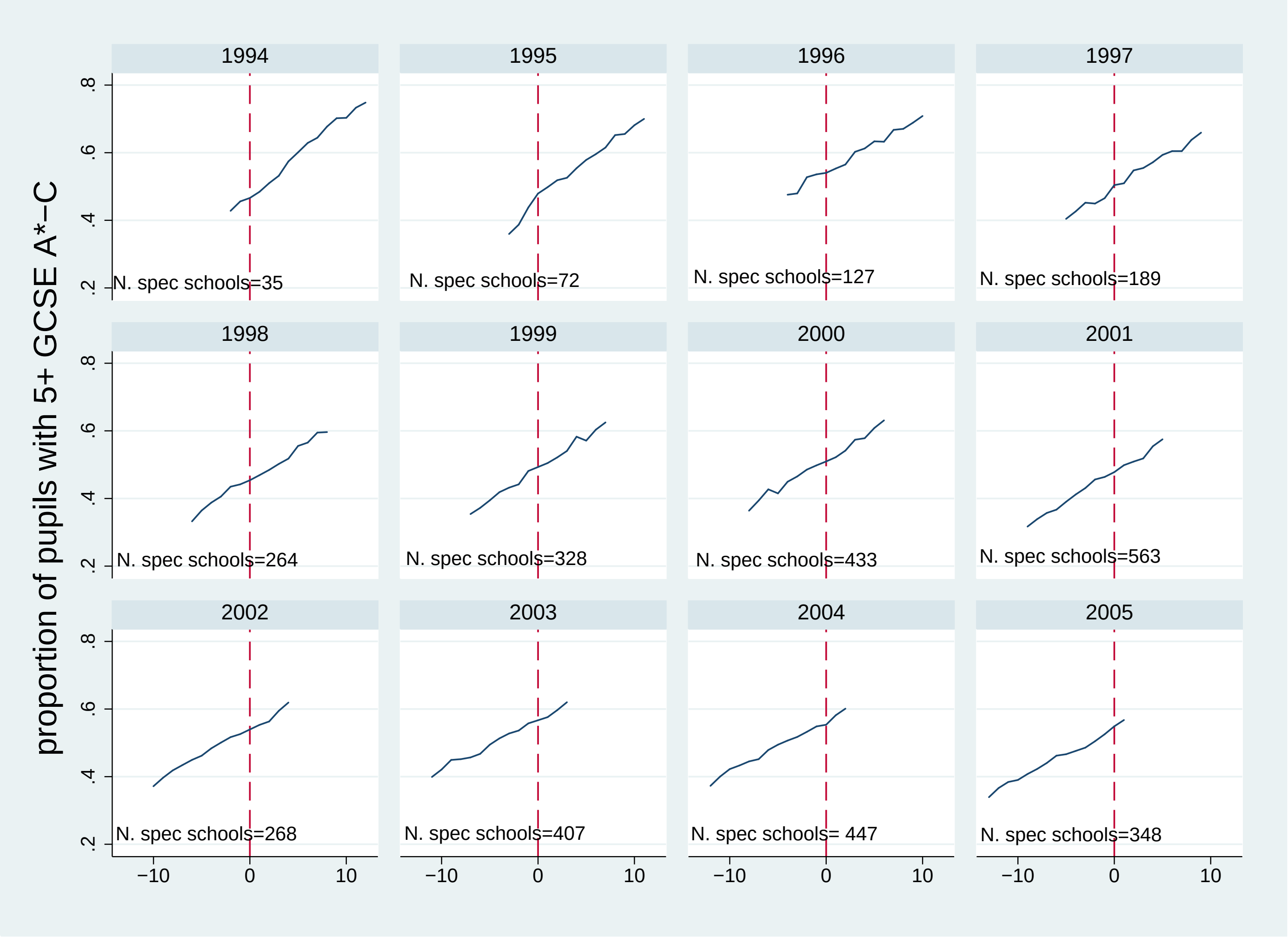


Table 1: Dependent variables

	NPD				YCS11-12	LSYPE		
	Full sample	*Ratio of Specialist*				Years of specialisation		
		Non-Specialist schools[a]						
		60-40	*80-20*	*100-0*		5+years	4 years	2 years
Gcsescore mean								
Non-spec	36.436	38.371	38.087	35.968	43.124	45.805	44.303	45.805
N	218,548	62,854	13,712	3,363	2,796	1,455	1,873	1,455
Spec	39.754	40.658	41.679	41.608	45.904	49.859	49.871	48.491
N	237,942	64,201	21,496	9,384	2,448	2,382	1,150	1,476
Gcse proportions								
Gcse A*-C <5								
Non-spec				32.05	54.19			
Spec				67.95	45.81			
N				5606	1,705			
Gcse A*-C >5								
Non-spec				21.93	1,872			
Spec				78.07	47.10			
N				7141	3,539			
Gcse A*-C <10								
Non-spec					55.73			
Spec					44.27			
N					4,536			
Gcse A*-C >10								
Non-spec					37.85			
Spec					62.15			
N					708			

[a] Within an educational district.

Table 2: The effect of the specialist schools policy on test score outcomes

	Full sample	*Ratio of Specialist Non-Specialist schools*[a]		
		60-40	*80-20*	*100-0*
Panel A: The effect on Gcsescore				
unmatched	3.176***	2.399***	3.733***	5.823***
	(0.053)	(0.100)	(0.191)	(0.350)
NN(1)	1.820***	1.833***	2.294***	3.220***
	(0.054)	(0.103)	(0.203)	(0.396)
NN with confounder	1.530***	1.605***	1.974***	2.768***
	(0.058)	(0.114)	(0.236)	(0.513)
kernel$_{(0.1)}$	2.99***	2.249**	3.598***	4.910***
	(0.497)	(0.073)	(1.086)	(0.188)
Post-matching				
	1.912***	1.958***	2.384***	2.436***
	(0.041)	(0.091)	(0.179)	(0.300)
Panel B: The effect on Gcsebin				
unmatched	0.072***	0.049***	0.083***	0.132***
	(0.001)	(0.002)	(0.005)	(0.010)
NN(1)	0.038***	0.034***	0.046***	0.057***
	(0.002)	(0.003)	(0.006)	(0.012)
NN with confounder	0.025***	0.023***	0.032***	0.043***
	(0.002)	(0.003)	(0.007)	(0.016)
kernel$_{(0.1)}$	0.067***	0.046	0.077**	0.107***
	(0.013)	(0.030)	(0.032)	(0.035)
Post-matching				
	0.040***	0.038***	0.047***	0.034***
	(0.041)	(0.091)	(.005)	(0.009)
Panel C: Standardized Bias				
before matching	6.44	4.044	16.559	15.142
	(4.440)	(2.411)	(2.930)	(9.146)
after matching	0.033	0.399	0.122	1.720
	(0.045)	(0.401)	(0.109)	(2.462)

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

Balancing Property and Common Support satisfied

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

[a] Within an educational district.

Table 3: Policy-off policy-on analysis

	Gcsescore	*Gcsebin*	*Gcsebin10*	*St.Bias*[a]
	(s.e.)	(s.e.)	(s.e.)	(s.e.)
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)

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

Balancing Property and Common Support satisfied

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

[a] 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.

Table 4: 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
	(s.e.)	(s.e.)	(s.e.)	(s.e.)	(s.e.)	(s.e.)
unmatched	4.053^{***}	7.823	5.550^{***}	11.179	2.686^{***}	9.748
	(0.609)	(6.240)	(0.699)	(9.423)	(0.674)	(6.017)
NN(1)	1.794^{**}	1.178	2.575^{**}	2.839	0.539	2.528
	(0.824)	(0.867)	(0.938)	(2.296)	(0.914)	(1.621)
NN $_{\text{with confounder}}$	1.401		2.262^{***}		0.010	
	(1.049)		(1.180)		(1.150)	
Kernel$_{(0.1)}$	1.816^{***}	0.884	2.792^{***}	1.075	0.427	0.863
	(0.564)	(0.731)	(0.634)	(0.696)	(0.621)	(0.583)

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

Balancing Property and Common Support satisfied

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

Table 5: 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*** (0.045)	0.237*** (0.047)	0.180*** (0.045)
NN	0.181*** (0.034)	0.181*** (0.037)	0.144*** (0.029)
NN with confounder	0.133*** (0.045)	0.113*** (0.054)	0.071* (0.038)
$\text{Kernel}_{(0.1)}$	0.180*** (0.0453)	0.161*** (0.058)	0.101** (0.042)
Panel B: pupils in schools specialising in subject n vs pupils non-specialist schools taking subject n			
unmatched	0.432*** (0.039)	0.460*** (0.033)	0.334*** (0.049)
NN	0.226*** (0.032)	0.231*** (0.034)	0.180*** (0.035)
NN with confounder	0.175*** (0.039)	0.185*** 0.043	0.101** (0.045)
$\text{Kernel}_{(0.1)}$	0.288*** (0.050)	0.318*** (0.051)	0.187 (0.060)

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

Balancing Property and Common Support satisfied

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

Note: the ratio Spec-Non-Spec schools within an educational district is 100:0

Table 6: The effect of the specialist school policy on test score gain

	NN(1) (s.e.)	$Kernel_{(0.1)}$ (s.e.)	$Strat_{(6)}$ (s.e.)
unmatched	0.039	0.088^{***}	0.099^{**}
	(0.034)	(0.029)	(0.047)
matched	0.027	0.075^{**}	0.073^{**}
	(0.050)	(0.036)	(0.034)

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

Balancing Property and Common Support satisfied

Analytical s.e. for NN

Bootstrap (500) for Kernel and Stratification

Note: the ratio Spec-Non-Spec schools within
an educational district is 100:0

Appendix

Table 7: The Rosenbaum sensitivity analysis

	Bounds M-H statistics				
	$e^\gamma = 1$	$e^\gamma = 1.25$	$e^\gamma = 1.50$	$e^\gamma = 1.75$	$e^\gamma = 2$
Panel A: NPD					
Gcsescore	4.027***	2.736-5.290***	1.655-6.290***	0.728-7.117***	-0.073-7.810
Gcsebin	8.249***	5.522-11.037***	3.322-13.373***	1.473-15.403*	0.040-17.210
Panel B: YCS11-YCS12					
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
Panel C: LSYPE					
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 $*$: 10% $**$: 5% $***$: 1% indicates treatment effect is not sensitive to selection bias.

Note: bounds computed with the methods NN for NPD, Kernel for YCS and LSYPE

Table 8: The effect of the specialist schools policy on test score outcomes controlling for expenditure per pupil, 1999-2003

	Full sample	*Ratio Specialist Non-Specialist schools*[a] *60-40*	 *80-20*	 *100-0*
Panel A: The effect on Gcsescore				
unmatched	3.176***	2.153***	3.750***	5.780***
	(0.053)	(0.111)	(0.211)	(0.382)
NN(1)	1.820***	1.655***	2.651***	3.611***
	(0.054)	(0.115)	(0.225)	(0.433)
Post-matching				
	1.272***	1.790***	2.842***	2.042***
	(0.046)	(0.101)	(0.205)	(0.337)
Panel B: The effect on Gcsebin				
unmatched	0.072***	0.043***	0.076***	0.124***
	(0.001)	(0.003)	(0.006)	(0.011)
NN(1)	0.038***	0.029***	0.049***	0.058***
	(0.002)	(0.003)	(0.007)	(0.013)
Post-matching				
	0.025***	0.037***	0.058***	0.021***
	(0.001)	(0.003)	(0.006)	(0.010)

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

Balancing Property and Common Support satisfied

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

[a] Within an educational district.