

The literacy hour ☆

Stephen Machin^a, Sandra McNally^{b,*}

^a *Department of Economics, University College London, Centre for the Economics of Education and Centre for Economic Performance, London School of Economics, United Kingdom*

^b *Centre for the Economics of Education and Centre for Economic Performance, London School of Economics, United Kingdom*

Received 7 February 2007; received in revised form 16 November 2007; accepted 19 November 2007

Available online 11 February 2008

Abstract

In countries like the UK and the US, a significant and challenging problem facing educators is how to ensure that future generations do not suffer from the severe basic skills problems that currently hinder a sizeable group of adults. We look at a primary school programme introduced into English schools, the literacy hour, to work out whether changing the structure and content of teaching can enhance literacy skills, thus acting as a tool to alleviate problems of low literacy. Our results point to a significant impact of the literacy hour with there being around a 2–3 percentage point improvement in the reading and English skills of primary school children affected by the introduction of the policy. The literacy hour fares well when compared to other policies in terms of cost effectiveness.

These findings are of strong significance when placed into the wider education debate about what works best in schools for improving pupil performance. The evidence reported here suggests that public policy aimed at changing the content and structure of teaching can significantly raise pupil achievement. The literacy hour therefore has practical implications for raising literacy standards in many countries.

© 2007 Elsevier B.V. All rights reserved.

JEL classification: I2

Keywords: The literacy hour; English; Reading

☆ We would like to thank Matthew Young at the Department for Education and Skills for making available the codes for identifying schools that took part in the National Literacy Project. We also thank Mike Treadaway of the Fischer Trust for providing the pupil-level data. We would like to thank the Director of the National Literacy Project, John Stannard, for a very useful discussion and much information. We also thank Pat Brown and Mary Smalley for providing background to the scheme in one of the London Local Education Authorities. We are very grateful to the Editor, two anonymous referees, Josh Angrist, Sami Berlinksi, Steve Gibbons, Rick Hanushek, Caroline Hoxby, Paul Johnson, Eric Maurin, David Neumark, Patrick Puhani, Anna Vignoles, Joan Wilson, Matthew Young and participants in numerous seminars and conferences. We would also like to thank Panu Pelkonen for valuable research assistance. We are very grateful to researchers at MetaMetrics, Inc. (Developers of the Lexile and Quantile Frameworks) for conducting an analysis of the reading demand of the reading and Mathematics tests using the Lexile Framework. In particular, we would like to thank Jack Stenner and Eleanor Sanford.

* Corresponding author.

E-mail address: s.mcnally1@lse.ac.uk (S. McNally).

1. Introduction

Literacy matters. Many people in different countries fail to reach even basic levels of literacy. This severely hampers their personal circumstances and lowers national productivity. The lower tail of the adult literacy skills distribution is particularly pronounced in some developed countries, notably the UK and the US.¹ How can we ensure that future generations of adults do not suffer from such problems? One way is by trying to provide a means whereby literacy levels and, by association, overall pupil educational performance can be raised through government education policy.

More generally, there is a substantial body of research investigating how best to raise pupil achievement. The economic literature primarily addresses the impact of changing school resources such as class size, teacher quality and measures of expenditure.² Educationalists address many of the same issues, but also investigate the efficacy of what is taught in particular subjects and how this is put across in the classroom in terms of such issues as time on task, frequency of monitoring pupils' work and grouping arrangements.³

Some of the research on reading appears to have influenced public policy. For example, in the US, research findings have been used as a rationale for the reading component of the No Child Left Behind Act of 2001.⁴ As discussed by Beard (2000), this research, as well as the experience of 'Success for All' in the US (Slavin and Madden, 2000, 2003), has been influential in the development of a national policy on literacy in England (the National Literacy Strategy). Both these policies involve the implementation of a framework stating what should be taught and specifying the structure of the lesson, as well as its duration.

An important question is whether policies that affect what and how a subject is taught can affect achievement in schools. This is the subject matter of this paper. We look at the introduction of a highly structured literacy hour that was introduced to English primary schools in the 1990s. The policy was introduced to try to alleviate low levels of reading and writing skills held by children in many schools by placing a much more rigid and structured form on the curriculum for English teaching than what was previously in place.⁵ Since the policy occurred in the mid-1990s it was a very early policy intervention in the UK education system which has been characterized by a whole host of education reforms in recent years (see Machin and Vignoles, 2005). Its early introduction is important since our estimates are not therefore contaminated by other policy interventions that may have occurred at the same time. We apply a research design that exploits the fact that some children were first exposed to the literacy hour for up to two years in a period when other children were not.

To our knowledge, this is the first economic study of whether a policy of this type actually works in raising standards of literacy, and the only one in a developed country. There is a paper which studies effects of changing pedagogy through randomized experiments in the very different developing economy setting of schools in India (see Banerjee et al., 2005). Looking at whether the content of teaching (i.e. what is taught) and the pedagogy of teaching (how it is taught) matter for student achievement is clearly an important, policy relevant question.

More specifically, we use a difference-in-difference framework, sometimes coupled with statistical matching methods, to carry out an analysis of what happened to reading and English achievement before and after introduction of the literacy hour to pupils in schools affected by the policy relative to that in an unaffected control group of schools. Our data enable us to conduct placebo tests (in the pre-policy period) to ensure that pupils in schools affected by the policy were not exhibiting different trends in student performance before the policy was implemented.

¹ Many people think of literacy problems in terms of developing countries but, for example, in the UK the Moser report (DfEE, 1999) identifies one in five adults as not being functionally literate. Numbers from the International Adult Literacy Survey of 1995 show countries like the UK and US have very dense lower tails of their adult literacy skill distributions (including amongst younger adults) whereas in other countries like Sweden and Germany hardly any adults are at these low levels.

² See, for example and amongst many others, Angrist and Lavy (1999), Card and Krueger (1992) or Hanushek (1997, 2003).

³ See the reviews by Sammons (1999) or Teddlie and Reynolds (2000) on educational research on school effectiveness, or Stainthorp (1996) for a review of evidence on what children need to learn to become skilled readers, and Scheerens (2000) and Creemers (1994) on classroom instruction. On the other hand, there are hardly any papers in economics looking at how subjects are taught. Exceptions are Angrist and Lavy's (2002) paper addressing the role of Computer Aided Instruction; the paper by Glewwe et al. (2000), which investigates the impact of flip charts in Kenyan classrooms; Angrist and Lavy's (2001) paper about the effect of in-service teacher training on achievement in Jerusalem elementary schools; and Rouse and Krueger's (2004) paper about the effect of an instructional computer program designed to improve literacy.

⁴ In particular the 'Success for All — Reading First' research: see the Success for All Foundation, "Success For All-Reading First: fulfilling the requirements of the Reading First Legislation" (<http://successforall.net/current/ReadingFirst/index.htm>).

⁵ See West and Pennell (2003) for a discussion of these low levels of achievement.

We are also able to look at the impact of the policy within ‘school performance’ sub-groups, i.e. according to whether or not the national School Inspectorate classified the school as needing ‘substantial improvement’ during their first round of inspections. The latter analysis matters so as to ensure we are not simply identifying an improvement in schools placed under pressure by participating in the policy. Although the ‘literacy hour’ was introduced in a context in which all schools were under pressure to improve, it is important to show that effects are not much higher in ‘bad schools’ that might be expected to respond to any school initiative. This helps shed light on whether we should think of the ‘literacy hour’ as mainly affecting the content and teaching of literacy or whether the policy is more akin to a general school improvement policy.⁶ We also address the potential of the ‘literacy hour’ to have spillover effects for teaching and learning in other subjects.

We report results showing significant improvements in reading and English achievement for children exposed to the literacy hour. There were also quantitatively bigger gains for boys than for girls (although estimates are not always statistically different). There are no ‘pre-policy’ effects and thus our estimates are unlikely to capture something that would have happened in the absence of the ‘literacy hour’. There is also no significant differential in the effectiveness of the ‘literacy hour’ according to the performance group of the school (as classified by the School Inspectorate), showing that the ‘literacy hour’ is effective in schools under greater and lesser outside pressure to improve.

We also argue that there is good reason to expect subject spillovers if the teacher applies the pedagogy encouraged by the literacy hour to his/her teaching of other subjects and if students are enabled to learn more effectively on account of improved literacy skills. Using an analysis conducted on our behalf, we are able to comment on the consequences of improved literacy skills for comprehension of the test in mathematics. There is also a possibility that the ‘literacy hour’ freed up teaching time for other subjects given the efficacy of this method compared to previous practice.

Finally, the policy seems to be successful in that it was not expensive to implement and thus highly cost effective. Hence, when one considers this kind of policy against other, much more expensive, alternatives – like class size reductions and raising teacher salaries – one sees that the literacy hour fares extremely well. Of course, one should realise that its success rests on improving literacy skills at the lower end of the achievement distribution and getting children to basic levels. On this basis, it seems to be a highly successful and desirable education policy.

The structure of the rest of the paper is as follows. In Section 2, we discuss the introduction of the literacy hour, paying a lot of attention to the way in which schools were selected and the implications of this for our research design. In Section 3, we describe the nature of testing in English schools, the data used, and we present some descriptive statistics. Section 4 provides treatment-control estimates of the impact of the literacy hour on primary school performance. We also test for differential pre-policy trends between the treatment and control group; examine whether there is evidence of differential policy effects for ‘good’ and ‘bad’ schools; explore the possibility of how unobserved pupil heterogeneity could affect our results in the absence of spillovers; assess whether there is evidence of any differential impact by gender; and discuss the possibility that the literacy hour had spillover effects on performance in other subjects. Then in Section 5, we assess the cost effectiveness of the policy. Conclusions are presented in Section 6.

2. English primary schools and the literacy hour

2.1. *The literacy hour*

Following much discussion about poor standards of English teaching, the literacy hour was introduced into all English primary schools during the school year 1998/99 in the context of the National Literacy Strategy. The national policy was based on the perceived success of the National Literacy Project (NLP), which was introduced to a sub-set of schools in certain Local Education Authorities⁷ (LEAs) in September 1996.

The NLP policy was not envisaged as a pilot at the time, and indeed was introduced by a different political administration to that which launched the national strategy. Also, the policy was introduced in a context in which there was considerable pressure on all schools to improve due to monitoring and accountability measures introduced in the

⁶ If one observed a stronger effect of the policy in ‘bad’ schools, this could just indicate genuine heterogeneity in the effect of the policy. However, if the policy only affected ‘bad’ schools, one might suspect that the mechanism of the policy was more to do with general school improvement, helping only those schools that were desperate to try anything.

⁷ There are 150 Local Education Authorities in England. They are responsible for the strategic management of local authority education services including planning the supply of school places, ensuring every child has access to a suitable school place, intervening where a school is failing its pupils and for allocating funding to schools.

early 1990s. For example, from 1995/96 onwards, primary school examination results were published in School Performance Tables. From 1993/94 a system of school inspections was implemented, whereby the regulatory body (OfSTED, the Office for Standards in Education) would inspect each school at four year intervals and make reports publicly available. If OfSTED gave a very poor evaluation of school performance, the school could be put on ‘special measures’ and ultimately shut down.⁸

The literacy hour is firmly based upon criteria aimed at improving standards of literacy highlighted in the National Curriculum that was introduced to schools in 1988. This sets out details of what must be taught, the standards that should be achieved at different stages of the education sequence and recommends a minimum teaching time for core subjects. The National Literacy Strategy, and the National Literacy Project before it, was aimed at raising standards of literacy in primary schools by improving the quality of teaching through more focused literacy instruction and effective classroom management. It can also be seen as an attempt to improve school management of literacy through target-setting linked to systematic planning and monitoring and evaluation.

Key components of the policy are a *framework for teaching*, which sets out termly teaching objectives for the 5–11 age range and provides a practical structure of time and class management for a daily literacy hour.⁹ With regard to the former, a range of texts are specified and teaching objectives are set out at three levels (text, sentence and word) to match the text types studied. The daily literacy hour is divided between 10–15 minutes of whole-class reading or writing; 10–15 minutes whole-class session on word work (phonics, spelling and vocabulary) and sentence work (grammar and punctuation); 25–30 minutes of directed group activities (on aspects of writing or reading); and a plenary session at the end for pupils to revisit the objectives of the lesson, reflect on what they have learnt and consider what they need to do next. The general structure and content of the hour is illustrated in Fig. 1.

Beard (2000) discusses how the literacy hour relates to findings from educational research. Of particular relevance are ‘structured teaching’ (e.g. making clear what has to be learnt; dividing material into manageable units; teaching in a well-considered sequence) and ‘effective learning time’ (e.g. as reported by Scheerens (1992) in his meta-analysis). Beard (2000) also notes that the literacy hour has been influenced by literacy schemes in the US and in Australia. The background to the ‘literacy hour’ and its content is discussed at length by Stannard and Huxford (2007).

2.2. Did the literacy hour change teaching practices?

Whilst the literacy hour introduction constituted a discrete, well-defined and systematic change in the teaching of literacy in primary schools, one might plausibly ask to what extent the quality of English teaching was sub-standard prior to its introduction. It is evident that the general standard of reading and writing was a serious concern, particularly in some LEAs.¹⁰ For example, an OfSTED report about the teaching of reading in Inner London primary schools included criticism of the following practices: free reading with little or no intervention by the teacher; too much time spent hearing individual pupils read; insufficient attention to the systematic teaching of an effective programme of phonic knowledge and skills (OfSTED, 1996). It was thought that standards in the teaching of reading varied widely from school to school, with many primary teachers not having had the opportunity to update their skills to take account of evidence about effective methods of teaching reading and how to apply them (Literacy Task Force, 1997a).¹¹

Stannard and Huxford (2007) provide similar insights into the nature of teaching before the ‘literacy hour’, and particularly stress the lack of whole-class teaching. They also note evidence from the School Curriculum and Assessment Authority and the Schools Inspectorate (OfSTED) estimating that schools allocated between 22 and 25% of their time to the teaching of English; but in a class of 30, this might only amount to direct teaching time (per pupil) of 3–5 minutes per week. Thus, a daily hour of continuous teaching time devoted to literacy with carefully planned

⁸ Parents could also act on a poor OfSTED inspection or a low ranking in the School Performance Tables by applying to admit their child to another school (and there are no restrictions on state schools to which they can apply, although criteria such as proximity to school are applied if there is over-subscription). Such ‘voting with feet’ materially affects schools since funding is tied closely to pupil numbers.

⁹ In this paper, we refer to these set of measures as ‘the literacy hour’.

¹⁰ Indeed, John Stannard, the Director of the National Literacy Project, told us that ‘in some LEAs teaching of literacy had fallen apart’. The person in charge of introducing NLP in Tower Hamlets LEA in East London was equally downbeat saying, prior to NLP, teachers used to ‘hear’ reading rather than ‘teach’ it and noted that there was a lot of ‘quiet reading time’.

¹¹ These concerns about reading in particular prompt us to consider the impact of the NLP on reading in specifically, as well as on overall performance in English.

Structure of the Literacy Hour

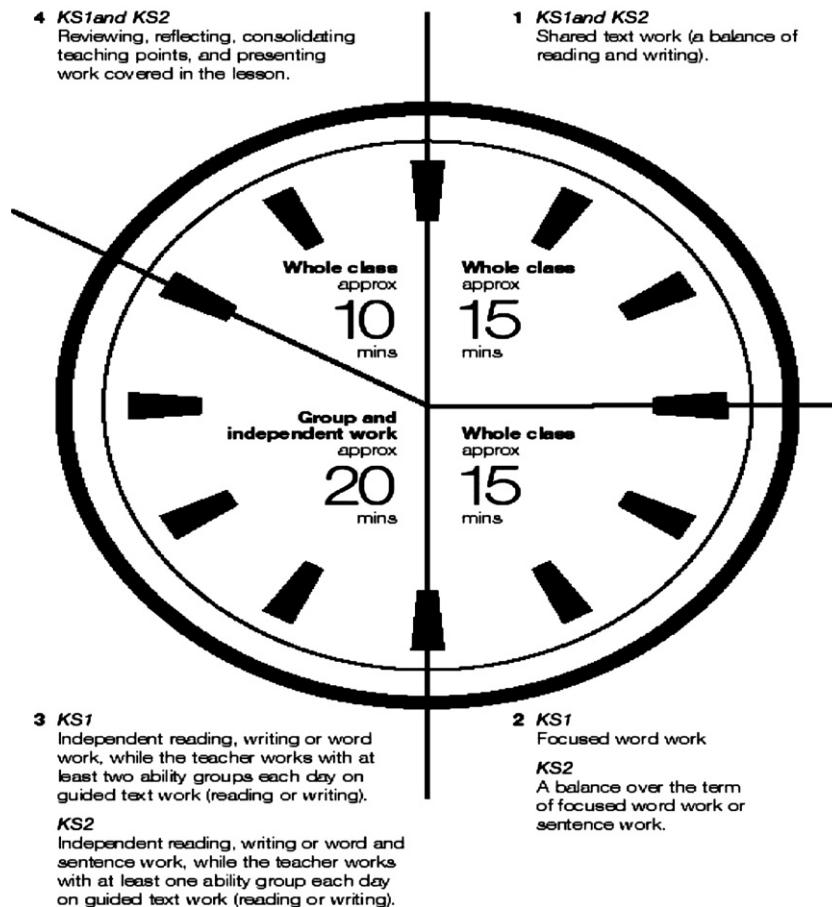


Fig. 1. Structure of the literacy hour. Source: Department for Education and Employment, 1998.

content and a clear structure was a radical departure from what went before. It represents a change in how literacy skills are taught rather than an increase in the time devoted to English.¹² Stannard and Huxford (2007) argue that a revolutionary feature of the NLP was its focus on intervening at the level of classroom instruction because prior to that, English policy makers had stressed that teachers should always take the lead on this.

In schools that were selected for the NLP, the programme was introduced to school staff through attendance at an induction day on Management of Literacy by the Headmaster and Chair of Governors, and a training week for designated key teachers, including English coordinators. There was also one school inset day devoted to NLP issues. There were conferences in each LEA for head teachers and a representative of the governing body. Thus the NLP was a fairly low-cost policy to implement and we will return to this later.

The Schools Inspectorate (OfSTED, 1998) evaluated the NLP by visiting 20% of the schools involved. Three visiting were made: two in 1997 during the spring and the autumn terms and the final one in the summer term of 1998. Over 300 literacy hours were observed and discussions were held with head teachers and staff, including the key teachers. The report says that the impact of the literacy hour on the quality of teaching, and on pupils' attitudes and interest in reading and writing has been very positive. It says that the NLP brought about significant improvements in

¹² Also, according to Stannard and Huxford (2007), the decision to recommend 1 hour per day was based on the working assumptions underpinning the balance of time for English in the National Curriculum.

the quality of teaching and has been the important catalyst in raising standards of literacy in the large majority of schools.

2.3. *The National Literacy Project*

The National Literacy Project was introduced in about 400 schools during the school years 1996/97 and 1997/98.¹³ The Office for Standards in Education (which published the above cited report about the teaching of literacy in inner city London) approached the then Secretary of State for Education to adopt the policy within Local Education Authorities (LEAs) where educational standards were low. However, as we will see below, not all such LEAs were selected into the NLP.

From a research design perspective the NLP is highly attractive in that it gives a setting where some children were exposed to the literacy hour for up to two years when children in similar schools were not. We can thus implement a treatment-control type evaluation, comparing what happened to pupil achievement before and after NLP introduction in affected and unaffected schools.

Of course, for this evaluation to be valid, we have to be very careful to ensure that affected and unaffected schools are similar to one another in the pre-policy period and indeed that there was no differential pre-treatment trend in outcomes measures in the treatment and comparison group. Thus, discussion of how the NLP was introduced is a vital ingredient of our analysis, as is the way we define our set of control schools for our empirical application.

2.4. *How were NLP schools selected?*

There was a two stage process in the selection of NLP schools. First, some Local Education Authorities were selected to participate, and second some schools within LEAs were chosen. At the first stage, the LEAs chosen were, in the main part, those perceived to have lowest pupil performance. In actuality, this turns out to be somewhat mixed, with a range of LEAs being chosen. Most, but clearly not all, are from the lower end of the national achievement distribution, and some LEAs with below average levels were not included. According to the Director of the National Literacy Project, the LEAs chosen for the NLP were selected (fairly arbitrarily) from a longer list.

It turns out that about 80% of NLP schools were located in LEAs in inner city, urban areas.¹⁴ This is where the most disadvantaged and poorly performing schools in England are concentrated (West and Pennell, 2003). Collectively, about 40% of primary schools within these LEAs were involved in the NLP. The remaining NLP schools were in eight LEAs in three counties.¹⁵ Only about 7% of primary schools in these counties were involved in the NLP. Since the policy was more actively promoted in the former areas (and partly for methodological reasons¹⁶) we focus our analysis on the effect of the NLP in urban schools. In the context of debate about poorly performing inner city schools in the UK, it is entirely appropriate to focus on schools in these areas.¹⁷

In terms of which schools were selected within LEAs, advice was given to choose schools most in need of the programme, but to achieve some balance between schools that were perceived to be low achieving and those with ongoing problems (e.g. poor leadership) and low achieving schools showing some signs of improvement.¹⁸ However, since the NLP was planned to last 5 years, a rolling programme was envisaged wherein schools would enter the programme in different waves. The planned cost of the NLP was £12.5 million over five years.

¹³ The NLP was also launched in an additional 112 'first' or infant schools, which we do not consider here because the treatment group in this case does not correspond to the same birth cohort. These schools cater for children up to between the age of 7 and 10 (depending on the school). The majority of children in England (over 80%) attend primary schools that cater for children up to and including the Key Stage 2 assessment at age 11.

¹⁴ Specifically, these inner city schools are located within several Local Education Authorities in London (see Appendix Table A1) and also Sandwell, Liverpool, Manchester, Sheffield, Newcastle and Bristol.

¹⁵ The three counties are Hampshire, Essex and Norfolk. The precise LEAs are as follows: Hampshire, Portsmouth, Southampton, Essex, Southend-on-Sea, Thurrock, Isle of Wight, Norfolk.

¹⁶ It is easier to define a suitable control group for schools in urban areas (as discussed below).

¹⁷ However, we note that the intervention seems to have been ineffective for the small number of treatment schools located in these rural areas. This could be as a result of genuine heterogeneity in the effect of the policy or (in our view more plausibly) because the implementation might have been weaker in these areas.

¹⁸ John Stannard, Director of the National Literacy Project 1996-98, personal communication.

This NLP policy is of considerable interest for evaluating the impact of structured literacy through the literacy hour, as it exposed children in some schools to the policy whereas children in other schools did not participate. To evaluate the policy, one needs to carefully define a counter-factual where children did not receive treatment. We have therefore spent considerable time defining the comparison group against which to benchmark the policy and in designing a methodology used to evaluate the NLP in this treatment-control setting.

2.5. Definition of control group

We compare pupil performance in literacy hour schools with non-literacy hour schools in several ways. First, we simply implement a standard difference-in-difference approach when the comparison group is all other non-NLP primary schools. We also refine this by carrying out statistical matching where we estimate a propensity score for being an NLP school based on observable characteristics in the pre-policy period; we trim the sample to exclude schools in the treatment group that do not have ‘common support’ with schools in the control group (and vice versa); we estimate ‘difference-in-difference’ regressions for schools in the remaining sample.

The second control group is based upon geographical matching of Local Education Authorities (LEAs). One reason for also taking this approach is that LEAs were involved in the administration of the NLP within the treatment schools. The *Literacy Task Force (1997b)* identified LEAs as the ‘first link in the chain’ of implementation, being expected to provide a strong lead to all their schools, profiling literacy as a priority for the LEA with publicity, producing visible targets and so on. It seems evident that all primary schools within an LEA taking part in the NLP may have (at least indirectly) benefited from the initiative — and not just those schools taking part in the project. For example, the NLP policy was well advertised within LEAs. Also, since the NLP was originally envisaged as a five-year rolling programme, school staff would have anticipated taking part officially at some point. Therefore, it is very questionable as to whether ‘non-NLP’ schools within NLP LEAs should be used as controls for the treatment schools.^{19,20} We thus identified LEAs that are in areas geographically adjacent to LEAs involved in the NLP. If there were multiple non-NLP areas, we chose the one with the closest educational performance indicator in the pre-policy period. This approach has similarities to the approach adopted in evaluations of recent UK area-level initiatives (such as the Educational Maintenance Allowance in *Dearden et al., 2005*, and the New Deal in *Blundell et al., 2004*).²¹ A list of NLP LEAs and their ‘matched areas’ are provided in Appendix Table A1.

Hence we have two groups of control schools: all non-NLP schools in England (13,573 schools); another using schools in adjacent non-NLP areas (529 schools). Within both approaches, we further refine the estimation sample by applying statistical matching methods.

3. Patterns of achievement in English, data and initial descriptive statistics

3.1. Testing in English primary schools

Following the introduction of the National Curriculum in 1988, testing throughout the school years has been an important feature of the English school system. Children are administered tests at ages 7, 11, 14 and 16, in what is known as Key Stages 1, 2, 3 and 4. Key Stages 1 and 2 take place in primary school and Key Stages 3 and 4 in secondary school. These national tests are externally set and marked. Key Stages 2 and 4 are particularly ‘high stake’

¹⁹ According to the Director of the NLP, John Stannard, maximum co-operation between schools and cross-fertilisation of ideas was encouraged within LEAs. Indeed there has always been an effort by LEAs to facilitate collaboration and co-operation between primary schools. Hence there would have been opportunities through meetings and informal networking for teachers working in NLP schools to impart information to other teachers working in the same LEA. However, a deliberate effort was made to contain the effect of the NLP within the selected LEAs. For example, there was no formal transfer of information about the NLP across LEA boundaries. Schools outside the LEAs involved would not have been able to obtain a copy of the framework from the national centre. This was to avoid a dilution of important messages expected to come out of the project.

²⁰ In fact it does turn out, as one would expect if these schools were used as controls, that the magnitude of the estimates reported below are tempered, showing there to be an indirect effect of the NLP on non participating schools in the LEA.

²¹ Note that this approach is not amenable to the minority of NLP schools that are in the counties or surrounded by semi-rural areas (i.e. Bristol). This is the practical reason for excluding NLP schools in the counties from our analysis. This has some similarities with the Educational Maintenance Allowance evaluation where there was one rural treatment area (Cornwall). Being unable to define a suitable control area, the researchers dropped this area from their analysis. As discussed above, the policy was more actively promoted in urban areas and these are precisely the schools of interest for which to evaluate such a policy.

tests for schools as results are published in national school Performance Tables. Since the literacy hour was introduced in primary schools we assess its impact on the Key Stage 2 tests that children take at the end of their time in primary school.

3.2. Data

The empirical analysis is based on administrative records of pupil-level achievement and school-level data. The former consists of detailed test score information on students at the end of Key Stage 2.²² The first available year of national Key Stage 2 data at pupil level is in 1995/96, which corresponds to the school year before the National Literacy Project was introduced. We also use the national Key Stage 2 data in the ‘policy on’ school years of 1996/97 and 1997/98, before the National Literacy Strategy was introduced nationwide. Although no pupil-level information is available for 1994/95, we have obtained this data at school-level. This extra year of data enables us to conduct a pre-policy ‘differences-in-differences’ analysis.

The pupil-level administrative files have detailed information on test scores, gender of the student and school codes.²³ This latter information allows the files to be matched up to national school-level data available in the School Performance Tables and the LEA and School Information Service (LEASIS). Available information includes measures of school outcomes (results, absences), inputs (e.g. pupil–teacher ratios), disadvantage (e.g. the percentage of students eligible for Free School Meals or identified as Special Educational Needs) and other school characteristics (e.g. school type).²⁴

With regard to outcome measures, we concentrate on two measures at the end of primary school: the percentile reading score and the percentage of students achieving level 4 or above in Key Stage 2 English. Since the marking scheme can change over time, we convert raw scores to percentile scores. The scores of various tests are aggregated and then converted to an overall ‘level’ (which is in a range of 2–6). The key indicator of policy interest is the percentage of students attaining level 4 and above at age 11, which is the standard deemed to be appropriate at this age (as outlined in the National Curriculum).

3.3. Descriptive statistics

Descriptive statistics for the outcome variables of interest are provided in [Table 1](#) (in [Appendix Table A4](#), we show summary statistics for other variables). The table shows measures of primary school achievement before and after the introduction of the policy for NLP schools and the two comparison groups: all non-NLP schools; and non-NLP schools in adjacent (urban) LEAs. The table shows various primary school measures — the mean percentile reading score; the percentage reaching level 3 or above in KS2 English; the percentage reaching level 4 or above (i.e. the expected standard); the percentage reaching level 5 or above. Years denote the end of school year — namely when children take their tests (so for example, 1996 refers to 1995/96 year).

Outcome measures are lower on average in NLP schools at each point in time than in either of the comparison groups. For example, in the pre-policy year, the mean percentile reading score was 37.56 in NLP schools, as compared to 46.34 in adjacent LEAs and 50.88 in all non-NLP schools. The percentage of students reaching levels 3, 4 and 5 in NLP schools was about 79, 39 and 5% respectively, as compared to 85, 52 and 10% for schools in adjacent LEAs and 88, 58 and 12% in all non-NLP schools. This is important since for valid inference to be drawn we need to standardise for baseline differences in the pre-policy period (to ensure we are comparing like with like). We are careful to investigate this in detail in our empirical modelling below.

With this in mind, one should be careful in reading the table, but a suggestive and interesting pattern emerges in terms of changes before and after NLP introduction. For reading scores, there is evidence of improvement in the NLP schools relative to the control schools. The mean reading score goes up by about 2 percentile points in the NLP schools relative to that in either of the two control groups, as shown in the final column which shows the difference-in-

²² The Key Stage 2 tests were first taken in the school year 1994/95.

²³ Unfortunately, we do not have more detailed information on pupil-level characteristics and we do not know their prior attainment (age 7 test scores).

²⁴ A full list of the detailed set of variables used as controls is provided in the notes to [Table 2](#).

Table 1
Mean outcomes for various samples

	Number of schools	Pre-policy, 1996	Post-policy, 1997–98	Change	Difference-in-difference, NLP — controls
<i>A. Percentile reading scores</i>					
NLP schools	269	37.56	39.90	2.34 (.53)	–
All non-NLP schools	13573	50.88	51.14	0.26 (.08)	2.08 (.59)
Non-NLP schools, adjacent LEAS	529	46.34	46.69	0.35 (.43)	1.99 (.67)
<i>B. Percent level 3 or above KS2 English</i>					
NLP schools	269	78.85	83.80	4.95 (.57)	–
All non-NLP schools	13573	88.20	91.15	2.95 (.07)	2.00 (.62)
Non-NLP schools, adjacent LEAS	529	85.38	88.94	3.56 (.38)	1.39 (.70)
<i>C. Percent level 4 or above KS2 English</i>					
NLP schools	269	38.96	49.16	10.20 (.82)	–
All non-NLP schools	13573	58.15	65.11	6.96 (.11)	3.24 (.91)
Non-NLP schools, adjacent LEAS	529	51.52	59.13	7.61 (.63)	2.59 (1.04)
<i>D. Percent level 5 or above KS2 English</i>					
NLP schools	269	4.94	8.96	4.02 (.42)	–
All non-NLP schools	13573	12.35	16.62	4.27 (.09)	–0.25 (.46)
Non-NLP schools, adjacent LEAS	529	10.14	13.25	3.11 (.42)	0.91 (.53)

Notes: standard errors, clustered on school, in parentheses.

difference estimate for the period surrounding NLP introduction.²⁵ A similar relative pattern of improvement is seen for achievement at level 3+ or level 4+ (which is higher when the comparison group is all other schools rather than schools in adjacent LEAs). However, at high measures of achievement – level 5 or above – the positive differential in NLP schools is either strongly reduced and statistically insignificant (when the comparison group consists of schools in adjacent LEAs), or removed altogether (when the comparison group consists of all other schools). Similar differences-in-differences are seen when we trim the sample (using propensity score matching methods) and these are reported in Table A3 of the Appendix (in Table A4, summary statistics are presented for other variables).

4. The impact of NLP on pupil achievement

In this section, we evaluate the NLP policy impact, looking at difference-in-difference models²⁶ that control for a large number of observable factors and for unobserved school heterogeneity. This is important owing to the different levels of pre-policy achievement in the NLP and control schools already discussed. The basic estimates are derived from the following model for pupil i in school s in year t :

$$A_{ist} = \alpha_s + \beta \text{NLP}_s * \text{Policy On}_t + \delta X_{ist} + \lambda Z_{st} + \pi T_t + \varepsilon_{ist} \quad (1)$$

where A is pupil achievement; X denotes a set of pupil characteristics; Z a set of school characteristics; T a set of year dummies; α_s denotes school fixed effects (which encompasses any time constant effect common to NLP schools) and ε is an error term. NLP is a dummy equal to one for NLP schools. This is interacted with a ‘Policy On $_t$ ’ variable, which is set equal to one for time periods when the NLP policy was in effect and zero in pre-policy periods. The coefficient β is the difference-in-difference estimate of the NLP policy. Since school fixed effects are included in the model, it measures within-school changes in achievement before and after NLP introduction in treatment schools relative to within-school changes in achievement in control schools.

²⁵ We discuss potential spillover effects on to other subjects later in the paper. The raw difference-in-difference estimate with regard to the mathematics test is between .55 and .87 (depending on the comparison group used) and is not statistically significant.

²⁶ The method is based on the Wald estimator and has been described and used in a number of early papers including Ashenfelter (1978) and Heckman and Robb (1985).

4.1. Estimated policy impacts

We estimate Eq. (1) and report the estimated coefficient on $NLP_s * Policy On_t$ for the percentile score in reading and overall performance in English, as measured by whether the pupil achieves level 3 or above, level 4 or above (the expected standard), and level 5 or above. These results are reported in Table 2 (panels A–D). We report results using four different samples: where the control group is all non-NLP schools (column 1); non-NLP schools in adjacent LEAs (column 2); and analogues to columns (1) and (2) where we have selected a sample of matched schools using propensity score matching techniques on the 1996, pre-policy data (columns 3 and 4). This is an attempt to further standardise, in a less parametric way, the set of treatment and control schools. The probit models used to generate them are shown in the Appendix (where, as already noted, we also replicate the descriptive statistics shown in Table 1 for the relevant sub-samples; the number of schools in each group is also reported). The basic method used is that of Heckman et al. (1997), where propensity scores are estimated and the sample then trimmed to exclude poorly matched schools.²⁷

For each outcome measure, results across the four samples are very similar. With regard to percentile reading scores, the difference-in-difference estimate is 2.384 when the control group is all non-NLP schools (column 1) and 2.154 when the control group consists of non-NLP schools in adjacent LEAs (column 2). It is notable that the difference-in-difference estimates are higher than in the simple descriptive table (Table 1) after the controls are added.²⁸ When using the matched samples of schools (in columns 3 and 4), the coefficients decline a little, but they are still of the same order of magnitude. For the trimmed sample of NLP schools and all non-NLP schools (column 3), the coefficient declines to 1.841. For the trimmed sample of NLP schools and non-NLP schools in adjacent LEAs, the coefficient declines to 2.027. The estimates are statistically significant in each case, showing an improvement in reading for children exposed to the literacy hour relative to those who were not.

When the outcome measure is whether the pupil achieves level 3 or above in English, the coefficients vary from 2.329 and 1.936 when the control groups consist of all non-NLP schools (column 1) and non-NLP schools in adjacent LEAs (column 2), to 1.999 and 1.859 in the corresponding matched samples (i.e. columns 3 and 4 respectively). The estimates are qualitatively similar (and a little higher in magnitude) when the outcome measure is the headline measure of whether the pupil achieves level 4 or above. In this case, estimates in columns 1 and 2 are 3.564 and 3.024, whereas they are 2.747 and 2.930 in the corresponding matched samples (columns 3 and 4). Hence, these estimates suggest that NLP raised the percentage of students achieving level 3 or above in Key Stage 2 English by about 2 percentage points and the percentage of students achieving level 4 or above by about 3 percentage points. All estimates are statistically significant.

One concern might be that the more rigid literacy hour teaching structure dampens prospects for higher ability children by holding them back. The final panel shows the estimated effect of NLP where the outcome measure is whether the pupil achieves level 5 or above. In this case the policy is shown to have had a negligible impact: the coefficient is small across all samples and the estimated effects are always statistically insignificant. Thus the policy did not have negative effects on the high ability children. Equally it did not benefit them as it did for the children with low and average literacy skills.

The results for reading and overall performance in English (up to and including level 4) are highly supportive of the hypothesis that the literacy hour, via the NLP policy, significantly improved the acquisition of literacy skills. The fact that the policy had no impact on the probability of achieving very high scores (i.e. level 5 or above) shows two things: firstly that the policy improves basic literacy, but it does not create star performers; secondly, the existence of the policy does not hold students back if they are at the upper end of the distribution: they continue to perform as well as they would have done in the absence of the policy.

²⁷ See Rosenbaum and Rubin (1983, 1984) for the initial statements on how to use propensity score matching as a means of reducing bias in observational studies designed to compare treatments and controls. Note that in this context, it is appropriate to match at school-level because the treatment is at school-level. Furthermore, we only have pupil-level data on gender and outcome variables.

²⁸ It is of interest to note that the 'baseline' difference between NLP and non-NLP areas reduces considerably after controls have been added (not reported in the tables). This can only be seen if one does not control for school fixed effects (since being an NLP school is a fixed characteristic). After controlling for observable school characteristics the baseline difference between NLP and non-NLP schools (in adjacent LEAs) reduces from 12 percentage points to less than 1 percentage point (which is not statistically significant).

Table 2
Basic results — NLP and primary school reading and English

	(1)	(2)	(3)	(4)
	Least squares/linear probability models		With matching	
Control schools	All non-NLP schools	Non-NLP schools, adjacent LEAS	All non-NLP schools	Non-NLP schools, adjacent LEAS
<i>A. Percentile reading scores^a</i>				
NLP* Policy On	2.384 (.491)	2.154 (.583)	1.841 (.495)	2.027 (.585)
<i>B. Percent level 3 or above KS2 English</i>				
NLP* Policy On	2.329 (.521)	1.936 (.591)	1.999 (.520)	1.859 (.594)
<i>C. Percent level 4 or above KS2 English</i>				
NLP* Policy On	3.564 (.770)	3.025 (.906)	2.747 (.774)	2.930 (.912)
<i>D. Percent level 5 or above KS2 English</i>				
NLP* Policy On	-.158 (.440)	.461 (.543)	-.447 (.444)	.287 (.542)
Control variables	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Number of pupils	1,623,972	98,071	1,262,993	96,926
Number of schools	13,942	798	10,064	786

Notes: standard errors (clustered on school) in parentheses; all specifications include year dummies; control variables are as follows (all are at school-level apart from gender of student): % eligible for Free School Meals; % non-white students; % students with Special Educational Needs, with statement and without statement; pupil–teacher ratio; number of pupils; whether religious school; % teachers who are not fully qualified; ratio of support staff to teachers; % teachers who are graduates and with particular class of degree; % female teachers; missing variable indicators.

^aSample size for reading in full sample is 1608581; sub-sample: 96837; full sample after matching: 1250618; sub-sample after matching: 95708. In the following tables based on pupil-level data, the sample sizes are also slightly lower when reading is the outcome measure.

4.2. Differential pre-policy trends between treatment and control groups?

The above analysis identifies the effect of NLP policy under the assumption that the coefficient on NLP_{*s*}* Policy On_{*s*} reflects the impact of the NLP policy alone. Our methodology helps to achieve this goal through inclusion of a long list of covariates (including school fixed effects), and by demonstrating that results are robust to estimation using different sets of control schools, even after trimming the sample to remove schools that do not have ‘common support’.

However, there would be a problem if NLP schools were on an upward trend prior to the NLP policy. For example a ‘regression to the mean’ argument would say that schools with below average performance in reading and English are likely to be on an upward path relative to other schools.²⁹ Since the NLP policy was not randomly assigned, and there was no explicit formula for participation, careful matching of treatment and control schools may not completely remove this concern.

To examine this possibility, we have estimated the same regression for the two pre-policy periods where we have data available (i.e. 1995 and 1996). We use data at school-level, since pupil-level data is not available prior to 1996. The only available outcome measure in the school-level data is the percentage of pupils achieving level 4 or above. Results for the pre-policy difference-in-difference are reported in panel C of Table 3. For comparison, we also show the ‘policy on’ effect of NLP in the school-level data in Panel B (i.e. replicating results in panel C of Table 2, except the data is from the School Performance Tables rather than pupil-level data for the years 1996–98). Again, we report results for four different samples: when the control group of schools consists of all non-NLP schools (column 1), non-NLP schools in adjacent LEAs (column 2), and the corresponding matched samples (columns 3 and 4 respectively). All estimates in Panel C show the ‘pre-policy’ NLP effects to be relatively small and statistically insignificant.³⁰ The ‘pre-

²⁹ Indeed, there is evidence of convergence of test scores in that there is a strong school-level lagged achievement effect on pupil achievement. The lagged achievement variables are entered in a flexible way, including a number of variables measuring proportions achieving different levels of English and Mathematics. In these specifications there is a strong effect from the ‘own’ school-level lagged achievement measure (e.g. lagged proportion achieving Level 4 English in a Level 4 English achievement equation), but also significant positive effects from the other proportions.

³⁰ The raw difference-in-difference effect is almost the same as that reported when controls are added.

Table 3
Exploring possible mean reversion — looking at pre-policy (1995–96) changes

	(1)	(2)	(3)	(4)
	Least squares/linear probability models		With matching	
Control schools	All non-NLP schools	Non-NLP schools, adjacent LEAS	All non-NLP schools	Non-NLP schools, adjacent LEAS
<i>A. Percent level 4 or above: KS2 English — pupil level, policy period</i>				
NLP*Policy On	3.564 (.770)	3.025 (.906)	2.747 (.774)	2.930 (.912)
Control variables	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
<i>B. Percent level 4 or above: KS2 English — school-level, policy period</i>				
NLP*Policy On	3.483 (.781)	2.918 (.921)	2.653 (.786)	2.828 (.927)
Control variables	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
<i>C. Percent level 4 or above: KS2 English — school-level, pre-policy period</i>				
NLP*Pre-Policy On	.685 (1.011)	.595 (1.253)	.818 (1.017)	.644 (1.263)
Control variables	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Number of schools	13,942	798	10,064	786

Notes: standard errors, clustered on school, in parentheses. The outcome variables for school-level regressions are from the School Performance Tables (and not based on aggregate individual-level data). The regressions are weighted by number of pupils in year group.

policy' NLP coefficient is 5 times smaller than the 'post-policy' coefficient in columns 1 and 2 and 3–4 times smaller in the matched samples.

This pre-policy test leads us to reject the hypothesis that the NLP effects estimated in Table 2 reflect 'regression to the mean' in NLP schools. It would appear that their relative improvement in performance would not have happened in the absence of the introduction of the NLP policy.

4.3. Differential effects for 'good' and 'bad' schools?

The next question is whether the effect of the NLP resembles a school improvement policy that would differentially benefit schools under pressure to improve or whether its positive effect is more likely to reflect the new content and pedagogy for teaching literacy, embedded in the 'literacy hour'. One argument against the former interpretation is that all schools were under pressure to improve in England at this time due to various measures to improve school accountability, including the publication of School Performance Tables.

Furthermore, the establishment of a statutory school inspection process by a regulatory body with power to put schools under 'special measures' (after which they could face closure) was an additional pressure for schools in the 1990s. Inspectors from the regulatory body (OfSTED) visit every school in England once every four years. The output of the inspection includes a report on the school, which is publicly available. We have obtained the overall grade assigned to schools during the first round of inspections (1994–98). These take into account the quality of education, leadership and management, standards and ethos of the school. We separate the schools into two groups: schools needing substantial improvement (which we label as unsatisfactory) and schools with any grade above that (very good, good, or needing some improvement — which we label as 'good/satisfactory').³¹ One would think that schools in the former category would be under particular pressure to improve given the 'accountability' environment in England at this time and the capacity for parents to act on it.

If the NLP operates as a school improvement policy then one might expect the NLP effect to be much bigger in schools labelled as 'unsatisfactory'. However, if the NLP raises performance mainly through a change in the content and pedagogy for teaching literacy, one would expect NLP policy to be effective for both groups.

³¹ Although OfSTED inspected all schools over this time period, there are missing composite grades in the OfSTED data set for about 16% of schools. We drop these schools from this part of the analysis. Of the available sample, about 4% of schools have an 'unsatisfactory' grade (7% of which are NLP schools; 32/215 NLP schools in this sample).

Table 4
Effect of NLP conditional on OfSTED grade (1994–98)

	(1)	(2)	(3)	(4)
	Least squares/linear probability models		With matching	
Control schools	All non-NLP schools	Non-NLP schools, adjacent LEAS	All non-NLP schools	Non-NLP schools, adjacent LEAS
<i>A. Percentile reading scores</i>				
NLP*Policy On* ‘Good’ or ‘Satisfactory’	2.512 (.611)	2.393 (.695)	1.964 (.619)	2.265 (.699)
NLP*Policy On* ‘Unsatisfactory’	3.337 (.899)	2.589 (1.075)	2.843 (.892)	2.505 (1.070)
<i>P</i> -value of test of no difference	.446	.869	.415	.840
<i>B. Percent level 3 or above KS2 English</i>				
NLP*Policy On* ‘Good’ or ‘Satisfactory’	2.478 (.629)	2.263 (.696)	2.142 (.624)	2.186 (.701)
NLP*Policy On* ‘Unsatisfactory’	3.186 (1.713)	2.867 (1.739)	2.940 (1.714)	2.800 (1.742)
<i>P</i> -value of test of no difference	.698	.740	.661	.736
<i>C. Percent level 4 or above KS2 English</i>				
NLP*Policy On* ‘Good’ or ‘Satisfactory’	3.539 (.934)	3.253 (1.059)	2.708 (.942)	3.141 (1.067)
NLP*Policy On* ‘Unsatisfactory’	4.115 (1.590)	3.512 (1.655)	3.429 (1.562)	3.412 (1.655)
<i>P</i> -value of test of no difference	.754	.888	.692	.883
<i>D. Percent level 5 or above KS2 English</i>				
NLP*Policy On* ‘Good’ or ‘Satisfactory’	.0417 (.540)	.679 (.645)	-.267 (.546)	.535 (.648)
NLP*Policy On* ‘Unsatisfactory’	.422 (.787)	.466 (.930)	.161 (.806)	.392 (.931)
<i>P</i> -value of test of no difference	.689	.838	.658	.891
Control variables	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Number of pupils	1,384,322	81,244	1,074,965	80,273
Number of schools	11,765	658	8537	648

Note: The number of schools used in these regressions is lower than in the corresponding regressions shown in Table 2. We dropped schools that were missing from the OfSTED data base. These are schools that will have had an inspection, but where the composite score is missing.

In Table 4, we show estimates of the same regressions as those reported in Table 2, except here we interact $NLP_s * Policy On_t$ with whether the school is classified as either ‘good/satisfactory’ or ‘unsatisfactory’. The difference in the effectiveness of the NLP between these two groups of schools is numerically similar and not statistically different for the outcome variables of interest: the reading score; the probability of achieving level 3 or above, level 4 or above in English. As with the main results, there is no discernible effect of the NLP policy on the probability of achieving level 5 or above, however the school is categorised. Although the coefficient $NLP_s * Policy On_t$ is a little higher for schools if they are categorised as ‘unsatisfactory’ compared to if they are classified as ‘good/unsatisfactory’, the differential is never big enough to be of concern (the coefficient for schools in the latter category is about 70–90% of what it is for ‘unsatisfactory schools’ in panels A–C, depending on the outcome measure and sample chosen). This suggests that the NLP policy is not acting differentially according to the degree to which schools are under pressure to improve. In this respect, it does not resemble a more general school improvement policy. It is more probable that the success of the policy reflects its impact on the teaching of literacy rather than through any other channel.

4.4. Unobserved pupil heterogeneity?

In the analysis reported so far there does still remain a question of unobserved pupil heterogeneity. We view this as not so important since we are comparing across cohorts of children in the same schools. Nonetheless the question does deserve some attention. One possible way of conditioning out unobserved pupil heterogeneity is to estimate the effect of the NLP on outcome measures after conditioning on analogous measures of performance in other subjects. We have data on Mathematics performance for the same children and can consider this. If performance in Mathematics were completely unaffected by the NLP but captured heterogeneity between pupils in NLP schools and control schools, then the estimates would show an effect of the NLP net of this pupil heterogeneity.

Table 5
Unobserved pupil heterogeneity — English relative to Maths

	(1)	(2)	(3)	(4)
	Least squares/linear probability models		With matching	
Control schools	All non-NLP schools	Non-NLP schools, adjacent LEAS	All non-NLP schools	Non-NLP schools, adjacent LEAS
<i>A. Percentile reading scores</i>				
NLP*Policy On	1.516 (.394)	1.113 (.480)	1.209 (.396)	1.052 (.479)
<i>B. Percent level 3 or above KS2 English</i>				
NLP*Policy On	1.631 (.433)	1.162 (.492)	1.413 (.435)	1.151 (.493)
<i>C. Percent level 4 or above KS2 English</i>				
NLP*Policy On	2.288 (.633)	1.647 (.787)	1.790 (.668)	1.603 (.791)
<i>D. Percent level 5 or above KS2 English</i>				
NLP*Policy On	-.407 (.392)	-.068 (.493)	-.549 (.396)	-.228(.493)
Control variables	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Number of pupils	1,623,972	98,071	1,262,993	96,926
Number of schools	13,942	798	10,064	786

Notes: standard errors, clustered on school, in parentheses.

Results are reported in Table 5 for the same outcome measures and samples used for our main results (Table 2). When the outcome measure is the reading score, the probability of achieving level 3 or above, level 4 or above in English (panels A–C), the coefficient on $NLP_s * Policy On_t$ is about 50–70% of its magnitude in the main regressions (Table 2). As before, there is no discernible impact of the policy on the probability of achieving level 5 or above (panel D). The regressions therefore show a positive and significant effect of the NLP policy even if one controls for unobserved pupil heterogeneity.

However, we argue in the next sub-section below that there are good reasons to expect a spillover effect of the NLP policy on performance in other subjects. In this case, an additional effect of the NLP may be to improve performance in Mathematics and one would not want to control for this in regressions. Another way of interpreting Table 5 is that the NLP has an additional impact on Mathematics, but this is not as large as the estimated impact for English (i.e. it is not large enough to nullify the estimated NLP effect for English in Table 5).³²

4.5. Spillover effects on to other subjects

The ‘literacy hour’ may have had spillover effects for performance in other subjects. One possibility is that the standard of teaching improves. In English primary schools, children are generally taught by the same teacher for most of their subjects. If the literacy hour changes the way teachers teach (for example, if they are inspired by the pedagogical practice encouraged by the policy), then children’s performance in other subjects may improve as well. Transferable aspects of the new pedagogy include the highly structured nature of lessons and the balance between group work and whole-class activity. The school inspectorate suggested that the National Literacy Strategy has raised the quality of teaching in the rest of the curriculum (OfSTED, 2002).

Another reason for positive spillovers is that improved language skills may lead to better performance in other subjects. For example, children need to be able to read and understand the mathematics test in order to do well in this subject.³³ A comparison has been conducted of the reading demand of the mathematics test relative to the reading test of 1996.³⁴ This has been done using the Lexile Framework, which measures text difficulty by such characteristics as

³² If we estimate regressions where the Maths test score is the dependent variable, the coefficient on $NLP * Policy On$ varies between .865 (.540) and 1.464 (.614). These are considerably lower than the corresponding estimates for the reading score (Table 2, panel A).

³³ Our reading of the mathematics tests (which are available on request) has convinced us on this point.

³⁴ As also stated in the acknowledgements, we are very grateful to researchers at MetaMetrics, Inc. (Developers of the Lexile and Quantile Frameworks) for conducting this analysis on our behalf.

word frequency and sentence length (which has been linked to reading comprehension — see [Stenner et al., 1983](#)). This analysis shows that the reading demand of the mathematics test is nearly 70% of what it is in the reading assessment (i.e. based on text difficulty). It is also possible to estimate that the median grade 5 reader in the US would have a forecast comprehension rate of 80% on the reading assessment and 94% on the mathematics assessment.³⁵ However, for boys at the lower part of the distribution (at the 25th percentile), the forecast comprehension rate for the reading and mathematics test is only 44% and 74% respectively. If reading ability increased by the magnitude of the estimate in Panel A of [Table 2](#), the new forecast comprehension rates would be 76% for mathematics and 46% for reading. This illustrates the potential for the literacy hour to lead to an improvement in the comprehension of the mathematics test. Whilst of course, reading comprehension should be more highly rewarded in the reading test, one would expect it to improve performance in other subjects since reading and understanding the questions are essential to doing well in examinations.

A further reason why there might be positive spillover effects is if the ‘literacy hour’ was a more efficient way to impart knowledge to children than what had been previous practice in many schools (i.e. discussed above: hearing children read one at a time). In this case, the ‘literacy hour’ might have freed up teacher time to devote to other subjects.

4.6. Gender differences

From a theoretical perspective, one may be interested in the impact of the literacy hour on gender differences in achievement. Boys have traditionally performed considerably worse at literacy-related activities, where there has been a gender gap in favour of girls for many years.³⁶ For example, in the school year before the literacy hour was introduced, 50% of 11 year old boys achieved the expected standard in English at the end of primary school, whereas 65% of girls achieved this standard (i.e. as defined in the National Curriculum). There are many (often hotly contested) reasons for why this gender gap exists, and there is a large literature on gender gaps in education.³⁷ One pertinent issue is whether the literacy hour had the potential to make inroads into this gender gap. If the ‘literacy hour’ has greater potential to benefit students at the lower end of the literacy distribution, then one might expect a greater impact for boys. Furthermore, the highly structured nature of the literacy hour may help boys differentially if one thinks that they have great problems with concentration and focus relative to girls.

Given the existence of sizable gaps in English achievement between boys and girls, this also matters from a policy perspective. Thus, the results reported in this section break down the NLP effect, estimating separate effects by gender. [Table 6](#) shows separate NLP effects for boys and girls for the same outcomes measures and samples as those reported above.³⁸ The table reveals some evidence of gender differences in the NLP policy impact at primary school. The NLP effect for boys is numerically larger than that for girls. However, estimates are not always statistically different from each other. In general, they are statistically different (at the 5 or 10% level) when using schools in adjacent LEAs as the relevant comparison group whereas they are not statistically different when using the full sample.

If we consider the point estimates, the literacy hour raised boys’ mean percentile reading scores by somewhere between 2.2 and 3 percentile points, whereas the effect is between 1.2 and 1.9 percentile points for girls. For boys, the NLP raised the probability of achieving level 3 or above by 2.5 to 3 percentage points and the probability of achieving level 4 or above by 3.3 to 4.6 percentage points. For girls, it raised the probability of achieving level 3 or above and level 4 or above, respectively, by 0.8–1.7 and 2.3–2.9 percentage points. As before, there is no discernible impact of the policy on the probability of achieving level 5 or above.

The fact that the estimates are only statistically significant using one of the two possible comparison groups means that we should not over-emphasise the impact of the literacy hour on the gender gap. However, tentatively, the analysis

³⁵ In this framework, reader ability and text difficulty are evaluated separately. Forecast comprehension rates are based on a combination of reader ability and text difficulty. Further details of the methodology are available on request.

³⁶ [Machin and McNally \(2006\)](#) show there to be a significant gender gap (favouring girls) in primary school reading abilities dating at least far back as 1980 using data on reading tests administered to children of the British Cohort Study, a birth cohort of all children born in a week of April 1970.

³⁷ A lot of this work, again, is within the education field (see [Maynard, 2002](#), [White, 1996](#), or [Millard, 1996](#)), but there is some more recent work in economics (like [Jacob, 2002](#)).

³⁸ The model is estimated over the pooled data. It allows for a separate literacy hour impact by gender. All variables (including fixed effects) are interacted with the gender variable.

Table 6
Gender gaps

	(1)	(2)	(3)	(4)
	Least squares/linear probability models		With matching	
Control schools	All non-NLP schools	Non-NLP schools, adjacent LEAS	All non-NLP schools	Non-NLP schools, adjacent LEAS
<i>A. Percentile reading scores</i>				
NLP*Policy On, boys	2.819 (.594)	2.964 (.692)	2.235 (.597)	2.812(.694)
NLP*Policy On, girls	1.948 (.579)	1.273 (.692)	1.463 (.581)	1.179 (.697)
<i>P</i> -value of test of no difference	.293	.085	.354	.097
<i>B. Percent level 3 or above KS2 English</i>				
NLP*Policy On, boys	2.944 (.742)	2.977 (.846)	2.545 (.748)	2.823 (.850)
NLP*Policy On, girls	1.750 (.586)	.842 (.658)	1.463 (.578)	.842 (.660)
<i>P</i> -value of test of no difference	.207	.047	.253	.066
<i>C. Percent level 4 or above KS2 English</i>				
NLP*Policy On, boys	4.162 (1.010)	4.601 (1.172)	3.310 (1.010)	4.495 (1.179)
NLP*Policy On, girls	2.918 (.919)	1.360 (1.074)	2.126 (.923)	1.280 (1.080)
<i>P</i> -value of test of no difference	.362	.042	.387	.045
<i>D. Percent level 5 or above KS2 English</i>				
NLP*Policy On, boys	-.256 (.458)	.350 (.567)	-.550 (.458)	.138 (.561)
NLP*Policy On, girls	-.270 (.626)	.316 (.770)	-.510 (.628)	.180 (.773)
<i>P</i> -value of test of no difference	.986	.972	.959	.965

Notes: as for Table 2.

lends support to the hypothesis that the literacy hour was more effective for boys and as such reduced the gender gap at primary school.

5. Measuring economic costs and benefits

This analysis has shown a significant impact of the literacy hour on reading and on English achievement. The question remains as to whether the policy was cost effective. Hence we now compare the per pupil costs of the policy with the economic benefits, as reflected in predicted labour market earnings.

The planned cost of the NLP was £12.5 million over 5 years. The main costs were 14 local centres (each costing about £25,000 per year) and literacy consultants in each participating Local Education Authority (about £27,000 per year for each consultant). Schools also received some funding for teacher training and resources, which was broadly the same for each school (though some account was taken of the pupil–teacher ratio). However, since the national roll out took place two years after the NLP was introduced, only the first two years are relevant. The total cost per annum was thus £2.5 million (or about £2.8 million in 2001 prices). We observe the number of students affected from pupil numbers in the schools within Cohorts 1 and 2 in 1997 and 1998 (i.e. 222,261 pupils in aggregate).³⁹ Hence the cost per pupil is £25.52 per annum. The cost is so low that it does not take a much of a benefit to offset it.

To estimate benefits of the policy, we first convert the impact of the policy on reading scores (i.e. 2.38 percentiles, as shown in Table 2, column 1) to an equivalent estimate in terms of standard deviations. This is calculated as 0.083 standard deviations.⁴⁰ Secondly, we estimate the impact of reading scores on future labour market earnings using the British Cohort Study. This is a panel survey of all those living in Great Britain who were born between 5th and 11th April 1970. We regress the log of labour market earnings (at age 30, in 2000) on age 10 percentile reading scores (from 1980).⁴¹ Results are shown in Table 7. We show three specifications. In column (1), controls are included for gender

³⁹ This includes infant schools.

⁴⁰ It is worth pointing out that the estimate may be higher for subsequent cohorts because they would have been exposed to more than 1 or 2 years of the literacy hour.

⁴¹ The reading test is a shortened version of the *Edinburgh Reading Test*, which is a test of word recognition. It examines vocabulary, syntax, sequencing, comprehension and retention (see Godfrey Thompson Unit, University of Edinburgh, 1978, or Plake and Impara, 2001, for more details).

Table 7
Earnings gains associated with age 10 reading skills, British Cohort Study

	(1)	(2)	(3)
	Basic specification	(1) Plus family background	(2) Plus highest qualification
Reading score at age 10 percentile ($\times 100$)	0.544 (0.024)	0.423 (0.035)	0.207 (0.036)
Controls	Yes	Yes	Yes
Family background	No	Yes	Yes
Highest qualification	No	No	Yes
Sample size	6587	3488	3488
Percent earnings impact of .091 increase in standard deviation	1.4	1.1	0.5

Notes: dependent variable is $\log(\text{weekly earnings})$ in 2001 prices; standard errors in parentheses; controls included in all specifications for gender and region; family background variables are $\log(\text{parental income at age 16})$, dummies for mother's and father's education; highest qualification are dummy variables for highest educational qualification achieved by age 30. The percent earnings impact of a .091 increase in the standard deviation (SD) of reading percentiles is calculated as $.091 * SD * [\exp(\text{reading score coefficient}/100) - 1]$.

and region; in (2) we add controls for family background; and in (3) we include dummy variables for the participant's highest educational qualification achieved by age 30. Since the latter variable is likely to partly capture the effect of the reading score, the effect of reading on labour market earnings in column (3) should be considered as a lower bound estimate (or even an under-estimate).

The estimates are 0.54, 0.42 and 0.20 in each of the respective columns and are always statistically significant, showing a higher standard of reading at age 10 to be associated with higher earnings at age 30. The earnings impact of a 0.083 increase in the standard deviation (SD) of reading percentiles is then calculated as $0.083 * SD * [\exp(\text{reading score coefficient}/100) - 1]$. This amounts to an annual sum of £179.06, £140.67 and £68.77 for each specification. Assuming that labour market participation occurs between the age of 20 and 65, and using a discount rate of 3%, the corresponding present discounted value of the cumulative effect of the literacy hour is estimated as £4995, £3924 and £1918.

There are, as with all such calculations, certain issues that may lead one to question these economic benefits. One clear example in this case is that the beneficiaries of the NLP literacy hour tended to be children from less well performing schools. On average these children are likely to be located further down the reading score distribution, yet the earnings effect from age 10 reading is assumed linear in the regressions in Table 7. Therefore we have considered economic benefits from parametric models where we allow separate effects on earnings for the top and bottom half of the reading score distribution and from non-parametric regressions (using the Nadaraya–Watson estimator). In both cases the linearity assumption seems reasonable: the bottom half effect for the column (2) specification in Table 7 is estimated to be 0.458 (standard error=0.100) as compared to the top half effect of 0.435 (0.084); a non-parametric regression is shown in Appendix Fig. A1.

Whichever way one looks at it, the benefits of the literacy hour seem to be sizeable and the costs are much smaller. Even if we take the smallest impact estimate from our analysis (the 1.84 percentile improvement in column (3) of Table 2, which corresponds to a 0.06 standard deviation increase), the economic benefits are measured in the range of £1375 to £3581.

The benefits of the policy (a 0.06 to 0.08 standard deviation in reading scores) seem comparable to more expensive programs like improving teacher quality. For example, Rivkin et al. (2002) suggest that having a teacher at the higher end of the quality distribution raises student achievement by at least 0.11 standard deviations. Krueger and Whitmore (2001) found a class size reduction in the STAR program to lead to an increase in later test scores of 0.13 standard deviations.⁴² However, the costs of such programs are more substantive. They are simply not comparable to this apparent low-cost literacy hour program of changing teacher practice. Of course, the financial costs of the program may not reflect its true resource cost. For example, it might be argued that the measured costs do not reflect any extra effort the teacher might have to put in to learning and implementing the new teaching method. On the other hand, such effort may be fully accounted for in the cost of training. Furthermore, there are reports of a very positive response by teachers

⁴² However, the study by Finn and Achilles (1990) is probably more comparable to our study in that they consider the short-term effects of the programme on test scores (also for the STAR experiment). They find estimates of between 0.2 and 0.3 standard deviations on various tests. The class size reduction is substantial, from an average of 22 students to about 15 students.

(e.g. Fisher and Lewis, 1999, or Smith and Whitely, 2000), who find the learning objectives and structure of the literacy hour to provide a clear focus for what they teach.

One might also argue that the literacy hour takes teaching effort and resources away from other subjects and that this indirect cost effect (via substitution) should be taken account of in a cost-benefit calculation. However, given the guidelines in the National Curriculum, it seems likely that literacy was being taught in some form before the policy, for (at least) a commensurate time period. The inefficient way in which literacy was taught before the introduction of the 'literacy hour' (discussed above) suggests that its introduction may have freed up teacher time for the teaching of other subjects. Therefore, the literacy hour represents a change in how reading and writing are taught and not an increase in the time devoted to the subject. Furthermore, as argued above, there are several potential mechanisms through which the literacy hour could have generated positive spillovers on performance in other subjects.

Hence, the literacy hour seems to be cost effective. It represented a change in the content and organization of how literacy was taught — this enhanced pupil performance, but was not a change that involved much diversion of resources. However, to implement such practices requires knowledge of what works in the teaching of literacy. As discussed above, ideas used to construct the literacy hour were based on experience in other countries and on research. The value of this information is not included in the costs, yet it is manifestly important in generating the benefits.

6. Conclusions

In this paper, we have considered the potential for a change in the content and structure of teaching to impact on pupil performance. Our analysis is facilitated by the introduction of a literacy hour to English primary schools, through the National Literacy Project (NLP), which introduced the literacy hour to around 400 English primary schools in 1997 and 1998. We adopt an explicit treatment-control group approach, investigating what happened to pupil achievement in schools exposed to the literacy hour before and after the policy was introduced relative to pupils in schools that were not subject to the policy.

We find that reading and English Key Stage 2 levels rose by more in NLP schools between 1996 and 1998. Having subjected our identification strategy to a number of robustness checks, we are confident that this constitutes an NLP effect. We show there to be no trend difference in pupil achievement in NLP relative to comparison schools in the pre-policy period. We show no significant differential effect of NLP policy conditional on whether the school belonged to a group likely to be under greater pressure to improve. We also present some evidence to suggest that spillover effects of the NLP policy to other areas of the curriculum are likely. Since significant gender gaps in English performance exist (in favour of girls), we consider whether the literacy hour had a differential impact by gender and report evidence from some specifications that, at age 11, boys benefited more than girls. Finally, we show the benefits of the literacy hour to exceed the costs of the policy by quite a large margin.

These findings are of considerable significance when placed into the wider education debate about what works best in schools for improving pupil performance. They are also important for education policies in countries which have problems with their levels of literacy skills. This is true in the developed world context we study and, interestingly, in the other economics paper we know of which looks at a literacy and numeracy programme in a developing country context (i.e. the evidence for schools in India, as reported in Banerjee et al., 2005). As suggested in the Introduction, the research remit of economists has tended to be rather narrow in comparison with educationalists in this area. Our approach shows that one of the areas receiving much less attention from economists in the past, namely looking at the content and organisation of what is taught, matters for English and reading.⁴³ This is particularly important given that it will almost certainly be the case that the same teachers were teaching literacy before and after the introduction of the literacy hour.⁴⁴ Indeed, as the effects we identify come from a government policy aimed at improving literacy, the evidence we report suggests that public policy aimed at changing literary instruction can significantly raise pupil achievement and can do so in a highly cost effective manner.

⁴³ Of course, one should recognize that changing teaching methods more generally may not necessarily operate in a similar manner to the literacy hour. For example, Cohen and Hill (2000) report that changing teacher practice in mathematics teaching in the US only happens slowly and partially, and they argue that the inertia in the process is probably driven by the lack of a knowledge base amongst mathematics teachers. Understanding better whether the change was facilitated more easily because the literacy hour is a government policy, or whether it is something specific to the ability to change teaching methods for particular subjects, is an important area for future research.

⁴⁴ Moreover, we have looked at teacher turnover before and after the introduction of the literacy hour in NLP schools versus the control schools and find no significant change occurring.

Appendix A

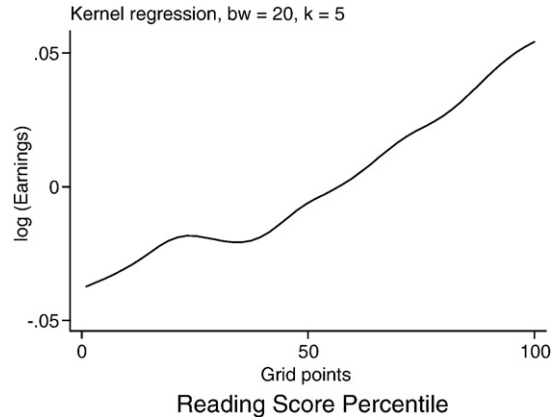


Fig. A1. Non-Parametric Earnings Regression (Nadaraya-Watson Estimator). Notes: This is the log(earnings)-reading score percentile relation from a specification comparable to column (2) of Table 7.

Table A1

NLP and 'Matched' Local Education Authorities for NLP Cities (close geographically and with a similar level of educational achievement in 1996)

NLP LEAs	Control LEAs
Inner London: Hackney, Islington, Lambeth, Southwark, Tower Hamlets, Newham, Waltham Forest	Inner London: Camden, Haringey, Lewisham, Wandsworth,
Sandwell	Walsall
Liverpool	Knowsley
Manchester	Rochdale
Sheffield	Rotherham
Newcastle	South Tyneside

Table A2

Probability of treatment (NLP=1), estimates for Table 2 specifications

	(1) Control group: all non-NLP schools	(2) Control group: non-NLP schools in adjacent LEAs
Percentile reading score	-0.016 (0.006)	-0.024 (0.011)
Percentile Maths score	0.009 (0.011)	-0.001 (0.016)
% achieving Level 4+ in English	0.088 (0.451)	0.142 (0.800)
% achieving Level 4+ in Maths	-0.576 (0.595)	-0.045 (0.919)
Prop. achieving Level 3+ in English	-0.261 (0.465)	-0.071 (0.880)
Prop. achieving Level 5+ in English	-0.202 (0.729)	-1.205 (1.241)
Prop. achieving Level 3+ in Maths	0.604 (0.552)	0.226 (0.902)
Prop. achieving Level 5+ in Maths	-0.305 (0.753)	-0.161 (1.181)
Religious school	0.071(0.072)	0.106 (0.131)
% eligible for free school meals	0.022 (0.002)	0.009 (0.004)
% special educational needs (no statement)	0.007 (0.003)	0.019 (0.006)
% special educational needs (with statement)	-0.023 (0.023)	0.008 (0.034)
Number of pupils/100	0.053 (0.028)	0.143 (0.055)
Pupil-teacher ratio	0.022 (0.011)	0.050 (0.018)
Prop. non-white	0.684 (0.150)	0.644 (0.281)
Prop. teachers not fully qualified	1.181 (0.901)	1.947 (1.846)
Ratio support staff to teachers	-1.279 (0.288)	-0.866 (0.551)
Prop. of teachers: graduates	0.069 (0.312)	0.781 (0.578)

(continued on next page)

Table A2 (continued)

	(1) Control group: all non-NLP schools	(2) Control group: non-NLP schools in adjacent LEAs
Prop. of teachers with — higher/1st/2nd degree	0.358 (0.311)	–0.193 (0.553)
Prop. female teachers	–0.258 (0.294)	–1.035 (0.480)
Constant	–3.105 (0.602)	–1.303 (0.906)
Observations	13371	796

Notes:

Probit model; Coefficients and standard errors reported.

All explanatory variables are 1996 values of school-level variables.

The following observations are dropped: NLP schools that are not in city areas; schools with missing information on pupils' reading scores or Maths scores (pupil-level data); schools that are missing from the Performance Tables in 1996.

Regression weighted by number of pupils in the school.

These regressions are used to predict the linear index of the propensity score for the sample of NLP schools and all non-NLP schools (column 1) and the sample of NLP schools and non-NLP schools in adjacent LEAs (column 2). Schools within the 'common support' are then selected for the difference-in-difference analysis that is reported in columns 3 and 4 of the tables in the text.

Table A3

Descriptive statistics for outcome variables in matched samples

	Number of schools	Pre-policy, 1996	Post-policy, 1997–98	Change	Difference-in-difference, NLP — controls
<i>A. Percentile reading scores</i>					
NLP schools, full sample	266	37.51	39.81	2.30 (.53)	–
NLP schools, adjacent LEAs	268	37.58	39.91	2.33 (.53)	–
Non-NLP schools, full sample	9798	48.57	49.46	0.89 (.09)	1.41 (.59)
Non-NLP schools, adjacent LEAs	518	46.08	46.56	0.48 (.43)	1.85 (.67)
<i>B. Percent level 3 or above KS2 English</i>					
NLP schools, full sample	266	78.86	83.72	4.86 (.57)	–
NLP schools, adjacent LEAs	268	78.89	83.83	4.94 (.57)	–
Non-NLP schools, full sample	9798	87.20	90.49	3.29 (.08)	1.57 (.62)
Non-NLP schools, adjacent LEAs	518	85.29	88.94	3.65 (.39)	1.29 (.70)
<i>C. Percent level 4 or above KS2 English</i>					
NLP schools, full sample	266	38.92	49.06	10.14 (.83)	–
NLP schools, adjacent LEAs	268	38.97	49.18	10.21 (.82)	–
Non-NLP schools, full sample	9798	55.11	63.01	7.90 (.13)	2.24 (.92)
Non-NLP schools, adjacent LEAs	518	51.28	58.98	8.70 (.63)	1.51 (1.04)
<i>D. Percent level 5 or above KS2 English</i>					
NLP schools, full sample	266	4.88	8.85	3.97 (.42)	–
NLP schools, adjacent LEAs	268	4.96	8.96	4.0 (.42)	–
Non-NLP schools, full sample	9798	10.58	15.14	4.56 (.10)	–0.59 (.46)

Table A4

Descriptive statistics for different groups in 1996

	Main samples			Matched sample: adjacent LEAs		Matched sample: all schools	
	(1) NLP schools	(2) Non-NLP schools in adjacent LEAs	(3) All non-NLP schools	(4) NLP schools	(5) Non-NLP schools in adjacent LEAs	(6) NLP schools	(7) Non-NLP schools
Religious school	.25	.31	.39	.25	.35	.25	.35
% eligible for free school meals	46.96 (17.82)	34.06 (18.38)	19.26 (16.17)	46.82 (17.71)	34.37 (18.08)	46.87 (17.47)	23.47 (16.48)
% special educational needs (no statement)	20.96 (11.19)	16.81 (10.17)	15.79 (9.72)	21.04 (11.15)	16.83 (10.00)	21.12 (11.11)	17.06 (10.02)

Table A4 (continued)

	Main samples			Matched sample: adjacent LEAs		Matched sample: all schools	
	(1) NLP schools	(2) Non-NLP schools in adjacent LEAs	(3) All non-NLP schools	(4) NLP schools	(5) Non-NLP schools in adjacent LEAs	(6) NLP schools	(7) Non-NLP schools
% special educational needs (with statement)	1.37 (1.77)	1.30 (1.51)	1.57 (1.74)	1.37 (1.78)	1.29 (1.46)	1.38 (1.78)	1.48 (1.61)
Number of pupils	327 (119)	292 (101)	250 (133)	328 (119)	293 (101)	328 (119)	278 (124)
Pupil–teacher ratio	21.24 (3.75)	21.11 (3.68)	20.54 (4.35)	21.22 (3.75)	21.14 (3.66)	21.22 (3.74)	21.50 (3.94)
Prop. non-white	.26 (.24)	.18 (.21)	.07 (.15)	.26 (.24)	.18 (.21)	.26 (.24)	.09 (.16)
Prop. teachers not fully qualified	.01 (.04)	.01 (.03)	.01 (.02)	.01 (.04)	.01 (.03)	.013 (.03)	.01 (.02)
Ratio support staff to teachers	.12 (.12)	.12 (.10)	.15 (.13)	.12 (.12)	.12 (.10)	.12 (.12)	.13 (.12)
Prop. of teachers: graduates	.56 (.17)	.52 (.16)	.48 (.22)	.56 (.17)	.52 (.16)	.56 (.17)	.52 (.19)
Prop. of teachers with — higher/1st/2nd degree	.41 (.17)	.38 (.17)	.34 (.20)	.41 (.17)	.38 (.17)	.41 (.17)	.38 (.18)
Prop. female teachers	.81 (.12)	.83 (.12)	.79 (.19)	.81 (.12)	.83 (.11)	.81 (.12)	.81 (.12)
No. schools	269	529	13573	268	518	266	9798

Note: This excludes outcome variables, where we report raw difference in difference estimates in Table 1 and Table A3.

In the regressions reported in the text, we always control for school fixed effects. Also, the 1996 values of all outcome variables (as well as Maths scores) are used for the statistical matching.

References

- Angrist, J., Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114, 533–575.
- Angrist, J., Lavy, V., 2001. Does teacher training affect pupil learning: evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics* 19, 343–369.
- Angrist, J., Lavy, V., 2002. New evidence on classroom computers and pupil learning. *Economic Journal* 112, 735–765.
- Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. *Review of Economics and Statistics* 60, 47–57.
- Banerjee, A., Cole, S., Duflo, E., Linden, L., 2005. Remedying education: evidence from two randomized experiments in India. *National Bureau of Economic Research Working Paper* vol. 11904.
- Beard, R., 2000. Research and the national literacy strategy. *Oxford Review of Education* 26, 421–436.
- Blundell, R., Dias, M., Meghir, C., Van Reenen, J., 2004. Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association* 4, 569–606.
- Card, D., Krueger, A., 1992. Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy* 100, 1–40.
- Cohen, D., Hill, H., 2000. Instructional policy and classroom performance: the mathematics reform in California. *Teachers College Record* 102, 294–343.
- Creemers, B., 1994. *The Effective Classroom*. Cassell, London.
- Dearden, L., Emmerson, C., Frayne, C., Meghir, C., 2005. Education subsidies and school drop-out rates. *Institute for Fiscal Studies, Working Paper* vol. 05/11.
- Department for Education and Employment, 1998. *The National Literacy Strategy: Framework for Teaching*. DfEE, London.
- Department for Education and Employment, 1999. *A Fresh Start: Improving Literacy and Numeracy*. DfEE, London.
- Finn, J.D., Achilles, C.M., 1990. Answers and questions about class size: a statewide experiment. *American Educational Research Journal* 27 (3), 557–577.
- Fisher, R., Lewis, M., 1999. Anticipation or trepidation? Teachers' views on the literacy hour. *Reading* 33, 23–28.
- Glewwe, P., Kremer, M., Moulin, S., Zitzewitz, E., 2000. Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *National Bureau of Economic Research, Working Paper*, vol. 8018.
- Godfrey Thompson Unit, University of Edinburgh, 1978. *Edinburgh Reading Test*. Hodder and Stoughton, Sevenoaks.
- Hanushek, E., 1997. Assessing the effects of school resources on student performance: an update. *Educational Evaluation and Policy Analysis* 19, 141–164.
- Hanushek, E., 2003. The failure of input-based schooling policies. *Economic Journal* 103, F64–F98.
- Heckman, J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65, 261–294.
- Heckman, J., Robb, R., 1985. Using longitudinal data to estimate age, period and cohort effects in earnings equations. In: Mason Feinberg, W., Feinberg, S. (Eds.), *Cohort Analysis in Social Research Beyond the Identification Problem*. Springer-Verlag, New York.

- Jacob, B., 2002. Where the boys aren't: non-cognitive skills, returns to schooling and the gender gap in higher education. National Bureau of Economic Research Working Paper vol. 8964.
- Krueger, A., Whitmore, D., 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from project star. *Economic Journal* 111, 34–63.
- Literacy Task Force, 1997a. *A Reading Revolution: How We Can Teach Every Child to Read Well*, London: The Literacy Task Force c/o University of London: Institute of Education.
- Literacy Task Force, 1997b. *The Implementation of the National Literacy Strategy*. Department for Education and Employment, London.
- Machin, S., Vignoles, A., 2005. What's the Good of Education? The Economics of Education in the United Kingdom. Princeton University Press.
- Machin, S., McNally, S., 2006. Gender and student achievement in English schools. *Oxford Review of Economic Policy* 21 (3), 357–372.
- Maynard, T., 2002. *Boys and Literacy: Exploring the Issues*. RoutledgeFalmer, London.
- Millard, E., 1996. *Differentially Literate: Boys, Girls and the Schooling of Literacy*. Falmer Press, London.
- Office for Standards in Education, 1996. *The Teaching of Reading in 45 Inner London Primary Schools*. A report by Her Majesty's Inspectors in collaboration with the LEAs of Islington, Southwark and Tower Hamlets. Ofsted, London.
- Office for Standards in Education, 1998. *The National Literacy Project: An HMI Evaluation*. Ofsted, London.
- Office for Standards in Education, 2002. *The Curriculum in Successful Primary Schools*. Ofsted, London.
- Plake, B., Impara, J. (Eds.), 2001. *The Fourteenth Mental Measurements Yearbook*. Buros Institute of Mental Measurements, Lincoln, NE.
- Rivkin, S., Hanushek, E., Kain, J., 2002. *Teachers, Schools and Academic Achievement*, Revised Version of Working Paper, vol. 6691. National Bureau of Economic Research. available at <http://edpro.stanford.edu/eah/eah.htm>.
- Rouse, C.E., Krueger, A.B., 2004. Putting computerized instruction to the test: a randomized evaluation of a 'scientifically based' reading program. *Economics of Education Review* 23, 323–338.
- Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
- Rosenbaum, P., Rubin, D., 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 516–524.
- Sammons, P., 1999. *School Effectiveness: Coming of Age in the Twenty First Century*. Swets and Zeitlinger, Lisses, The Netherlands.
- Scheerens, J., 1992. *Effective Schooling: Research, Theory and Practice*. Cassell, London.
- Scheerens, J., 2000. *Improving School Effectiveness*. IIEP, Paris.
- Stainthorp, R., 1996. Teaching reading in the primary classroom. In: Croll, P., Hastings, N. (Eds.), *Effective Primary Teaching: Research Based Classroom Strategies*. David Fulton, London.
- Stannard, J., Huxford, L., 2007. *The Literacy Game: The story of the National Literacy Strategy*. Routledge, London.
- Stenner, A.J., Smith, M., Burdick, D.S., 1983. Towards a theory of construct definition. *Journal of Educational Measurement* 20, 305–315.
- Slavin, R., Madden, N., 2000. Research on achievement outcomes of success for all: a summary and response to critics. *Phi Delta Kappan* 82, 59–66.
- Slavin, R., Madden, N., 2003. *Success for all/roots and wings: 2003 summary of research on achievement outcomes*. Johns Hopkins University, Center for Research on the Education of Students Placed at Risk, Baltimore.
- Smith, C., Whitely, H., 2000. Developing literacy through the literacy hour: a survey of teachers' experiences. *Reading* 34, 34–38.
- Teddle, C., Reynolds, D., 2000. *The International Handbook of School Effectiveness Research*. Falmer Press, London.
- West, A., Pennell, H., 2003. *Underachievement in Schools*. RoutledgeFalmer, London.
- White, J., 1996. Research on English and the teaching of girls. In: Murphy, P., Gipps, C. (Eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*. Falmer Press, London.