Studies in Educational Evaluation 67 (2020) 100939

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Absolute effects of schooling as a reference for the interpretation of educational intervention effects

Hans Luyten [a,*], Christine Merrell [b], 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, Netherlands
[b] School of Education, Durham University, Durham, UK

ARTICLE INFO

Keywords:
Absolute effect of schooling
Regression-discontinuity
Effect size
Primary education
Northern Ireland

ABSTRACT

Knowledge of the absolute effects of schooling provides a useful reference for the interpretation of the effectiveness of educational interventions. We use discontinuities in test scores between the oldest pupils in one birth cohort and the youngest in the next to assess the absolute effects of schooling. Our study includes 90 % of all pupils in year-groups 4–6 of primary education (ages 7–10) in Northern Ireland. Assignment to year-groups is strictly determined by date of birth in Northern Ireland. This creates a situation which parallels randomized controlled experimentation. The findings support the view that the guidelines suggested by Cohen (in 1969) may be overly ambitious when evaluating the effectiveness of educational interventions.

1. Introduction

Most educational research is driven by an ambition to improve school practices and, thereby, pupils' outcomes. Numerous studies address the effects of specific educational approaches, comparing them to control conditions, which typically involve "business as usual". Mainstream educational research thus focuses on *relative* effects. In contrast, research aiming to assess the *absolute* effects of schooling is rare. It might be thought that assessing the effects of schooling itself would involve a control group of students that receive no education at all. However, even without any formal education, children acquire some knowledge and skills. Language acquisition presents an illustrative example; before entering school, many children have already acquired vocabulary and are learning the rules of syntax. Pupils may make a lot of progress over the course of a school year, but it cannot be concluded that schooling is the sole cause. As a result, total annual learning gains paint an imperfect picture of the actual effect of one year of schooling. For some outcome measures the contribution of non-school factors to learning gains will be much larger than for others. The aim of this paper is to assess the contribution of schooling to pupils' total learning gains.

Well-founded knowledge of the absolute effects of schooling provides a valuable frame of reference for the interpretation of relative effects of additional educational interventions. In most cases, standardized effect sizes are used to express the impact of an educational intervention. This metric expresses the difference between the pupils in the intervention group versus those in a comparison group in terms of standard deviations. E.g., an effect size of 0.50 implies that pupils in the intervention group score half a standard deviation above the control group average. But it remains difficult to understand what such effect sizes imply in real-world contexts. To assist with their interpretation, effect sizes are often "translated" to years or months of learning. If the effect of an educational intervention is equal to one third of the total annual learning gain, it may seem reasonable to equate the intervention effect to four months of learning. However, Baird and Pane (2019) point out that such a translation assumes linear growth over time which is somewhat questionable. Still, comparing intervention effects to total annual learning gains seems useful to put effect sizes into perspective. This also applies when an alternative measure such as the "Improvement Index" (U.S. Department of Education, 2014) is used. This index expresses the difference between the percentile rank of an average pupil in the experimental group and the percentile rank of an average counterpart in the control group. For example, an effect size of 0.50 implies that an average pupil in the experimental group would score at 69th percentile in the control group (i.e., a 19 percentile gain).

The guidelines for the interpretation of effect sizes as suggested by Cohen (1969) are still widely applied. In Cohen's categorisation, effect sizes around 0.20 (8 percentile gain) are considered small, while effects sizes of 0.50 and 0.80 represent medium and large effects (19 and 29 percentile gains respectively). Kraft (2019) argues that these interpretations are overly ambitious with regard to educational

* Corresponding author.
E-mail address: j.w.luyten@utwente.nl (H. Luyten).

https://doi.org/10.1016/j.stueduc.2020.100939
Received 9 June 2020; Received in revised form 15 October 2020; Accepted 21 October 2020
Available online 9 November 2020

interventions. Effects found in randomized controlled trials are usually much more modest in the field of education. Kraft proposes that effect sizes of to 0.20 or higher should be considered large. Effects ranging from 0.05 to 0.19 (2–7.5 percentile gains) may be considered medium. Kraft's recommendation is based on a meta-analysis of 1,942 effects from 742 studies on randomized controlled trials in education that use standardized tests as outcome measures. The median effect size in this meta-analysis is 0.10 (4 percentile gain). One in four effects is zero or even negative.

It is important to take into account that annual learning gains vary across the school career and outcome measures. The benchmarks developed by Bloom, Hill, Rebeck Black, and Lipsey (2008) are informative in this respect. They report annual gains for four subjects (reading, math, science and social studies) from U.S. nationally normed tests as effects sizes (Bloom et al., 2008;). In general, learning gains become less as pupils grow older. Reading gains are especially large in the early years of schooling and then tend to decrease more strongly than annual math gains. It should be noted that even for a single subject the results vary to some extent across different tests (Bloom et al., 2008;). Furthermore, it is important to note that the benchmarks by Bloom et al. (2008) represent *total* annual learning gains. That is, both the gains that can be attributed to schooling and non-school factors (like maturation and informal learning).

Therefore, more detailed knowledge of the absolute effects of schooling is useful for putting research findings about the effects of educational interventions into perspective. It also seems likely that the contributions of schooling to total annual learning gains vary both across year-groups and outcome measures. In the present paper, discontinuities in test scores between the oldest pupils in one birth cohort and the youngest in the next are used to assess the absolute effect of schooling.

2. How can the impact of schooling be estimated?

Assessing the proportion of pupils' learning gains which can be attributed to schooling is challenging. Randomized controlled trials are not an option. Ethical issues aside, a control group that receives no schooling would be forbidden by law in most countries. Nevertheless, a rigorous and scientific approach is feasible. This is illustrated in the present paper, which reports research findings that relate to primary education in Northern Ireland, where assignment to year-groups is almost completely determined by date of birth. Over 99 % of the pupils are in the expected year-group by their date of birth. (for more details see the supplementary materials, Section 1). As a result, pupils that differ by only a few days in age are placed in adjacent year-groups. This creates a borderline group that closely resembles random assignment to different treatments. It seems safe to assume that, apart from a minute difference in age, pupils on either side of the cut-off date are similar in all other respects. Obviously, children make considerable gains in knowledge and skills during their time at school (Baird & Pane, 2019; Bloom et al., 2008). But, as noted above, it would be a mistake to attribute *all* learning gains during the school years to schooling.

Although educational researchers have considered several options for assessing the absolute effect of schooling, it can be argued convincingly that a comparison of (nearly) same age children in adjacent year-groups yields the most conclusive results (Ceci, 1991). This approach has been adopted in a limited number of studies. Estimates of the contribution that schooling makes are based on the difference in test scores between the oldest pupils in one year-group and the youngest in the next. The approach is known as regression-discontinuity (RD) and strongly draws on the use of cut-off points to determine assignment to different treatments (Imbens & Lemieux, 2008; Trochim, 1984). In the present case, it involves a statistical modelling of the relationship between age and test scores for pupils in adjacent birth cohorts. A sudden "jump" (discontinuity) in the relationship at the cut-off date provides compelling evidence for an effect of schooling. It is particularly difficult to come up with plausible explanations for the discontinuities other than assignment to year-groups. A comparison between pupils born on either side of the cut-off date amounts to a close approximation of the classic randomized controlled trial, which counts as the gold standard for assessing causal effects.

As with randomized experiments, RD does not require any statistical corrections for confounding variables. In RD the cut-off point mimics random assignment to different "treatments". Randomization ensures that different groups will be equivalent on any characteristic except the assigned treatment. The same goes for pupils on either side of the cut-off date. It is extremely unlikely that achievement-related background characteristics like aptitude, learning motivation, family income, educational level of the parents or ethnicity suddenly change at the cut-off date. As a result, statistically controlling for the impact of such variables will not affect the estimated effect of schooling. That would only be the case if pupils on either side of the cut-off date differ in these respects.

Most alternative approaches to assess the absolute effect of schooling as discussed by Ceci (1991) focus on unusual situations such as changes in compulsory schooling regulations, suspended schooling during wartime or absence of schooling in isolated regions. Clearly, these approaches are not suited for routine application.

Another viable approach is known "seasonality of learning". Here, the effect of schooling is estimated by comparing learning rates during the school year to learning rates during the summer holidays. The underlying assumption is that during the school year learning rates are affected by both school factors and non-school factors, while during the summer holidays only non-school factors have an effect (Heyns, 1978). It is important to point to both practical complications and conceptual problems with regard to this approach. The first practical complication is that it requires longitudinal data. Collecting such data comes with specific challenges like correct matching of data collected at different points in time and avoiding selective attrition. In addition, it is more time-consuming than an approach that requires only cross-sectional data (like RD). Second, timing of the tests is crucially important. Ideally one would want to test pupils on the very last day of school before the start of the summer holidays and re-test them at the first day of the new school year. In practice, this is hardly ever feasible. As a result, additional corrections are nearly always needed. On a conceptual level, the validity of this approach revolves around the assumption that the summer holidays can be considered a viable alternative for the control condition in a randomized experiment. This comes down to assuming that the impact of non-school factors (like parental support) during the summer holidays gives an accurate estimate of their effect in the absence of schooling. It may also be the case that the summer holidays are an opportunity for closer contact between parents and children at a level that is not feasible for a prolonged period. In short, RD presents a more straightforward and closer approximation of a classic randomized experiment.

3. Prior research

All previous studies into the effect of schooling based on RD, that we have been able to identify, made use of sampled data, whereas the present study covers 90 % of an entire population (see Section 2 of the supplementary materials for more details on population coverage; see Section 5 for a list of prior RD studies on the absolute effects of schooling).

An important complication in previous research has been that cut-off dates are applied with some degree of flexibility, especially for children with birth dates close to the cut-off date. Pupils who are (believed to be) slow learners are more likely to end up in a lower year-group than expected given their date of birth and particularly talented pupils are relatively often assigned to a higher year-group (Cahan & Cohen, 1989). The usual approach for dealing with this complication has been to exclude delayed and accelerated pupils. Inevitably, this limits the conclusions of the research findings, specifically because high and low performing pupils with birth dates close to the cut-off date are excluded. In some studies, estimation of the age-achievement relationship is based on

pupils born at least two months after the cut-off date (e.g., Cahan & Cohen, 1989; Cahan & Davis, 1987; Wang, Ren, Schweizer, & Xu, 2016). Thus the estimated relation is solely based on birth months with small percentages of delayed and accelerated pupils. However, the estimated effect of schooling (i.e., the discontinuity between the oldest pupils in the lower year-group vs. the youngest pupils in the higher year-group) is then based on extrapolations of the age-achievement relationship. In such cases, discontinuities are *inferred* rather than observed. In Northern Ireland, deviations from assignment in line with the cut-off date are extremely rare. Moreover, the small number of pupils who were outside of their expected year-group have *not* been excluded in the present study. The analyses focus on the effect of *providing* schooling. In prior studies researchers usually tried to assess the effect of *receiving* schooling. Rather than focussing on differences between year-groups, our approach focusses on differences between the age cohorts on either side of the cut-off date. This amounts to an intention-to-treat (ITT) analysis, which uses the initial treatment assignment and not the treatment that is eventually received. In medical research this approach is frequently applied to take into account the point that the effect of a treatment may be impeded, if not all eligible individuals actually receive or complete the treatment (Hollis & Campbell, 1999). A similar situation arises in education, when school careers are delayed. In that case, some children do not receive the "treatment" to which they are eligible given their date of birth. Likewise, pupils placed in a year-group higher than expected for their age receive a more advanced treatment than most of their same age peers. ITT renders the findings more suitable for comparisons across education systems. It is conceivable that some educational systems produce strong year-group effects, while, at the same time, a high percentage of the school careers is delayed. In such cases, the prevalence of delayed careers may offset the year-group effects to some extent. The frequency of delayed schooling varies considerably across countries (Eurydice, 2011), but as noted earlier it is very low in Northern Ireland. This implies that the overlap between providing and receiving schooling is almost perfect.

Empirical studies of the effects of schooling that are based on regression discontinuity consistently show positive effects. Most studies indicate that the effects of schooling outweigh age effects (e.g., Cahan & Cohen, 1989; Cliffordson, 2010; Luyten, 2006), but there is considerable variation in the size of the effects. This may be due to variation in outcome measures, age ranges and educational systems. Findings that relate to the United Kingdom show relatively large effects of age and relatively modest effects of schooling (Luyten, 2006; Luyten, Merrell, & Tymms, 2017). Most studies show that at least some part of the total learning gain attained during the school career is related to age. Only a limited number of studies address the variation in schooling effects at different stages of school (Cliffordson, 2010; Luyten et al., 2017; Wang et al., 2016). These findings suggest declining effects of schooling as the school career progresses. In the first years of primary education, the effect of schooling is found to be particularly strong for reading. In later stages of the school career, the effect of schooling tends to be larger for mathematics (Luyten et al., 2017). Nearly all prior research has been focussed on the effect of receiving schooling. Studies focussing on the effects of providing schooling have yielded relatively modest effects. These studies suggest that the effects sizes of one year schooling in primary education range from 0.20 to 0.40 (Luyten et al., 2017; Webbink & Gerritsen, 2013).

4. Context of the present study: Primary education in Northern Ireland

Compulsory education in primary schools in Northern Ireland begins with the Foundation Stage for children aged 4–6 years (year-groups 1 and 2) and covers 7 years. The curriculum in primary education is set out in six areas of learning: Language and Literacy; Mathematics and Numeracy; The Arts; The World Around Us (encompassing geography, history, science and technology); Personal Development and Mutual Understanding; and Physical Education (CCEA, 2007). At age 11, pupils move on from primary school. Northern Ireland has a tradition of selective education for students of 11 years upwards. Although the selective system of secondary education has been subject to review since the late 1990s (Gallagher & Smith, 2000), secondary schools continue to be allowed to select pupils on the basis of their academic ability.

5. Research population

The findings we report relate to 59,113 pupils from 775 schools in the school-year 2011–12. This constitutes 90 % of all pupils enrolled in year-groups 4–6. Even though data on pupils in year-group 7 are available, we decided not to include findings about that cohort in the present paper. The upcoming transition to secondary education makes it difficult to interpret the discontinuities unmistakably as effects of schooling. Private tutoring and preparation for the selective secondary school system increases substantially in the final year of primary school. In other words, discontinuities between year-group 6 and 7 may pick up effects of private tutoring and additional preparation for secondary school as well as general schooling. For more details on selection and deselection of data records, see the supplementary materials, Section 4. A number of records from cohorts 4–6 are excluded, because they cannot be linked to identifiable pupils. Additional data analyses tested whether the discontinuities are different for this group. The findings show that estimates of the discontinuities are virtually the same if we include the initially excluded records.

On 2 July 2011 (the cut-off date that determines assignment to year-groups), the ages of the pupils ranged from 7 years (the youngest pupils in year-group 4) to 10 years (the oldest pupils in year-group 6). The data analysis focusses on birth cohorts of pupils, i.e. those pupils who are expected to be in a certain year-group given their date of birth. Pupils born before 2 July (exactly halfway the year) are nearly always assigned to a higher year-group than the ones born at 2 July or later.

The present secondary data analyses conducted in the present study had received ethical approval from the School of Education ethics committee at Durham University.

6. Measurements: the InCAS assessments

Data were collected in 2011 in the first months after the summer vacation (mostly in September and October; for more details, see the supplementary materials, Section 1) using the InCAS assessment. InCAS (Interactive Computerised Assessment System) was developed and run by the Centre for Evaluation and Monitoring (CEM) at Durham University (Merrell & Tymms, 2007). It is an on-line adaptive assessment designed for children aged 5–11 years. The software takes the child's age at the time of assessment as the starting point to select an appropriate first question. In response to the child's right and wrong answers, they are presented with easier or more difficult items; a series of stopping rules are used to decide when to stop each section and move on to the next. In this way, each child was presented with an assessment that was appropriate for their ability and motivating because they did not spend excessive amounts of time attempting questions that were either too simple or beyond their reach. All primary schools were required by the Northern Ireland Department of Education to assess their pupils' reading and general mathematics attainment in year-groups 4, 5, 6 and 7 on a mandatory basis so as to provide schools with detailed information at pupil level to target their teaching and monitor progress, and to provide parents with an overview of their children's learning. Pupils' scores were not collated centrally by the Department or used for school accountability purposes.

The reading assessment consisted of three separate tests (word recognition, word decoding and reading comprehension). Pupils completed the word recognition and decoding sections, and then on the basis of their scores on those sections proceeded to the comprehension section or not, if their scores were very low. The general maths assessment consisted of multiple-choice questions which were reflective of the school curriculum. Nearly half the schools also administered optional tests on spelling, picture vocabulary, nonverbal ability and mental

arithmetic. InCAS was adapted for use in Northern Ireland, ensuring that the content in the reading, spelling and mathematics sections was appropriate for the school curriculum, and that the picture vocabulary items were culturally appropriate. Curriculum and educational experts from the Council for the Curriculum, Examinations and Assessment Authority (CCEA), Northern Ireland and school teachers examined the assessment content alongside the results from extensive trials to validate the assessment prior to its mandatory use. For further information about the validity of InCAS, including correlations between InCAS and other assessments, see Merrell and Tymms (2007). In addition to the test scores, the key variables in the analyses are the pupils' ages and their birth cohorts. The date of testing was also included in the analyses. In this paper, we focus on the (mandatory) reading and general mathematics tests. As an example of the procedure consider the word recognition test; pupils hear a high or medium frequency word being read aloud to them using computer sound files, in this case a voice with a widely understood Northern Irish accent. The word is repeated within a sentence to put it in context. They must then select the target word from a choice of five words on screen. The word decoding test involves nonsense and unfamiliar words. After hearing the word, pupils must select the target word from a choice of five words on screen. For the comprehension test, the pupil reads through a passage and, when given a choice of three words, must select the word that fits into the sentence most appropriately. The general mathematics test covers the following topics: informally and formally presented number problems, measures, shape and space, handling data. In all sections, when a pupil has answered a certain number of questions incorrectly, the software stops the section and moves onto the next. It is important to note that the InCAS assessment yields vertically equated test scores. This means that scores can be mapped on to a common scale, even though pupils do not take identical tests. Our method of data-analysis (RD) requires comparable scores in adjacent groups.

7. Data analysis

The primary aim of the data-analysis is to assess discontinuities in the age-achievement relationship at the cut-off dates that determine assignment to year-groups. Regression analysis and multilevel analysis (in SPSS, version 25) are used for this purpose. Each discontinuity denotes the effect of being in the older vs. the younger cohort. The discontinuities are expressed as effect sizes (Cohen's d) and also in terms of the improvement index. We also report the discontinuities as percentages of the difference between the mean scores of two adjacent cohorts. The analyses take into account variations in the date of testing. It is also taken into account that the discontinuities may vary between schools and that the effect of age may vary from one cohort to the next. The following discontinuities are assessed:

- pupils born just after 1 July 2003 (nearly always assigned to year-group 4) versus those born just before 2 July 2003 (nearly always assigned year-group 5)
- pupils born just after 1 July 2002 (nearly always assigned to year-group 5) versus those born just before 2 July 2002 (nearly always assigned to year-group 6)

Discontinuities that are expressed as Cohen's d can be compared to effects in other fields of research. Cohen's d is defined as the difference between two groups divided by the pooled standard deviation.[1] If the difference in mean test scores between adjacent cohorts is small, the discontinuity can never be large. Therefore, expressing discontinuities as percentages of differences between cohorts may be more informative in some respects.

A number of different statistical models were fitted to the data. For a detailed account of the findings we refer to the supplementary materials, Section 3, which reports findings on the effect of control variables and variance of discontinuities between schools. The main text of this paper focuses exclusively on the discontinuities between cohorts. The main message from the supplementary materials is that different statistical models yield highly similar estimates of the discontinuities at the cutoff dates. The supplementary materials (Section 3) also provides evidence to show that the standards for regression-discontinuity designs as described by Schochet et al. (2010) are met.

All analyses relate to comparisons between two adjacent cohorts (cohort 5 vs. 4; cohort 6 vs. 5). Three different models are fitted to estimate the cohort effects (i.e. discontinuities between the oldest pupils in the first cohort vs. the youngest in the second cohort). In order to test the robustness of the findings, the most extensive model is also fitted to a subset of the pupils. In that case only the pupils born 13 weeks before and 13 weeks after the cut-off date are included (i.e. birth dates ranging from 2 April to 30 September). The other analyses comprise two entire birth cohorts (i.e. pupils born within a year either side of the cut-off date).

The first model is a straightforward OLS regression model and represents the standard regression-discontinuity model (see Eq. (1)). Separate regression analyses are conducted for every outcome measure. The test-scores are modelled as a function of age and birth cohort. The cohort amounts to a binary variable. The zero value denotes the younger cohort and pupils in the older cohort get the score one. Age is recoded as a linear variable from -1 to 1. The oldest pupils in the younger cohort get a zero score. Younger pupils get a negative score and the older ones get a positive score. As a result, the intercept in the statistical output denotes the score of the oldest pupils in the younger cohort. The cohort effect denotes the discontinuity in the relationship between age and test scores at the cut-off date. In addition to the main effects of age and birth cohort a product-term of both variables is included. This denotes the interaction of age and birth cohort. It accounts for the possibility that the effect of age may differ from one cohort to the next. In contrast to the next equations, Eq. (1) only includes a single residual (e_i). This denotes to what extent each individual score (pupil i) deviates from the statistical model.

The second model (see Eq. (2)) is a multilevel model and takes the nesting of pupils within schools into account (Bryk & Raudenbush, 1992). Deviations from the model are decomposed into a school specific component (u_{oj}) and an individual component (e_{ij}). In this case, e_{ij} represents to what extent each pupil (i) deviates from the mean score in his/her school (j). In turn, the school means are modelled as deviations (u_{oj}) from the intercept (β_{00}). Residual variance is estimated at two distinct levels: the school level (intercept variance) and the individual level (residual variance). Models like these are referred to as random intercept models.

Eq. (3) presents a further extension. The cohort effects estimated when fitting this model are the ones referred to in Table 2. The model takes into account that cohort effects may vary between schools. This is denoted by an additional residual term (u_{1j}). Models that allow the effect of an explanatory variable to vary across groups are known as random slopes models. The cohort effect is decomposed into a general component (β_{10}) and a school-specific deviation (u_{1j}). Fitting this model entails the computation of three residual variances: the intercept variance, the cohort variance and the residual variance. In addition, the correlation between the school specific intercept and cohort effect is estimated. A positive correlation indicates positive cohort effects when the intercept is high (i.e. high test-scores at a school). A negative correlation implies positive cohort effects when the test-scores at a school are low. Variation of cohort effects across schools may be due to differences in effectiveness between schools. But it may also reflect differences in background

[1] Cohen's d is based on the assumption that the standard deviations in both groups are identical. In practice, researchers need to choose either the standard deviation in the control or experimental group. We opted for the "compromise" of the pooled standard deviation, mainly because the distinction between control group and experimental group is somewhat contrived in the present case.

variables between cohorts within schools (Perry, 2017). It seems quite unlikely that for an entire population, adjacent cohorts differ substantially with regard to achievement-related background characteristics like aptitude, learning motivation, family income, educational level of the parents or ethnicity, although at the school level such differences between cohorts may occur by chance every now and then.

Finally, the third model also takes into account the assessment date. Separate effects are estimated for the average assessment date per school and year-group (measured in weeks) and individual deviations from this average. A positive effect on test-scores is expected when a school administers the test relatively late. But if individual pupils took the test at a later date than their classmates, this may point to exceptional circumstances, like illness or delayed enrolment. The equations denoting the three models are specified below.

$$Y_i = \beta_0 + \beta_1\, coh_i + \beta_2\, age_i + \beta_3\, age_i \times coh_i + e_i \quad (1)$$

$$Y_{ij} = \beta_{00} + \beta_{10}\, coh_{ij} + \beta_{20}\, age_{ij} + \beta_{30}\, age_{ij} \times coh_{ij} + u_{0j} + e_{ij} \quad (2)$$

$$Y_{ij} = \beta_{00} + (\beta_{10} + u_{1j})\, coh_{ij} + \beta_{20}\, age_{ij} + \beta_{30}\, age_{ij} \times coh_{ij} + \beta_{40}\, ads_{ij} + \beta_{50}\, (adp_{ij} - ads_{ij}) + u_{0j} + e_{ij} \quad (3)$$

Where:

Y = pupil test score (four outcome measures in this study)

coh = binary variable denoting pupil birth cohort (zero denotes the younger cohort)

age = pupil age in years (zero score denotes the oldest pupils in the younger cohort)

adp = assessment date: week the pupil took the test (zero stands for week 35)

ads = assessment date: school average per year-group (zero stands for week 35)

β_0/ β_{00}=intercept (the predicted score if all explanatory variables equal zero)

$\beta_1 - \beta_3$=regression coefficients denoting the effects of the independent variables (OLS)

$\beta_{10} - \beta_{50}$=regression coefficients denoting the effects of the independent variables (multilevel)

u_{0j}=school specific deviation from the intercept

u_{1j}=school specific deviation from the cohort effect

e_i/e_{ij}=residual term (at pupil level)

i=subscript denoting that a score or residual term relates to individual pupils

j= subscript denoting that a score or residual term relates to schools

8. Findings

Table 1 shows the descriptive statistics (mean score, number of pupils, standard deviation) per cohort for each test. The cohort labels (4–6) point to the corresponding year-groups (e.g., nearly all pupils in cohort 4 are in year-group 4). On every test, the mean scores increase from cohort 4–6. The main question, however, is whether discontinuities can be detected at the cut-off points.

Fig. 1 presents a visual impression of the discontinuities. The graphs display the relationship between test scores and age at 2 July 2011. Each square in the graphs represents the mean score for the pupils born in a given week (on average 379 pupils). Most discontinuities are easy to discern by eye. But, for word decoding, this is harder than for general math. It appears that for word decoding, pupils in the older cohorts get higher scores mainly because they are older. It does not seem that schooling makes large contributions to increases in word decoding; by the ages covered in this analysis they have largely acquired the skills to be able to decode unfamiliar words although they get better at it as they grow older. For all four measures, the graphs clearly show a positive relationship between age and the test scores. Within every cohort the older pupils score higher.

For more precise (i.e. numerical) estimates of the effects of schooling, we turn to the statistical analyses. The main outcomes of the analyses with regard to the discontinuities between birth cohorts are reported in Table 2. This table shows the effects of being in an older vs. a younger cohort (i.e. the effects of schooling) at the point of discontinuity. The contribution of schooling across the cohorts is reported as well (for more details see the supplementary materials, Section 3). We find substantial effects of schooling, but less than half of the differences between average scores per cohort can be attributed to schooling.

Table 2 shows the differences in mean test scores between adjacent cohorts and the total difference across three cohorts. The differences are expressed as the INCAS metric, Cohen's d and the Improvement Index. The discontinuities are expressed accordingly. In addition, they are expressed as percentages of the total difference between cohorts. The discontinuities in the bottom row denote the sums of the discontinuities between cohorts 4–5 and cohort 5–6. In terms of effect sizes, the total gains observed in Northern Ireland are larger than the gains reported by Bloom et al. (2008). This goes both for mathematics and reading.

Age accounts for the larger part of the differences between cohorts. None of the discontinuities exceeds 50 %, when expressed as a percentage of the total difference between cohorts. Still, the discontinuities between the oldest pupils in a cohort and the youngest in the next cohort provide distinct evidence for the contribution that schooling makes to the learning gains between the age of 7 and 10 in Northern Ireland. The sizes of the contributions vary considerably, both across outcome measures and between different stages of the school career. In terms of effect sizes, the largest contribution relates to general math (0.74 in total or 27 percentile gain). The contribution for word decoding is the smallest (0.33 or 13 percentile gain). Word recognition (0.55 or 21 percentile gain) and reading comprehension (0.45 or 17 percentile gain) take up the middle ground. For all three reading measures the total growth is smaller when cohorts 6 and 5 and compared. The contribution of schooling to this declining growth falls away as well. For general math, both the total growth and the contribution of schooling is more constant. The discontinuities between cohorts 4 and 5 range from 0.20 to 0.42 (or 8–16 percentile gain) and from 0.13 to 0.32 (or 5–12 percentile gain) between cohorts 5 and 6.

Table 1
Descriptive statistics per cohort.

Cohort	Range of birthdates	Descriptive statistics	General Math	Word Recognition	Reading Comprehension	Word Decoding
4	1 July 2004 2 July 2003	Mean	7.828	7.580	7.563	7.751
		N	19,445	19,738	18,607	19,753
		SD	0.896	1.778	1.761	1.869
5	1 July 2003 2 July 2002	Mean	8.799	8.896	8.861	8.925
		N	19,216	19,526	18,651	19,519
		SD	1.024	1.692	1.773	1.925
6	1 July 2002 2 July 2001	Mean	9.696	9.881	9.864	9.936
		N	19,264	19,623	18,811	19,620
		SD	1.183	1.624	1.727	2.030
		Total N	57,925	58,887	56,069	58,892

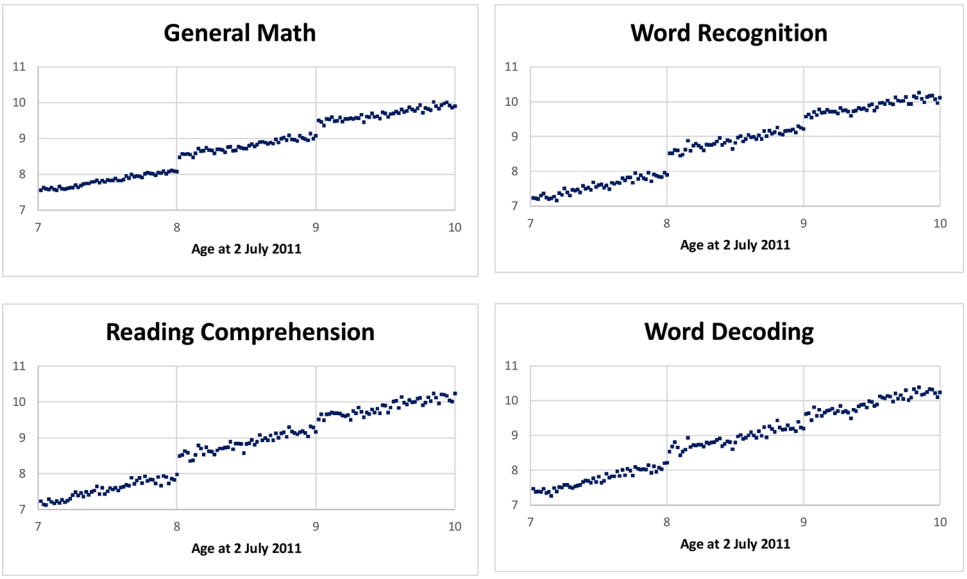

Fig. 1. Test score by age; averages by week of birth.

Table 2
Differences and discontinuities between adjacent cohorts.

COHORTS	MEASURE	Differences between cohorts			Discontinuities between cohorts			
		INCAS metric	Cohen's d	Improvement index (percentile gain)	INCAS metric	Cohen's d	Improvement index (percentile gain)	Discontinuity as percentage of total difference
5 vs. 4	General math	0.971	1.01	34	0.402	0.42	16	41%
	Word recognition	1.316	0.76	28	0.586	0.33	13	45%
	Reading comprehension	1.298	0.73	27	0.514	0.29	11	40%
	Word decoding	1.174	0.62	23	0.386	0.20	8	33%
6 vs. 5	General math	0.897	0.81	29	0.354	0.32	13	39%
	Word recognition	0.985	0.59	22	0.354	0.21	8	36%
	Reading comprehension	1.003	0.57	22	0.285	0.16	6	28%
	Word decoding	1.011	0.51	19	0.249	0.13	5	25%
Total (6 vs. 4)	General math	1.868	1.82	47	0.756	0.74	27	40%
	Word recognition	2.301	1.35	41	0.940	0.55	21	41%
	Reading comprehension	2.301	1.31	40	0.799	0.45	17	35%
	Word decoding	2.185	1.13	37	0.635	0.33	13	29%

9. Conclusion and discussion

The current study is the first in which test scores that cover almost an entire population were analysed to assess the absolute effects of schooling. The findings relate to an education system in which deviations from assignment to year-groups by date of birth are extremely rare.

The findings indicate that the effects of schooling may vary considerably from year to year and also across outcomes. Still, some (tentatively) inferred general trends from prior research are confirmed by the present study. The effects of schooling get smaller as the school career progresses and are larger for math than reading. The curriculum is probably an important factor as well. Reading skills typically receive much attention in the first years of primary schooling.

So far, nearly all research that has made use of RD to assess the absolute effect of schooling has aimed to estimate the effect of *receiving* schooling. The approach applied in the present study (ITT), actually assesses the effect of *providing* schooling. However, in this case the overlap between the effects of receiving and providing schooling is nearly perfect. The strict assignment of pupils to year-groups by date of birth in Northern Ireland ensures that over 99 % of all pupils are in the expected year-group. In most other education systems, cut-off dates are applied more flexibly. This complicates the potential of RD to assess the effect of receiving schooling. Low and poor performing pupils with birth-dates close to the cut-off are most likely to end up in higher and lower year-groups than most of their same-age peers. Nearly all RD studies have so far aimed to assess the effect of receiving one year of schooling. Typically researchers exclude delayed and accelerated pupils from their analyses. But, the risk of biased estimates increases as the prevalence of delayed and accelerated school career grows. In such situations, the most viable alternative may be to apply an ITT approach and focus on cohort effects. Thus the effects of providing schooling can be assessed.

Cut-off dates that determine assignment to school-years apply in nearly every education system but the flexibility with which they are applied varies. It can be argued that the effect of providing schooling is more relevant from a policy perspective than the effect of receiving schooling. Quite often it is beyond the policy makers' reach to ensure that each and every individual in the target population actually receives the services made available. A focus on the effects of providing schooling, as illustrated in the present paper, may not only be less complicated to realize as a research approach. It also appears more relevant from a policy perspective.

Knowledge of absolute effects of schooling is important as a frame of

reference. If the effect of an educational intervention equalled the contribution that one year schooling makes, this may be viewed as a great success. The present study shows that these contributions (i.e., discontinuities between cohorts) may be not be particularly large according to the guidelines that were suggested by Cohen in 1969.

Although Cohen's guidelines are still widely used to interpret effect sizes, they do not seem appropriate for interpreting the effects of educational interventions. They are based on a limited number of psychological experiments, conducted in tightly controlled laboratory conditions over fifty years ago. Rigorous educational research mostly shows effects that would qualify as "small" (Cheung & Slavin, 2016; Lortie-Forgues & Inglish, 2019). An effect size of .50 would qualify as medium, but in educational research over 90 % of randomized trials with standardized outcome measures show effects that are (much) smaller (Kraft, 2019;). Kraft (2019) convincingly argues that effect sizes equal to .20 may be considered large in educational research. Only thirty percent of randomized trials with standardized outcomes measures show effects of this size or larger. Following the guidelines suggested by Kraft, two of the cohort effects reported in this study are medium in size. The remaining six can be considered large.

The present study confirms that the guidelines suggested by Cohen are hardly appropriate for interpreting the effects of educational interventions. In the age ranges addressed, the effects of a whole year schooling on math and reading would be (very) small to (nearly) medium according to Cohen's guidelines. It seems more appropriate to adopt the guidelines suggested by Kraft; if the effect of an intervention is close to .20, this may be similar to the effect of one year of schooling. Such a result should be considered to be of educational importance.

[1]Cohen's d is based on the assumption that the standard deviations in both groups are identical. In practice, researchers need to choose either the standard deviation in the control or experimental group. We opted for the "compromise" of the pooled standard deviation, mainly because the distinction between control group and experimental group is somewhat contrived in the present case.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.stueduc.2020.100939.

References

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher, 48*(4), 217–228, 2019.

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*(4), 289–328.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. SAGE.

Cahan, S., & Cohen, N. (1989). Age versus schooling eff ;ects on intelligence development. *Child Development, 60*(5), 1239–1249.

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), 1–12.

CCEA. (2007). *Key stages 1 & 2, Age 6-11*. Retrieved from Council for the Curriculum Examinations and Assessment http://ccea.org.uk/curriculum/key_stage_1_2/overview.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27*(5), 703–722.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292.

Cliffordson, C. (2010). Methodological issues in investigations of the relative effects of schooling and age on school performance: The between-grades regression discontinuity design applied to Swedish TIMSS 1995 data. *Educational Research and Evaluation, 16*(1), 39–52.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). New York: Academic Press.

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

Gallagher, T., & Smith, A. (2000). *The effects of the selective system of secondary education in Northern Ireland*. Retrieved from https://www.education-ni.gov.uk/sites/default/files/publications/de/gallagherandsmith-mainreport.pdf (Department of Education).

Heyns, B. (1978). *Summer learning and the effects of schooling*. New York: Academic Press.

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal, 319*(7211), 670–674.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615–635.

Kraft, M. (2019). *Interpreting effect sizes of education interventions. (EdWorkingPaper: 19-10)*. Retrieved from Annenberg Institute at Brown University http://www.edworkingspapers.com/ai19-10.

Lortie-Forgues, H., & Inglish, M. (2019). Rigorous large-scale RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48*(3), 158–166.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: Regression-discontinuity applied to TIMSS-95. *Oxford Review of Education, 32*(3), 397–429.

Luyten, H., Merrell, C., & Tymms, P. (2017). The contribution of schooling to learning gains of pupils in Years 1 to 6. *School Effectiveness and School Improvement, 28*(3), 374–405.

Merrell, C., & Tymms, P. (2007). Identifying reading problems with computer-adaptive assessments. *Journal of Computer Assisted Learning, 23*(1), 27–35.

Perry, T. (2017). Inter-method reliability of school effectiveness measures: A comparison of value-added and regression discontinuity estimates. *School Effectiveness and School Improvement, 28*, 22–38 (2017).

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *What works clearinghouse*. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rd.pdf.

Trochim, W. M. K. (1984). *Research design for program evaluation, the regression-discontinuity approach*. Beverly Hills, CA: SAGE Publications.

U.S. Department of Education. (2014). *What Works Clearinghouse: Procedures and standards handbook (Version 3.0)*. Institute of Education Sciences.

Wang, T., Ren, X., Schweizer, K., & Xu, F. (2016). Schooling effects on intelligence development: Evidence based on national samples from urban and rural China. *Educational Psychology, 36*(5), 831–844.

Webbink, D., & Gerritsen, S. (2013). *How much do children learn in school? International evidence from school entry rules*. CPB Discussion paper 255. The Hague: CPB Netherlands Bureau for Economic Policy Analysis.