International Journal of Research & Method in Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cwse20

A critical evaluation of regression discontinuity studies in school effectiveness research

Adrian Simpson

To cite this article: Adrian Simpson (09 Oct 2024): A critical evaluation of regression discontinuity studies in school effectiveness research, International Journal of Research & Method in Education, DOI: 10.1080/1743727X.2024.2412730

To link to this article: https://doi.org/10.1080/1743727X.2024.2412730

INTERNATIONAL JOURNAL OF RESEARCH & METHOD IN EDUCATION
https://doi.org/10.1080/1743727X.2024.2412730

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

A critical evaluation of regression discontinuity studies in school effectiveness research

Adrian Simpson

School of Education, Durham University, Durham, UK

ABSTRACT

School start regulations allocate children born immediately either side of a given date to different life paths: those slightly older starting school a full year earlier. School effectiveness literature exploits this to estimate causal effects described as 'the absolute effect of schooling' or 'the effect of an additional year's schooling', using the logic of regression discontinuity (RD). This paper examines the causal arguments and assumptions underpinning RD, noting particularly the importance of the causal description. It highlights concerns with describing causes in terms of school effectiveness including failure to consider the alternative treatment pathway; presence of other post-allocation causal factors and potential discontinuities at allocation. The paper notes that these can be overcome by using wider causal descriptions but at the expense of no longer identifying school effectiveness.

ARTICLE HISTORY
Received 10 February 2023
Accepted 5 August 2024

KEYWORDS
Effect of schooling; regression-discontinuity; causality; methodologies

1. Introduction

Many personal milestones align with age: driving, voting, military service, receiving benefits, etc. One rite of passage does not: starting school. Compulsory schooling often involves children born across 12-month periods starting together. For example, those starting school in Northern Ireland in September 2007 included children born between 2nd July, 2002 and 1st July, 2003 (Luyten *et al.* 2020).

This gives an apparent opportunity to separate influences which are normally hard to disaggregate, in the hope of identifying effects of schooling (e.g. Cahan and Davis 1987, Alexander and Martin 2004, Cliffordson and Gustafsson 2008, Luyten *et al.* 2020). These studies responded, in part, to flaws in earlier research which, in finding only small proportions of test-score variance attributable to schools, maintained that schooling was not particularly effective. Such logic is flawed: it measures differential impact of schools, not 'the absolute effect of schooling as compared to no schooling' (Madaus *et al.* 1980, p.50). Low variance results as easily from schools playing similar, large causal roles as from them playing little role. So, school effectiveness research looked for alternative measures for 'absolute effects of schooling' on achievement (Cahan and Davis 1987, Luyten 2006, Ali and Heck 2012).

Achievement has many influences including schooling, biological maturation, socio-economic factors, parental behaviours etc. Deciding whether factors are causal requires excluding other influences. Disaggregating four effects – biological age, age-in-grade, length of schooling and age at

CONTACT Adrian Simpson adrian.simpson@durham.ac.uk School of Education, Durham University, Durham, DH1 3LE, UK

starting school – is difficult. Experimenters cannot assign participants to states, and studies cannot vary only one factor, since they are unavoidably interrelated (Crawford *et al.* 2014).

Nonetheless, school effectiveness studies aim to identify cause, and the cutoff induced by school start regulations teases the promise of alternatives to randomized controlled trials (RCTs) for rigorous identification. Studies vary in aim and approach. Some view 'effects of schooling' as of direct interest (e.g. Angrist and Krueger 1991), while others investigate how effects vary across factors, including the type of task (Cahan and Cohen 1989) or country (Marchionni and Vazquez 2019). Commonly, studies compare 'effects of schooling' with 'effects of age' (Artman and Cahan 1993, Crone and Whitehurst 1999, Luyten *et al.* 2020).

They exploit children of similar age being in different grades to invoke regression discontinuity (RD) arguments. The image in Figure 1 is typical – an outcome is measured against age. Children on either side of the cutoff (the dashed vertical line) were born close in time, but started school one year apart; those between cutoffs were born up to a year apart but started school together. The discontinuity at the cutoff purportedly represents the 'effect of an additional year of schooling' (Cahan and Davis 1987, p.6) and the difference between youngest- and oldest-in-grade represents the 'effect of chronological age' (op cit).

This paper critically examines RD in school effectiveness research. It is not intended as a systematic review, instead exploring how RD logic is exploited in these studies, the extent to which the logic is followed and the resulting reliability of this field.

The next section describes RD's causal logic, distinguishing three types of potential cause. Thereafter, the paper examines how the existing school effectiveness regression discontinuity (SERD) literature follows that logic. Briefly touching on concerns with matching at allocation, the main focus is on two key issues in SERD studies: partial causal descriptions and unevaluated post-allocation causes. While highlighting the unreliability of existing SERD results, the paper concludes with suggested reinterpretations of the studies' results and recommendations for future practice in RD use more widely.

2. Causal logic

SERD studies make causal claims: a difference in outcomes being caused by 'effects of schooling' (e.g. Cahan and Davis 1987, Alexander and Martin 2004, Cliffordson and Gustafsson 2008, Luyten *et al.*

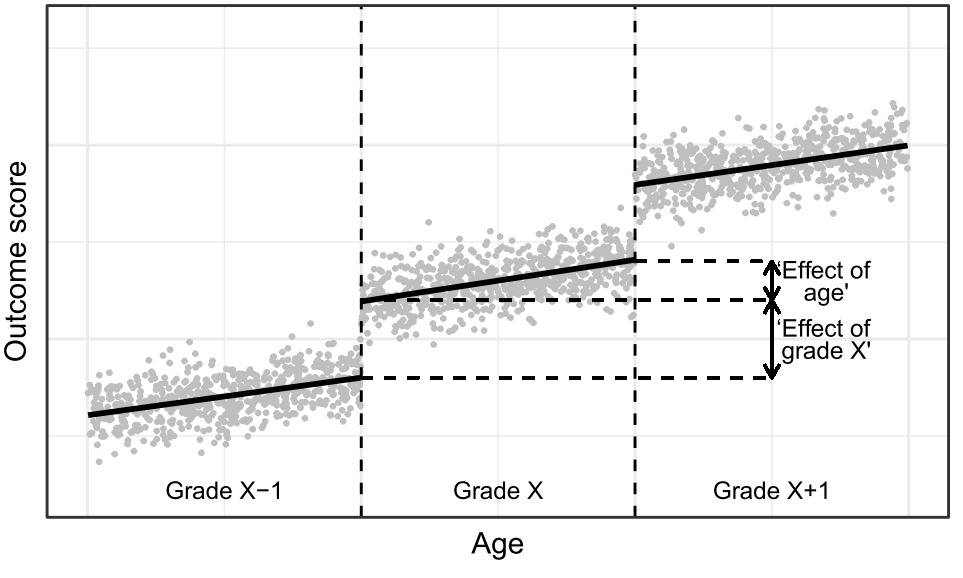

Figure 1. Simulated, typical figure illustrating effects from RD school effectiveness studies.

2008), 'absolute effects of schooling' (Cahan and Davis 1987, Luyten 2006, Ali and Heck 2012) or a 'grade level effect' (Luyten 2006).

Ascribing cause involves excluding or accounting for other causes. Between-group studies can be viewed as simple causal graphs as illustrated in Figure 2. Some mechanism allocates participants to groups, the study describes treatments for each group and outcomes are measured. The study can ascribe the cause of the difference in outcomes to the described difference in treatments only if it can discount causes arising from the allocation mechanism and causes arising post-allocation which might impact on outcomes (unless captured in the described difference in treatments or its consequences).

RCTs' involve random allocation mechanisms. These do not eliminate allocation/outcome relationships but give probability distributions for mean differences in outcomes if random allocation played the sole causal role. If the measured outcome difference is large for that distribution and there are no other post-allocation, outcome-related, between-group differences then, subject to well-defined error risks, the described difference in treatments is warranted to have played a causal role in the difference in outcomes.

Regression discontinuity studies rely on different allocation mechanisms. Ideally, groups are formed of participants infinitesimally close, on either side of some point (the 'cutoff') on a variable (the 'forcing variable'). Provided there are no other outcome-related discontinuities at this cutoff on allocation and no other outcome-related, between-group differences occur post-allocation, then the cause of any difference in outcomes can be ascribed to the described difference in treatments. This difference appears as a jump in the outcome/forcing variable graph (as in Figure 1). While a brief outline suffices for a critical evaluation of the SERD literature, other papers explore the logic and assumptions of RD in more depth (e.g. Imbens and Lemieux 2008, Bloom 2012). Nonetheless, an example may illustrate the key points.

Suppose a very large number of students take a diagnostic mathematics test. The score is the forcing variable: those scoring below 50 get assigned to a remedial curriculum, the remainder are assigned to the usual curriculum. The outcome is a subsequent post-test (Figure 3). The RD argument is that not only are mathematical abilities of the group scoring 49.999…and that scoring 50 infinitesimally close, so are all other factors: height, IQ, parental income etc. Otherwise, whatever

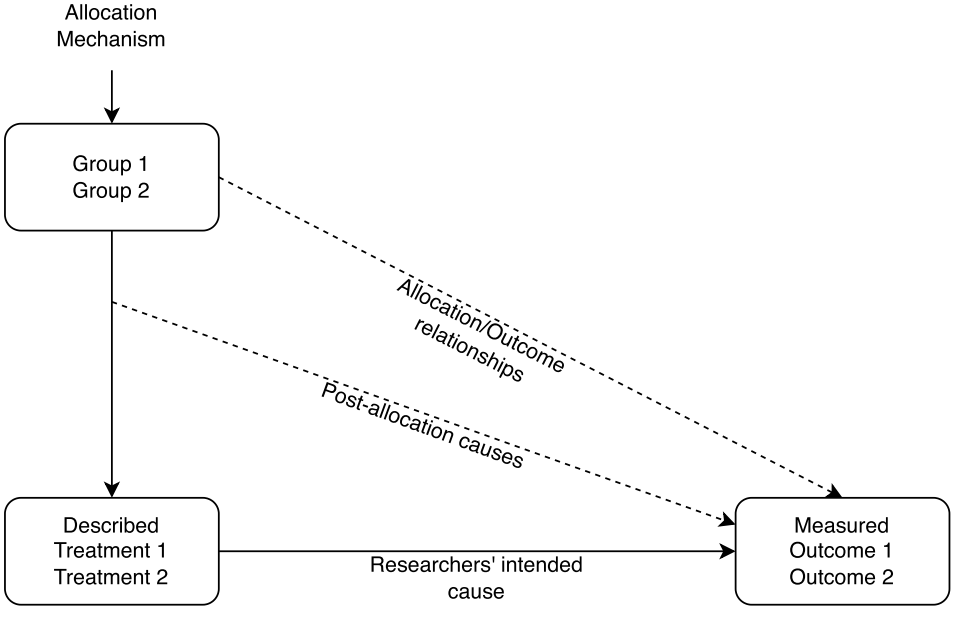

Figure 2. A simplified causal path diagram.

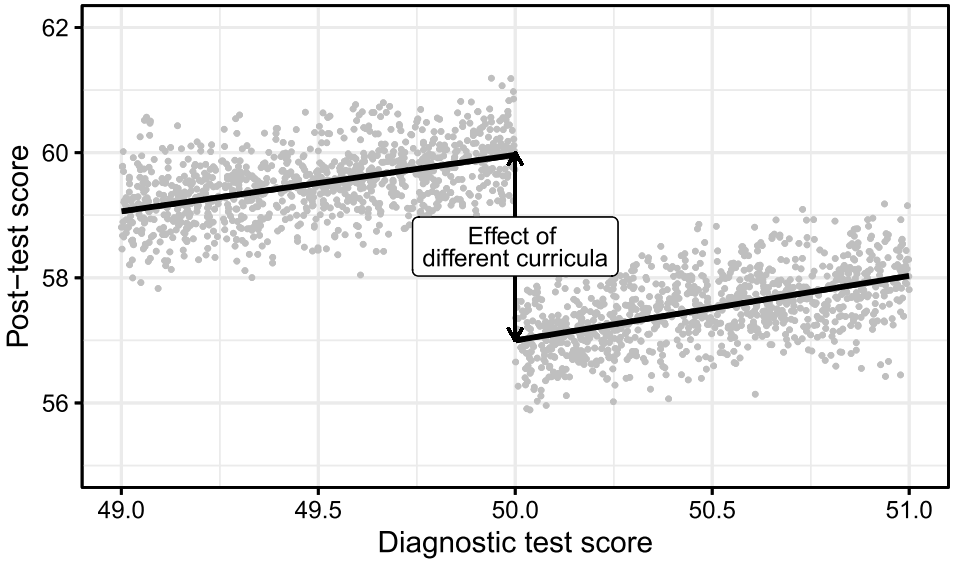

Figure 3. Illustration of an RD study. Participants are allocated to groups by a score on the 'forcing variable' (diagnostic test) and the post-test/diagnostic score relationship is examined for the discontinuity at the cutoff.

mechanism connects, say, IQ and diagnostic test scores would somehow jump at exactly 50. Even given an expected IQ/diagnostic score relationship, it is prima facie absurd it would jump at that point. It follows from the definition of continuity that any continuous variable, unrelated to the forcing variable, will be equal in the limit approaching the cutoff from the left and from the right.

Then, if the only post-allocation, outcome-related, between-group difference is the curriculum, the cause of any discontinuity in post-test scores occurring at 50 in the post-test/diagnostic-test score graph can be ascribed to the difference in curricula for participants at the cutoff.

For even this idealised argument to hold, each causal route needs careful consideration. If the diagnostic test does not, in fact, match the two groups – perhaps the scorer adjusts marks for students near 50 which results in an imbalance in some outcome-related factor – then a relationship between allocation and outcome at the point of assignment cannot be eliminated. On the other hand, if the study does not eliminate other potential outcome-related causes (such as the difference between teachers, non-compliance or absence at testing) ascribing cause to 'difference in curricula' will be unwarranted.

That is, causal description matters. Widening the description to 'effects of assignment to remedial or normal curricula as implemented in this school' encompasses more post-allocation causes. Differences in teachers, non-compliance or absence are consequences of *assignment* to the implemented curricula. This wider causal ascription is better warranted, but is less informative: the research no longer identifies the difference in curricula as causal.[1]

However, assuming the assignment does indeed match groups infinitesimally close to the cutoff and all other post-allocation causes have been eliminated except those captured by the described difference in treatments, RD logic ensures that the cause of the difference in outcomes can be ascribed to the described difference in treatments. Unlike RCTs, RDs identify only a local average treatment effect (LATE) – it applies only to those at the cutoff. In the simple mathematics curriculum example, one would not expect differences in curricula to have constant effects across the ability range: of itself, RD logic says little about effects for those with scores away from 50.

Naturally, real-world implementations of RD logic cannot rely on infinite samples with continuous measures. As findings apply only at the cutoff, researchers model the relationship between forcing and outcome variables in some window above and below the cutoff and calculate limits from above and

below to estimate any discontinuity at this point (Imbens and Lemieux 2008). This leads to another important assumption for RD studies: that the model accurately predicts the values of the outcome at the cutoff. This involves researchers balancing the amount of data in the window against model complexity. Too little data, in too narrow a window, results in noise swamping the functional relationship, giving spurious predictions. Wider windows can need more complex models to accurately reflect the relationship but increase the risk of noisy estimates (Gelman and Imbens 2019).

There are recommendations for balancing analytic choices (e.g. Ludwig and Miller 2007, Fan and Gijbels 2018), but researchers should report extensive robustness checks. First, using different models and windows to evaluate the impact on estimates; second, checking for spurious discontinuities at non-cutoff forcing variable values where no jump should occur.

3. SERD studies

One of the first school effectiveness studies to invoke RD arguments is Cahan and Davis (1987). Using a pre-existing large scale survey of Israeli elementary schools, it examines mean reading and mathematics scores for children born in different months across grades 1 and 2 examining both the discontinuity at the grade boundary as 'the effect of one year of schooling' (p.9) and the difference in mean scores between the oldest and youngest in class as 'the effect of the one year difference in chronological age' concluding 'about 2/3 of the difference between grade levels in mean achievement is due to additional schooling' (p.9) for both subjects.

A more recent example looks at data from around 90% of children in grades 4–6 in Northern Ireland taking a number of mathematics and literacy tests (Luyten *et al.* 2020). In addition to more complex models, the conclusions compare grade boundary jumps across years and outcomes ('The effects of schooling get smaller as the school career progresses and are larger for math than reading', p. 6), as well as comparing the size of grade-boundary discontinuities to the total difference between cohorts ('We find substantial effects of schooling, but less than half of the differences between average scores per cohort can be attributed to schooling', p.5).

These SERD studies promise much for policy makers. For example, in examining how 'added year effects' varies with a variety of school factors, Heck and Moriyama (2010) suggests 'improvement-focused school leadership directly affected subsequent school instructional practices and, in turn, instructional practices affected added-year outcomes' (p.377). Examining 'age effects' and 'schooling effects' in urban and rural settings in China leads Wang *et al.* (2016) to suggest that 'schooling contributes more to rural children's intelligence development than to urban children's' (p.840) proposing the need to 'guarantee quality education for all children, especially those from disadvantaged environments' (p. 841).

Critical to such policy suggestions is the validity of the causal ascription – that the discontinuity is indeed an effect of an additional year of schooling and its size is a measure of the size of that effect. This is warranted only if other causal pathways have been eliminated.

4. Half-described causes

Much of this paper focusses on causes related to the allocation mechanism and those post-allocation causes which are not captured by the described difference in treatments. Before examining those, it is important to note that many SERD studies are notable for describing only half of the cause – that is, only one of the two treatments.

It would be incorrect to describe any jump in the idealised mathematics diagnostic test example as 'the effect of the remedial curriculum', even if all other causes had been eliminated. Instead, the cause should be described as 'the difference between remedial and normal curricula'. The groups scoring 49.999…and 50 do not differ only in the former receiving the remedial curriculum: the latter receive the normal curriculum. Both activities need to be included in the causal description unless one treatment clearly plays no causal role on the outcome.

For SERD, consider two timelines for children born just before and after midnight on the cutoff date (e.g. between 1st and 2nd July 2003 for one discontinuity considered in Luyten *et al.* 2020). From birth until the older attend school (September 2007) it may be plausible to maintain that outcome-related activities are identical in expectation. Then the older children go to school, passing through multiple grades until the outcome test (Autumn 2011). The younger children may not go to school at the same point, but they still undertake outcome-related activities: they have a year's additional pre-school. Taking students in Luyten *et al.* (2020) as an example, at the time of the test, the older have passed through four years and two months of pre-school activity and grades 1, 2, 3 and 4. The younger have passed through five years and two months of pre-school activity and grades 1,2 and 3 (see Figure 4).

Describing the cause solely in terms of schooling will not suffice. The difference is not only that one group has had an additional year in school: the other has had an additional year's pre-school. Nonetheless, SERD literature routinely captures only one half of the treatment: treatments are described variously as the 'absolute effect of schooling' (Cahan and Cohen 1989, p.3), 'effect of an extra year of schooling' (Marchionni and Vazquez 2019, p.15), ' "schooling" vs "no schooling" ' (Luyten 2006, p.397), ' "schooling up to grade x" versus "schooling up to grade $x+1$" ' (Cahan and Davis 1987, p.10), 'the effect of one grade' (Alexander and Martin 2004, p.409), 'the absolute effect of one year of schooling' (Heck and Moriyama 2010, p.387), 'extra schooling' (Angrist and Krueger 1991, p.1010) or 'the effect of second-grade schooling' (Crone and Whitehurst 1999, p. 611). The younger group's activity is omitted from the description.

The error of taking the discontinuity as an absolute effect is further problematic for studies comparing school effectiveness in different contexts or for different outcomes. For example, Cahan and Cohen (1989) compares discontinuities across measures of subdomains of intelligence, noting, for example 'the clear distinction between the verbal and nonverbal tests in terms of the magnitude of the effect of schooling' (p.1245). This conclusion is unwarranted until the impact of additional pre-school activity on these areas is addressed. More accurately, the study might have concluded that the effect of the difference between an additional year's schooling and an additional year's pre-school activity is smaller for nonverbal subdomains. Heck and Moriyama (2010) associates differences in discontinuities between schools with features such as types of leadership, resulting in conclusions justified only if the effectiveness of additional pre-school activities does not vary between schools. Marchionni and Vazquez (2019) and Luyten (2006) compare discontinuities across countries, maintaining the effect of an extra year's schooling differs between countries; this is justified only if pre-school activities do not differ between countries. It is unlikely such assumptions hold.

In addition to half-causal descriptions, Figure 4 highlights a further issue: the younger group passes through each grade of schooling, but does so one year later. The argument that any jump in outcome score at cutoff can be ascribed to the difference between grade 4 and an additional year of pre-school holds only if the effect of each grade on the outcome does not change over time. For example, undertaking grade 1 in the 2007/8 school year would need to have an identical impact as undertaking grade 1 in the 2008/9 school year. This bold assumption is rarely explored, let alone justified, in the SERD literature.

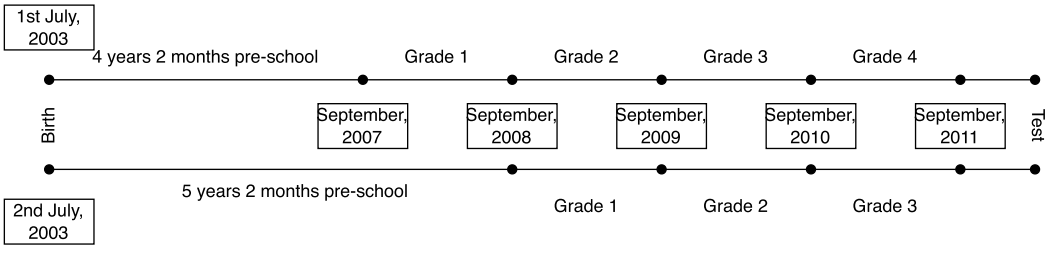

Figure 4. A timeline for a SERD study (based on Luyten *et al.* 2020) showing the educational pathways of two groups born either side of the cutoff.

Both concerns touch on the importance of causal descriptions and the need to ensure that descriptions capture every post-allocation, between-group difference either directly or as a causal consequence.

Luyten *et al.* (2020) at one point describe their findings as 'the effects of being in an older vs. a younger cohort (i.e. the effects of schooling) at the point of discontinuity' (p.5). However, 'the effects of being in the older vs. younger cohort' is a very different causal description to 'the effects of schooling'. The differences resulting from one group having an additional year's schooling, the other group an additional year's pre-school and the latter group receiving schooling at each grade one year later are all causal consequences of 'being in an older vs. a younger cohort' but are *not* causal consequences of 'the effects of schooling'. Comparing two effects to argue, say, that being in an older vs. a younger cohort has a larger impact in one country than another may be warranted; but the argument that the effect of schooling is larger is not.

As noted earlier, widening causal descriptions comes at the cost of being less informative: we can no longer identify discontinuities with effects of schooling, nor draw comparative conclusions about school effectiveness in different subjects or places.

5. Allocation/outcome relationships

Notwithstanding these initial concerns about half-causal descriptions, SERD studies' validity also relies on examining other causal pathways: whether allocation/outcome relationships are eliminated by the RD design and whether post-allocation causes are accounted for in the described difference in treatments.

RD requires groups matched on all factors at the cutoff prior to treatment. This allows the allocation process to be excluded as a cause of subsequent differences in outcomes. In the mathematics diagnostic test example, while we might expect a relationship between IQ and the forcing variable, we would not expect it to jump at exactly 50; so groups infinitesimally above and below 50 will be matched on IQ. However, a scorer systematically changing marks near the cutoff might imbalance groups, so the scorer's actions are a potential cause of difference in outcomes at the time of allocation. So consideration needs to be given to discontinuous processes near the cutoff at allocation.

In SERD studies, the forcing variable is birthdate[2] with the school start cutoff assigning those born either side to two different life pathways (which apparently only diverge some years later when the older start school). The idea is that the two groups of new-borns with adjacent birthdates should not differ on any factor. Many SERD papers exclude the possibility of relationships between factors and birthdate by assuming them away; explicitly claiming birth is random (e.g. Cahan *et al.* 2008, Jabr and Cahan 2015, Dicks and Lancee 2018). This is certainly false: there is a long-established literature on birth seasonality. For example, Buckles and Hungerman (2013) show clear seasonal changes in the educational attainment of mothers against date of giving birth. However, for the idealised RD, the issue is not birth being random across the analytic window, but whether factors impacting birth and outcome are discontinuous at cutoff (De la Cuesta and Imai 2016).

At first glance, one might expect that relationships between birthdate and other factors would be smooth but this may not be correct.

Gelman *et al.* (2013) analyses daily births in the US between 1969 and 1988. As well as smooth shallow trends in the number of births, peaking in September, declining in May, there are many large discontinuities. Births are much less common at weekends and on holidays, with a particular drop at Christmas where the rate is 20% below average even after accounting for other trends. Such phenomena have been seen in other countries and in more recent data (e.g. Martin *et al.* 2018). For example, Figure 5 shows the number of births in the US by date in 2015.

These trends have been increasing over time. Martin *et al.* (2018) argues they result from large reductions in induced births and elective caesareans away from standard weekdays and working hours. As elective caesareans relate to socio-economic factors (Fairley *et al.* 2011) two groups of children born on adjacent days may differ on some relevant factors. Thus, using school start cutoff may not result in matching at birth. For example, natality data from the US National Vital Statistics System

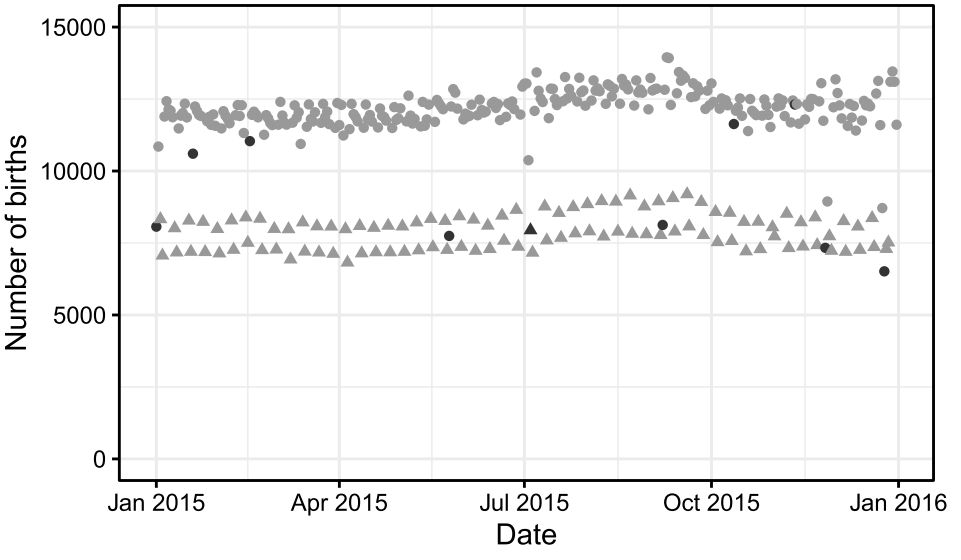

Figure 5. Number of births for each day in the US in 2015. Weekdays are shown as circles, weekends as triangles; US federal holidays in bold. Data from the US National Vital Statistics System's natality birth data, provided in Pruim *et al.* (n.d).

(National Center for Health Statistics 2015, the basis of Figure 5) show a substantial relationship between maternal educational attainment and birth weekday: there is a 6.7% (95% CI [6.1, 7.2]) relative risk reduction for giving birth on a weekend if the mother's educational level is high school graduation or above.

Real-world studies do not directly compare groups born either side of the cutoff directly: they model the forcing/outcome variable relationship from above and below the cutoff to see if left and right limits match. Thus, provided the effects of these birth factors are small and particularly if repeated large disturbances are far from the cutoff or well balanced on either side, they will appear as noise and may introduce only small biases. However, there are large jumps around holidays as well as weekends, and many countries with multiple major holidays near year-end also use calendar years to determine school start dates. Studies here may be undermined by these birth discontinuities (e.g. Alexander and Martin 2004, Heck and Moriyama 2010, Ali and Heck 2012).

6. Post-allocation causes

Notwithstanding concerns about birth discontinuities, the main source for biases in SERD literature relates to post-allocation factors which go uncaptured in the causal description. While the SERD literature tends to describe only one half of the cause ('the effect of an additional year of schooling') and implicitly assume that the effect of a grade is independent of the calendar year in which it is undertaken, this section puts these issues to one side. That is, this section presumes a cause described more completely (e.g. 'the difference between an additional year of schooling and an additional year of pre-school') and accepts the bold assumption about the consistency of the impact of grade across time.

Even given these assumptions, RD analysis still needs to be alert to factors which are not captured by the intended causal description – those which are potentially different on either side of the cutoff, but which are not consequences of the difference between an additional year of schooling and an additional year of pre-school.

There are three which are particularly problematic: missing data; non-compliance and differential behaviours.

6.1. Missing data

In the mathematics RD example, the study sample and the measured sample are easy to identify and there are well-known processes for dealing with data which go missing between group allocation and testing (Little and Rubin 2019).

In SERD studies, the issue is more problematic (Lipsey *et al.* 2015). Group allocation happens at birth: years pass before the groups born either side of cutoff diverge to different schooling pathways and more years pass before testing. In an isolated country with no immigration/emigration, the groups born before or after the cutoff constitute the study sample and (assuming no birth discontinuities) are matched in expectation at that point. Provided all progress along their assigned life pathway and take the test, there are no missing data issues to consider.

In the real world, however, children move in and out of school catchment areas, attend different types of schooling and miss tests for a wide variety of reasons. If these reasons relate to forcing and outcome variables, they may bias SERD estimates.

Considering missing data requires historical conditional thinking: subject to concerns in Section 5, RD designs may ensure groups are matched at birth, but SERD studies require the groups at measurement *would have been* matched at birth – that is, the measured sample is somehow representative of the ill-defined study sample created years earlier. Only if data at the cutoff have gone missing 'completely at random' (in the sense of Little and Rubin 2019) from a study sample defined at birth can impacts from missing data (or additional data from those moving into an area) be ignored.

Within the SERD literature, there is a subset of studies using 'cutoff designs' (e.g. Bisanz *et al.* 1995) which rely on small, opportunistic samples. Such studies typically use a very small number of children (e.g. 20 for Morrison *et al.* 1995; 70 for Naito and Miura 2001) born in a window of a couple of months around the cutoff. These children are repeatedly tested across a period of schooling. The argument is that this design compares subjects who are close in age but whose schooling lags a year. However, these are a very small proportion of the children assigned at birth to the two groups and are unlikely to be representative of those matched at birth. There is little more reason to believe groups selected in this way will be matched on all factors than for any other opportunistic group allocation approach.

More commonly, studies use larger samples defined by sets of schools. For example, Kyriakides and Luyten (2009) uses six Cypriot secondary schools, while Crone and Whitehurst (1999) involves nine Head Start centres. Nevertheless, it remains a bold assumption that children born just before and after cutoff have an equal likelihood to attend these schools or centres independent of any outcome-related factors.

Many other studies use very large samples or make explicit claims for being representative. Angrist and Krueger (1991) uses microdata from the US census; Luyten *et al.* (2020) conducts testing with 90% of the entire set of mainstream primary schools in Northern Ireland and Cahan and Cohen (1989) uses nearly the entire set of Hebrew language primary schools in Jerusalem (data reused in Artman and Cahan 1993, Cahan and Artman 1997, Cahan and Elbaz 2000).

Even with near population-level data, mechanisms underpinning absence need scrutiny. Most SERD studies with large samples use mainstream schools. If school entry is ability-based, or educational systems disproportionately remove students with performance at either extreme (to specialist schools, private schools or homeschooling), there will be outcome-related, systematic differences between groups which are not consequences of the described difference in treatments. Indeed, any absence from testing related to achievement-in-grade (thus relative-age-in-grade) will result in a discontinuity at cutoff irrespective of the described difference in treatments.

Missing data are detectable if they cause sample size imbalance. For example, given ability/age-in-grade relationships, ability-based grade retention policies allow fewer young-in-grade students to progress to later grades. This can be tested by looking for regression discontinuities in sample size at cutoff (McCrary 2008). However, the use of McCrary's test is uncommon in SERD literature (Marchionni and Vazquez 2019 being a rare example) and only detects imbalanced mechanisms. Systems which under-sample both extremes of performance within a cohort in roughly equal

proportion would pass McCrary's test. For example, given an age effect, if higher or lower scoring students left mainstream schooling in roughly equal number, sample sizes either side of cutoff would remain similar but discontinuities in scores at the cutoff would result irrespective of the described difference in treatments.

Some SERD studies use inappropriate missing data tests. For example, Luyten *et al.* (2020) cites What Works Clearinghouse standards (WWC 2019) where attrition below 20% is acceptable provided differential attrition is below 5%. However, these are RCT standards: for RD designs, the concern is not differential attrition in the window, but differential attrition at cutoff – particularly attrition systematically different either side.[3] Figure 6 illustrates the effect of omitting the top and bottom 2.5% of each grade in a sample of students in which there is an age effect but no additional schooling/pre-school effect. That is, even seemingly small amounts of missing data can result in the characteristic SERD jumps at grade boundaries irrespective of any 'effect of schooling'.

6.2. Non-compliance

Non-compliance raises similar concerns as missing data, but also recalls the issue of causal descriptions.

In the mathematics example, if some students at the cutoff do not receive their assigned curriculum and, particularly, if non-compliance relates to forcing and outcome variables, then ascribing the cause of a discontinuity to the difference in curricula is unwarranted. For example, if students just below the cutoff are demotivated by being assigned to a remedial curriculum and do not attend (or take the regular class), their outcome scores are not the result of receiving the curriculum to which they were assigned.

Again, widening the causal description from 'receiving different curricula' to 'being assigned to different curricula' (a so-called 'intention to treat' description) can address this. Non-compliance resulting from, say, demotivation is then a consequence of wider causal description. Widening descriptions comes at the cost of no longer identifying the cause from amongst the curricular, motivational or other factors which follow from 'being assigned to different curricula'.

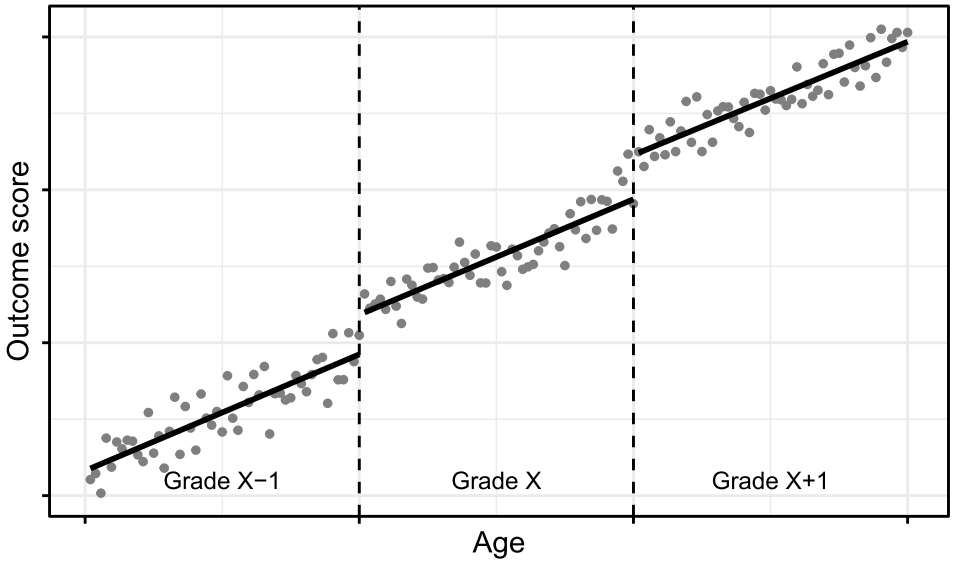

Figure 6. Simulation of 4000 students born across three years, grouped by week of birth, where there is an effect for age and some random noise, but no effect for schooling (over the effect of additional pre-school) where the top and bottom 2.5% of each grade is omitted.

Non-compliance in SERD studies appears most readily as 'misallocation' – students appearing in grades other than that assigned by the school start cutoff and their date of birth. This is usually from deliberate delay or acceleration. Many such mechanisms would be expected to be related to both the forcing variable and the outcome. For example, parents and school systems delay or accelerate using 'school readiness'. Given readiness/age-in-cohort relationships, disproportionate numbers of old-in-cohort high achievers get accelerated and disproportionate numbers of young-in-cohort low achievers are delayed. These would result in discontinuities irrespective of an effect of the difference between an additional year's schooling and an additional year's pre-school.

The proportion misallocated varies between contexts: 2% of Alexander and Martin's (2004) Australian sample; 10% of Cahan and Cohen's (1989) Israeli sample and 50% of Tiumeneva and Kuzmina's (2015) Russian sample are not in the grade determined by birthdate. However, even low levels of misallocation are concerning: Cahan and Cohen's misallocated 10% is not evenly spread across age-in-cohort. Around 40% of those born just before cutoff are in the grade below that expected, with the proportion tailing off further away.

When examined at all, the issue is sometimes dismissed as too small to substantially impact estimates and analysis proceeds with misallocated students omitted: Luyten (2006) restricts investigation to countries with under 5% misallocation. However, given expected strong misallocation/age-in-cohort relationships, small proportions of misallocation substantially bias results in much the same way as the missing data situation illustrated in Figure 6.

Some papers address the issue using the Cahan and Davis (1987) method of not just omitting misallocated children but excluding all those born one month before and after cutoff, where most misallocated lie (Cahan and Cohen 1989, Cahan and Noyman 2001, Alexander and Martin 2004). Linear regression based only on the other 10 months then predicts outcomes for those at the cutoff. The extent to which this approach addresses bias depends on the extent of misallocation outside the omitted months. In Cahan and Cohen's data, November and December are deleted, but around 10% of those born in January and even around 5% in August are misallocated, so substantial impacts of misallocation remain.[4]

Luyten *et al.* (2017, 2020) are unusual in taking an 'intention to treat' approach to address this problem. These studies retain misallocated students' data. Their description of the difference in treatments is not 'receiving schooling' but 'providing schooling'. Parents of children born before the cutoff are provided with the option of sending their children to school in September as an alternative to additional pre-school activity, though can choose to delay receiving it. In this case, misallocation is a consequence of the described treatment and thus does not undermine the validity of the causal conclusion.

Lipsey *et al.* (2015) note that 'intention to treat' does not help address the difficulties caused by an ill-defined underlying study sample and missing data. It can, however, address non-compliance, albeit with the cost that the wider description means the study does not estimate the effect of an additional year of schooling, but the effect of the *provision* of an additional year of schooling (against an additional pre-school year). Moreover, despite their emphasis on the intention to treat approach, Luyten *et al.* (2017, 2020) both nonetheless draw conclusions about the 'effects of schooling'.

6.3. Differential behaviours

The difference between those born either side of cutoff is not just schooling pathway but also their identification as youngest and oldest in their classes. Parents and teachers systematically treat the youngest and oldest in class differently. Indeed, if there is a relative-age/achievement relationship within a class, teachers treating higher and lower achievers differently also treat youngest and oldest in class differently.

Regimes allowing flexibility in school start date clearly demonstrate parents treat children very differently either side of the cutoff, resulting in very different levels of non-compliance in different countries (Eurydice 2011). Parents of children born just before cutoff are more likely to delay schooling and (to a lesser extent) parents of those born just after are more likely to accelerate schooling

(Graue and DiPerna 2000). The existence of willingness to alter a child's entire educational path is strong evidence that parents believe young-in-grade children are disadvantaged. So, even in regimes which do not permit delay or acceleration, parents are likely to treat youngest-in-grade differently from oldest-in-grade. Any additional focus parents place on getting children 'school ready' will impact differently on the youngest amongst those about to start school.

Further evidence of the way in which relative-age results in very different treatment can be seen in other studies of school start date discontinuities and relative-age effects. There is considerable literature demonstrating age-in-grade as a strong predictor of special needs placement (e.g. Dhuey and Lipscomb 2010). These relative-age effects result in discontinuities across grades. For example, Figure 7 shows the relationship between month of birth and having a diagnosis of ADHD across grades 4 and 5 (Schwandt and Wuppermann 2016). As this appears across different school systems, independent of when the school start date appears within the calendar year (e.g. Elder 2010, Morrow *et al.* 2012, Zoëga *et al.* 2012) and is not seen in children not yet in school, this is unlikely to be a birth seasonality effect. The accepted mechanism for such phenomena is that teachers and parents compare behaviour of young-in-class children to the grade norm, rather than the age norm, and disproportionately refer those children for diagnosis. Whatever the underlying mechanism, despite similar grade boundary discontinuities to SERD diagrams, it would be inappropriate to ascribe the discontinuity in Figure 7 (and a relative increase in the risk of ADHD diagnosis of around 30%) to 'the absolute effect of schooling' or even to the difference between a year's additional schooling against pre-school activity.[5]

If relative-age effects resulting from differential teacher and parental behaviour leads to effective additional educational support for those young-in-grade, that alone would create a discontinuity for educational outcomes.

Widening causal descriptions can help. While parents and teachers behaving differently according to relative age is not a causal consequence of 'the effect of schooling' (or even the more appropriate 'effect of a year's additional schooling or pre-school'), it is a consequence of 'starting school earlier or later'. Again, though, this comes at the cost that the RD analysis alone does not identify what it is about early or late school start and its consequences which impact on outcomes.

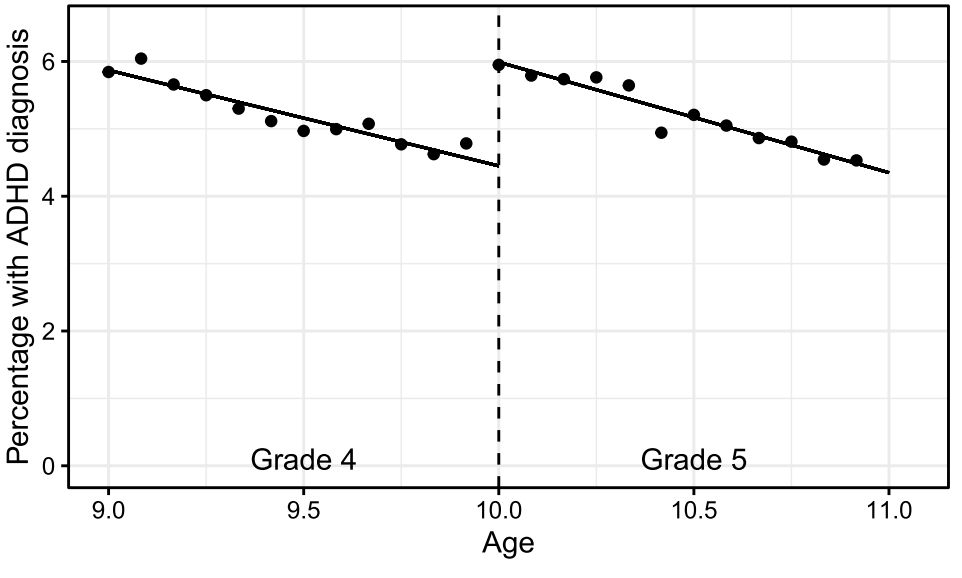

Figure 7. The relationship between age and ADHD diagnosis on June 2010 for those in grades 4 and 5 in German states with school cut off date 30th June (redrawn from Schwandt and Wuppermann 2016).

7. Conclusions

7.1. Summary

RD logic is impeccable, provided the assumptions hold. It is potentially as powerful as the RCT in establishing causal relationships. Nonetheless, as with RCTs, RD studies need to ensure that the described difference in treatments is the only post-allocation difference and any imbalances between groups resulting from the allocation mechanism are addressed.

SERD studies, by definition, attempt to evaluate the effectiveness of schooling. Studies claim they measure the 'absolute effect of schooling' (Cahan and Cohen 1989, p.3), 'effect of an extra year of schooling' (Marchionni and Vazquez 2019, p.15), ' "schooling" vs "no schooling" ' (Luyten 2006, p.397), ' "schooling up to grade *x*" versus "schooling up to grade $x+1$" ' (Cahan and Davis 1987, p.10), 'the effect of one grade' (Alexander and Martin 2004, p.409), etc. They purport to do so by measuring jumps in the age/achievement relationship at the school start date cutoff. However, these jumps can be larger or smaller for reasons unrelated to effectiveness of schools. Larger jumps occur when pre-schools are less effective; when there is less effective teaching for a given grade in the subsequent year; when more lower- or higher-achieving children attend non-mainstream schools or are otherwise missing from the measured sample; when more children are accelerated or delayed; when there is a larger bias in the way parents or teachers treat children on the basis of achievement or relative-age-in-class; or even, perhaps, when school start date occurs closer to a period with more major public holidays.

Positive discontinuities in age/achievement relationships can occur when we would expect schooling to have no (or even a negative) role. In Figure 7, it would be absurd to argue that an additional year of schooling results in a 30% relative risk increase in ADHD, rather than conclude the jump results from teachers' and parents' responding to deviation from age-in-grade norms rather than age norms.

7.2. Recommendations for future practice

RD designs are thus not appropriate for measuring school effectiveness. SERD papers which interpret their results as 'the effect of schooling' and compare these effects across cognitive areas (e.g. Cahan and Cohen 1989) or countries (e.g. Marchionni and Vazquez 2019) may particularly misdirect policy. For example, Luyten (2006) found a seemingly large proportion of schools without a positive discontinuity in some countries, concluding that 'for several countries, one extra year of schooling did not always yield a positive effect on achievement' (p.422). Such a claim may lead to very different policy responses compared with suggesting the effect of the provision of schooling early or late may not always be positive.

One alternative is to widen causal descriptions, changing the aim and interpretation of the research. For determining and measuring 'the effect of starting school early or late' at the cutoff, RD does hold promise. Being treated differently on the basis of relative age, missing an additional year of pre-school and experiencing different teaching are all consequences of starting school early for children born just before the cutoff and are captured by this wider causal description. Better still, 'provision of schooling early or late' also captures non-compliance (Luyten *et al.* 2020).

So, while the SERD literature as a whole should be considered unfit for its intended purpose, there are recommendations which might aid interpretation of the underlying studies and help future research. These include

- Phrasing effects as differences in treatments.
- Avoiding descriptions such as 'the effect of schooling', 'the absolute effect of schooling', 'the effect of grade X' etc.
- Considering intention to treat phrasing (the provision of difference in treatments) or modelling non-compliance with fuzzy RD designs.

- Using good robustness and sensitivity tests to evaluate findings (e.g. different analytic windows, a variety of regression models and tests for discontinuities at other points).
- Considering the effects of birth discontinuities (and particularly avoiding analyses in situations where school start cutoff is near the start of the calendar year until more is known about the relationships between elective birth and outcome-related factors such as parental education and socio-economic status).
- Avoiding small, opportunistic samples.
- Examining how the measured sample might relate back to the implicit study sample which was formed at the moment of birth.
- Evaluating the impact of missing data – McCrary's test is useful but insufficient for checking the nature of missingness as a confound.

This paper has shown how critical it can be to describe causes carefully; but also that wider, more accurate descriptions are less informative. RD as used in the SERD literature cannot answer the question of whether schools are effective nor how school effectiveness varies with other factors. With care, RD may be able to assess the effect of provision of starting school earlier or later for those born at the school start date cutoff. Nevertheless, this comes at the cost of no longer being able to support parents, practitioners and policy makers by identifying which of the many factors resulting from starting school earlier or later cause differences in outcomes.

Notes

1. This widening of causal description applies equally to RCTs. If participants are assigned at random to two different curricula taught by two different teachers, the RCT cannot identify the cause between the difference in curricula and the difference in teachers. The wider description – 'the effect of assignment to treatments as implemented in this school' – is warranted but less informative.
2. A strand of research purports to use regression discontinuity methods to measure the effect of age by comparing children in a given cohort whose birth*days* (rather than birth*dates*) are close (e.g. Bernardi 2014, Dicks and Lancee 2018, Marcenaro-Gutierrez and Lopez-Agudo 2021). That is, by comparing those born on, say, 2nd July 2003 and 1st July 2004, they aim to assess the effect of age. Such research assumes that two groups born almost exactly one year apart should be equivalent on all outcome-related factors except age. This argument is not valid.
3. Indeed, the data in the supplementary material for Luyten *et al.* (2020) suggest sample size discontinuities.
4. A rarely used alternative to deleting or ignoring data is modelling misallocation using fuzzy RD designs. The logic of the RD is that allocation to treatment at cutoff needs to be discontinuous: sharp designs rely on allocation probability moving from 0 to 1 at cutoff, fuzzy designs allow any discontinuous shift in probability. The relationships between forcing variables and allocation are modelled and estimates from that model are used for the main RD argument (see Marchionni and Vazquez 2019)
5. Of course, none of the authors of the ADHD studies make any such claim

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Adrian Simpson http://orcid.org/0000-0002-3796-5506

References

Alexander, J.R., and Martin, F., 2004. The end of the reading age: grade and age effects in early schooling. *Journal of school psychology*, 42 (5), 403–416.

Ali, E., and Heck, R.H., 2012. Comparing the contexts of middle-grade schools, their instructional practices, and their outcomes: a regression discontinuity approach. *NASSP bulletin*, 96 (2), 93–118.

Angrist, J.D., and Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings?*The quarterly journal of economics*, 106 (4), 979–1014.

Artman, L., and Cahan, S., 1993. Schooling and the development of transitive inference. *Developmental psychology*, 29 (4), 753–759.
Bernardi, F., 2014. Compensatory advantage as a mechanism of educational inequality: a regression discontinuity based on month of birth. *Sociology of education*, 87 (2), 74–88.
Bisanz, J., Morrison, F.J., and Dunn, M., 1995. Effects of age and schooling on the acquisition of elementary quantitative skills. *Developmental psychology*, 31 (2), 221–236.
Bloom, H.S., 2012. Modern regression discontinuity analysis. *Journal of research on educational effectiveness*, 5 (1), 43–82.
Buckles, K.S., and Hungerman, D.M., 2013. Season of birth and later outcomes: old questions, new answers. *Review of economics and statistics*, 95 (3), 711–724.
Cahan, S., et al., 2008. The differential effects of age and first grade schooling on the development of infralogical and logico-mathematical concrete operations. *Cognitive development*, 23 (2), 258–277.
Cahan, S., and Artman, L., 1997. Is everyday experience dysfunctional for the development of conditional reasoning? *Cognitive development*, 12 (2), 261–275.
Cahan, S., and Cohen, N., 1989. Age versus schooling effects on intelligence development. *Child development* 6 (5): 1239–1249.
Cahan, S., and 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.
Cahan, S., and Elbaz, J.G., 2000. The measurement of school effectiveness. *Studies in educational evaluation*, 26 (2), 127–42.
Cahan, S., and Noyman, A., 2001. The kaufman ability battery for children mental processing scale: a valid measure of "pure" intelligence? *Educational and psychological measurement*, 61 (5), 827–840.
Cliffordson, C., and Gustafsson, J.E., 2008. Effects of age and schooling on intellectual performance: estimates obtained from analysis of continuous variation in age and length of schooling. *Intelligence*, 36 (2), 143–152.
Crawford, C., Dearden, L., and Greaves, E., 2014. The drivers of month-of-birth differences in children's cognitive and non-cognitive skills. *Journal of the royal statistical society. series A (statistics in society)*, 177 (4), 829–860.
Crone, D.A., and Whitehurst, G.J., 1999. Age and schooling effects on emergent literacy and early reading skills. *Journal of educational psychology*, 91 (4), 604–614.
De la Cuesta, B., and Imai, K., 2016. Misunderstandings about the regression discontinuity design in the study of close elections. *Annual review of political science*, 19, 375–396.
Dhuey, E., and Lipscomb, S., 2010. Disabled or young? Relative age and special education diagnoses in schools. *Economics of education review*, 29 (5), 857–872.
Dicks, A., and Lancee, B., 2018. Double disadvantage in school? Children of immigrants and the relative age effect: a regression discontinuity design based on the month of birth. *European sociological review*, 34 (3), 319–333.
Elder, T.E., 2010. The importance of relative standards in adhd diagnoses: evidence based on exact birth dates. *Journal of health economics*, 29 (5), 641–656.
Eurydice, 2011. Grade retention during compulsory education in Europe: regulations and statistics., Tech. rep., European Commission, Education, Audiovisual and Culture Executive Agency.
Fairley, L., Dundas, R., and Leyland, A.H., 2011. The influence of both individual and area based socioeconomic status on temporal trends in caesarean sections in Scotland 1980-2000. *BMC public health*, 11, 1–10.
Fan, J., and Gijbels, I., 2018. *Local polynomial modelling and its applications*. New York: Routledge.
Gelman, A., et al., 2013. *Bayesian data analysis*. 3rd ed. New York: Chapman and Hall/CRC.
Gelman, A., and Imbens, G., 2019. Why high-order polynomials should not be used in regression discontinuity designs. *Journal of business & economic statistics*, 37 (3), 447–456.
Graue, M.E., and DiPerna, J., 2000. Redshirting and early retention: who gets the "gift of time" and what are its outcomes? *American educational research journal*, 37 (2), 509–534.
Heck, R.H., and Moriyama, K., 2010. Examining relationships among elementary schools' contexts, leadership, instructional practices, and added-year outcomes: a regression discontinuity approach. *School effectiveness and school improvement*, 21 (4), 377–408.
Imbens, G.W., and Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. *Journal of econometrics*, 142 (2), 615–635.
Jabr, D., and 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 (3), 441–466.
Kyriakides, L., and 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 (2), 167–186.
Lipsey, M.W., et al., 2015. The prekindergarten age-cutoff regression-discontinuity design: methodological issues and implications for application. *Educational evaluation and policy analysis*, 37 (3), 296–313.
Little, R.J., and Rubin, D.B., 2019. *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.
Ludwig, J., and Miller, D.L., 2007. Does head start improve children's life chances? Evidence from a regression discontinuity design. *The quarterly journal of economics*, 122 (1), 159–208.

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., and 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.
Luyten, H., Merrell, C., and Tymms, P., 2020. Absolute effects of schooling as a reference for the interpretation of educational intervention effects. *Studies in educational evaluation*, 67, 100939.
Luyten, H., Peschar, J., and Coe, R., 2008. Effects of schooling on reading performance, reading engagement, and reading activities of 15-year-olds in England. *American educational research journal*, 45 (2), 319–342.
Madaus, G.F., Airasian, P.W., and Thomas, K., 1980. *School effectiveness*. New York: McGraw-Hill.
Marcenaro-Gutierrez, O.D., and Lopez-Agudo, L.A., 2021. Too late or too soon for school? The impact of school entry age. *Journal of research on educational effectiveness*, 14 (2), 309–352.
Marchionni, M., and Vazquez, E., 2019. The causal effect of an extra year of schooling on skills and knowledge in Latin America: evidence from PISA. *Assessment in education: principles, policy & practice*, 26 (4), 489–515.
Martin, P., et al., 2018. Timing of singleton births by onset of labour and mode of birth in nhs maternity units in England, 2005–2014: a study of linked birth registration, birth notification, and hospital episode data. *PloS one*, 13 (6), e0198183.
McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: a density test. *Journal of econometrics*, 142 (2), 698–714.
Morrison, F.J., Smith, L., and Dow-Ehrensberger, M., 1995. Education and cognitive development: a natural experiment. *Developmental psychology*, 31 (5), 789–799.
Morrow, R.L., et al., 2012. Influence of relative age on diagnosis and treatment of attention-deficit/hyperactivity disorder in children. *CMAJ*, 184 (7), 755–762.
Naito, M., and Miura, H., 2001. Japanese children's numerical competencies: age-and schooling-related influences on the development of number concepts and addition skills. *Developmental psychology*, 37 (2), 217–230.
National Center for Health Statistics, 2015. Natality data set. Available from: https://www.nber.org/research/data/vital-statistics-natality-birth-data [Accessed 7 August 2024].
Pruim, R., Kaplan, D., and Horton, N., n.d. mosaicdata: project mosaic (mosaic-web. org) data sets. Available from: https://github.com/ProjectMOSAIC/mosaicData R package version 0.14.0 [Accessed 7 August 2024].
Schwandt, H., and Wuppermann, A., 2016. The youngest get the pill: ADHD misdiagnosis in Germany, its regional correlates and international comparison. *Labour economics*, 43, 72–86.
Tiumeneva, Y.A., and Kuzmina, J.V., 2015. The difference that one year of schooling makes for Russian schoolchildren. *Russian education & society*, 57 (4), 214–253.
Wang, T., et al., 2016. Schooling effects on intelligence development: evidence based on national samples from urban and rural China. *Educational psychology*, 36 (5), 831–844.
WWC, 2019. What Works Clearinghouse standards brief for attrition. Available from: https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_brief_attrition_080715.pdf [Accessed 7 August 2024].
Zoëga, H., Valdimarsdóttir, U.A., and Hernández-Díaz, S., 2012. Age, academic performance, and stimulant prescribing for ADHD: a nationwide cohort study. *Pediatrics*, 130 (6), 1012–1018.