Routledge
Taylor & Francis Group

School Effectiveness and School Improvement
An International Journal of Research, Policy and Practice

ISSN: 0924-3453 (Print) 1744-5124 (Online) Journal homepage: http://www.tandfonline.com/loi/nses20

The contribution of schooling to learning gains of pupils in Years 1 to 6

Hans Luyten, Christine Merrell & Peter Tymms

To cite this article: Hans Luyten, Christine Merrell & Peter Tymms (2017): The contribution of schooling to learning gains of pupils in Years 1 to 6, School Effectiveness and School Improvement, DOI: 10.1080/09243453.2017.1297312

To link to this article: http://dx.doi.org/10.1080/09243453.2017.1297312

Published online: 28 Feb 2017.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT, 2017
http://dx.doi.org/10.1080/09243453.2017.1297312

Routledge
Taylor & Francis Group

OPEN ACCESS

The contribution of schooling to learning gains of pupils in Years 1 to 6

Hans Luyten[a], Christine Merrell[b] and Peter Tymms [b]

[a]Department of Research Methodology, Measurement and Data Analysis (OMD), Faculty of Behavioural, Management and Social Sciences (BMS), University of Twente, Enschede, The Netherlands; [b]Centre for Evaluation and Monitoring (CEM), Durham University, Durham, UK

ABSTRACT

By means of a regression-discontinuity approach with multiple cut-off points, the effects of age and schooling on learning gains in English primary schools are estimated. The analyses relate to over 3,500 pupils in 20, predominantly independently funded, schools and focus on 4 different learning outcomes. In order to take into account delayed and accelerated school careers, an intention-to-treat analysis was applied. The findings reveal substantial effects of schooling, which in line with previous studies in English primary education account for about 40% of the total learning gains. The year-to-year gains show a declining trend as the school career progresses. The analyses produce evidence for both decreasing effects of schooling on achievement and a weakening age–achievement relationship in the higher years of primary education.

ARTICLE HISTORY
Received 2 May 2016
Accepted 15 February 2017

KEYWORDS
Effect of schooling; regression discontinuity; intention-to-treat analysis; primary education; independent schools

Introduction

All over the world, children spend many hours in school. In most countries, formal education in schools is compulsory from a young age (often 5 or 6) until at least the early teenage years. In primary education, language and mathematics make up a large part of the school curriculum, and there can be no doubt that children make much progress in these respects during the primary school period. Available evidence further suggests that the learning gains decline as the school career progresses.

Bloom, Hill, Rebeck Black, and Lipsey (2008) report outcomes that express the annual growth in the United States for language, maths, science, and social studies as effect sizes (based on Cohen, 1988). For language and maths, the yearly gains are approximately 1.00 in the first years of primary education. This implies that the average pupil in Year 2 scores one standard deviation above the average in Year 1. In later years, this growth gradually declines. In the final years of primary school, annual growth has already declined by more than half, and in the final years of high school it is 0.20 or less. These findings can be used as benchmarks for assessing the impact of educational interventions (Lipsey et al., 2012). For example, an effect size of 0.20 would be considered small, following the broad

CONTACT Hans Luyten j.w.luyten@utwente.nl

guidelines suggested by Cohen (1988), but in many cases such an effect would already equal half the annual learning gain or even more.

Even though children show substantial learning gains over the primary school years, one cannot assume that all progress is caused by schooling. At least some part of the learning gains during primary education occurs independently from schooling. Children not only learn through formal schooling but acquire knowledge and skills due to non-school factors as well (e.g., their home environment). In that respect, the annual gains reported by Bloom et al. (2008) reflect growth due to non-school factors as well as schooling. This may be a reason to set expectations about the impact of educational interventions at even more modest levels. Educational interventions may help to improve key factors like the quality of classroom instruction, but the impact of non-school factors will be more difficult to manipulate.

It is also important to note that it is unclear to what extent the decreasing growth in learning gains reflects declining effects of schooling. It may be the result of a declining age effect; as learning curves tend to flatten over time, we expect that a declining age effect is a major factor in the decrease of annual learning gains. It is however, conceivable that the effect of schooling remains constant over the years and that declines in annual learning gains are solely due to a decreasing age effect. It is also possible that both the effect of schooling and age declines in the later years. Perhaps the impact of instruction is strongest in the first years, while its effects are less profound later on. Maybe changes in the school curriculum are a relevant factor. In the first years of primary schooling, there is an emphasis on the development of basic skills in language and mathematics, whereas in the later years there is more focus on applying those skills in more specialist knowledge in subjects such as science, history, and geography. Further, the findings reported by Bloom et al. (2008) suggest that it may be inappropriate to apply the same benchmarks for evaluating the effects of educational interventions at all stages of the educational career. However, if decreases in annual learning gains are solely due to declining age effects with the schooling effects being stable across the school years, this conclusion may be reconsidered. There is little reason to set different benchmarks at different stages of the school career, if the effects of schooling remain constant as the school career progresses and decreases in annual learning gain are solely due to declining age effects.

The present paper reports the findings of a study aimed at an empirical investigation of the contribution schooling makes to children's learning gains during the primary school period (Years 1 to 6) in the English education system. The main research question we address is:

Do effects of schooling decline in the later years of primary education?

The present paper explicitly aims to contribute to a strand of research in which the effects of schooling are estimated by comparing same-age children in different year groups. This approach has been applied in a considerable number of studies (e.g., Cahan & Cohen, 1989; Cahan & Davis, 1987; Cliffordson, 2010; Gormley & Gayer, 2005; Kyriakides & Luyten, 2009; Luyten, 2006; Stelzl, Merz, Ehlers, & Remer, 1995). Estimates of the schooling effect are based on the difference in achievement between the oldest pupils in one year and the youngest in the next. The approach is usually referred to as regression discontinuity and strongly draws on the use of cut-off dates to determine

assignment to year groups. To our knowledge, only the study by Cliffordson (2010), which relates to Swedish pupils in Years 6, 7, and 8 (ages 12, 13, and 14, respectively), has addressed the possibility that the effects of schooling may vary at different phases in the school career.

Our study focuses on the effects of schooling on learning gains in language, maths and general cognitive ability. The dependent variables are four different measures that are part of the Interactive Computer Adaptive System (InCAS) assessment (Merrell & Tymms, 2007), which was developed by CEM (Centre for Evaluation and Monitoring at Durham University, UK) (Tymms & Coe, 2003) to assess the learning gains of primary school children. Three of these measures relate to skills that are typically taught in school (Reading, General Maths, and Mental Arithmetic). The fourth measure (Developed Ability) reflects skills that may largely be acquired outside the school (vocabulary) and that are not explicitly included in the primary school curriculum (non-verbal pattern recognition). With regard to this fourth measure, we expect to find that schooling contributes relatively little to learning gains.

Our main research question is whether effects of schooling decline in the later years of primary education. More specifically, our data analyses address the following questions:

(1) Is there evidence for a decreasing effect of schooling on learning gains?
(2) Is there evidence for a weakening age–achievement relationship as the school career progresses?
(3) What is the impact of schooling on learning gains expressed as an effect size?
(4) What is the relative contribution of schooling as a percentage of total learning gains?

In the next section, we briefly summarize findings from prior research on the effects of schooling that capitalizes on cut-off dates for assigning pupils to year groups. The basic principles of regression discontinuity are outlined and also the main requirements that must be met when this approach is used for assessing the effects of schooling. Complications that arise when the approach is applied in these studies are discussed as well. We will explicitly explain why intention-to-treat analysis is an appropriate approach to deal with some of the main complications. In the present case, an alternative method is not feasible, because it would require background information on the pupils that is not present in our dataset. In the next section, the methods of analysis applied in the present study will be outlined. After that, we present our findings in the results section. The paper concludes with a discussion section in which we also address the main limitations of the present study (i.e., with regard to sample size and power, internal validity and external validity).

Prior research on the effects of schooling using regression discontinuity: basic principles main findings, requirements, and complications

Basic principles

A major challenge for any research that aims to estimate the impact of schooling on young children is the fact that nearly everyone attends school. As a result, an equivalent

control group of children that receives no schooling is absent. A valid method to deal with this challenge is to compare (nearly) same-age children in different year groups (Ceci, 1991). This approach (usually referred to as regression discontinuity) strongly draws on cut-off dates. In most education systems, such cut-offs are applied to determine assignment to year groups. If they are strictly adhered to, children with minimal differences in age are assigned to adjacent year groups. This creates a situation very close to a randomized experiment, as children that are similar in each and every aspect, apart from a minute difference in age, end up in a higher or lower year group. Therefore, the effect of being assigned to a higher year group can be separated from the relationship between age and achievement.

In the simplest case, pupils from two adjacent school years are compared. The data analysis comes down to a straightforward regression analysis with two main explanatory variables: age and year group. The dependent variable is usually a test score, but it may also be a non-cognitive measure (e.g., self-efficacy or attitudes toward school). If only two years are involved, the year group variable amounts to a dichotomy. In that case, the effect of the year group in the regression analysis actually expresses the difference between the oldest pupils in the lower year and the youngest in the higher year. This is the case, because the analysis controls for the effect of age. In this paper, we report the findings of a study that covers a wider range of year groups.

It is important to note that with regard to this particular application of regression discontinuity, it is reasonable to assume that the individuals on either side of the cut-off are similar on relevant background characteristics (e.g., aptitudes, motivation, gender, family background, ethnicity). This may not be a safe assumption in other cases of regression-discontinuity research. For example, if one wants to assess the effects of a *voluntary* preschool programme for 4-year-olds on language skills, it is quite likely that certain parents are more eager to enrol their children for the programme. These parents may also be more motivated and/or able to stimulate their children's learning (Gormley & Gayer, 2005). In the present case, however, we are dealing with children in adjacent years of *compulsory* schooling. In this case, we can expect the background of the pupils to be similar in each and every year, as schools draw their pupils from the same population year after year. This assumption has hardly ever been tested in educational research, but a study covering over 80% of all primary schools in The Netherlands shows that over a 5-year period pupil backgrounds are highly consistent from year to year (Luyten & De Wolf, 2011). In their secondary analysis on the 1995 TIMSS data, Webbink and Gerritsen (2013) found indications of non-equivalence between adjacent school years in secondary education, but corrections for bias in this respect hardly affected the estimated effects of schooling (Webbink & Gerritsen, 2013, p. 19). Note that in this case not all pupils from different years attended the same schools. In the present study, only schools were included that tested their pupils in all six years. Moreover, there is no reason to assume that date of birth is related to relevant background variables. This would imply that in certain months more talented children are born, or more high-socioeconomic-status (SES) children, or more girls, and so forth. With regard to the present study, it is also relevant that the schools included are nearly all independent schools that attract specific populations. Some of the schools provide boarding facilities, some are religious schools, and some are single-sex schools, and all charge substantial fees. It stands to reason to expect that the pupil backgrounds in such schools will be particularly similar from one year to the next.

Main findings

Research based on regression discontinuity invariably shows positive and generally substantial effects of schooling on cognitive measures (Luyten, 2015). Most studies indicate that the impact of schooling on differences in achievement between year groups outweighs the impact of age, although some exceptions have been reported. Jabr and Cahan (2015) report relatively small effects of schooling for pupils in Palestinian schools on the West Bank and in Israeli State Arab schools (but not for pupils in Israeli State Jewish schools), and Luyten (2006) reports relatively small effects of schooling for England compared to several other countries. The earliest regression-discontinuity studies to assess the impact of schooling on learning gains were conducted by developmental psychologists (e.g., Baltes & Reinert, 1969; Cahan & Cohen, 1989). Work by researchers in the field of education (e.g., Crone & Whitehurst, 1999; Luyten, 2006) and human capital economics (e.g., Cascio & Lewis, 2006; Gormley & Gayer, 2005; Webbink & Gerritsen, 2013) is more recent.

The basic regression-discontinuity approach can be extended to a wider range of years as long as it relates to pupils in adjacent years. There are a few examples of this. The study by Cahan and Cohen (1989) covers children in Years 4, 5, and 6 of primary education in Israel. The studies by Kyriakides and Luyten (2009) and Jabr and Cahan (2015) cover even wider ranges of school years. So far, in nearly all studies that cover more than two adjacent school years, the effect of one year of schooling has been modelled as a linear effect. In other words, the effect of schooling is assumed to be constant across the entire range of years. The study by Cliffordson (2010) is the only exception that we are aware of. Her findings show a linear relationship between age and mathematics achievement for Swedish pupils in Years 6, 7, and 8 (ages 12–14) and a declining effect of schooling. The effect of being in Year 8 versus 7 appears to be smaller than the effect of being in Year 6 versus 7.

Requirements

Here, we focus on the requirements that relate specifically to research that makes use of regression discontinuity to assess the effects of schooling. General requirements that should be met with regard to the internal validity of the regression-discontinuity approach, as specified by Schochet et al. (2010), are addressed in Appendix 1, which also provides further details on these requirements with regard to the current study.

An important advantage of regression discontinuity as a method to assess the effects of schooling is that it actually sets rather few requirements. Perhaps the most challenging requirement is that information on student scores in at least two adjacent year groups is available and, most of all, that the scores of pupils in different year groups are comparable. In studies that cover a wide range of years, administering the same test to all pupils would hardly be sensible. Items that are appropriate for pupils in the final years of primary education will be too difficult for younger pupils. Items appropriate for pupils in the first years will be trivial for the older ones. To deal with this complication, one can make use of vertically equated test scores. This enables us to express scores from children throughout primary school on a common scale (Verhelst, 2010). Scores based on different tests can be made comparable either if the tests have a number of items in

common or if some pupils have taken both tests. Performance in a specific domain (such as reading or maths) can be scaled in such a way that scores from School Years 1 to 6 become comparable. Scores from Years 1 and 2 can be made comparable through overlap in items. These will be relatively difficult items for Year 1 pupils and relatively easy items for Year 2 pupils. The same principle can be applied for Year 2 and Year 3, and so forth. In this case, the overlap in items will include the more difficult items for Year 2 (not included in the Year 1 and Year 2 overlap) and the relatively easy ones for Year 3. Through indirect comparability, scores that cover the entire primary school period can be made comparable.

In addition to vertically equated scores, information on the pupils' birthdates is required (at least month and year of birth). Third, the cut-off date that determines assignment to school years must be known. Preferably, the same cut-off date should apply to all pupils in the education system that is studied. For example, a serious complication arises when the cut-off date varies among regions or even individual schools within the education system. Fourth, information on each pupil's actual year group is needed. This information can also be used to assess how strictly the cut-off is applied. Thus, the percentages of delayed and accelerated school careers can easily be determined. A very convenient feature of the approach is that cross-sectional data suffice. If one accepts the assumption that pupil backgrounds do not vary between years and that month of birth is unrelated to relevant background variables like talent, motivation, SES, and gender, no background data or pre-test scores are required.

Complications (and solutions)

A first complication when applying regression discontinuity may be variation of the cut-off date by regions within countries, although in many countries a nation-wide cut-off date applies. Examples of countries with variation in cut-offs between regions are Australia, Germany, the United Kingdom, and the United States. The most sensible approach in those cases seems either to focus on a region where a single cut-off date is applied or to run separate analyses per region and synthesize the findings in a subsequent stage. The present study relates to schools in England. Across schools within this part of the UK, the same cut-off date (1 September) is applied, and exceptions to this rule are particularly rare.

Another important complication in studies applying a regression-discontinuity approach to assess the effect of schooling is the adherence to the cut-off date. If the cut-off date was applied strictly to assign pupils to year groups (i.e., leaving no room for exceptions), each year would contain exclusively pupils born within a 1-year age range, and all children born within this age-range would be in the same year group. In that case, the correspondence between age cohorts and year groups would be perfect. Estimating the effect of schooling would then be quite straightforward. One would only need to assess the relationship between age and achievement, and the effect of schooling would be equal to the difference in mean achievement between both years after adjusting for the age–achievement relationship. Within year groups, one may expect an advantage for the older pupils, but the difference in achievement between pupils in adjacent year groups that cannot be attributed to age can be conceived as the schooling effect.

However, in nearly every education system cut-off dates are applied with some flexibility. Pupils do not always end up in the "right" year group, especially as grade retention and (to a lesser extent) grade skipping are common phenomena in many education systems. As a result, delayed and accelerated school careers occur to some extent, but this varies greatly across education systems (Eurydice, 2011; Organisation for Economic Co-operation and Development [OECD], 2010, pp. 61–68). As the present study relates to primary education in England, the prevalence of pupils assigned to a "wrong year" is quite limited (for more details, see Appendix 2). In many other systems, the prevalence of non-standard school careers potentially presents a serious problem for the assessment of the effects of schooling. In each year group, delayed pupils are the oldest ones and accelerated students are the youngest. The main reason for delay in school career is (perceived) lack of learning aptitudes, and the main reason for acceleration is usually the opposite (exceptional aptitudes). Ignoring this complication will produce an underestimation of the relationship between age and achievement, because the oldest pupils (the delayed ones) will not get particularly high tests scores but the young, accelerated pupils will score quite high. As a result of underestimating the age–achievement relationship, the effect of schooling will be overestimated.

The often applied "solution" of eliminating pupils with non-standard careers from the analysis does not solve this problem. Non-standard school careers are most frequent among pupils with birthdates close to the cut-off (Luyten & Veldkamp, 2011). The youngest pupils in each cohort run the highest risk of being retained, and acceleration occurs most frequently among the oldest. Excluding these pupils from the analysis would still result in an underestimation of the age–achievement relationship. Thus, the effect of schooling will again be overestimated. Relatively young and less talented pupils will be excluded, and the same goes for relatively old and highly talented pupils. The complication of flexible cut-off dates may not be too much of a problem in systems with very small percentages of non-standard school careers (like the English case), but in other systems it presents a major problem.

Most prior research that capitalizes on cut-off dates to assess the effects of schooling has been conducted in education systems with relatively low percentages of delayed and accelerated school careers (less than 5%). In the majority of these studies, the researchers chose to exclude the pupils with non-standard careers from the analysis. This probably did not have any strong effects on their findings, as only small percentages were excluded. Still, excluding these pupils is a somewhat crude way to deal with pupils that are assigned to a "wrong" year, and in the end it is incorrect. A more appropriate method would be to apply an instrumental variables method (Angrist & Krueger, 2001). This would produce an unbiased assessment of the effect of being assigned to a higher year, but the analysis would require that the factors determining delay and acceleration are taken into account. Therefore, a wide range of background variables (including prior achievement scores) should be controlled for in the analysis. The findings would demonstrate the effect of being assigned to a higher year, but an important aspect of the effectiveness of an education system is also determined by the amount of delayed and accelerated school careers; for example, suppose that in a given system the effect of being assigned to a higher year is very strong, while at the same time a large percentage of school careers are delayed. Should this system be considered more effective than a system where the year group effect is smaller, while the percentage of delayed school careers is close to zero?

A straightforward method to deal with the complication of flexible cut-off dates is to conduct an intention-to-treat (ITT) analysis (Hollis & Campbell, 1999). Intention to treat is frequently applied in medical research to take into account that the impact of a treatment may be hampered if a substantial percentage of the intended patients is not reached or does not complete the treatment. Likewise in education, some pupils may not get the "treatment" for which they are eligible (given their date of birth) because their school career is delayed, whereas accelerated pupils get a more advanced treatment than most others in their age cohort. Applying this approach when assessing the effect of schooling simply means that the analysis focuses on the year group a pupil "should" be assigned to, given his/her date of birth, rather than the actual year. In other words, a pupil's birth cohort will be the main explanatory variable of interest. The relationship between age and achievement will be assessed, and at the crossing from one cohort to the next we expect to find an extra increase in achievement. An advantage of this approach is that the effects of delayed and accelerated careers are taken into account. Most likely, delay in school careers leads to a disadvantage in comparison to same-age counterparts with a standard career (Hattie, 2009, pp. 97–99), and acceleration may give the pupils involved an advantage in comparison to their same-age peers. To the extent that delay and acceleration are common phenomena in an education system, its effects will be taken into account in an intention-to-treat analysis. This approach seems especially useful for comparing the effects of schooling across education systems as the prevalence of non-standard school careers is likely to affect the overall effectiveness of education systems.

A practical advantage of intention-to-treat analysis (ITT) compared to the instrumental variables method (IV) in the context of assessing the effect of schooling is that IV would require information on the factors that are related to delay and acceleration (like aptitudes, motivation, family background, gender). In the present case such information is absent. Apart from outcome measures, the dataset only contains information on the pupils' birthdates, their year groups, schools, and dates of testing. One should realize that the intention-to-treat approach requires that all pupils from the age cohorts are included in the analysis. This may present a serious complication in systems where cut-off dates are applied rather flexibly, but in the present case it is not a serious problem; first of all, because non-standard careers are very rare in English primary education, but also because we cover a wide range of school years. As a result, most of the delayed and accelerated pupils are still included in our sample. The only exception is delayed pupils who should be in Year 1 (but are still in the previous class) and accelerated pupils that should be in Year 6 (but have already left primary education).

Method

This section starts with a description of the dataset that was analysed. Next, we present more details with regard to our method of analysis. General requirements that should be met with regard to the internal validity of the regression-discontinuity approach, as specified by Schochet et al. (2010), are briefly addressed in this section as well. Further details on these requirements with regard to the study reported here are provided in Appendix 1.

Dataset

We make use of data that were collected during the period 1 September to 30 November 2012. Out of a much larger sample of English primary schools (350 schools; 35,226 pupils) that had used InCAS (the Interactive Computer Adaptive System, developed to assess the learning gains of English primary school children), only those schools were selected that administered the assessment in all years from 1 to 6 in the autumn period. A large number of schools (118; 16,392 pupils) was excluded, because they administered the InCAS assessment in other months. Of the remaining 232 schools (18,834 pupils), most did not administer InCAS with all years; only 20 schools (3,634 pupils) met this criterion and were included in the analyses.

All but one of the 20 schools selected turned out to be independent schools and not representative of the demographics of England as a whole. Parents pay substantial fees to send their children to independent schools. Our knowledge about the pupil backgrounds is quite limited, as these schools do not publish data on percentages of children receiving free school meals, average results on statutory examinations in the primary years, and so forth, on either the Department for Education website or their own websites. We can report that 6 schools are single-sex schools (5 girls' schools) and 11 schools admit pupils from the ages 4 (in some cases as young as 2) through 18 or 19. Five of the schools are religious schools (3 Roman Catholic; 2 Church of England), and 6 schools provide boarding facilities, while also having day students who return home every day after school. In the final section of this paper, we discuss to what extent our findings can be generalized to a wider population. The basic question in this respect is to what extent the effects of schooling may be different for pupils of different backgrounds.

The four dependent variables in this study derive from the InCAS assessment. Through Rasch scaling, the test scores have been equated across Years 1 to 6 and given age equivalent scores as linear transformation of logits. The first measure, Reading, includes word recognition (the pupil hears a word and is asked to choose the correct written version out of five options), word decoding (the pupils must decide which of five written options matches a nonsense word), comprehension (the pupils must fill in the missing word in a sentence), and spelling. The second measure, General Maths, relates to general mathematical comprehension (e.g., interpreting graphs or positioning numbers on a number line). The third measure, Mental Arithmetic, relates to addition, subtraction, multiplication, and division. The final measure is called Developed Ability and includes picture vocabulary (the pupil hears a word and is asked to point to the picture on the computer screen that represents the word) and non-verbal ability (recognizing patterns).

The two main independent variables are the age of the pupils (on 1 September 2012) and their age cohort. The age scores are based on year and month of birth. Age is recoded so that the youngest pupils (born in August 2006) get a zero score. The age cohort is based on pupil age as well. The first cohort comprises pupils with birthdates from September 2006 to August 2007. The second cohort comprises pupils born from September 2005 to August 2006, and so forth. The cohort score is also recoded in such a way that pupils in the first cohort get a zero score.

In the analysis, we also control for date of testing. Pupils took the tests in the period from 1 September until 30 November 2012. The date of testing has been recoded so that its value ranges from 0 to 1. Pupils who took the test on 1 September get a zero

Table 1. Descriptive statistics per cohort on four outcome measures.

Cohort	Range of birthdates		Reading	Mental Arithmetic	General Maths	Developed Ability
1	Aug. 2007	Mean	5.37	4.89	6.54	4.66
	–	*N*	485	515	510	493
	Sept. 2006	*SD*	1.29	1.66	.82	2.12
2	Aug. 2006	Mean	7.14	6.63	7.40	6.71
	–	*N*	526	519	527	508
	Sept. 2005	*SD*	1.66	1.60	.86	2.19
3	Aug. 2005	Mean	8.72	8.04	8.52	8.79
	–	*N*	563	570	571	571
	Sept. 2004	*SD*	1.67	1.49	1.05	1.94
4	Aug. 2004	Mean	9.88	9.31	9.45	10.07
	–	*N*	621	620	629	628
	Sept. 2003	*SD*	1.60	1.42	1.14	1.96
5	Aug. 2003	Mean	10.77	10.20	10.36	11.21
	–	*N*	644	645	648	651
	Sept. 2002	*SD*	1.54	1.30	1.30	1.86
6	Aug. 2002	Mean	11.53	11.09	11.31	12.18
	–	*N*	600	593	615	611
	Sept. 2001	*SD*	1.46	1.39	1.29	1.77
1–6	Aug. 2007	Total *N*	3439	3462	3500	3462
	–	Pooled *SD* across	1.54	1.48	1.08	1.97
	Sept. 2001	cohorts[1]				

[1]The pooled standard deviations are used to calculate effect sizes (see Tables 2 and 5), as it seems reasonable to assume that the standard deviations of the four measures remain essentially the same across years. For General Maths, the standard deviation appears to increase over time, but the opposite applies for Developed Ability. On average, no clear trend can be discerned.

score, and the ones who took the test on 30 November get the score 1. Test scores on at least one of the four outcome measures were available for the large majority (97.2%) of the 3,634 pupils that were enrolled in Years 1 to 6 of the schools included in the analysis. A small percentage of the pupils (2.8%) had no score on any of the tests. For 91.7% of the pupils, scores on all four outcomes measures were available. With regard to the remaining pupils, 4.2% had scores on three outcome measures and 1.3% had scores on one or two outcome measures. Table 1 reports the descriptive statistics (mean, number, and standard deviation) per cohort for each outcome measure.

For the large majority of pupils in English primary education, birth cohort and year group coincide. Pupils with delayed or accelerated school careers are extremely rare. In the present sample, 98.1% of the pupils were within their expected years (see Appendix 2). Only 0.6% of the pupils were delayed and 1.3% were accelerated. Still, this implies that a (very) small percentage of the target population is not included in the dataset, namely, the delayed pupils from the first age cohort and the accelerated ones from the sixth cohort. Given the small percentages of delayed and accelerated school careers, this indicates that approximately 0.4% of all pupils from the six cohorts are not included (about 5 pupils from the first cohort and 10 from the sixth).

Analysis

First of all, we report gross annual gains from Year groups 1 to 6 for the four outcome measures. These are expressed as effect sizes as defined by Cohen (1988). The differences in mean achievement scores between two adjacent cohorts are divided by the pooled standard deviation across all six years (assuming that the standard deviations

essentially remain the same across years; see note below Table 1). These findings will serve as a basis for interpretation of the main findings.

The effect of schooling is then estimated by means of a multilevel regression analysis with age and birth cohort as explanatory variables. Age presents the "forcing variable", as it (largely) determines assignment to school years through the cut-off dates (Imbens & Lemieux, 2008). Multilevel analysis is applied to take into account the clustering of pupils within schools. The effects of both age and age cohort on learning outcomes are modelled as a quadratic function, so that curvilinear relationships (especially declining effects of schooling in the later years) can be detected. We expect to find positive effects of the linear terms and negative effects of the quadratic terms. This would imply declining effects of schooling (i.e., age cohort) and a weakening age–achievement relationship. Using this parametrization, it is also possible to detect radically different patterns, for example, increasing effects of schooling or even u-shaped and inverse u-shaped patterns. Gelman and Imbens (2014) suggest using only linear and quadratic specifications of the forcing variable and advise against higher polynomial functions as this may produce misleading results due to overfitting. In the analyses, we control for date of testing, as one may expect that pupils that took the test at a later date are likely to get somewhat higher scores.

Equation (1) describes the statistical model that is fitted to the data:

$$Y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 age_{ij}^2 + \beta_3 coh_{ij} + \beta_4 coh_{ij}^2 + \beta_5 td_{ij} + u_{0j} + e_{ij} \quad (1)$$

where:

Y_{ij} = outcome score of pupil i in school j (four outcome measures in this study);
age_{ij} = pupil's age (zero score stands for 6 years and zero months on 1 Sept. 2012);
coh_{ij} = pupil's age cohort (zero stands for the youngest cohort);
td_{ij} = date the pupil took the test (zero stands for 1 Sept.; one stands for 30 Nov.);
β_0 = intercept (the predicted score if age, coh, and td equal zero);
β_1 – β_5 = regression coefficients denoting the effects of the independent variables;
u_{0j} = school-specific deviation from the intercept;
e_{ij} = pupil-level deviations from the fitted scores.

The model is fitted using the SPSS software Version 23. The model fitted is a random intercepts model with fixed slopes. This implies that the average level on the outcomes is allowed to vary across schools, while the effects of the independent variables (age, cohort, and test date) are fixed (i.e., these effects are not allowed to vary across schools). The output will show estimates for the intercept (β_0) and the five regression coefficients (β_1 – β_5) and also residual variance at the school level (variance of u_{0j}) and individual level (variance of e_{ij}). Due to the way of coding the explanatory variables, the intercept (β_0) expresses the (fitted) mean score for the youngest pupils in Cohort 1 that took the test on 1 September. From the fitted scores, we can infer the relation between age and outcome scores. We expect to find discontinuities in the outcome score between the oldest pupils in one cohort and the youngest in the next cohort. These discontinuities will be reported using both the InCAS scores and as effect sizes (Cohen's *d*). The cohort effects (linear and quadratic) denote the sizes of these discontinuities. However, as noted in the introduction, the main question is whether the discontinuities decline in the upper cohorts (i.e., a

declining effect of schooling). Declining discontinuities indicate declining effects of schooling. Decreasing annual gains, as reported by Bloom et al. (2008), do not necessarily imply declining effects of schooling. Our analyses might reveal a weakening age–achievement relationship with stable cohort effects across the entire school career. As noted earlier, it is conceivable that decreases in annual learning gains are solely due to a weakening age–achievement relationship. The major goal of our analyses is to find out if the effect of schooling declines when controlling for a curvilinear age–achievement relationship.

In order to illustrate the results of the analyses, the estimated relations between age and outcomes scores (for each measure) are displayed in a number of graphs. Thus, the discontinuities between the oldest pupils in each year and the youngest in the next can be visualized. It should be noted that the effects found in the analyses apply only at the cut-offs at which the breaks in schooling age occurs. Generalizations or extrapolations away from the cut-off date (1 September) would go beyond the data. Finally, we report the effects of schooling from Cohorts 1 to 6 expressed as effect sizes (following Cohen's, 1988, definition) and as percentages of the total learning gains from Years 1 to 6.

The statistical model that is fitted to the data differs from the standard regression-discontinuity model. In the standard situation, only one discontinuity is estimated that denotes the effect of assignment to the treatment versus control group. In the present study, the number of groups involved and consequently the number of discontinuities is considerably larger. Moreover, we specifically address the question whether the effects decline in the later stages of the primary school career. Our statistical model is designed to answer precisely this question and requires the estimation of only two cohort effects (linear and quadratic). An alternative approach is to assess the cohort effects separately for each transition from one cohort to the next. This involves the estimation of five discontinuities per outcome measure. Findings from analyses based on a standard regression-discontinuity model will be reported briefly in the Results section. More details are provided in Appendix 3.

Findings

In Table 2, the annual gains are reported both using the InCAS scores and as effect sizes (Cohen's *d*), that is, the differences between one cohort and the next that can be inferred from the figures in Table 1 are divided by the pooled standard deviation across years. This implies that the year-to-year gains are expressed in terms of standard deviations. For example, Table 1 shows a difference in reading between Cohorts 3 and 4 that is equal to 1.16 (9.88 – 8.72). Divided by the pooled standard deviation (1.54), this gives an effect of .75.

Table 2. Progress compared to the previous cohort (InCAS scale and Cohen's *d*).

	Reading		Mental Arithmetic		General Maths		Developed Ability	
Cohorts	InCAS	Cohen's *d*	InCAS	Cohen's *d*	InCAS	Cohen's *d*	InCAS	Cohen's *d*
Cohort 2 vs. 1	1.77	1.15	1.74	1.18	0.86	0.80	2.05	1.04
Cohort 3 vs. 2	1.58	1.03	1.41	0.95	1.12	1.04	2.08	1.06
Cohort 4 vs. 3	1.16	0.75	1.27	0.86	0.93	0.86	1.28	0.65
Cohort 5 vs. 4	0.89	0.58	0.89	0.60	0.91	0.84	1.14	0.58
Cohort 6 vs. 5	0.76	0.49	0.89	0.60	0.95	0.88	0.97	0.49
Total	6.16	4.00	6.20	4.19	4.77	4.42	7.52	3.82

In general, the findings in Table 2 are in line with those reported by Bloom et al. (2008). Overall, we find decreasing annual growth, although the pattern for General Maths deviates from the main trend. From Years 1 to 6, the learning gains are large, as they amount to at least four standard deviations for Reading, Mental Arithmetic, and General Maths. For the last measure, Developed Ability, which is not as strongly aligned to the school curriculum, the gain is just slightly less.

The outcomes of the multilevel regression analyses are reported in Table 3. Visual displays of the relation between age and age cohort as estimated by the multilevel models are presented in Figures 1 to 4. These figures show for each outcome measure both the average scores per age and the fitted scores per age. In most respects, the multilevel analyses yield similar results for all four outcome measures. The linear age–achievement relationship is consistently positive and statistically significant. For two measures (Mental Arithmetic and Developed Ability), the quadratic term for age is significant (at the .05 level, non-directional) and negative. This indicates a weakening age–achievement relationship. The same relationship holds for reading, but in this case the quadratic term is only significant at the .05 level in a one-tailed test. For General Maths, the quadratic term is clearly non-significant, and the relationship with age is linear. The visualisations of the relationship between age and achievement are shown in Figures 1 to 4.

Table 3. Multilevel analyses – age and schooling coefficients.

	Reading			Mental Arithmetic			General Maths			Developed Ability		
Fixed effects	Coeff.	*SE*	Sign.	Coeff.	*SE*	Sign.	Coeff.	*SE*	Sign.	Coeff.	*SE*	Sign.
Intercept	5.030	.202	.000	4.178	.168	.000	6.576	.139	.000	4.365	.259	.000
Age – linear	1.053	.175	.000	1.492	.171	.000	.416	.122	.001	1.760	.200	.000
Age – squared	−.048	.025	.055	−.100	.025	.000	.027	.018	.118	−.096	.028	.001
Cohort – linear	.862	.163	.000	.408	.158	.010	.507	.114	.000	.552	.185	.003
Cohort – squared	−.087	.027	.001	−.016	.026	.547	−.030	.019	.115	−.061	.030	.045
Assessment date	−.211	.246	.392	.105	.239	.661	−.414	.188	.028	−.417	.339	.218
Variances (residual)												
Pupil level	2.042	.049	.000	1.991	.048	.000	1.030	.025	.000	2.632	.064	.000
School level	.314	.110	.004	.116	.051	.010	.138	.050	.005	.598	.204	.003
Explained (total)	64.7%			67.5%			69.4%			66.7%		

Note: Significance levels relate to non-directional tests (i.e., two-tailed).

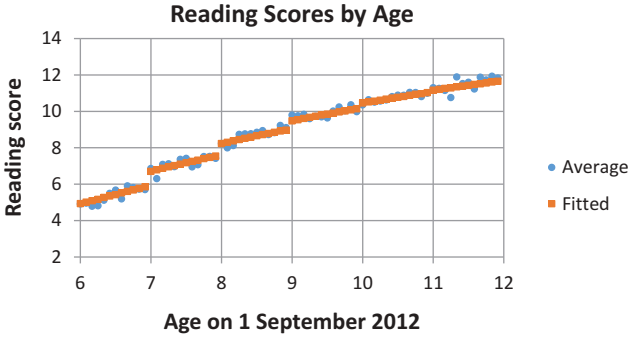

Figure 1. Reading by age; fitted scores and averages by age (month of birth).

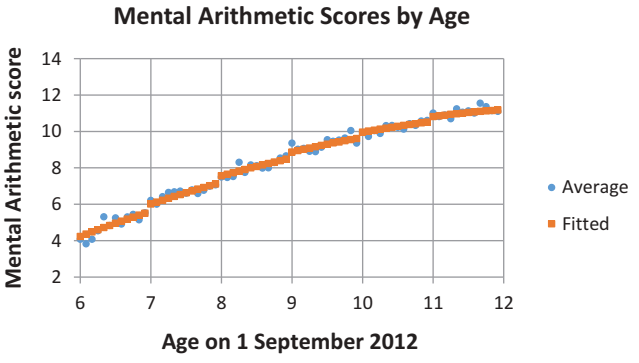

Figure 2. Mental arithmetic by age; fitted scores and averages by age (month of birth).

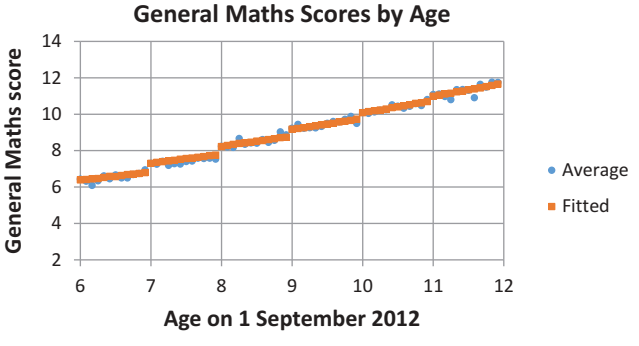

Figure 3. General maths by age; fitted scores and averages by age (month of birth).

Figure 4. Developed ability by age; fitted scores and averages by age (month of birth).

The cohort effects express the difference between the oldest pupils in a lower cohort versus the youngest in a higher cohort. Thus, they reflect the effect of schooling on the achievement measures. The linear cohort effects are all positive and statistically significant as well. Two quadratic cohort effects are significantly negative (suggesting decreasing effects). These relate to Reading and Developed Ability. For Mental Arithmetic and General Maths, the quadratic cohort effects are not significant. The cohort effects appear as discontinuities in Figures 1 to 4. These visualisations show larger discontinuities at the

early stages of the school career. The effect of the assessment date is statistically significant only for General Maths. Contrary to expectations, the effect is negative. This implies that, for this outcome, lower scores were attained by pupils that took the test relatively late in the period from early September until the end of November. The percentage explained in the total variance indicates that in this dataset, age and schooling are associated with about two thirds of the total variance. The percentages of explained variance are obtained by comparing the school- and pupil-level variances as reported in Table 3 to the variances when fitting the zero models (see Appendix 4).

Numerical details with regard to the discontinuities between adjacent cohorts are provided in Table 4. First of all, the fitted scores of the oldest pupils in one cohort and the youngest in the next are reported. The differences between both scores amount to the discontinuities that are shown in Figures 1 to 4. These discontinuities express the effects of schooling. Table 4 shows that the discontinuities (i.e., cohort effects) get smaller in the later phases of the primary school career. This goes for all four outcome measures, but the trend is most clearly apparent for Reading and Developed Ability. The trend is more moderate for Mental Arithmetic and General Maths. With regard to these latter measures, the evidence for declining cohort effects lacks statistical significance, as the multilevel analyses do not produce significant quadratic cohort effects (see Table 3). For both Reading and Developed Ability, the discontinuities between Cohorts 5 and 6 come close to zero. The differences between the youngest pupils in Cohort 6 and the oldest in Cohort 5 can be attributed almost entirely to age. At this stage of the school career, the effect of schooling on growth for these two measures appears to be quite modest. Additional analyses (see Appendix 3) indicate that the cohort effects are not significant in four instances (Reading, Cohorts 5–6; Mental Arithmetic, Cohorts 4–5; Developed ability, Cohorts 4–5 and Cohorts 5–6).

The discontinuities reported in Table 4 can also be expressed as effect sizes and as percentages of the progress from one cohort to the next. The results are displayed in Table 5. First of all, the discontinuities from Table 4 are repeated, and in the adjacent columns they are expressed as effect sizes (Cohen's *d*) and as percentage of the total progress between cohorts. The effect sizes are computed by dividing the discontinuities by the pooled standard deviations reported in Table 1 (1.54 for Reading, 1.48 for mental Arithmetic, 1.08 for General Maths, 1.97 for Developed Ability). The percentages of total progress are computed by dividing the discontinuities by the differences between

Table 4. Fitted scores and discontinuities at the cut-off points.

	Reading		Mental arithmetic		General maths		Developed ability	
Cohort and age (years-months)	fitted scores	disc.	fitted scores	disc.	fitted scores	disc.	fitted scores	disc.
Cohort 1; 6–11	5.85		5.51		6.79		5.70	
Cohort 2; 7–0	6.70	.85	6.01	.50	7.30	.51	6.32	.62
Cohort 2; 7–11	7.54		7.11		7.75		7.68	
Cohort 3; 8–0	8.23	.69	7.56	.45	8.22	.47	8.18	.50
Cohort 3; 8–11	8.97		8.48		8.73		9.36	
Cohort 4; 9–0	9.48	.51	8.87	.39	9.17	.44	9.73	.37
Cohort 4; 9–11	10.14		9.61		9.72		10.73	
Cohort 5; 10–0	10.46	.32	9.96	.35	10.08	.36	10.96	.23
Cohort 5; 10–11	11.03		10.51		10.69		11.79	
Cohort 6; 11–0	11.16	.13	10.82	.31	10.98	.29	11.84	.05
Total		2.50		2.00		2.07		1.77

Table 5. Discontinuities as effect sizes (Cohen's *d*) and percentages of progress between cohorts.

	Reading			Mental Arithmetic			General Maths			Developed Ability		
	disc.	*d*	perc.	disc.	*d*	perc.	disc.	*d*	perc.	disc.	*d*	perc.
Coh. 1 vs. 2	.85	.55	48.0%	.50	.34	28.7%	.51	.47	59.3%	.62	.31	30.2%
Coh. 2 vs. 3	.69	.45	43.7%	.45	.30	31.9%	.47	.44	42.0%	.50	.25	24.0%
Coh. 3 vs. 4	.51	.33	44.0%	.39	.26	30.7%	.44	.41	47.3%	.37	.19	28.9%
Coh. 4 vs. 5	.32	.21	36.0%	.35	.24	39.3%	.36	.33	39.6%	.23	.12	20.2%
Coh. 5 vs. 6	.13	.08	17.1%	.31	.21	34.8%	.29	.27	30.5%	.05	.03	5.2%
Total	2.50	1.62	40.6%	2.00	1.35	32.3%	2.07	1.92	43.4%	1.77	.90	23.5%

cohort means (see Table 2; e.g., difference between Cohorts 1 and 2 for reading is 1.77; therefore, .85/1.77 = 48.0%).

When expressed as effect sizes, 15 of the 20 cohort effects are in between .20 and .50. According to Cohen's guidelines (1988), this would be within the range from small to medium. The only case showing a somewhat larger effect size relates to the start of primary education (Reading, Cohort 1 vs. 2). Three of the four cases with a smaller effect relate to Developed Ability (Cohort 3 vs. 4, Cohort 4 vs. 5, and Cohort 5 vs. 6). The remaining case showing a small effect relates to Reading (Cohort 5 vs. 6). The effect sizes for Developed Ability are the smallest for each and every cohort. The cumulative effect over all cohorts is less than one standard deviation (.90) for this outcome. For the other three outcome measures, which are more closely aligned to the school curriculum, the effect over all cohorts ranges from 1.35 to 2.00.

When the effect of schooling is expressed as a percentage of entire progress from one cohort to the next, the figures show that for each measure the schooling effects account for a considerable portion but still less than half of the entire gains. The percentage is smallest for Developed Ability (23.5%) and largest for General Maths (43.4%). The percentages also indicate that in the final phase of primary education, the cohort effects account for a more modest part of the learning gains in Reading and Developed Ability (17.1% and 5.2%, respectively).

Limitations and discussion

Before discussing the main findings of our study, we will address three limitations of the present study. The first one relates to the sample size and statistical power. The other two relate to the external and internal validity of the study.

Sample size and power

At the start of this project, we expected to work with a huge dataset that would include tens of thousands of pupils and a few hundred schools. Consequently, we assumed sample size and statistical power to be issues of little relevance. In the end, the analysis was restricted to a (very) small subset of the original dataset. Although the resulting dataset can hardly be considered small as it still includes over 3,500 pupils, it needs to be acknowledged that only 20 primary schools were involved in the analyses. Considering the main purposes of this study (estimating the effects of schooling and age–achievement relationships), this is not a major problem. Due to the small number of schools, estimates of the standard errors of the school-level variance may be biased, but

the regression coefficients and their standard errors are estimated without bias in multilevel analysis (Maas & Hox, 2005, p. 90). In the current study, the regression coefficients denoting the cohort effects and age–achievement relationships are the main parameters of interest.

As the size of a sample declines, so does the statistical power (i.e., the probability to detect effects of a certain size) of the data analysis (Cohen, 1988). With regard to the present study, it can be concluded on the basis of the outcomes (especially the standard errors reported in Table 3) that relatively small effects of schooling can still be detected (assuming a .05 significance level, one-tailed and .80 power). The dataset that was analysed allows for detection of linear cohort effects that correspond to effect sizes (Cohen's *d*) between .20 and .30. See Appendix 5 for details.

External validity

An important limitation of the present study is that the pupils included in this study predominantly attended independent schools and cannot be considered a representative sample of all English pupils. Therefore, it may seem questionable whether our findings can be generalized to the English pupils in primary education as a whole. Independent schools charge substantial fees, which only wealthy parents can afford. As family background is clearly related to school achievement (e.g., Coleman et al., 1966; Strand, 1999), one can safely assume that the average scores of the pupils in our sample exceed the national average. The main research question in the present study, however, relates to the effects of schooling. The fact that pupils with high-income parents generally get higher scores on educational tests does not necessarily imply that the effect of schooling is stronger for these pupils. Actually, we *expect* that the effects of schooling are not radically different for advantaged and disadvantaged pupils. Even though elaborate sociological theories have been formulated stating that pupils from disadvantaged backgrounds are bound to profit less from formal schooling than more advantaged ones (e.g., Bourdieu & Passeron, 1990), consistent findings from empirical research on comparisons between progress during the school year versus the summer vacation (when schools are not in session) indicate that disparities in learning gains mainly develop during the summer period (Alexander, Entwisle, & Olson, 2007; Downey, von Hippel, & Broh, 2004). When schools are in session, pupils from advantaged and disadvantaged backgrounds have been found to progress at a similar pace. In the end, a larger and especially a more representative dataset will be required in order to arrive at more precise estimates of the effects of schooling at various stages in the educational career.

Internal validity

Schochet et al. (2010, pp. 2–3) list the following criteria a study should meet in order to qualify as a regression-discontinuity study:

(1) "Treatment assignments are based on a forcing variable; units with scores at or above (or below) a cut-off value are assigned to the treatment group while units with scores on the other side of the cut-off are assigned to the comparison group".

(2) "The forcing variable must be ordinal with a sufficient number of unique values".
(3) "There must be no other factor confounded with the forcing variable".

Appendix 1 provides detailed information to show that the present study meets these criteria, although there is a problem regarding the third criterion. An additional requirement with respect to this criterion is that equivalence should be demonstrated on key covariates at the cut-off of the forcing variable. As our sample lacks information on pupil backgrounds (apart from their schools, year groups, and dates of birth), it is impossible to provide empirical evidence for such equivalence. However, it seems safe to assume that pupils in adjacent years are highly similar with regard to relevant background variables. It seems particularly unlikely to find a sudden change in background variables like IQ, SES, gender, ethnicity, and motivation at the cut-off date. Empirical research that has addressed this issue indicates a strong consistency of pupil backgrounds between different school years (e.g., Luyten & De Wolf, 2011). Even when there are indications of differences between years, corrections for bias in this respect hardly affect the estimated effects of schooling (Webbink & Gerritsen, 2013, p. 19).

In the present study, the sample includes mainly pupils from independent schools. These schools charge substantial fees, some provide boarding facilities, some are single-sex schools, and some are religious schools. As a result, we can expect that the backgrounds of the pupils will be similar in each and every year.

Discussion of the main findings

This article presents findings regarding the learning gains from Years 1 to 6 of children in English primary schools on four outcome measures. Through application of a regression-discontinuity approach, the specific contribution of schooling to the cognitive gains could be estimated. Regression discontinuity was combined with intention-to-treat analysis to take into account that some school careers are delayed or accelerated. Our analyses focused on the differences in test scores between the oldest pupils in one age cohort and the youngest in the next. We specifically focused on the phenomenon of declining growth as previously reported by Bloom et al. (2008). More specifically, our analyses addressed the question if decreasing learning gains can be attributed to declining effects of schooling.

First, the present study shows weakening age–achievement relationships for three of the four outcome measures in our analyses. General Maths is the exception. General Maths is rather different from Reading and Mental Arithmetic in that new concepts and procedures are gradually introduced and learned. With regard to Reading and Developed Ability, we found convincing (i.e., statistically significant) evidence for a declining effect of schooling. Together with a positive linear relationship, negative quadratic terms of age and schooling indicate decreasing effects of schooling and weakening age–achievement relationships. The additional analyses that are based on comparisons between two adjacent cohorts (see Appendix 3) indicate significant effects in most cases. Only 4 out of 20 effects do not reach statistical significance at the .05 level, and all four cases of non-significant effects relate to the final phase of the primary school career (Cohort 4 vs. 5 or Cohort 5 vs. 6). Especially for Reading and Developed Ability, the progress which children make in the final years of primary education can

hardly be accounted for by schooling. It may be that by this age, many children have learned to read and the focus is then on using those skills to study more specialised areas of the curriculum. Similarly, Developed Ability may continue to increase but is not the specific focus of schooling by the end of the primary years. It goes without saying that future research into the association between progress in reading and developed ability and schooling would be highly relevant for evaluating the impact of educational interventions in the final years of primary school. For Mental Arithmetic and General Maths, the data analysis also revealed negative, but statistically non-significant, quadratic effects of schooling. We consider these results as additional but tentative support for the supposition that the effect of schooling on learning gains decreases in the later years of primary education. Still, the analyses clearly indicate that, even if the schooling effects decrease for Mental Arithmetic and General Maths, the decline is quite modest.

The additional analyses also indicate limited sensitivity of the estimated cohort effects to the range of birthdates around the cut-off that is used to fit the models. The analyses reported in Tables 3 to 5 all relate to the entire 6-year age range (birthdates from September 2001 until August 2007), whereas the analyses in Appendix 3 (Table A4) all relate to a 2-year age range.

Our findings point in the same direction as the ones presented by Cliffordson (2010) on maths in Swedish education from Years 6 to 8. If the conjecture that the effects of schooling decline in higher years is correct, it would imply that the benchmarks for assessing the impact of educational interventions presented by Bloom et al. (2008) are still quite ambitious. In general, these benchmarks are more lenient than the broad guidelines suggested by Cohen (1988). Still, they are based on the gross learning gains pupils make from one year to the next. The present study confirms the conclusion, already drawn in dozens of prior studies (e.g., Cahan & Cohen, 1989; Cahan & Davis, 1987; Gormley & Gayer, 2005; Luyten, 2006) that not all growth made during the school age can be attributed to schooling. A considerable part of the learning gains in English primary education can be attributed to schooling, but it should also be noted that schooling accounts for less than 40% of the total gains. Findings from other countries tend to show somewhat larger percentages (Luyten, 2006). Still, the present study shows that children make large learning gains during primary education. The total gain from Years 1 to 6 adds up to about 4 standard deviations, which is a large amount by all means. Expressed as an effect size (Cohen's *d*), the impact of schooling on learning gains from Years 1 to 6 is 1.62 for Reading, 1.35 for Mental Arithmetic, and 2.07 for General Maths (see Table 5). These are large effects, also according to the general guidelines proposed by Cohen (1988), even if they account for less than 40% of the total learning gain.

The findings presented by Bloom et al. (2008) are based on gross annual learning gains and indicate that the general guidelines suggested by Cohen (1988) may be overly ambitious for evaluating the impact of educational interventions. For example, a rise in test scores equal to 0.20 of a standard deviation may already represent nearly half the annual learning gain in the later years of primary education. Still, this would be a "small effect" according to the Cohen guidelines. Even the benchmarks based on Bloom et al. (2008) look ambitious as not all learning gain can be attributed to schooling (in the present study, it appears to be less than 40%). Using gross annual learning gains as benchmarks seems problematic, as they capture gains that are both the result of school and non-school factors. It remains unclear what actually causes the learning gains beside

formal schooling. In our view, benchmarks that are based on the learning gains that can be attributed to the effect of schooling would be preferable, although in practice it may quite challenging to set such empirically based benchmarks.

Bloom et al. (2008) also show that learning gains decline as the school career progresses. This may suggest that it is appropriate to set lower standards with regard to the effects of educational interventions in later phases of the school career. This conclusion may be premature if the declining learning gains are solely due to a weakening age–achievement relation and not to declining effects of schooling. To our knowledge, the present study is the very first to show that not only gross annual learning gains in basic skills decline in the later years of primary education, but that this also applies to the effects of schooling. Thus far, this issue has only been addressed by Cliffordson (2010) with regard to Swedish students in Years 6 to 8. Initially, the effect of 1 year of schooling may amount to .50, but in the final years, the effects decrease to near zero for some measures. Our findings also show that both the size of the effects and their decline may vary considerably across outcome measures. In other words, when setting benchmarks for the effect of educational interventions, it is not only important to consider the phase of the educational career but also the specific measure.

The findings with regard to Developed Ability deserve some specific discussion. This measure stands out from the other three, as it relates to skills that may largely be acquired outside school (vocabulary and pattern recognition). In line with our expectations, we found the smallest schooling effect (Cohen's $d = 0.90$) for this measure. However, the schooling effects for this measure are still substantial, which indicates that they are not confined to knowledge and skills that are explicitly included in the school curriculum. Similar conclusions were drawn nearly three decades ago in the study by Cahan and Cohen (1989), in relation to primary education in Israel. Their analyses provide compelling evidence for the effects of schooling on general IQ scores, including subtests on topics such as figure classification and figure analogies. These are topics that receive hardly any attention in school curricula. Even more surprising is the beneficial effect of schooling on obesity in the United States (von Hippel, Powell, Downey, & Rowland, 2007). This conclusion is based on a comparison of growth in children's body mass index (BMI) during the summer vacation versus kindergarten and first grade. The study shows considerably faster growth rates during the summer vacation. Findings like this indicate that schooling can affect pupils in unexpected ways. On the one hand, the present study and many others show that not all learning gains in skills that are generally considered part of the core curriculum (especially language and mathematics) can be attributed to schooling. On the other hand, we find compelling evidence for effects of schooling on aspects of cognitive and even physical development that are usually not seen as primary goals of education. Apparently, schooling may set processes in motion that produce unanticipated outcomes, and its impact does not stop at promoting basic cognitive skills.

Considering the large number of studies on educational effectiveness (for a recent overview, see Reynolds et al., 2014), the impact of schooling on the learning gains of children in general has only been addressed in a limited number of educational studies. Educational effectiveness research has traditionally focused on comparing different teaching methods and identifying promising levers for further improvement of education. The frequently mentioned phrase "school effect" actually refers to the percentage of variance in student achievement scores situated at the school level. The famous

Coleman report (Coleman et al., 1966) provided an estimate of this "effect" for the first time. The modest amount (15%) was considered as disappointingly small at the time, but has been confirmed in hundreds of studies (Scheerens & Bosker, 1997). It needs to be emphasized that this figure expresses variation in achievement scores between schools. It is perfectly possible that little variation between schools goes together with a large contribution of education to learning gains in a country (and vice versa). Note also that the contributions of schooling reported in the present study (in comparison to studies in other education systems) clearly exceed the 15% "school effect".

Acknowledgments

We thank the reviewers for their very helpful and insightful comments, which have been important in the production of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Hans Luyten is an associate professor of education at the Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioural, Management and Social Sciences of the University of Twente, Enschede, The Netherlands, and an honorary professor at the Centre for Evaluation and Monitoring (CEM) at Durham University, Durham, UK. His research interests include longitudinal studies both at the individual student level (growth curve analysis) and higher (trends at school and system level), international comparisons, educational disadvantage, and the development of methodologies for assessing the effect of schooling on student development.

Christine Merrell is the Director of Research and Development at the Centre for Evaluation and Monitoring (CEM), and Professor in the School of Education, Durham University, Durham, UK. Her research interests are assessment development and monitoring the progress of children through primary school, the development of children in the early years and the prediction of their later attainment, including children who are deaf and hearing impaired, the academic attainment and progress of severely inattentive, hyperactive, and impulsive young children, and ways to help them succeed in the classroom.

Peter B. Tymms is Director of the iPIPS project in the School of Education, Durham University, Durham, UK. iPIPS is an international project designed to study children starting school around the world. His main research interests include monitoring, assessment, performance indicators, ADHD, reading, and research methodology.

ORCID

Peter Tymms http://orcid.org/0000-0002-7170-2566

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review*, *72*, 167–180. doi:10.1177/000312240707200202

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. doi:10.1257/jep.15.4.69

Baltes, P. B., & Reinert, G. (1969). Cohort effects in cognitive development as revealed by cross-sectional sequences. *Development Psychology*, *1*, 169–177. doi:10.1037/h0026997

Bloom, H. S., Hill, C. J., Rebeck Black, A., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289–328. doi:10.1080/19345740802400072

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in education, society and culture*. London, UK: Sage.

Cahan, S., & Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development*, *60*, 1239–1249. doi:10.2307/1130797

Cahan, S., & Davis, D. (1987). A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal*, *24*, 1–12. doi:10.3102/00028312024001001

Cascio, E. U., & Lewis, E. G. (2006). Schooling and the Armed Forces Qualifying Test: Evidence from school entry laws. *Journal of Human Resources*, *41*, 294–318.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology*, *27*, 703–722. doi:10.1037/0012-1649.27.5.703

Cliffordson, C. (2010). Methodological issues in investigations of the relative effects of schooling and age on school performance: The between-grade regression discontinuity design applied to Swedish TIMSS 1995 data. *Educational Research and Evaluation*, *16*, 39–52. doi:10.1080/13803611003694391

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: Government Printing Office.

Crone, D. A., & Whitehurst, G. J. (1999). Age and schooling effects on emergent literacy and early reading skills. *Journal of Educational Psychology*, *91*, 604–614. doi:10.1037/0022-0663.91.4.604

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, *69*, 613–635. doi:10.1177/000312240406900501

Eurydice. (2011). *Grade retention during compulsory education in Europe: Regulations and statistics*. Brussels, Belgium: European Commission/Eurydice.

Gelman, A., & Imbens, G. (2014) *Why high-order polynomials should not be used in regression discontinuity designs* (NBER Working Paper No. 20405). Cambridge, MA: National Bureau of Economic Research.

Gormley, W. T., Jr., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's Pre-K program. *Journal of Human Resources*, *40*, 553–558.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Abingdon, UK: Routledge.

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal*, *319*, 670–674. doi:10.1136/bmj.319.7211.670

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635. doi:10.1016/j.jeconom.2007.05.001

Jabr, D., & Cahan, S. (2015). Between-context variability of the effect of schooling on cognitive development: Evidence from the Middle East. *School Effectiveness and School Improvement*, *26*, 441–466. doi:10.1080/09243453.2014.944546

Kyriakides, L., & Luyten, H. (2009). The contribution of schooling to the cognitive development of secondary education students in Cyprus: An application of regression discontinuity with multiple cut-off points. *School Effectiveness and School Improvement*, *20*, 167–186. doi:10.1080/09243450902883870

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Frey, K., Cole, M. W., … Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: Regression-discontinuity applied to TIMSS-95. *Oxford Review of Education*, *32*, 397–429. doi:10.1080/03054980600776589

Luyten, H. (2015). Schooling: Total impact of. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (Vol. 21, 2nd ed., pp. 125–127). Oxford, UK: Elsevier.

Luyten, H., & De Wolf, I. (2011). Changes in student populations and average test scores of Dutch primary schools. *School Effectiveness and School Improvement*, *22*, 439–460. doi:10.1080/09243453.2011.591614

Luyten, H., & Veldkamp, B. (2011). Assessing effects of schooling with cross-sectional data: Between-grades differences addressed as a selection-bias problem. *Journal of Research on Educational Effectiveness*, *4*, 264–288. doi:10.1080/19345747.2010.519825

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample size for multilevel modeling. *Methodology*, *1*, 86–92. doi:10.1027/1614-1881.1.3.86

Merrell, C., & Tymms, P. (2007). Identifying reading problems with computer adaptive assessments. *Journal of Computer Assisted Learning*, *23*, 27–35. doi:10.1111/j.1365-2729.2007.00196.x

Organisation for Economic Co-operation and Development. (2010). *PISA 2009 results: What makes a school successful? – Resources, policies and practices* (*Volume* IV). Paris, France: Author.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, *25*, 197–230. doi:10.1080/09243453.2014.885450

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs*. Retrieved from What Works Clearinghouse website https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rd.pdf

Stelzl, I., Merz, F., Ehlers, T., & Remer, H. (1995). The effect of schooling on the development of fluid and crystallized intelligence: A quasi-experimental study. *Intelligence*, *21*, 279–296. doi:10.1016/0160-2896(95)90018-7

Strand, S. (1999). Ethnic group, sex and economic disadvantage: Associations with pupils' educational progress from Baseline to the end of Key Stage 1. *British Educational Research Journal*, *25*, 179–202. doi:10.1080/0141192990250204

Tymms, P., & Coe, R. (2003). Celebration of the success of distributed research with schools: The CEM centre, Durham. *British Educational Research Journal*, *29*, 639–653. doi:10.1080/0141192032000133686

Verhelst, N. (2010). IRT models: Parameter estimation, statistical testing and application in EER. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 183–218). Abingdon, UK: Routledge.

von Hippel, P. T., Powell, B., Downey, D. B., & Rowland, N. J. (2007). The effect of school on overweight in childhood: Gain in body mass index during the school year and during summer vacation. *American Journal of Public Health*, *97*, 696–702. doi:10.2105/AJPH.2005.080754

Webbink, D., & Gerritsen, S. (2013). *How much do children learn in school? International evidence from school entry rules* (CPB Discussion Paper No. 255). The Hague, The Netherlands: CPB Netherlands Bureau for Economic Analysis.

What Works Clearinghouse. (2016). *WWC standards brief for attrition*. Retrieved from http://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_attrition_080715.pdf

Appendix 1. Standards for regression discontinuity designs

According to Schochet et al. (2010, pp. 2–3) a study qualifies as a regression-discontinuity study if it meets the following criteria:

(1) "Treatment assignments are based on a forcing variable; units with scores at or above (or below) a cutoff value are assigned to the treatment group while units with scores on the other side of the cutoff are assigned to the comparison group".
(2) "The forcing variable must be ordinal with a sufficient number of unique values".
(3) "There must be no other factor confounded with the forcing variable".

They note that noncompliance with treatment assignment is permitted as long as the What Works Clearinghouse (WWC) randomized control trial (RCT) standards for attrition are met. The forcing variable must include at least four unique values below and at least four unique values above the cut-off. It is also important that there is no systematic manipulation of the forcing variable. It is conceivable that in some cases the scores on the forcing variable may be manipulated, so that certain individuals are made just eligible for the treatment group (e.g., when the score on a test serves as the forcing variable).

With regard to Criterion 2, it is clear that in the present study the forcing variable is an interval variable with 72 unique values (12 per year). Regarding Criterion 1, Figure A1 illustrates the close relation between the forcing variable (age/month of birth) and assignment to year groups. The discontinuities at the cut-off dates are unmistakable. Nearly all pupils born in the month before the cut-off (September) are in a lower year than the ones born after the cut-off. Given the very low percentages of delayed and accelerated pupils (0.6% and 1.3%, respectively), over 98% of the pupils in our sample are in the "right" year given their date of birth. In English primary education, the 1 September cut-off-date is applied with great rigour when assigning pupils to school years.

Figure A2 shows the frequency of pupils per age category (month of birth). It is important to note that this figure shows no signs of discontinuities at the cut-offs that are larger than discontinuities at other points. All in all, Figure A2 fails to show a clear relation between pupil age and frequency in the sample. If scores near the cut-off value were manipulated, this would produce notable discontinuities in frequency near the cut-off dates (i.e., at ages 7, 8, 9, 10, and 11). In the present case, it does not seem likely that scores on the forcing variable (dates of birth) have been manipulated to affect assignment to school years.

Two approaches may establish if Criterion 3 is met. The first approach involves demonstrating equivalence at the cut-off value on relevant covariates. The alternative is to show that there are no indications of discontinuities away from the cut-off that correspond to alternative interventions.

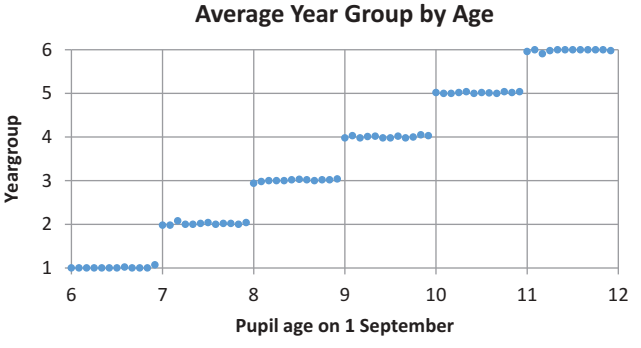

Figure A1. Relation between age and year group.

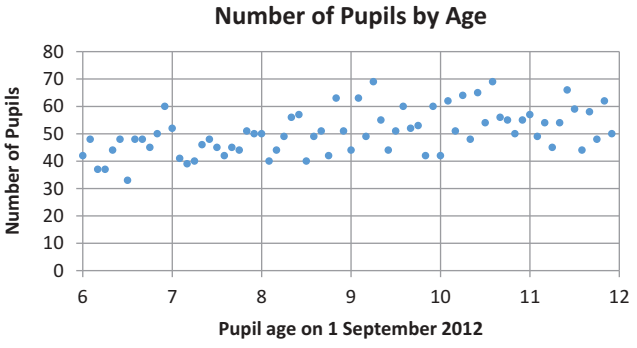

Figure A2-2. Frequency of pupils per age category (month of birth).

The first approach is not feasible in the present sample as no information on pupil backgrounds is available (apart from their schools, year groups, and dates of birth). However, in the present case, it seems safe to assume that the pupils in adjacent years are highly similar with regard to relevant background variables. It seems far-fetched to expect a relation between month of birth and relevant background variables like IQ, SES, gender, ethnicity, and motivation. The few studies that provide empirical evidence on this matter (e.g., Luyten & De Wolf, 2011) indicate that pupil backgrounds in different school years are highly similar. Webbink and Gerritsen (2013) report significant differences between years, but in their dataset pupils from different year groups did not always attend the same school. However, even in that case, corrections for sample bias hardly affected the estimated effects of schooling (Webbink & Gerritsen, 2013, p. 19). In the present study, only schools were included that had tested their pupils in all years from 1 to 6. Moreover, the sample consists of pupils in independent schools that attract specific populations. In schools like this, pupil backgrounds are probably particularly similar from one year to the next.

Figures A3-1 to A3-4 show the average scores on the four outcome measures by age. The plots show some discontinuities at the cut-off dates (especially for the early years), but no discontinuities away from the cut-off dates.

Table A1 reports the attrition rates in the present study per cohort and outcome measure. Test scores are available for the large majority (95.4%) of the pupils enrolled in Years 1 to 6 in the schools that were included in the analysis. According to the attrition standards set by What Works Clearinghouse (2016), it is important to consider two types of attrition: overall attrition (attrition for all participants) and differential attrition (differences in attrition between the intervention and comparison groups). The combination of both attrition rates determines the risk of biased results.

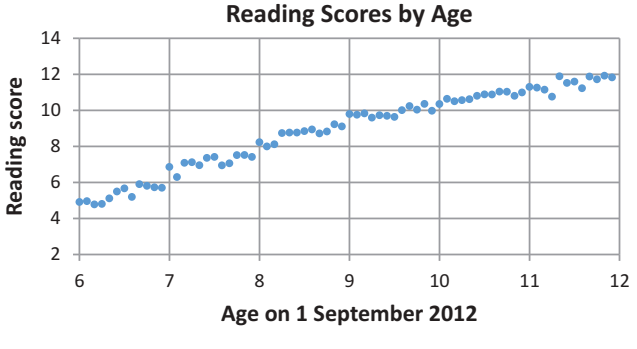

Figure A3-1. Reading scores by age.

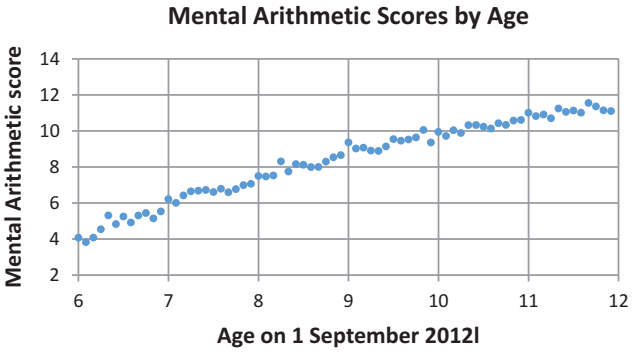

Figure A3-2. Mental arithmetic scores by age.

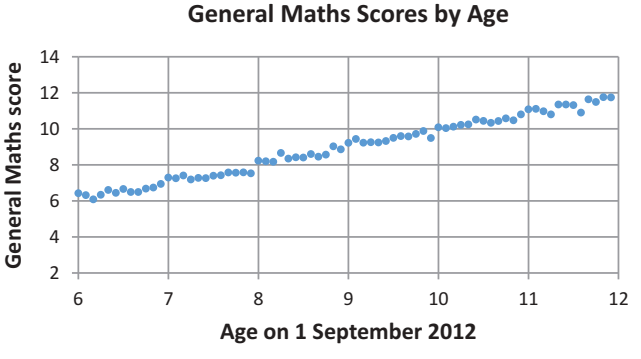

Figure A3-3. General maths scores by age.

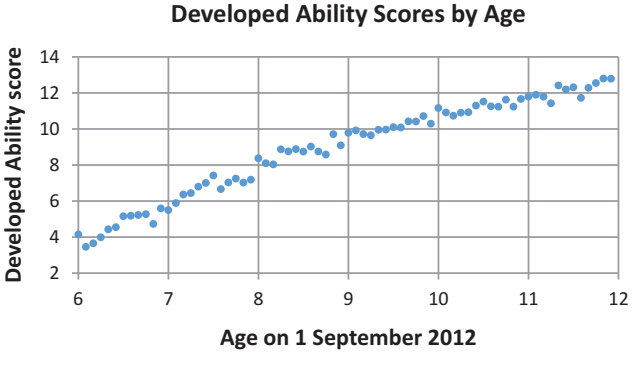

Figure A3-4. Developed ability scores by age.

Table A1. Overall attrition and attrition by treatment.

	Reading	Mental Arithmetic	General Maths	Developed Ability	Average
Cohort1	10.2%	4.6%	5.6%	8.7%	7.4%
Cohort2	3.1%	4.4%	3.0%	6.5%	3.9%
Cohort3	4.9%	3.7%	3.6%	3.6%	4.1%
Cohort4	3.3%	3.4%	2.0%	2.2%	2.7%
Cohort5	4.0%	3.9%	3.4%	3.0%	3.2%
Cohort6	7.1%	8.2%	4.8%	5.4%	6.8%
Total	5.4%	4.7%	3.7%	4.7%	4.6%

Even with high overall attrition rates (up to 65%), the risk of bias may be low if the differential attrition rates are very close to zero. In the present study, the overall attrition rates are very low as they range from 3.7% for General Maths to 5.4% for Reading. This implies that differential attrition rates up to 10% are still acceptable. The highest differential attrition rate in the present study is 7.1% and relates to Reading in Cohort 1 versus Cohort 2, which implies that attrition does not seem a likely cause for potential bias. This conclusion is based on the assumption that a liberal attrition standard is appropriate. If a more conservative standard would be applied, differential attrition should not exceed 6%. The conservative standard is used in cases when attrition is likely to be related to the intervention (e.g., a voluntary high school dropout prevention programme). In our view, the liberal standard is appropriate in the present study. However, with the exception of the differential attrition between the first two cohorts with regard to Reading, the conservative standard is met as well.

Appendix 2. School careers by month of birth and cohort

Tables A2 and A3 provide some more details on the prevalence of delay and acceleration in school careers by month of birth and cohort. First of all, both tables clearly show that delay and acceleration are extremely rare in English primary education. More than 98% of the pupils in

Table A2. School careers by month of birth.

	Delayed	On Track	Accelerated	*N*
September	0.3%	96.0%	3.7%	312
October	0.0%	98.7%	1.3%	282
November	0.0%	98.6%	1.4%	328
December	0.6%	98.7%	0.6%	303
January	0.3%	97.8%	1.9%	304
February	0.4%	98.2%	1.4%	274
March	0.6%	98.5%	0.9%	303
April	0.0%	99.0%	1.0%	287
May	0.3%	99.0%	0.7%	326
June	1.8%	97.4%	0.8%	318
July	0.7%	98.7%	0.7%	287
August	2.8%	96.5%	0.7%	310
Total	0.6%	98.1%	1.3%	3634

Table A3. School careers by cohort.

	Delayed	On Track	Accelerated	*N*
Cohort 1	0.0%	99.1%	0.9%	540
Cohort 2	0.4%	98.0%	1.7%	543
Cohort 3	0.8%	97.6%	1.5%	592
Cohort 4	0.9%	97.5%	1.6%	642
Cohort 5	0.3%	97.8%	1.9%	671
Cohort 6	1.2%	98.8%	0.0%	646
Total	0.6%	98.1%	1.3%	3634

our sample are on track. Table A2 shows that delay and acceleration are most frequent among pupils born in the months on either side of the cut-off date. August-born pupils are more frequently delayed than pupils born in any other month, and acceleration occurs most frequently among September-born pupils.

Table A3 does not indicate much variation among cohorts. The zero percentages of delay in Cohort 1 and acceleration in Cohort 6 require some extra comment. The delayed pupils from Cohort 1 are missing from the sample because they are not yet in Year 1 of primary school, and the same goes for accelerated pupils from Cohort 6. Our sample only includes pupils in Years 1 to 6, and consequently a small number of pupils from Cohorts 1 and 6 are missing. The number of missing pupils is probably very small. Considering the percentages of delay and acceleration in the other cohorts, we estimate that the number of delayed pupils missing from Cohort 1 amounts to about 5 (less than 1%) and that the number of accelerated ones missing from Cohort 6 amounts to about 10 (1.5%). It seems unlikely that such small numbers have a substantial impact on the findings.

Appendix 3. Estimating cohort effects with standard regression discontinuity

The findings reported in the results section are based on a statistical model that differs from the standard regression-discontinuity model. In the standard situation, only one discontinuity is estimated that denotes the effect of assignment to the treatment versus control group. In the present study, the number of groups involved and consequently the number of discontinuities is considerably larger. To check the robustness of the findings, the standard regression-discontinuity model is fitted to the data for five pairs of cohorts (1–2; 2–3, etc.) for each outcome measure. Equation (2) describes this model.

$$Y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 coh_{ij} + \beta_3 age_{ij} \times coh_{ij} + \beta_4 td_{ij} + u_{0j} + e_{ij} \quad (2)$$

Cohort amounts to a binary variable. It is customary to recode the treatment variable (cohort) to zero and one. The forcing variable (age) is centred on the cut-off date. As a result, β_0 (the intercept) denotes the predicted outcome of the oldest pupils in the first cohort, and β_2 denotes the difference in outcome between the oldest pupils in the first cohort and the youngest in the second cohort. The inclusion of an interaction term of age with cohort denotes that the age–achievement relationship may differ between both cohorts. In this case, β_1 expresses the age–achievement relationship in the first cohort, and β_3 indicates to what extent the age–achievement relationship is stronger or weaker in the second cohort. This analysis produces a number of cohort effects (β_2), which can be compared to the discontinuities that result from fitting the model described by Equation (1).

The findings are reported in Table A4. Only the fixed regression coefficients are reported (variances of u_{0j} and e_{ij} are not reported). The table shows that all the intercepts significantly differ from zero, which is hardly surprising. More important, the age coefficients are also significant in each of the 20 analyses (at least at the .05 level in a two-tailed test). All cohort effects are positive, but they are not always significant at the .05 level. Most of the non-significant effects relate to Developed Ability and/or to the older cohorts. The cohort effects are clearly not significant in four instances (Reading, Cohorts 5–6; Mental Arithmetic, Cohorts 4–5; Developed ability, Cohorts 4–5 and Cohorts 5–6). In three instances, the cohort effect is significant at the .05 level in a one-tailed test but not so in a non-directional test (Developed Ability, Cohorts 1–2 and Cohorts 3–4; General Maths Cohorts 5–6). The interaction effects of age with cohort are not significant in most cases. Only two interactions are significant at the .05 level (two-tailed). This means that from one cohort to the next, the age–achievement relationships do not differ significantly. This seems to conflict with the findings presented in the results section, which indicate weakening age–achievement relationships for three outcome measures. On the other hand, Table A4 shows smaller age and cohort coefficients in the later stages of the primary school career in general. In most cases, the coefficients of test

Table A4. Cohort effects obtained with standard regression discontinuity (Equation [2]).

	Reading			Mental Arithmetic			General Maths			Developed Ability		
	Effect	*SE*	Sign.	Effect	*SE*	Sign.	Effect	*SE*	Sign.	Effect	*SE*	Sign.
β_0 (Intercept)	5.535	.272	.000	5.360	.267	.000	6.740	.151	.000	5.673	.408	.000
β_1 (Age)	1.143	.218	.000	1.624	.230	.000	.577	.116	.000	1.866	.274	.000
β_2 (Cohorts 1–2)	.908	.178	.000	.657	.189	.001	.430	.095	.000	.425	.224	.058
β_3 (Age × Cohort 2)	−.366	.304	.229	−.807	.327	.014	−.229	.164	.161	−.329	.387	.395
β_4 (Test date)	.655	.410	.113	.589	.436	.179	.303	.255	.238	.135	.666	.840
β_0 (Intercept)	6.918	.321	.000	7.029	.272	.000	7.654	.184	.000	8.069	.405	.000
β_1 (Age)	.741	.233	.001	.824	.220	.000	.354	.131	.007	1.592	.258	.000
β_2 (Cohorts 2–3)	.615	.191	.001	.383	.179	.033	.514	.106	.000	.572	.210	.007
β_3 (Age × Cohort 3)	.342	.325	.294	.256	.306	.404	.355	.182	.052	−.522	.356	.143
β_4 (Test date)	1.193	.533	.026	.148	.500	.768	−.033	.331	.921	−.588	.748	.433
β_0 (Intercept)	9.078	.304	.000	8.433	.253	.000	8.745	.190	.000	9.606	.338	.000
β_1 (Age)	1.058	.215	.000	1.070	.197	.000	.697	.144	.000	1.039	.230	.000
β_2 (Cohorts 3–4)	.435	.170	.011	.382	.156	.014	.360	.114	.002	.328	.182	.071
β_3 (Age × Cohort 4)	−.657	.299	.028	−.367	.274	.181	−.246	.200	.219	−.233	.319	.467
β_4 (Test date)	.196	.549	.722	.295	.496	.554	.316	.361	.383	−.243	.671	.718
β_0 (Intercept)	10.613	.274	.000	9.563	.223	.000	9.961	.198	.000	11.193	.316	.000
β_1 (Age)	.419	.197	.034	.731	.181	.000	.442	.153	.004	.831	.215	.000
β_2 (Cohorts 4–5)	.354	.163	.030	.174	.149	.243	.320	.126	.011	.253	.177	.153
β_3 (Age × Cohort 5)	.201	.279	.471	.029	.256	.911	.250	.218	.250	.017	.304	.955
β_4 (Test date)	−1.356	.515	.009	.221	.435	.614	−.564	.386	.148	−1.233	.614	.046
β_0 (Intercept)	11.563	.265	.000	10.721	.217	.000	10.941	.216	.000	11.794	.317	.000
β_1 (Age)	.603	.190	.002	.803	.182	.000	.674	.166	.000	.835	.207	.000
β_2 (Cohorts 5–6)	.081	.155	.604	.296	.150	.048	.231	.135	.087	.028	.169	.870
β_3 (Age × Cohort 6)	.094	.269	.728	−.386	.259	.135	.091	.233	.696	.206	.293	.482
β_4 (Test date)	−1.457	.515	.006	−.503	.444	.265	−.717	.439	.106	−.291	.621	.640

Note: Significance levels relate to non-directional tests (i.e., two-tailed).

date are not significant. This is in accordance with the findings presented earlier, which only show a significant (but negative) effect) for one of the outcome measures (General Maths). In the analyses reported in Table A4, three of the exceptions relate to reading. Two of these significant effects are negative, which implies that pupils score low if they take the test relatively late. The fourth exception relates to developed ability and again to a negative effect.

The results reported in Table A4 can be compared to the results based on Equation (1) as reported in Table 4. In general, the results from Equations (1) and (2) are quite similar. This is shown in the Figures A4-1 to A4-4, which show the cohort effects based on Equations (1) and (2). The effects from Equation (1) show a smoothed pattern compared to the effects based on Equation (2), but both effects reveal a similar, downward trend. The cohort effects appear to decline near the end of the primary school career.

Appendix 4. Zero models

Table A5 shows the findings when a zero multilevel model is fitted to the data. The main purpose of these models in the present study is that they serve as a baseline for calculating the percentages of variance explained of the models described by Equation (1). These percentages are reported in Table 3.

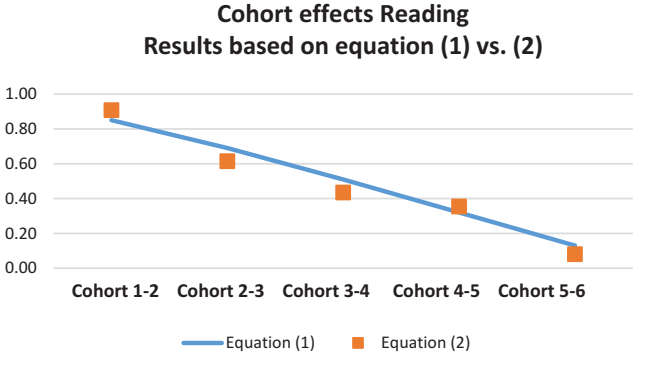

Figure A4-1. Cohort effects reading; Equation (1) vs. (2).

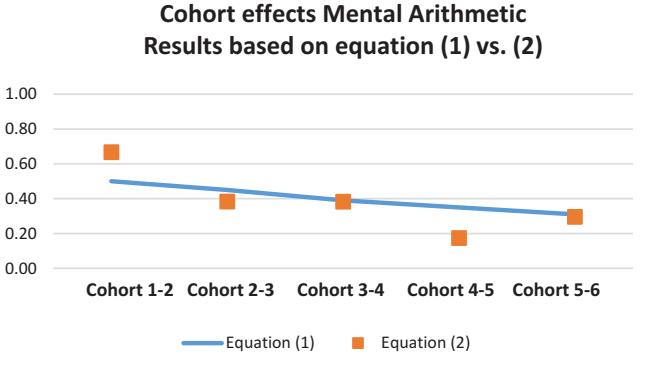

Figure A4-2. Cohort effects mental arithmetic; Equation (1) vs. (2).

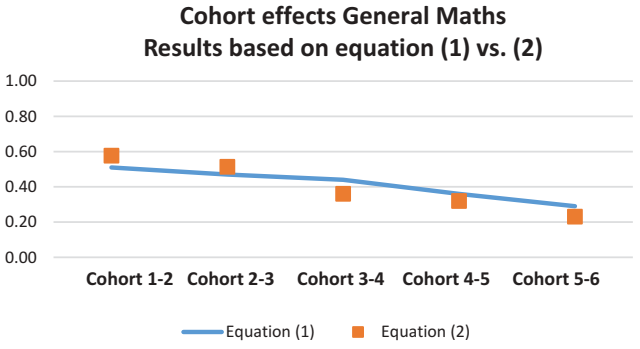

Figure A4-3. Cohort effects general maths; Equation (1) vs. (2).

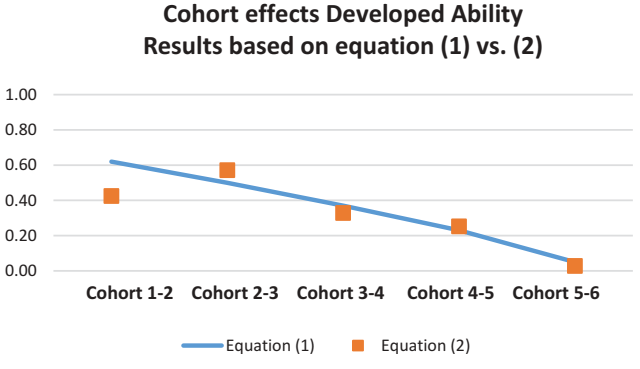

Figure A4-4. Cohort effects developed ability; Equation (1) vs. (2).

Table A5. Pupil- and school-level variance.

	Reading	Mental Arithmetic	General Maths	Developed Ability
Pupil Variance	6.080	6.228	3.531	8.480
School Variance	.600	.258	.280	1.232
Total Variance	6.680	6.487	3.811	9.713

Appendix 5. Statistical power

Based on the standard errors reported in Table 3, it is possible to calculate what effects the analyses were able to detect given a certain level of significance (α) and statistical power (1 – β). It can be concluded that the dataset that was analysed allowed for the detection of linear cohort effects that correspond to effect sizes (Cohen's *d*) between .20 and .30 at the .05 significance level (one-tailed) and .80 power.

For example, the analyses show a standard error for coefficient β_3 (denoting the linear cohort effect) equal to .163 (see Table 3). One can calculate the discontinuity that can be detected with .80 power and .05 significance, by multiplying the standard error by 2.50. This gives a discontinuity of .408 (see Table A6). The quantity 2.50 is the sum of 1.65 and .85, which are the *z* values associated with the .05 and .20 probability levels (α and β, respectively). The discontinuity

Table A6. Linear cohort effects detectable at .05 significance (one-tailed) and .80 power.

Outcome	Standard error (see Table 3)	Discontinuity	*SD* Outcome (see Table 1)	Effect size (Cohen's *d*)
Reading	0.163	0.408	1.54	0.26
Mental Arithmetic	0.158	0.395	1.48	0.27
General Maths	0.114	0.285	1.08	0.26
Developed Ability	0.185	0.463	1.97	0.23

obtained in this way can be expressed as an effect size (Cohen's *d*) by dividing it by the standard deviation of the outcome measure involved (see Table 1). For Reading, the pooled standard deviation equals 1.54. Therefore, it can be concluded that the detectable effect size equals .26 (.408/1.54). Table A6 provides numerical details four all four outcome measures.