

Changing How Literacy Is Taught: Evidence on Synthetic Phonics[†]

By STEPHEN MACHIN, SANDRA McNALLY, AND MARTINA VIARENGO*

A significant number of people have very low levels of literacy in many OECD countries. This paper studies a national change in policy and practice in England that refocused the teaching of reading around “synthetic phonics.” This was a low-cost intervention that targeted the pedagogy of existing teachers. We evaluate the pilot and first phase of the national rollout. While strong initial effects tend to fade out on average, they persist for those with children with a higher initial propensity to struggle with reading. As a result, this program helped narrow the gap between disadvantaged pupils and other groups. (JEL I21, I24, I28)

Learning to read and write is an essential skill for modern life, yet a surprising fraction of adults in OECD countries have not yet mastered the basics. For example, about 15 percent of adults cannot fully understand the instructions on a bottle of aspirin.¹ These literacy problems are especially serious in England, where younger adults perform no better than older ones (Kuczera, Field, and Windisch 2016). In this context, it is unsurprising to see that basic literacy skills generate significant wage and employment gains in the labor market (Vignoles 2016). Poor literacy also drives low social mobility, since children from disadvantaged backgrounds are more likely to start school with lower literacy skills.

*Machin: Department of Economics and Centre for Economic Performance, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom (email: s.j.machin@lse.ac.uk); McNally: School of Economics, University of Surrey, Centre for Economic Performance and Centre for Vocational Education Research, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom (email: s.mcnally1@lse.ac.uk); Viarengo: Department of Economics, The Graduate Institute, Geneva and Harvard University, Center for International Development, Chemin Eugène Rigot 2, Geneva 1202, Switzerland (email: martina.viarengo@graduateinstitute.ch). We would like to thank Simon Brown, Marilyn Joyce, Michele Mann, Winter Rogers, Helen Walker, and Edward Wagstaff of the Department for Education for data and detailed information about the policy evaluated in this paper. We thank the NPD team at the Department for Education and Jon Johnson and Rachel Rosenberg of the Institute of Education for provision of data. We thank participants at conferences hosted by CESifo Economics of Education, the European Association for Labour Economics, the Association of Education, Finance and Policy, the Fondazione Rodolfo Debenedetti; and seminars at the Centre for Economic Performance LSE, the University of Sheffield, the Institute of Education, Lancaster University, and the IFAU in Uppsala. In particular, we would like to thank two very helpful referees, Sandra Black, David Figlio, and John Van Reenen for helpful comments. We thank Andrew Eyles for excellent research assistance. We are grateful to the Economic and Social Research Council for supporting this work. Also, Viarengo gratefully acknowledges the support received from the British Academy and the Royal Society in the framework of the Newton International Fellowship.

[†]Go to <https://doi.org/10.1257/pol.20160514> to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

¹The results of PIAAC (OECD 2013) show that 15.5 percent of adults have a proficiency of “level 1” or below. See online Appendix Table 2.2. http://skills.oecd.org/documents/SkillsOutlook_2013_Chapter2.pdf.

How can the situation be improved? There is a solid evidence base that teachers, and teaching methods, can matter both for literacy (e.g., Jacob 2017, Machin and McNally 2008, and Slavin et al. 2009) and for learning outcomes more generally (e.g., Aaronson, Barrow, and Sander 2007; Araujo et al. 2016; Chetty, Friedman, and Rockoff 2014a, b; and Hanushek, Rivkin, and Kain 2005). There is also a small, but growing, literature showing how interventions can boost teacher skills. For example, Taylor and Tyler (2012) show that teacher evaluation has an effect on teacher (and student) performance, and that the effect persists beyond the evaluation year.

In this paper, we investigate an intervention that changes how teachers teach. Specifically, teachers were trained to teach literacy using “synthetic phonics.” This technique involves a focus on sounding out letters and then blending sounds together to form words. While other studies have evaluated phonics-based programs in smaller tests (e.g., as summarized by Slavin et al. 2009 and, for example, the recent experimental evaluation by Jacob 2017), this study evaluates a pilot and a national program, where we can test both the change in pedagogy and teacher training together.² Put differently, our estimates directly measure the input of a pedagogy reform when implemented at scale, under a realistic policy setting. We show that the policy is low cost and effective, both in the pilot and in the first phase of the national rollout.

How reading should be taught in schools is hotly debated amongst education experts.³ Historically, experts have divided between proponents of “whole language” approaches versus supporters of “phonics” approaches. In essence, the “whole language” approach introduces language through context (e.g., through stories, picture books, etc.) whereas the “phonics” method builds literacy more systematically by connecting spelling patterns to sounds, assembling the building blocks of the language before stories are introduced. Only relatively recently has “systematic phonics” instruction been advocated in English-speaking countries, beginning in 2000 by the US National Reading Panel (National Reading Room 2000).⁴

In England, a review commissioned by the government in 2006 (Rose 2006) recommended “synthetic phonics” and was subsequently implemented in all schools. Relative to other countries, the policy adopted was narrower (Wyse and Goswami 2008) because it advocated a more extreme view of how exactly phonics should be taught (known as “synthetic phonics”) and then obliged all schools to implement the approach. However, a change in guidance alone was not deemed sufficient to ensure the adoption of the new approach in English schools. The government employed teaching consultants to disseminate good practice throughout local authorities and to give more intensive guidance to a small number of schools each year within

²Differences in the details of what programs consist of and how they are implemented make it difficult to compare our findings with earlier studies (which are generally much smaller scale; focus is on pedagogy rather than training; the counterfactual is also different from that considered here).

³See Mike Baker’s synopsis around the time of the 2005 controversy: <http://news.bbc.co.uk/1/hi/education/4493260.stm>.

⁴As defined by Torgerson, Brooks, and Hall (2006), systematic phonics means the teaching of letter-sound relationships in an explicit, organized, and sequenced fashion, as opposed to incidentally or on a “when-needed” basis. It may refer to systematic synthetic or systematic analytic phonics. The sounding out of letters and then blending to form words characterizes synthetic phonics. Analytic phonics involves inferring sound-symbol relationships from sets of words which share a letter and a sound.

local authorities. The staggered nature of training in the new approach across local authorities allows us to evaluate the effect of this reform on student achievement. Specially, we apply a difference-in-difference approach where we compare schools that were first exposed to intensive training from literacy consultants to schools that were exposed to this treatment in subsequent phases of the rollout (in other local authorities).

To briefly preview our results, we find there to be an instant effect of the program at age five, similar in magnitude to the initial effect of lower class size in Project STAR (Krueger 1999, Krueger and Whitmore 2001). These effects persist to age seven, even for cohorts exposed to the program at treatment schools but after the literacy consultant had finished working with the school. By the end of primary school (age 11), average test scores no longer differed significantly in treatment schools. This “fade out” is similar to other studies (e.g., Deming 2009, and Chetty et al. 2011) where effects reemerge in adulthood, although of course we have no direct evidence on long-term effects in this setting. However, the effects persist for those children with a higher initial propensity to struggle with reading, for instance those from disadvantaged backgrounds and/or those who are non-native speakers of English. Some of these children start school behind other students in measures of literacy. The program drives convergence in reading skills across the distribution. A back-of-the-envelope calculation suggests that the effect sizes for these disadvantaged groups are high enough to justify the costs of the policy.

This paper contributes to the literature in three ways. First, we estimate how teacher skills can be improved and the effect this has on students in the short and medium term. Such policies have not often been rigorously evaluated, especially in a context of both a pilot and a national policy. This is useful because it shows that, if effective, this is a low-cost intervention likely to be scalable. Second, our evidence contributes to discussions and research that focus upon the relative merits of targeted versus universal interventions.⁵ Third, this evidence shows a low-cost way to reduce inequalities within the education system between groups with a higher and lower propensity to start out their schooling as struggling readers.

The rest of the paper is structured as follows. In Section I, we explain the English education system, our data, and how phonics has been used in schools before and after the policy change in the mid-2000s. In Section II, we outline our conceptual framework and empirical strategy. In Section III, we discuss our results, firstly in the context of an “event study” for five-year-olds, then based on an analysis of program effects as relevant cohorts progress through the school system (at age 5, 7, and 11). Then we evaluate whether the policy has a heterogeneous effect depending on

⁵There are sizable bodies of research that discuss the advantages and disadvantages of targeted and universal interventions in education, and these are particularly prominent in the areas of preschool provision and of education policies in developing countries (see, for example, Education Sector 2006 on preschool and Behrman 2010 on development research). Areas of particular importance stressed in targeted versus universal debates are whether higher or lower economic returns accrue under each, differences in program quality, the implications for total costs, the extent of displacement of private spending, and administrative costs that determine eligibility under targeting. In our setting, where phonics were first targeted and then rolled out nationally, a particular focus of interest is on whether a differential impact on pupil performance is seen (i.e., whether returns differ). However, even in the national program, only a minority of schools were targeted for intensive support by the literacy consultant (which is what we evaluate).

whether the student is classified as disadvantaged or a non-native English speaker. We consider effects on other outcome measures and also whether the policy affected the number of teachers, the pupil-teacher ratio, or pupil enrollment. We also conduct a robustness check to examine whether results at age 11 are sensitive to imputation. We conclude in Section IV.

I. The English Education System

A. Assessment and Data

In England, children must start school the September after they turn 4 years old, and there is essentially no grade repetition. Primary schools last up to age 11, after which children make the transition to secondary schools where they stay up to (at least) the age of 16, which was the end of compulsory education up to very recently (and during the time of this study). In this paper, we focus on the primary phase of education.

The national curriculum in England is organized around “Key Stages.” In each “Key Stage,” there are various goals made out for children’s learning and development, and it ends with a formal assessment: in primary school, this is the Foundation Stage at age 5 and Key Stage 1 at age 5–7 and Key Stage 2 at age 7–11. At ages five and seven, pupils are assessed by their own teachers. However, there is extensive guidance on how the assessment should be made, and it is moderated. At age 11, assessments are set and marked externally.

At age five, we focus on the assessment of communication, language, and literacy. The first year for which this information is produced is 2003. Between 2003 and 2006, the assessment was only done for a 10 percent child-level sample.⁶ From 2007 onwards, all children in England have been assessed in this way. At age seven, we focus on teacher assessments for reading, although we do examine whether there are effects on other subjects (described in Section IIID). Students are given a “level” (i.e., there is no test score as such). We transform national curriculum levels achieved in reading, writing, and mathematics into point scores using Department for Education point scales. At age 11, we use national tests of English and math.

The National Pupil Database (NPD) is a census of all pupils in the state system in England available from 2002 onwards. This covers the vast majority of primary age children, as only a small percentage attend private school at this age. We use data between 2003 and 2012 and exclude a small number of independent and special schools from the analysis.⁷

The NPD gives information on all the assessments previously described and basic demographic details of pupils—such as ethnicity, deprivation (measured by whether

⁶Between 2002–2003 and 2005–2006 local education authorities (LEAs) were only required to submit a 10 percent sample of complete individual pupil FSP data. Since 2006–2007, LEAs have submitted complete individual pupil FSP data. The proportion of students in treatment schools is identical when we move from the 10 percent sample to the whole cohort. Furthermore, the characteristics of students in the 10 percent sample are similar to those pertaining to the whole cohort if we consider characteristics available in census data (such as ethnicity, gender, whether English is spoken as a first language, and eligibility for free school meals).

⁷We start our analysis from 2003 onwards because this is the year the age 5 assessment was introduced.

they are eligible to receive free school meals), gender, and whether or not English is their first language. As we know the school attended, we can control for school fixed effects in our analysis—and we can track students if they change schools.

For the period covered by our study, schooling was organized at the local level into local education authorities (of which there are 152). Schools are largely self-governing and the main functions of the local authority are in building and maintaining schools, allocating funding, providing support services, and acting in an advisory role to the head teacher regarding school performance and implementation of government initiatives. The Department for Education have provided us with details of the local authorities and schools involved in the initial phonics pilot and how support was phased in across local authorities and schools in subsequent years. We describe this later and give more detail in section C of the online Appendix. First, we discuss the use of phonics in schools.

B. The Use of Phonics in Schools

There are two main approaches to learning the alphabetic principle: synthetic phonics and analytic phonics. The former is used in Germany and Austria and is generally taught before children are introduced to books or reading. It involves learning to pronounce the sounds (phenomes) associated with letters “in isolation.” These individual sounds, once learnt, are then blended together (synthesized) to form words. By contrast, analytic phonics does not involve learning the sounds of letters in isolation. Instead, children are taught to recognize the beginning and ending sounds of words, without breaking these down into the smallest constituent sounds. It is generally taught in parallel with, or sometime after, graded reading books, which are introduced using a “look and say” approach.⁸ One of the reasons the debate between educationalists is so divisive is because those advocating “synthetic phonics” argue this should be taught before any other method. The other side argue that one size does not fit all, and it is possible to teach other aspects of reading at the same time.⁹

Up to 2006, the English literacy strategy recommended analytic phonics as one of four “searchlights” for learning to read in the National Literacy Strategy (in place since 1998). However, a review of this approach was prompted by a study in a small area of Scotland (Clackmannanshire), which claimed very strong effects for children taught to read using synthetic phonics (Johnston and Watson 2005). The outcome of the review was the “Rose Report” (Rose 2006), after which government guidelines were updated to require the teaching of synthetic phonics as the first and main strategy for reading. There was a detailed program called “Letters and Sounds: Principles and Practice of High Quality Phonics,” which teachers were expected

⁸Children are typically taught one letter sound per week and are shown a series of alliterative pictures and words, which start with that sound, e.g., car, cat, candle, caste, and caterpillar. When the 26 initial letter sounds have been taught, children are introduced to final sounds and to middle sounds. At this point, some teachers may show children how to sound and blend the consecutive letters in unfamiliar words.

⁹See also the paper by Kearney and Levine (2016) presenting evidence of a positive impact on educational outcomes through the early school years (including reading) for children in the United States who were exposed to the television program Sesame Street.

to follow (Primary National Strategy 2007). This is summarized (as in Wyse and Goswami 2008) in the online Appendix (Box A1).

At the same time as the review was taking place (before it was published), there was a pilot in 172 schools and nurseries that was principally to give intensive training to teachers on the use of synthetic phonics in early years. After the Rose report, training was rolled out to different local authorities (LA). The LAs were given funding for a literacy coordinator who would work intensively in about ten schools per year but also disseminate best practice throughout the LA by offering courses. The program was rolled out gradually to different local authorities—only reaching all local authorities by the school year 2009–2010. Thus, it was not anticipated that all schools would update their early years' teaching overnight, even though the government guidelines had changed.¹⁰

More specifically, "The Early Reading Development Pilot" was introduced in 2005 to test out the pace of phonics teaching and, in terms of timing, ran alongside the Rose Review.¹¹ This involved 18 local authorities (LAs) and 172 schools and settings in the school year 2005–2006.¹² The "Communication, Language, and Literacy Development Programme" (which we refer to as the national program) was launched in September 2006 to implement the recommendations of the Rose Review. This national program replaced the pilot in the original 18 LAs and a further 32 LAs were invited to join, each receiving funding for a dedicated learning consultant. The next phase of the national program was introduced from April 2008. This involved another 50 LAs. Then the last third of LAs (i.e., another 50) joined the national program in April 2009. The essential model of support was similar across the pilot and the national program. We evaluate them separately because the pilot was implemented before national guidelines had been changed (which potentially affects all schools).

Each local authority received £50,000 to support the appointment of a specialist consultant with a further £15,000 to allocate to schools. The consultant was to provide intensive support to at least ten schools per year, before moving on to another ten schools in a subsequent year. In these schools, the consultant worked mainly in the Reception year (first year of school, when children are aged 4–5) and Year 1, but also in Year 2 and nursery. The support involved an initial audit and assessment visit to help schools get started on the program. This included drawing up an action plan, making observations and detailed assessments of children. In a second visit, the consultant would model or co-teach the adult-led activity or the discrete teaching session and help teachers and practitioners to plan further learning and teaching opportunities over the following few weeks. At this and subsequent visits, the consultant would work with teachers, practitioners, and leadership teams to review children's learning and identify the next steps for teaching. The consultant also provided

¹⁰In 2010, a government spokesman implied that the "Communication, Language, and Literacy Programme" was necessary to enable schools to make the necessary changes: <http://www.theguardian.com/education/2010/jan/19/phonics-child-literacy>.

¹¹It was requested by Andrew Adonis, the then Minister of State for education, in response to the findings of the Select Committee on the teaching of early reading.

¹²As some preschool settings were involved (i.e., nurseries), we have fewer primary schools than this in our data—roughly 160 schools. However, it has been confirmed that the Reception year in these primary schools was the main initial focus for this policy.

support to other schools in the LA, usually through the provision of courses which were offered to all schools.

C. Selection of Schools and Local Authorities

The selection of local authorities and schools into the pilot and subsequent national rollout was not done in a systematic way according to specific criteria. Online Appendix section C details what we were able to learn about this process from the Department of Education. Of use here is that they tried to incorporate a mix of local authorities across different regions. Thus, we shouldn't think of the program as being concentrated in particular areas. The decision about what schools to target for intensive support was made by each local authority and again not made according to specific, measurable criteria (see online Appendix Section C). Although the guidance suggests that LAs should look for schools in need of improvement for literacy, it also suggests the need to look for potential to deliver (i.e., they were advised not to target schools where the quality of teaching in early years was not satisfactory).

We are interested in evaluating the effects of the pilot (i.e., schools in 18 LAs that were initially selected for intensive treatment) and schools in the first phase of the national rollout (i.e., again those selected for intensive treatment) relative to a control group. The fact that we do not have clear, transparent criteria for selection of schools for "intensive support" means that it is useful to use selected schools in subsequent phases of the national rollout (in other local authorities) to define the control group. Schools in the treatment and control groups were deliberately selected for "intensive support"—and thus have more in common (for the purposes of evaluating this policy) than all those schools that were not selected.¹³

The pilot schools received treatment in 2005–2006, and the first phase of the national program received treatment in 2006–2007. Schools in the control group received the treatment between 2008–2009 and 2009–2010. Further details on the description of the pilot, first phase, and control groups are shown in Table 1, and the actual LAs involved are listed in online Appendix Table A1.

Our approach is a difference-in-difference analysis, comparing outcomes before and after the policy was introduced (conditional on other attributes of schools and pupils). The credibility of the methodology rests on whether these groups show parallel trends in outcome variables pre-policy (later, we show that they do) rather than whether they match closely based on observable characteristics at a point in time.

In Table 2, we show key characteristics of different groups of schools in the pre-pilot year (2004–2005). This is designed to understand the selection process of local authorities and schools. Columns 1–6 show the following groups: all schools; schools in the pilot; nonselected schools in the 18 pilot local authorities; schools in the first phase of the national rollout (within 50 local authorities); schools that were

¹³ Other reasons for not using nonselected schools in treated local authorities as a control group is that the literacy consultant was supposed to disseminate best practice throughout the local authority, as discussed in Section IB. When we do use these schools as a control group, estimated effects are smaller but, for the most part, qualitatively similar to the current analysis. Results are available on request.

TABLE 1—DESCRIPTION OF GROUPS

Groups	Phonics program	Local authority coverage	Entry	Birth cohort of students first exposed to program	Year of age 5 assessment	Year of age 7 assessment	Year of age 11 assessment
Treatment group 1	Pilot, EDRp	Schools in 18 LAs	2005–2006	2001	2006	2008	2012
Treatment group 2	First phase, CLLD	Schools in same 18 LAs + 32 new LAs	2006–2007	2002	2007	2009	2013
Control group		Schools in next 50 LAs	2008–2009 and 2009–2010	2004	2009	2011	2015
		Schools in next 50 LAs	2009–2010	2005	2010	2012	2016

Notes: Schools in the first 50 LAs (i.e., pilot and first phase) did come into the scheme in subsequent years. These schools are not included in the analysis.

TABLE 2—PRE-POLICY (2005) SUMMARY STATISTICS FOR GROUPS OF SCHOOLS

	All primary schools (1)	Treatment group 1: Pilot (2)	Nonselected schools in 18 LAs of pilot (3)	Treatment group 2: First phase (4)	Nonselected schools in 50 LAs of first phase (5)	Control group (6)	<i>p</i> -value: (2)–(6) (7)	<i>p</i> -value: (4)–(6) (8)
Age 5 communication, language, and literacy score	0.000	-0.091	-0.014	-0.364	-0.006	-0.250	0.049	0.006
Age 7 reading score	0.000	-0.091	-0.059	-0.286	-0.023	-0.196	0.002	0.000
Proportion entitled to free school meals	0.181	0.263	0.230	0.340	0.210	0.273	0.563	0.000
Proportion native English speakers	0.880	0.817	0.860	0.756	0.884	0.823	0.814	0.000
Proportion white British/Irish	0.791	0.694	0.763	0.641	0.776	0.722	0.348	0.000
Number of schools	16,429	164	2,264	523	5,500	1,007	1,171	1,530

Notes: Treatment (columns 2 and 4) and control groups (column 6) in bold. The age 5 and age 7 scores are standardized to have a mean of 0 and a standard deviation of 1.

not selected for intensive support as part of the first phase of the national rollout, though within the same 50 LAs; and schools in subsequent phases of the national rollout for the other 100 local authorities that entered the program between 2008 and 2010. Thus, columns 2 and 4 show statistics for the two treatment groups of interest (the pilot and the first phase of the national rollout), and column 6 shows statistics for the control group.

We show summary statistics for our main outcome variables at age five and seven.¹⁴ They are the communication, language, and literacy score at age five and the reading score at age seven (as described in Section IA). Both measures are standardized to have a mean of zero and a unit standard deviation. We also show three important demographic variables:¹⁵ the proportion of children eligible to receive

¹⁴ In the analysis, we link age 7 outcomes to age 11 outcomes for students in the treatment and control group, respectively. The policy only applies to children during Key Stage 1—and some children move school between Key Stages 1 and 2 (i.e., between age 7 and 11).

¹⁵ Apart from outcome variables measured at age 5 and 11, all summary statistics relate to children of age 7 in 2005 (the pre-pilot year).

free school meals (an indicator of socio-economic disadvantage); the proportion of native English speakers; and the proportion of children who are classified as “White British or Irish.”

We learn from the table that within the two treatment groups (i.e., columns 2 and 4), schools selected for the treatment are (on average) lower performing than other schools within the local authorities of interest (i.e., as shown in columns 3 and 5). They also tend to include a higher proportion of disadvantaged children, a lower proportion of native English speakers, and a lower proportion of children classified as “White British/Irish.” If we consider the local authorities selected for the treatment based on their schools not selected for intensive support in the first year (i.e., columns 3 and 5), they do not look too different from the national average (column 1) on most of the reported indicators, although they are a little more disadvantaged (particularly the pilot local authorities). The control group (column 6) is a lot more similar to schools in the treatment groups (columns 2 and 4) compared to schools that were not selected for intensive support in treatment local authorities (columns 3 and 5) and to the overall sample. However, there are still significant differences at baseline between treatment and control groups, and it will be important to establish that there is no differential pre-trend in outcome variables. We show this in the context of an “event study” in Section III (see Figure 1) and in a regression context (e.g., online Appendix Table A2). These approaches show very clearly that the parallel trends assumption is reasonable, and there is no pre-policy differential effect of being in a treated school before the policy was introduced. Before we show these and other findings, we next turn to explain the conceptual framework and empirical strategy.

II. Conceptual Framework and Empirical Strategy

In our empirical analysis, we make use of the differential timing of the phasing-in of intensive support to schools as a “natural experiment” to identify the causal effect of teacher training in the new pedagogical approach. As previously discussed, we use two treatment groups of schools whose teachers were trained to deliver phonics teaching: the initial schools in the pilot that were set up to inform the Rose Review; and schools in the first phase of the national rollout. The control group consists of schools that were selected for intensive support as soon as their local authorities were incorporated to the national rollout (three years after the pilot and two years after the first phase).

Denoting schools ever treated by phonics exposure and control schools by a binary indicator variable *Phonics* (equal to 1 for treatment schools and 0 for control schools), we estimate the impact of the intervention in the following difference-in-difference equation:

$$(1) \quad A_{ist} = \alpha_1 \text{Phonics}_s \times I(t \geq p) + \alpha_2 Z_{ist} + \alpha_3 X_{st} + \tau_t + \gamma_s + \varepsilon_{ist}$$

where $I(t \geq p)$ is an indicator function equal to one for time periods after time p when the phonics programs were introduced, and zero prior to that. This research design enables us to estimate the effect of training teachers in the new pedagogical

approach (*Phonics*) on student attainment (A) for student i , in school s affected by the treatment at a given time t . The coefficient of interest is the difference-in-difference coefficient estimate α_1 . The specification in equation (1) includes school fixed effects (γ_s), which control for the baseline effect of being a “treated school” and any other school-level characteristics that do not change over time (including time invariant skills of the teacher workforce). We control for a set of time dummies (τ_t) and a vector of personal/family characteristics (Z) which includes gender and ethnicity, whether he/she is a native speaker of English, whether he/she is eligible to receive free school meals (an indicator of poverty), and whether he/she receives a statement of Special Educational Needs. We also control for a vector of peer characteristics in the student’s school (X), namely the percentage of students in the year group according to each of the previously named personal characteristics.

We estimate variants of equation (1) under the (plausible) assumption that the phonics treatment is the only relevant time-varying factor that affects the treated schools relative to the control schools. In fact, the phased introduction makes it highly unlikely that another shock to teaching skills occurred at the same time, and thus, we have a coherent research design for studying what is a relatively unusual policy in that it is inexpensive but has significant potential to reduce literacy inequalities in the early years of school.

Since we are interested in estimating effects as the affected cohorts age (through their schooling), we set most regressions up as interactions with birth cohorts rather than year. Thus, we estimate α_1 when the treatment cohort is at age 5, 7, and 11 relative to control cohorts. For the pilot, this is the cohort of children born in 2001 whereas for the first phase of the national rollout, this is the cohort of children born in 2002. The treatment was initially focused on the youngest age group but could have an effect on multiple age groups within the same year (i.e., children aged between five and seven). The cohort of children born in 1998 is completely unaffected at any stage. However, we show a full set of treatment \times cohort interactions for those born between 1998 and 2001 (and 2002 when analyzing the effect of the national rollout).

Finally, we look at heterogeneity by selecting the 1998 birth cohort and the two main “treatment” cohorts of interest (2001 for the pilot; 2002 for the national program). To do so, we estimate the following, more general, difference-in-difference specification containing interactions of the policy impact with pupil-level variables of interest:

$$(2) \quad A_{ist} = \beta_{11} \text{Phonics}_s \times I(t \geq p) \times D_{ist} + \beta_{12} \text{Phonics}_s \times I(t \geq p) \times (1 - D_{ist}) \\ + \beta_2 Z_{ist} + \beta_3 X_{st} + \tau_t + \gamma_s + \varepsilon_{ist}.$$

In equation (2), the pupil characteristic of interest in terms of heterogeneity is represented as D . More precisely, we estimate whether there is a differential treatment effect ($\beta_{11} \neq \beta_{12}$) according to whether the student is classified as: being eligible to receive free school meals and a native English speaker. Again, we estimate this regression as the student progresses through the school system (at ages 5, 7,

and 11). We set the regressions up such that the treatment effect is separately identified for each group (i.e., “free school meal” and “non-free school meal” children; native and non-native speakers of English). In a final specification, we estimate the full set of two-way interactions with treatment.

III. Results

A. Event Study

We first consider whether the policy had an effect by looking at results from an “event study” of five-year-olds. They were the initial target of the intensive support in schools, and there is no ambiguity about the year in which we should start to see an effect. It should be the two different years in which the policy was introduced to the two treatment groups: schools in the pilot and schools in the first phase of the national rollout. Furthermore, we should expect the effects to decline once the control group schools receive the treatment.

Having estimated equation (1), the estimated coefficient for the treatment effect (α_1) and the associated 95 percent confidence interval are plotted in Figure 1 for the pilot treatment versus control and the first-phase treatment versus control.¹⁶ The regression estimates that underlie the figure are shown in online Appendix Table A2. The dependent variable is the standardized score for communication, language, and literacy at age five. The figure shows zero effect for the two available pre-policy years for pilot versus control and the three available years for first phase versus control. However, as soon as the treatment is introduced, the effect jumps to over 0.2 standard deviations in both cases. Note that the year “ t ” is different for the pilot and first-phase groups, yet the effect sizes are similar (and the control group is the same). Furthermore, the pilot treatment stays high (at least 0.2 standard deviations) for each year until the control group receives the treatment (at $t + 3$), when the effect size falls and is no longer statistically different from zero. The pattern is similar for the first-phase treatment, except that the effect size does not fall as quickly when the control group enters the program at $t + 2$ (and also remains statistically different from zero).¹⁷ However, within two years of the control group entering, the effect size is halved.

The fact that the treatment effect stays high up until the control schools enter the program (and for some time after that in first-phase schools) shows that any effect of the program is not simply down to the presence of the literacy consultant in the school. The intensive support was only on offer for one year (except in cases where schools had difficulties). Thus, the effect sizes reflect the effect of the training and not the presence of the trainer.

¹⁶Equation (1) shows a difference-in-difference specification. Figure 1 shows estimated coefficients from an analogous strategy in an “event study” context where treatment is time-varying.

¹⁷We identify the effect of the policy through the staggered nature of the intervention. Inclusion and exclusion of time-varying school and pupil characteristics makes little or no difference to estimated effects of the treatment. When we include a measure of the number of teachers (as an attempt to proxy potential teacher turnover), this makes no difference to the results.

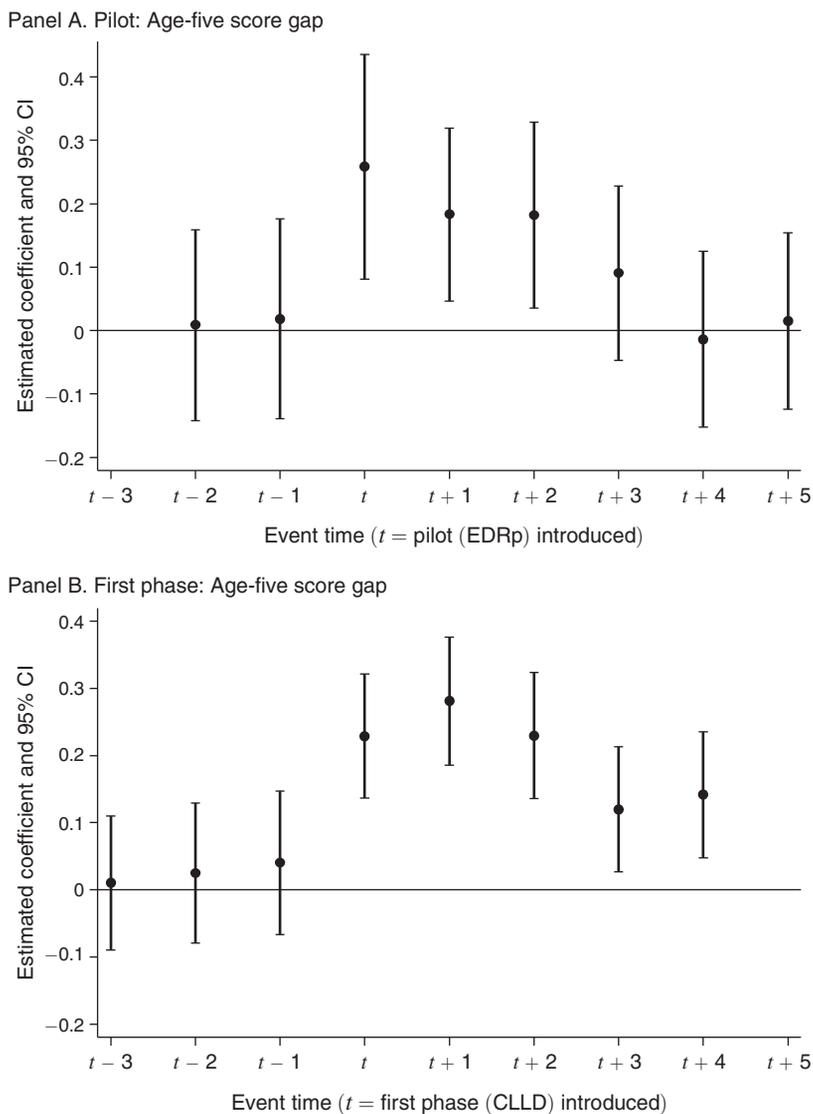


FIGURE 1. AGE-FIVE READING SCORES—TREATMENT \times YEAR COEFFICIENTS
(controlling for all observable variables)

B. Main Results by Cohort

Table 3 shows estimated effects of the policy for the pilot treatment (panel A) and the first phase of the national rollout (panel B) relative to the control group for different birth cohorts as they progress through the school system. The omitted category is the 1998 birth cohort. In each case, the cohorts fully exposed to the treatment throughout their entire early phase of primary education (i.e., age 5–7) and observable at age 11 are the 2001 cohort (for the pilot) and the 2002 cohort (for first phase). However, other birth cohorts are partially treated. For example,

TABLE 3—PILOT AND FIRST-PHASE TREATMENT ON READING, AGES 5, 7, AND 11

	Age 5 (1)	Age 7 (2)	Age 11 (3)
<i>Panel A. Pilot versus control</i>			
Treatment × 1999 birth cohort (Treatment: Only age 7)	0.005 (0.077)	0.037 (0.021)	0.003 (0.028)
Treatment × 2000 birth cohort (Treatment: Age 6–7)	0.072 (0.081)	0.040 (0.025)	−0.001 (0.027)
Treatment × 2001 birth cohort (Full treatment: Age 5–7)	0.298 (0.094)	0.075 (0.024)	−0.018 (0.031)
Additional controls	Yes	Yes	Yes
Sample size	17,279	191,342	163,272
Number of schools	1,185	1,217	1,217
	(4)	(5)	(6)
<i>Panel B. First phase versus control</i>			
Treatment × 1999 birth cohort (No treatment: Placebo)	0.009 (0.050)	−0.015 (0.015)	−0.024 (0.019)
Treatment × 2000 birth cohort (Treatment: Only age 7)	0.015 (0.053)	0.031 (0.016)	−0.016 (0.018)
Treatment × 2001 birth cohort (Treatment: Age 6–7)	0.033 (0.054)	0.046 (0.017)	0.021 (0.019)
Treatment × 2002 birth cohort (Full treatment: Age 5–7)	0.217 (0.047)	0.073 (0.017)	0.019 (0.019)
Additional controls	Yes	Yes	Yes
Sample size	82,495	309,769	268,565
Number of schools	1,568	1,598	1,598

Notes: Baseline is the 1998 birth cohort (who undertook the ages 5, 7, and 11 assessments in 2003, 2005, and 2009, respectively). Each column shows selected coefficients from a separate regression. The dependent variable at age 5 in columns 1 and 4 is the (teacher assessed) standardized score in communication, language, and literacy. The dependent variable at age 7 in columns 2 and 5 is the (teacher assessed) standardized score in Key Stage 1 reading. The dependent variable at age 11 in columns 3 and 6 is the pupil's (externally assessed) standardized test score in reading. The 2001 and 2002 birth cohorts (in bold) are the first cohorts to have received the treatment throughout their education for the pilot and first phase, respectively. For the pilot, the 2000 birth cohort received the treatment in Year 1 (at age 6). The 1999 birth cohort received the treatment in Year 2 (at age 7). For the first phase, the 2001 cohort received the treatment in Year 1 (at age 6). The 2000 birth cohort received the treatment in Year 2 (at age 7). Additional controls are: year dummies; school fixed effects, student gender, ethnicity; whether speaks English as an additional language; whether eligible to receive free school meals, whether receives a statement of special educational needs; and percent of students in the year group by: gender, ethnicity, whether speaks English as an additional language, whether eligible to receive free school meals, whether receives a statement of special educational needs. Standard errors are clustered by school. Untreated groups are in italics.

the cohort born in 2000 is potentially affected from the age of 6 if in a pilot school and at the age of 7 if in a school receiving the intervention in the first year of the national rollout. The cohort born in 1999 might be affected at the age of 7 if in a pilot school. Since the mandate of the literacy consultant was to focus mainly on the younger children, one would not expect the effect to be as strong on those students first exposed to the program at age six and seven compared to those first exposed at age five.

We look at effects at the ages of 5, 7, and 11. In each case, the dependent variable is the standardized test score and so the reported estimates can be viewed in units of

a standard deviation, σ . The data for those undertaking Key Stage 1 assessments at age 7 are linked to the same individuals' assessments at age 11. Thus, we follow the student exposed to the "treatment" whether or not he/she changes school between the age of 7 and 11.¹⁸ In any school, the "treatment" is only defined by what happens between the age of five and seven.

Focusing on the results for the cohort that receives the treatment throughout their early schooling and observable at age 11 (i.e., the 2001 cohort for the pilot and the 2002 cohort for the first phase of the national program), Table 3 shows that the initial effect on age 5 results is very high (as also shown in Figure 1). As columns (1) and (4) show, it is close to 0.30σ for the pilot and 0.22σ for first-phase schools. By the age of seven, the effect of the policy has reduced by at least two-thirds (although the test score is more coarsely defined at age seven and therefore not exactly comparable to that at age five). However, it is still of a reasonable size of about 0.07σ for both the pilot and first-phase schools and is statistically significant. However, at age 11, the results suggest an effect that is close to zero.

The age 7 results (i.e., columns 2 and 5) show an effect which seems to increase with successive cohorts (although not always significantly). The reason may be because earlier cohorts were fully exposed to the policy and because the mandate of the literacy consultant was to focus mainly on the younger age groups. For the pilot, the effect goes from 0.04σ to 0.04σ to 0.08σ from first exposure to the program at ages 7, 6, and 5, respectively. For first-phase schools, the effect goes from 0.03σ to 0.05σ to 0.07σ at these same ages. This also suggests an impact of the program on children when the intensive support actually stops (as it was only supposed to last one year in treatment schools; after which the literacy consultant would move on to other schools in the local authority). Thus, we can also infer that the effect is coming from training in the use of the program—not from the fact of having a consultant come to the school. However, the effect never persists to age 11.

A final insight from Table 3 comes from an additional placebo test for the first phase: did the policy appear to have an effect for cohorts to which it was not exposed?¹⁹ Of course, this might indicate differential trends in treatment and control schools. The coefficients shown in panel B of Table 3 for the "effect" of the treatment for the cohort born in 1999 reflect such a placebo test. This cohort could not have been affected by the intervention because the program was only introduced to their schools when they were eight-years-old. At age five and seven, the coefficients are close to zero and statistically insignificant, suggesting no evidence of differential pre-policy trends.

From this analysis, we see that average effects of the intervention are similar across schools participating in the pilot and in the first phase of the national rollout. Indeed, it was not clear a priori whether effects should be stronger in one group or in the other. We also found no significant differential effect according to whether the

¹⁸We do not do this between the age of 5 and 7 because the age 5 test score is only available for a 10 percent sample of schools between 2003 and 2006. Instead, treatment and control schools are separately merged to the age five and seven data. However, as a robustness check, we have followed the same (smaller) age 5 sample through time and estimated effects at 5, 7, and 11. The results (available on request) are qualitatively similar.

¹⁹Results of "pre-policy" exposure are also shown to be zero in Figure 1 and Table A1.

Schools Inspectorate had previously classified schools as “outstanding,” “good,” or “satisfactory.”²⁰

C. Heterogeneous Effects

We next consider whether the policy has a heterogeneous effect. We might expect any effects of the program to be stronger for pupils with characteristics that are likely to make them lower achieving on average in reading when they first go to school (like being from a low income background, or not speaking English as a first language). We can look at this at age of school entry using the Millennium Cohort Study (MCS). This longitudinal study began in the years 2000 and 2001 and follows around 20,000 children from birth.²¹ We look at the age-five wave to study test score differences between groups at about the time of school entry.

Online Appendix Table A3 shows regressions of age-five cognitive test scores (measuring “naming vocabulary,” “pattern construction,” and “pattern similarity”) on indicators of whether MCS cohort members are eligible for free school meals and whether their home language is not English.²² As the estimates show, both of these groups enter school at age five with significantly lower test scores, especially in vocabulary skills. The difference in the vocabulary score for native and non-native speakers of English is close to 1 standard deviation whereas it is about 0.6 standard deviations for those from poor and nonpoor family backgrounds (as measured by eligibility to receive free school meals). This vocabulary deficit at time of school entry clearly places children with these characteristics at a significant literacy disadvantage especially if such deficits hold them back. Other measures of cognitive ability (pattern construction and pattern similarity) also show large and significant differences between these groups—but the gap is much smaller than that for vocabulary skills. So it is interesting to ask whether intensive training in the use of phonics has a differential impact across these groups, both in terms of when they were first exposed to the program and at later ages.

In Table 4, we examine the impact of the treatment for the group most strongly impacted by the policy (i.e., receiving the treatment from age five onwards) relative to the control group. Thus, the first three columns show impacts for the 2001 cohort relative to the 1998 cohort for pupils in pilot schools and the next three columns show impacts for the 2002 cohort relative to the 1998 cohort for pupils in the first phase of the national rollout. In each case, we show heterogeneous effects of the 2 treatments at ages 5, 7, and 11 by estimating equation (2).

The upper panel A compares the effect of the treatment for native and non-native English speakers. For non-native English speakers, the effect size is stronger at age five for the pilot (though not statistically different from the effect for native English

²⁰The intervention seemed to work in all these groups apart from schools classified as “unsatisfactory.” However, there were very few schools within this group.

²¹See Hansen, Joshi, and Dex (2010) for more detail on the MCS data and a range of studies of cohort members up to age five.

²²Precise definitions of the three tests are given in the descriptive review of the age five (third wave) of the MCS in Jones and Schoon (2008). They are aimed to capture cognitive skills at age in verbal, pictorial reasoning, and spatial abilities (as in Elliott, Smith, and McCullough 1996 or Hill 2005).

TABLE 4—HETEROGENEITY IN ESTIMATED TREATMENT EFFECTS ON READING BY NON-NATIVE SPEAKER STATUS AND FREE SCHOOL MEALS ELIGIBILITY

	Pilot versus control (cohorts 1998 and 2001)			First phase versus control (cohorts 1998 and 2002)		
	Age 5 (1)	Age 7 (2)	Age 11 (3)	Age 5 (4)	Age 7 (5)	Age 11 (6)
<i>Panel A. Speech nativity</i>						
Native speaker	0.225 (0.083)	0.052 (0.024)	−0.045 (0.031)	0.211 (0.050)	0.061 (0.018)	0.001 (0.020)
Non-native speaker	0.567 (0.277)	0.134 (0.051)	0.045 (0.063)	0.201 (0.081)	0.113 (0.028)	0.068 (0.032)
<i>p</i> -value	0.194	0.115	0.155	0.906	0.055	0.035
<i>Panel B. Free school meals</i>						
Free school meals	0.290 (0.182)	0.135 (0.019)	0.064 (0.050)	0.207 (0.067)	0.136 (0.023)	0.062 (0.026)
Non-free school meals	0.306 (0.107)	0.042 (0.023)	−0.061 (0.031)	0.221 (0.051)	0.045 (0.018)	−0.002 (0.020)
<i>p</i> -value	0.934	0.024	0.009	0.833	0.000	0.000
<i>Panel C. Speech nativity and free school meals</i>						
Native speaker and free school meals	0.270 (0.183)	0.096 (0.046)	0.011 (0.052)	0.182 (0.078)	0.104 (0.025)	0.042 (0.028)
Native speaker and non-free school meals	0.217 (0.088)	0.038 (0.024)	−0.061 (0.032)	0.222 (0.054)	0.042 (0.020)	−0.017 (0.021)
Non-native speaker and free school meals	0.300 (0.406)	0.216 (0.077)	0.181 (0.087)	0.221 (0.108)	0.195 (0.038)	0.099 (0.041)
Non-native speaker and non-free school meals	0.671 (0.272)	0.093 (0.054)	−0.031 (0.066)	0.205 (0.100)	0.095 (0.030)	0.070 (0.035)
<i>p</i> -value: Native, FSM = Native, non-FSM	0.781	0.217	0.167	0.628	0.013	0.032
<i>p</i> -value: Non-native, FSM = Non-native, non-FSM	0.350	0.122	0.014	0.904	0.012	0.464

Notes: Under each panel (A, B, and C), results are shown from separate regressions where personal characteristics of pupils are interacted with birth cohort dummies and treatment status. Each column shows selected coefficients from regressions A, B, and C, respectively. The reported coefficients show the interaction between treatment, birth cohort, and personal characteristic of the student. The comparison group is “non-treated.” See Table 3 for details on the control variables. Standard errors are clustered at the level of the school.

speakers) whereas it is similar for these two groups for the first phase of the national rollout. However, at age 7, a difference has emerged in both cases—the estimated effect is at least twice as large for non-native speakers (*p*-values from tests of the difference in the estimated treatment effects for native and non-native speakers are 0.12 and 0.06 for the pilot and first phase, respectively). By age 11, the coefficient is positive for non-native English speakers—but only statistically significant for the first-phase cohort. With regard to the latter, the effect size is 0.07σ , and this is statistically different from that estimated for native English speakers (for whom we see no effect).

The middle panel B shows effects of the treatment for disadvantaged students and other students (based on their eligibility for free school meals). The effect sizes are similar at age five. However, we see differences at age seven for both the pilot and the first phase of the national rollout. Disadvantaged students benefit more from the

program than other students in each case. The differences are statistically significant and similar in both cases. Whereas the effect for more advantaged students (i.e., non-free school meals) is 0.04σ and 0.05σ for pupils in pilot and first-phase schools, respectively, it is 0.14σ (in both cases) for students eligible to receive free school meals. By the time students get to age 11, the effect size for disadvantaged students is 0.06σ in both cases. However, this is only statistically significant for the pupils in first-phase schools. For non-disadvantaged students, the pilot is shown to have a negative effect (of 0.06σ , which is significant at the 10 percent level); whereas for first-phase students, there is a zero effect. It is difficult to know what to make of the former (especially in view of the fact that they appeared to benefit at age seven). However, this counterintuitive result is not upheld when we compare pilot schools to control schools for the next cohort, which is shown in Table 5 (discussed further later).

Finally, in panel C, we show effects from specifications where we estimate two-way interactions between disadvantaged status and whether the student is a native speaker of English. We show estimates of the treatment on four groups: native English speakers and eligible to receive free school meals; native English speakers and not eligible to receive free school meals; non-native English speakers and eligible to receive free school meals (i.e., the most “disadvantaged group”); and non-native English speakers who are not eligible to receive free school meals. These regressions show that for both the pilot and first-phase schools, effect sizes are strongest for the most disadvantaged group (i.e., non-native English speakers AND eligible to receive free school meals) at both the age of 7 and 11. In both cases, the treatment increases test scores by around 0.20σ at age 7. With regard to effects estimated at age 11, the treatment increases scores by 0.18σ for the pilot and by 0.10σ for first-phase schools. With regard to first-phase schools, the effect persists to age 11 for only 1 other group: non-native speakers who are not eligible to receive free school meals (raising scores by 0.07σ). However, for the pilot, there remains a negative coefficient estimated for one group (i.e., native students who are not eligible to receive free school meals). It is difficult to know what to make of this estimate, though we do not find it when we consider effects for the next cohort (discussed next).

It is of interest to check whether these effects persist in the pilot group for the 2002 birth cohort. This enables us to look at the effects for a group who entered treatment schools the year after they had received intensive support (as a result of the pilot). This is reported in Table 5. The estimates are very similar to those reported for the 2001 birth cohort (i.e., Table 4, columns 1 to 3). Interestingly, the negative effect for native English speakers at age 7 (and English speakers who are not eligible to receive free school meals in panel C) that we found for the 2001 cohort goes away for the 2002 cohort. Furthermore, the high effect estimated for non-native English speakers who are eligible to receive free school meals is exactly the same for this cohort relative to the control group. The treatment raises the age 11 score by 0.18σ whether we consider the 2001 or 2002 birth cohort. This finding is important because it shows that the treatment effect has not been driven only for those students who were in the school at the same time as the literacy consultant. Longer term effects are present for disadvantaged groups even though there was no “intensive support” delivered by literacy consultants at any time when they were in the school.

TABLE 5—HETEROGENEITY IN ESTIMATED TREATMENT EFFECTS ON READING BY NON-NATIVE SPEAKER STATUS AND FREE SCHOOL MEALS ELIGIBILITY, DIFFERENT COHORTS FOR PILOT VERSUS CONTROL

	Pilot versus control (cohorts 1998 and 2002)		
	Age 5 (1)	Age 7 (2)	Age 11 (3)
<i>Panel A. Speech nativity</i>			
Native speaker	0.149 (0.064)	0.069 (0.026)	0.021 (0.033)
Non-native speaker	0.107 (0.145)	0.055 (0.048)	0.039 (0.056)
<i>p</i> -value	0.767	0.768	0.754
<i>Panel B. Free school meals</i>			
Free school meals	0.108 (0.124)	0.103 (0.043)	0.094 (0.049)
Non-free school meals	0.158 (0.069)	0.043 (0.026)	−0.007 (0.032)
<i>p</i> -value	0.711	0.133	0.030
<i>Panel C. Speech nativity and free school meals</i>			
Native speaker and free school meals	0.122 (0.122)	0.065 (0.049)	0.053 (0.053)
Native speaker and non-free school meals	0.160 (0.070)	0.069 (0.027)	0.012 (0.036)
Non-native speaker and free school meals	0.103 (0.290)	0.193 (0.066)	0.184 (0.082)
Non-native speaker and non-free school meals	0.121 (0.151)	0.000 (0.054)	−0.026 (0.056)
<i>p</i> -value: Native, FSM = Native, non-FSM	0.776	0.924	0.445
<i>p</i> -value: Non-native, FSM = Non-native, non-FSM	0.957	0.010	0.005

Notes: See Table 4 notes. Columns 1–3 are analogous to Table 4, columns 1 to 3, except that the 2002 cohort (rather than 2001) is being compared to the 1998 cohort.

In online Appendix Tables A4 and A5, respectively, we show evidence on two other sources of heterogeneity. Firstly, we estimate four-way interactions (between language and free school meal status) where we substitute the variable “native English speaker” with whether or not the students’ first language is based on the Latin script (Table A4). Secondly (in Table A5), we show results when we estimate four-way interactions by gender.

The Latin script is the basis for the largest number of alphabets of any writing system and is the most widely adopted in the world. However, one might hypothesize that a more structured approach to learn the English language is particularly important for those who have even more reliance on schools for learning the essential building blocks of the language. As this information is only derivable from 2009 onwards, we use the information when estimating effects for pupils of age 11. The results in Table A4 show effects that are similar to when we used the “non-native speaking” indicator, although they are a little higher. For students whose language does not use the Latin script AND who are disadvantaged, the treatment effect at age 11 is 0.21σ and 0.13σ for the pilot and first phase, respectively. For the latter, there is

an estimated effect even for these students if they are not classified as disadvantaged (0.09σ), but this is not the case for the pilot where there is no effect of the treatment.

In Table A5, we show the four-way interactions from our main specification for boys and girls, respectively, at ages 5, 7, and 11. The standard errors are larger (as we are splitting the sample) but produce results that are qualitatively similar and not systematically different for boys and girls. Results for the pilot suggest that effects are stronger for girls at age 11, but the opposite is true for the first phase of the national rollout.

D. Effects on Other Outcomes

We have also investigated whether the phonics treatment has any impact on other subjects at age 7 and age 11. We would expect this intervention to have an effect on other subjects, both because the same teacher teaches all subjects within a given year group (in primary school) and because subjects other than reading require competency in literacy. There is an obvious relationship between reading and writing with regard to the teaching of language. With regard to math, earlier work suggests a strong relationship between reading demands of tests in math and reading. Specifically, an analysis done on the age 11 reading and math test showed that the reading demand of the math test (based on text difficulty) is nearly 70 percent of what it is in the reading assessment (Machin and McNally 2008).

We show results for reading, writing, and math at age 7 and for reading, English, and math at age 11.²³ This is shown in Table 6. The results at age seven show that effect sizes are larger (though not statistically different) for writing than for reading and also show the pattern of increasing effects for cohorts exposed younger (and for longer) to the new way of teaching reading. The results are also positive for math. Results at age 11 show no overall effect of the treatment on reading, English, or math. Table 7 shows estimates of heterogeneous effects for math according to student characteristics (in an analogous way to that shown for reading in Table 4). The results are broadly consistent with those for reading except they tend to fade out (or at least not be statistically significant) for the pilot by age 11 whereas they remain for those exposed to the first year of the national rollout.

The finding that a pedagogical intervention directed at literacy also has an impact on math was found in earlier work for England where “the literacy hour” was introduced to primary schools (Machin and McNally 2008). This relates to another pilot intervention that subsequently became national policy but much earlier (starting in September 1996 in pilot schools and becoming national policy in 1999). In this case the effect on math was two-thirds of the effect on literacy. In many school-level interventions (such as US charter schools), effects are higher for math than for English. In a review of the US charter school literature, Eyles, Hupkau, and Machin (2016) highlight bigger effects for math than for English in 11 out of 12 papers

²³We only have an overall English mark up to 2012 (and not a separate writing test). The writing test was changed about this time, and we have no separate writing or English test that can be used in 2013. Thus, we can estimate the effect of the pilot on English but not the first phase of the national rollout (i.e., the relevant cohort did their Key Stage 2 tests in 2013).

TABLE 6—HETEROGENEITY BY SUBJECT

	Age 7 reading (1)	Age 7 writing (2)	Age 7 math (3)	Age 11 reading (4)	Age 11 English (5)	Age 11 math (6)
<i>Panel A. Pilot versus control</i>						
Treatment × 1999 birth cohort (Treatment: Only age 7)	0.037 (0.021)	0.052 (0.024)	0.043 (0.022)	0.003 (0.028)	0.032 (0.030)	0.002 (0.025)
Treatment × 2000 birth cohort (Treatment: Age 6–7)	0.040 (0.025)	0.057 (0.027)	0.045 (0.027)	−0.001 (0.027)	0.009 (0.030)	−0.006 (0.024)
Treatment × 2001 birth cohort (Full treatment: Age 5–7)	0.075 (0.024)	0.093 (0.027)	0.056 (0.027)	−0.018 (0.031)	0.010 (0.028)	−0.028 (0.026)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample size	191,342	191,325	191,330	163,270	162,448	163,293
Number of schools	1,217 (7)	1,217 (8)	1,217 (9)	1,217 (10)	1,217 (11)	1,202 (12)
<i>Panel B. First phase versus control</i>						
Treatment × 1999 birth cohort (No treatment: Placebo)	−0.015 (0.015)	−0.016 (0.016)	0.014 (0.016)	−0.024 (0.019)	—	0.022 (0.018)
Treatment × 2000 birth cohort (Treatment: Only age 7)	0.031 (0.016)	0.052 (0.017)	0.045 (0.017)	−0.016 (0.018)	—	0.006 (0.017)
Treatment × 2001 birth cohort (Treatment: Age 6–7)	0.046 (0.017)	0.055 (0.019)	0.052 (0.018)	0.022 (0.019)	—	0.017 (0.018)
Treatment × 2002 birth cohort (Full treatment: Age 5–7)	0.073 (0.017)	0.092 (0.019)	0.061 (0.019)	0.019 (0.019)	—	0.026 (0.019)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample size	309,769	309,751	309,737	268,563	—	269,018
Number of schools	1,598	1,598	1,598	1,598		1,598

Notes: See Table 3 notes. Columns 1 and 4 are reproduced from Table 3, and columns 2, 3, 5, and 6 use alternative subject scores.

(the estimates for math are between 0.10σ and 0.35σ ; while for English, they are between 0.05σ and 0.20σ). Thus, if our results reflected a successful school-level intervention (rather than a pedagogical change directed specifically at the teaching of literacy), we would be unlikely to find the pattern of results observed in this paper.

E. Other Potential Mechanisms

We have interpreted our results as reflecting changes in pedagogy, which were taught to teachers in treatment schools. However, it is possible that the intervention might have affected teachers by influencing their retention (or decision to leave), influenced school organization in some other way, or impacted on pupil enrollments. To examine this hypothesis, we estimate a simple difference-in-difference regression at school level where the dependent variable is the log number of teachers in the first column, the (log) teacher/pupil ratio in the second, and the new pupil entry rate in the third. Results are shown in Table 8. For both the pilot and first phase, the columns (1), (2), (4), and (5) show no differential change either in teacher numbers or the teacher/pupil ratio between treatment and control schools, either at the time when the policy was introduced or in previous time periods. This provides some

TABLE 7—HETEROGENEITY IN ESTIMATED TREATMENT EFFECTS ON MATH BY NON-NATIVE SPEAKER STATUS AND FREE SCHOOL MEALS ELIGIBILITY, DIFFERENT COHORTS FOR PILOT VERSUS CONTROL

	Pilot versus control (cohorts 1998 and 2001)			First phase versus control (cohorts 1998 and 2002)		
	Age 5 (1)	Age 7 (2)	Age 11 (3)	Age 5 (4)	Age 7 (5)	Age 11 (6)
<i>Panel A. Speech nativity</i>						
Native speaker	0.196 (0.085)	0.028 (0.026)	-0.046 (0.028)	0.174 (0.052)	0.038 (0.020)	-0.007 (0.020)
Non-native speaker	0.411 (0.333)	0.139 (0.061)	0.035 (0.060)	0.114 (0.088)	0.128 (0.030)	0.104 (0.030)
<i>p</i> -value	0.489	0.068	0.206	0.488	0.002	0.000
<i>Panel B. Free school meals</i>						
Free school meals	0.103 (0.187)	0.104 (0.049)	-0.002 (0.046)	0.167 (0.070)	0.095 (0.026)	0.072 (0.025)
Non-free school meals	0.267 (0.121)	0.029 (0.026)	-0.043 (0.027)	0.156 (0.055)	0.045 (0.020)	-0.002 (0.020)
<i>p</i> -value	0.374	0.083	0.367	0.889	0.029	0.002
<i>Panel C. Speech nativity and free school meals</i>						
Native speaker and free school meals	0.140 (0.180)	0.063 (0.050)	-0.030 (0.052)	0.158 (0.079)	0.046 (0.028)	0.038 (0.028)
Native speaker and non-free school meals	0.208 (0.090)	0.017 (0.027)	-0.051 (0.029)	0.178 (0.058)	0.034 (0.021)	-0.025 (0.021)
Non-native speaker and free school meals	0.001 (0.434)	0.192 (0.085)	0.051 (0.077)	0.165 (0.119)	0.194 (0.041)	0.140 (0.041)
Non-native speaker and non-free school meals	0.581 (0.349)	0.116 (0.065)	0.033 (0.063)	0.093 (0.110)	0.113 (0.032)	0.104 (0.033)
<i>p</i> -value: Native, FSM = Native, non-FSM	0.718	0.347	0.697	0.798	0.643	0.019
<i>p</i> -value: Non-native, FSM = Non-native, non-FSM	0.184	0.355	0.791	0.627	0.052	0.383

Note: See Table 4 notes.

support for the hypothesis that the intervention did not directly affect the composition of teachers or cause changes in student-teacher ratios. Furthermore, as shown in columns (3) and (6) of Table 8, there is no relationship between the timing of the intervention and the entry rate of new pupils (which is also true if we look at the exit rate of pupils or an overall measure of turnover). Thus, the policy did not lead to a change in pupil composition that could bias the earlier results.

F. Robustness

We have also checked whether estimated results at age 11 are sensitive to imputation of missing values on test scores where we know that the reason the children have not been entered for the test is because they are working “below the required level.”²⁴ This applies to about 4 percent (for the pre-policy cohort)—and is no different

²⁴Following the intuition of the bounds analysis of Angrist, Bettinger, and Kremer (2006), we present different imputations to check robustness to different assumptions.

TABLE 8—PILOT AND FIRST-PHASE TREATMENT AND SCHOOL CHARACTERISTICS

	log(teachers) (1)	log(teacher/pupils) (2)	New pupil entry rate (3)
<i>Panel A. Pilot</i>			
Treatment × 2004	0.001 (0.014)	−0.001 (0.006)	0.002 (0.005)
Treatment × 2005	0.013 (0.018)	0.005 (0.007)	0.004 (0.006)
Treatment × 2006	0.009 (0.020)	0.002 (0.008)	0.007 (0.006)
Treatment × 2007	0.029 (0.021)	0.008 (0.008)	−0.000 (0.005)
Treatment × 2008	0.028 (0.023)	0.006 (0.008)	−0.002 (0.006)
School fixed effects	Yes	Yes	Yes
Sample size	7,061	7,061	6,962
Number of schools	1,217	1,217	1,200
	(4)	(5)	(6)
<i>Panel B. First phase</i>			
Treatment × 2004	0.003 (0.010)	−0.000 (0.004)	0.000 (0.004)
Treatment × 2005	0.009 (0.011)	0.003 (0.004)	0.001 (0.004)
Treatment × 2006	0.002 (0.013)	−0.002 (0.005)	0.006 (0.004)
Treatment × 2007	0.005 (0.014)	−0.001 (0.005)	−0.002 (0.004)
Treatment × 2008	0.013 (0.014)	0.002 (0.005)	0.001 (0.004)
Treatment × 2009	0.015 (0.016)	0.003 (0.005)	−0.005 (0.004)
School fixed effects	Yes	Yes	Yes
Sample size	10,807	10,807	10,683
Number of schools	1,598	1,598	1,588

Notes: School-level regressions where school-level data on teachers, the teacher-pupil ratio, and the entry rate of new pupils have been merged with the sample of treatment and control schools. The difference-in-differences regressions include interactions between treatment and year, year dummies, and school fixed effects. The highlighted terms indicate years in which the intervention may have had an effect on treatment schools (as the intervention year was 2006 in the case of the pilot schools and 2007 in the case of first-phase schools). Standard errors are clustered at the school level.

between the treatment and control groups. We try two approaches: where we assign missing values to the lowest score given at the school that the student attended at this age (Imputation A); where we assign missing values to the mean values at the school that the student attended at this age (Imputation B). Results are reported in online Appendix Table A6. Specifically the Table shows average results for the whole cohort (replicating the analysis reported for Table 3) and when we interact the treatment for native/non-native speakers of English and eligible/non-eligible for free school meals (replicating the analysis in panel C of Table 4). The original results are compared with results in which missing values have been imputed.

We learn that the imputation has no implications for average results—they all suggest an effect, which is close to zero and not statistically significant. In the bottom panel, we show that results are very comparable when we examine whether the treatment has a heterogeneous effect. The only result that changes is that the impact of the pilot on the group classified as “non-native and free school meals” declines from 0.18σ to 0.13σ (under Imputation A)—making it closer to that estimated for the first phase (of about 0.10σ).

IV. Conclusion

A widespread, well established finding in the economics of education research literature is that good teachers matter. However, a critical, but currently much less studied question, is whether “good teaching” can be taught. Our empirical analysis shows that intensive training in the use of a “new pedagogy” produced strong effects for early literacy acquisition amongst young school children. We are able to provide convincing evidence of causal effects from the introduction of synthetic phonics in English primary schools because of the way in which training was staggered across different local authorities (and hence different schools). Indeed, we show similar effects from a pilot and the first phase of the national rollout which followed. Moreover, effects of the interventions become much smaller or cease completely in subsequent waves of the national rollout, suggesting that the targeted and large-scale rollout had beneficial effects on the literacy of primary age school children.

The initial effects are sizable and comparable to the early effects of project STAR in reducing class size (which found an effect of 0.19 standard deviations after 1 year from an average class size reduction of about 8 students). However, the policy is of far lower cost, as it involves employing a literacy consultant working with 10 schools per year to deliver intensive support as well as arranging for dissemination and training opportunities throughout the local authority. If effects only reflected the active involvement of the literacy consultant, one would not expect effects to persist for young students. The fact that effects are observed for younger students in years *after* the literacy consultant had been at the school (at least up until the control group entered the program) suggests that the training and not the presence of the trainer explains the treatment effect. It appears that training in the use of the new pedagogy really benefits measures of reading attainment (as well as writing) for young people.

However, most students learn to read eventually. This is the simplest explanation for why we do not see any overall effect of the intervention by age 11. There may of course be (unmeasured) benefits of learning to read well at an earlier age. It might also be that there are longer term effects that become apparent at a later stage (e.g., as happened in Project STAR and shown by Chetty et al. 2011).

The most interesting finding here is that there are long-term effects at age 11 for those with a high probability of starting their school education as struggling readers. Specifically, the results suggest that there is a persistent effect for those classified as non-native English speakers and economically disadvantaged (as measured by free school meal status). The effect persists for these children who enter school with significant literacy deficits and is at least 0.10 standard deviations on the reading test at age 11. This is impressive given that the phonics approach is only actively taught

up to the age of seven. There is an educational theory suggesting that it is more difficult to improve the reading scores for students who speak nonstandard English at home (discussed by Fryer 2014), and this is a potential reason for why so many interventions appear to have a much smaller impact on reading than on math. In the light of this, it is important to find persistent positive effects from this intervention for those who start their education as struggling readers. Without a doubt the effect is high enough to justify the fixed cost of a year's intensive training support to teachers. Furthermore, it contributes to closing literacy gaps based on disadvantage and (initial) language proficiency by family background.

This policy reduced literacy inequalities in primary school at relatively low cost. That such a policy was administered by local authorities takes on an added significance given the radical and far-reaching school policies underway in England in which all schools have been encouraged to become academy schools, which operate entirely outside of local authority control.²⁵ If this happens, it is unclear what role local authorities may play in schooling, but it will certainly be massively diminished, and perhaps nonexistent. The kind of policy we have studied in this paper would not be feasible in such a decentralized system. This has wider ramifications and relevance for other countries that are considering whether to decentralize their schooling systems.

REFERENCES

- Aaronson Daniel, Lisa Barrow, and William Sander.** 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25 (1): 95–135.
- Angrist, Joshua, Eric Bettinger, and Michael Kremer.** 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96 (3): 847–62.
- Araujo, M. Caridad, Pedro Carneiro, Yannú Cruz-Aguayo, and Norbert Schady.** 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131 (3): 1415–53.
- Behrman, Jere.** 2010. "Investment in Education—Inputs and Incentives." In *Handbook of Development Economics*, Vol. 5, edited by Dani Rodrik and Mark Rosenszweig, 4883–4975. Amsterdam: North-Holland.
- Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan.** 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126 (4): 1593–1660.
- Chetty, Raj, John N. Friedman, and Jonah Rockoff.** 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–2632.
- Chetty, Raj, John N. Friedman, and Jonah Rockoff.** 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79.
- Deming, David.** 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1 (3): 111–34.
- Education Sector.** 2006. "Universal or Targeted Preschool? A Debate between W. Steven Barnett and Bruce Fuller." https://www.air.org/sites/default/files/publications/Preschool%20Debate_0.pdf.
- Elliot, Colin, Pauline Smith, and Kay McCullouch.** 1996. *The British Ability Scales II (BASII)*. Windsor Berkshire: NFER-Nelson Publishing Company.
- Eyles, Andrew, Claudia Hupkau, and Stephen Machin.** 2016. "Academies, charter and free schools: Do new school types deliver better outcomes?" *Economic Policy* 31 (87): 453–501.

²⁵ On the English academy schools program see, for example, Eyles and Machin (2015) and Eyles, Machin, and McNally (2017).

- Eyles, Andrew, and Stephen Machin. 2015. "The Introduction of Academy Schools to England's Education." Centre for Economic Performance Discussion Paper 1368.
- Eyles, Andrew, Stephen Machin, and Sandra McNally. 2017. "Unexpected school reform: Academisation of primary schools in England." *Journal of Public Economics* 155: 108–21.
- Fryer, Roland. 2014. "Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments." *Quarterly Journal of Economics* 129 (3): 1355–1407.
- Hansen, Kirstine, Heather Joshi, and Shirley Dex. 2010. *Children of the 21st Century: The First Five Years*. Bristol: Policy Press.
- Hanushek, Eric, Steven Rivkin, and John Kain. 2005. "Teachers, Schools and Academic Achievement." *Econometrica* 73 (2): 415–58.
- Hill, Vivian. 2005. "Through the Past Darkly: A Review of the British Ability Scales Second Edition." *Child and Adolescent Mental Health* 10 (2): 87–98.
- Jacob, Brian. 2017. "When evidence is not enough: Findings from a randomized evaluation of Evidence-Based Literacy Instruction (EBLI)." *Labour Economics* 45: 5–16.
- Johnston, Rhona, and Joyce Watson. 2005. *The Effects of Synthetic Phonics Teaching on Reading and Spelling Attainment: A Seven Year Longitudinal Study*. Scottish Executive Education Department, Edinburgh, February.
- Jones, Elizabeth M., and Ingrid Schoon. 2008. "Child Cognition and Behaviour." In *Millennium Cohort Study Third Survey: A User's Guide to Initial Findings*, edited by Kirstine Hansen and Heather Joshi, 118–44. London: Centre for Longitudinal Studies.
- Kearney, Melissa S., and Phillip B. Levine. 2016. "Early Childhood Education by MOOC: Lessons from Sesame Street." National Bureau of Economic Research (NBER) Working Paper 21229.
- Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.
- Krueger, Alan B., and Diane M. Whitmore. 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 (468): 1–28.
- Kuczera, Małgorzata, Simon Field, and Hendrickje Catriona Windisch. 2016. *Building Skills for All: A Review of England*. Paris: Organisation for Economic Co-operation and Development (OECD).
- Machin, Stephen, and Sandra McNally. 2008. "The Literacy Hour." *Journal of Public Economics* 92: 1441–62.
- Machin, Stephen, Sandra McNally, and Martina Viarengo. 2018. "Changing How Literacy Is Taught: Evidence on Synthetic Phonics: Dataset." *American Economic Journal: Economic Policy*. <https://doi.org/10.1257/pol.20160514>.
- National Reading Panel. 2000. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. Washington, DC: National Institute of Child Health and Human Development.
- Organisation for Economic Co-operation and Development (OECD). 2013. *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. Organisation for Economic Co-operation and Development (OECD). Paris, November.
- Primary National Strategy. 2007. *Letters and Sounds: Principles and Practice of High Quality Phonics*. British Department of Education. London, September.
- Rose, Jim. 2006. *Independent Review of the Teaching of Early Reading*. British Department for Education and Skills. London, March.
- Slavin, Robert E., Cynthia Lake, Bette Chambers, Alan Cheung, and Susan Davis. 2009. "Effective Reading Programs for the Elementary Grades: A Best-Evidence Synthesis." *Review of Educational Research* 79 (4): 1391–1466.
- Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *American Economic Review* 102 (7): 3628–51.
- Torgerson, Carole J., Greg Brooks, and Jill Hall. 2006. *A Systematic Review of the Research Literature on the Use of Phonics in the Teaching of Reading and Spelling*. London: British Department for Education and Skills.
- United Kingdom Literacy Association (UKLA). 2010. *Teaching Reading: What the Evidence Says*. United Kingdom Literacy Association (UKLA). Leicester, October.
- Vignoles, Anna. 2016. "What is the economic value of literacy and numeracy? Basic skills in literacy and numeracy are essential for success in the labor market." IZA World of Labor Working Paper 229.
- Wyse, Dominic, and Usha Goswami. 2008. "Synthetic phonics and the teaching of reading." *British Educational Research Journal* 34 (6): 691–710.