UiO University of Oslo

The older the better? Relative age and grade effects on Norwegian national test performance

Oscar Skovdahl Jørstad

Master thesis

Master of Science in Assessment, Measurement and Evaluation

Centre for Educational Measurement

Faculty of Educational Sciences

University of Oslo

Spring, 2020

Popular abstract

In this study, we investigate the relationship between students' birth month and school performance. Previous studies suggests, when everything else is equal, that students born earlier in the year tends to perform better in school than students born later in the year. This is known as the relative age effect (RAE). We use data from Norwegian national tests to investigate the relationship between students' birth month and test scores. There are national tests in reading and numeracy for grade 5, 8 and 9, and English for grade 5 and 8. We aimed to estimate: (1) How much scores on national tests averagely increases for students born in December to January. (2) If RAE affects genders differently. (3) The ratio of RAE to the effect of having attended school for one additional year in grade 8/9. (4) Whether RAE decreases in older grade years.

We confirm that the older the students are, the better they perform. There are no gender differences in RAE on national tests. We found that being one year older has a larger effect on national test performance than having spent one additional year in school. RAE decreases over grade years, meaning that the difference in performance between the yougest and oldest declines with age. Students' birth month is one of the reasons why students' school performance differs.

Therefore, teachers needs to be sensitive to that some students might be lagging behind in school, because they are younger when tested in school, and need more time to mature.

Acknowledgements

Firstly, I would like to express a special thanks to my supervisors - Professor Rolf Vegar Olsen and doctoral research fellow Henrik Galligani Ræder. Your knowledge, availability and support has been outstanding, and this thesis would not have been possible without your guidance. Collaborating with you have been a real pleasure and your ways of conveying your knowledge has been inspirational.

Thanks to dr. Alexandra Niculescu and associate professor Stefan Schauber who have inspired and supported my work throughout the whole process. Your course on how to communicate scientific work and personal feedback is highly appreciated.

Thanks to Hilde Olsen and Marthe Akselsen at the Norwegian Directory of Education and Training. You have provided me with all the necessary resources to be able to do this research. It has been rewarding and insightful to collaborate with you.

Thanks to my peers for being fantastic colleagues throughout the whole Master program. You have made me a better student of this subject, and I am grateful to have you among my friends.

Thanks to my family and friends who supports me unconditionally. I can never express how grateful I am for you.

Abstract

Previous studies has found that the youngest students perform more poorly, on average, than older peers in school. This phenomenon is known as the relative age effect (RAE). In Norway, the age difference within a grade year can be up to 12 months. All Norwegian students participates annually in national reading, numeracy and English tests in 5th, 8th and 9th grade, which tests students' basic curriculum skills. We apply population data on national tests to study RAE. Specifically, we aim to investigate: The linear effect of RAE on national tests, the ratio of RAE to the grade effect in grade 8/9 in numeracy and reading and lastly how RAE changes from grade 5 to 9 in numeracy**.**

This study applies ordinary least square regression to estimate the linear effect of RAE, sharp regression discontinuity design to estimate RAE and grade effect on grade 8/9 in reading and numeracy. We incorporate vertical linking of the numeracy tests to investigate how RAE changes over grade years. In accordance with previous findings, we confirm that RAE is strongest in grade 5 and declines over grade years. We found that RAE has a larger impact than the grade effect on national reading and numeracy tests in grade 8/9. These results suggests that RAE has a stronger impact on national test proficiency than the amount of years spent in school at this point of the educational track.

Keywords: Relative age effect, grade effect, vertical linking, national tests, regression discontinuity design, ordinary least squares regression

Introduction

Does students relative age position within a grade year have an effect on their school performance? Currently, there exists a large body of literature which finds that relatively older students in a grade year tend to outperform their relatively younger peers on school outcomes, given that all other factors are held constant. This phenomena is referred to as the relative age effect (RAE). RAE can further be defined as the extent to which a students' relative age within a grade year is related to performance at the time of testing in school (Bedard & Duhey, 2006). RAE is a phenomenon with serious implications for school performance. RAE has been studied in academic, economical, mental health and sports settings. The findings usually suggests that older peers perform substantially better, at the time of testing, compared to the youngest peers within a given cohort. RAE is also strongest in the lower grade years in formal schooling (Black, Devereux, & Salvanes, 2008; Olsen & Björnsson, 2018). There are several ways to understand age in the context of studying its relationship to students' school performance. First, chronological age can be understood as a representation of students' relative age position within a grade year. Students' relative age position within a grade year could be considered as one of the many factors that can be used to identifying top and bottom performers in school. Second, when comparing students' performance across grade years, it is worth noting that students' age is correlated with the grade year they attend. Therefore it is reasonable to assume that a one year difference in age, or more, can to a certain extent explain differences in students' educational outcomes. The reason is that older students usually have spent more years in school. Third, age can also be understood as a function of when students enter primary schools. Therefore, it can be used as a measure of how many years they have attended formal schooling. The exception to the latter point concerns individual cases with deferred or accelerated school start.

In all countries, school start is strongly related to students' birth date. Usually, a particular date is used as a criteria to decide when students enter primary school. The consequence is then that students that are born just one day after the cutoff-point will start school a year later than students born on the day of the cutoff-date. According to OECD (2018), starting age for compulsory education differs across the member countries from 3 years of age (i.e., Mexico, Israel and Hungary) to 7 years of age (i.e., Sweden, Estonia and Finland). The starting age in Norway is 6, which is the most common starting age for OECD countries[1]. To be specific, the oldest students in Norwegian classrooms are born at 1st January and the youngest are born 31st December. Furthermore, there are very small degrees (i.e less than 1 percent annually) of deferred/accelerated school start (Cools,

[1] To be more precise, the Norwegian school year commences in mid-August. Children enroll into the first year of elementary school the year they turn 6 years of age. This means that the age at school start can be between 5 years and 8 months to 6 years and 8 months.

Schøne, & Strøm, 2017). In addition, re-sitting grade years in Norway is also very unusual. Given this strict practice and almost perfect relationship between age and grade, Norway provides a perfect system to study RAE and grade effects. In other countries this is much more complex to study, given that school starting policies may be quite flexible and retention/promotion is more frequently applied (Olsen & Björnsson, 2018).

It should also be mentioned that the estimation of RAE is important for studying or evaluating other features of educational policy than those under scrutiny in this thesis. For instance, in order to evaluate questions that considers an ideal age for school start, the appropriateness of flexible school start for younger students, whether students born at certain times of the year has increased risk of poorer school performance or what impact spending a certain number years in school has on school outcomes. Furthermore, RAE is interesting to study with clearly established cutoff-points, because it allows for investigation of how RAE changes over two adjacent grade years. Students born on the cutoff-date will be the youngest students in their cohort, and students born the day after the cutoff-date will be the oldest in their respective cohort. The difference in age might be as small as one day, but the difference in number of years spent in school will then be one whole year. Ultimately, in all of these contexts it is necessary to adjust for relative age effects, because students age is embedded as in all of these types of studies.

In the present study, we aim to estimate RAE and the grade effect (i.e the effect of having attended school for one additional year relative to the comparison group) on school performance. Students' age and grade year is strongly related, therefore we aim to separate students' age from their grade year to compare the impact of these components on school performance among grade 8 and grade 9 students. We also aim to investigate RAE in grade 5. However, since there are no adjacent grade years to grade 5 in the available data we cannot estimate the grade effect for grade 5. This study aims to contribute to research questions concerning whether RAE or the grade effect has the largest impact on school performance.

Literature review

RAE and school performance

Previous studies has found various results in terms of the impact of RAE on school performance. For example, in the Norwegian context, Olsen and Björnsson (2018) investigated the relationship between RAE and performance in large scale assessments (i.e PISA and TIMSS) over the last 20 years. The results suggest that the older the students were at the time the tests were conducted, the better they performed. RAE showed to be robust over the last 20 years of PISA and TIMSS assessments. Further, there are similar findings from various countries suggesting that the youngest students' school performance is affected by RAE. In addition to this, the youngest students are also less likely to enter higher education than their older peers. Support for these claims has been found in Italy (Ponzo & Scoppa, 2014), England (Crawford, Dearden, & Greaves, 2013), Germany (Puhani &

Weber, 2008), Canada and United States (Bedard & Duhey, 2006), Spain and France (González-Vallinas, Librero, Peiró, & San Fabián, 2019; Pedraja-Chaparro, Santín, & Simancas, 2015). Solli (2017) found, using Norwegian student data, that the oldest students within grade year cohorts has significantly higher GPA's than their youngest peers by the time they graduate from primary school (i.e., 10th grade). In addition, the oldest students are more likely to graduate from upper secondary school by age 19 and more likely to enroll directly into university or college after graduation from upper secondary school. In the same study, the relationship between students socio-economic status (SES) and RAE on school performance is also investigated. The findings suggests that the impact of being born late within a year cohort affects children with low SES-background stronger than children with high SES-background (Solli, 2017). A possible explanation for this finding is that students with higher SES-background have parents that tends to intervene faster when their children's school performance drops, compared to students with lower SES-background (Buckles & Hungerman, 2013; Crawford, Dearden, & Greaves, 2011; Currie, 2009).

Other factors that has shown to be associated with RAE and school performance is the degree of deferred school start. This includes the need for more educational support among the relatively youngest students. Several studies has found that a disproportionate number of the youngest students within grade years are referred to special education interventions, and needs more time in school to catch up with their older peers (Black, Devereux, & Salvanes, 2008; Sharp, 1995; Sykes, Bell, & Roderio, 2009; Wilson, 2000). In addition, Solli (2017) found that among the deferred children in the Norwegian education system, 20% of those children are born in November and 55% of the children that defer school start are born in December. Although deferred school start in Norway is rather unusual, more boys than girls delay school start.

With regards to gender differences on school performance in Norway, girls tend to outperform boys in all subjects by the end of primary school (grade 10) apart from physical education and, to some extent, mathematics (Statistics Norway, 2018). In terms of GPA by grade 10 (i.e final year of lower secondary school), girls obtain 0.7 higher GPA scores in the Norwegian subject and 0.2 higher GPA scores in mathematics than boys. These gender differences are smaller on exam grades and remains smaller in upper secondary school and higher education (Statistics Norway, 2017). However, on standardised tests (e.g., large scale assessments) the results are slightly different. In PISA, boys score higher in mathematics than girls. These results also reflects their performance on national tests. National tests are measures of basic skills that are central to the curriculum in all subjects. The results from national tests shows that boys tends to score higher in numeracy and English, whereas girls tends to score higher in reading (Stoltenbergutvalget, 2019).

More related to the present study is the findings concerning gender differences in RAE, which has recently been investigated on Norwegian national tests. In numeracy, Aune and colleagues (2018) found evidence to suggest there is a larger RAE for girls than boys. More specifically, they found that there is less variation in boys' than girls' numeracy scores, when controlling for birth month (Aune, Ingvaldsen, Vestheim, Bjerkeset, & Dalen, 2018). A similar trend was found for national reading test results across 5th, 8th and 9th grade students. However, in reading boys have a larger RAE than girls (Vestheim, Husby, Aune, Bjerkeset, & Dalen, 2019). An interesting remark on these points is that there seems to be unclarity from other standardized tests whether RAE differs between genders. Regarding large scale assessment results in Norway, Olsen and Björnsson (2018) found no gender differences in RAE for 4th and 8th graders on TIMSS, however there is a larger RAE for boys than girls in grade 10 on PISA. This finding is interesting because it indicates that the interaction between RAE and gender fluctuates across test scales. A plausible explanation regarding a larger RAE for boys could reflect that biological and cognitive mechanisms related to maturation has larger intra-sex variation in males than females (Lehre, Lehre, Laake, & Danbolt, 2009). The gender differences in RAE may be explained by differing maturation rates for boys and girls.

RAE in athletic performance

The effect of relative age has been extensively studied in the context of sports and athletic performance. RAE is of large interest to consider in these contexts as age is strongly related to biological maturation of necessary physical attributes for athletic performance. These attributes relates to greater height, muscular strength, speed and, to a certain extent, body mass. These attributes are beneficial for athletes in most sports and are usually more present in the oldest individuals, especially at youth levels. As a consequence, coaches might perceive the tallest, fastest and strongest athletes to be the more advantageous performers in their pool of athletes which are most likely athletes born within the first quartile of the sporting season (Cobley, Baker, Wattie, & McKenna, 2009). It is very common in sport to apply an organizational strategy such as annual age-grouping of athletes to define cut-off points for team selection. This is similar to cut-off dates used to assign students to grade years. Therefore, RAE might explain why relatively older athletes are more favored for promotion at youth levels. This claim is supported by findings on the relation between RAE and performance in baseball, soccer and ice hockey. The results shows that individuals born within the first quartile of the sporting season are overrepresented at various age groups and levels of performance. Hence, older athletes in the youth system has a better opportunity to acquire more play-time as they are more likely to be selected for matches and thus gets more experience in competition. This also includes technical advantages and more access to play at higher levels of competition and coaching (Baker & Horton, 2004; Barnsley & Thompson, 1988; Côté, Baker, & Abernethy, 2007; Helsen, van Winckel, & Williams, 2005;Musch & Grondin, 2001; Sherar, Baxter-Jones, Faulkner, & Russell, 2007; Thompson, Barnsley, & Stebelsky, 1991; Ward & Williams, 2003; Wattie, Schorer, & Baker, 2015). Athletic performances is not the type of performance that is studied in the present paper. However, these studies are relevant and interesting to consider, because many types of sports applies systems with, at times, extremely strong selection mechanisms based on performance. This suggest that school systems with performance-based selection could lead to (unfair) selection favoring the relatively older students.

RAE in psychological literature

There are many different psychological features that has been studied in relation to RAE. In this section we will present some of the findings which has an impact on students cognitive functioning and their well-being. Firstly, many families in various countries tend to delay school start for children that are not seemingly ready for formal schooling. Delaying school start might result in long-term benefits for students mental health. Using Danish register data, Dee and Sievertsen (2018) found that individuals that delayed school start by spending one additional year in kindergarten displayed a strong reduction in symptoms related to inattention and hyperactivity around age 7. In addition, a recent study in Florida investigated the relationship between school starting age and cognitive development. The results suggested that starting school later has a positive effect on school performance due to additional time to develop cognitively before formal testing (Dhuey, Figlio, Karbownik, & Roth, 2019).

Other findings related to mental health and RAE concerns individuals that commits suicide. Salib and Cortina-Borja (2006) found, using English and Welsh data on suicidal attempts, that individuals born in the spring and early summer (i.e. the youngest age quartile according to the British school starting age policies) had a 17 % increased risk of committing suicide, compared to individuals born in the other seasons of the year. A similar result was found in the US, where a disproportionally large number of individuals who were born in the second half of the year they were eligible for school start committed suicide between 1979 and 1992 (Thompson, Barnsley, & Dyck, 1999). In addition to suicidal attempts, there is also evidence to suggest that relatively younger peers in a classroom are overrepresented with mental disorders such as ADHD, mood disorders (i.e major depression disorder and bipolar disorder) and schizophrenia (Chen, et al., 2016; Disanto, et al., 2012; Fuller, Rawlings, Ennis, Merrill, & Flores, 1996; Morrow, et al., 2012; Rihmer, et al., 2011; Tochigi, Okazaki, Kato, & Sasaki, 2004). These findings should be emphasized in debates concerning youths mental health. RAE is certainly not a cause of these disorders, but may be interpreted as an indicator for which individuals that may be more prone to develop mental disorders.

Impact of RAE and grade effect on intelligence test performances.

Children’s intellectual performance is strongly related to their cognitive development, which in turn is strongly related to their chronological age. Educational psychologists has studied whether additional years of schooling can improve performances on intelligence measures. These studies raise the issue of whether intelligence scores increases simply because the students gets older, or if the increase in intelligence scores is due to students spending additional years in school (Cahan & Cohen, 1989). Cliffordson and Gustafsson (2008) utilized Swedish military enlistment test scores to study the effect of chronological age and length of schooling on various aspects of intellectual performances. They found that length of schooling is a considerably stronger predictor of intelligence than chronological age. The results suggested that IQ increased by 2.7 points, on average, for each added year of schooling (the age effect was close to zero). This especially concerns students enrolled in the most academically oriented education programs before they enroll into higher education (Balke-Aurell, 1982; Cliffordson & Gustafsson, 2008; Carlsson, Dahl, Öckert, & Rooth, 2014; Lund & Thrane, 1983).

RAE and teacher expectancy effects

Students capacity for succeeding in school, as perceived by their teachers, could potentially have a relationship with the students' relative age. This is especially the case in the lower years of the educational track when maturity differences between the oldest and youngest student is larger (Sharp, 1995). Teachers' perception of student behavior in the classroom could have implications for how they expect their students to perform on tests in school. Weinstein, Marshall, Sharp and Botkin (1987) claimed that teachers tends to label younger students as more immature, relative to the other students in the classroom. Immaturity is associated with relatively less developed attention spans and interpersonal skills for cooperating with older peers. Support for this claim has been found in the US, which raises a concern regarding the extent that teachers takes students maturity differences into consideration in the assessments of their students (May, Kundert, & Brent, 1995; Rubie-Davies, 2006; Sykes, Bell, & Roderio, 2009). Teacher expectancy effects are important to consider. The literature suggests that when teachers have high expectations to students they are more likely to perform better on assessments. Therefore, it is conceivable that students labelled as immature by teachers are not only the relatively youngest students in class, but may thus also be perceived as less capable of scholastic success, in contrast to the oldest students which are more likely to display relatively more mature behavior in class (Rosenthal & Jacobson, 1968; Rubie-Davies, Flint, & McDonald, 2012; Weinstein, Marshall, Sharp, & Botkin, 1987).

The Norwegian school context and national tests

An inherent methodological issue related to RAE is the limitations for comparing the effects across countries, especially when considering educational outcomes. Countries have differing school starting policies. Differing school starting policies concerns school starting age and number of annual school admissions. The relationship between relative age and school outcomes is affected by such differences in policies. One consequence could for instance be that students of the same age attend different grade years in some countries (i.e. countries with a bi-annual school admission policy), while this rarely happens in other countries (i.e. countries with an annual school admission policy). Further issues regarding comparisons between countries concerns the proportion of deferred students. This issue also concerns students with accelerated school start. Accelerated students will be among the youngest peers within a grade cohort, but because of their accelerated school start they would tend to score unusually high on cognitive scores relative to the youngest peers of their cohorts (Luyten, Merrell & Tymms, 2017). In turn, this causes issues with homogeneity in research designs which aims to compare the relationship between relative age and school performance across countries. In Norway, chronological age and grade year is almost perfectly coinciding. The policies for re-sitting grade years in compulsory school is very strict and the occurrences are almost non-existent. Non-compliance to the enrollment policies in Norway requires an expert assessment of whether a given student is too immature to begin school at the intended year (Solli, 2017).

In the present study we aim to investigate how RAE impacts students' performance on national tests. National tests is one of the Norwegian education systems key instruments to provide information about overall student achievement. The test scores are used as a pedagogical tool to inform schools about their students' basic skills in numeracy, reading and English. The tests also serve as a basis for formative assessments during the school year and for quality improvement in all parts of the Norwegian education system, including research (Hovdhaugen, 2016; Tveit, 2014). Participation in national tests are compulsory for all 5th, 8th and 9th grade students. The exceptions to this rule concern students with various special education needs and language difficulties among other reasons[2]. The national tests are low-stakes tests for the students. However, in some schools/municipalities the stakes of the tests are higher because the results are available to the public (Elstad, 2009). The results have no impact on their admission to higher grade years in the primary school system or applications for schools to higher levels of education, such as upper secondary school. National tests are conducted annually in the first semester (i.e. fall) of the school year. The assessments in numeracy and reading are conducted for all three aforementioned grades, but the English test only includes grade 5 and 8. All national tests are computer-based. 8th and 9th graders receive the same test in numeracy and reading. In reading and numeracy tests, the students are assigned 90 minutes for completion of the tests. For the English tests, the students are assigned 60 minutes for completion of the test. Scores on all the tests are divided into mastery levels which is characterized by various degrees of competency. In grade 5 there are three mastery levels and in grade 8 and 9 there are five mastery levels (Directory for Education and Training, 2017; Ræder, Olsen, & Blömeke, 2020).

[2] The percentage of students that were exempt from national tests in 2018/2019 ranged from 3.3% to 5.9% across all tests and grade years. The percentage of students that did not participate in national tests in 2018/2019 ranged from 1.1% to 1.9% across all tests and grade years. Similar results can be found for previous school years (Directory of Education, 2020).

The present study

In the present study we focus on RAE mostly, and the grade effect to a certain extent. The grade effect is of interest to study in the context of RAE when students in adjacent grade years with the same test are investigated. The reason is that this allows for estimating how large the effect of attending school for one additional year is on test performance. At the same time, this allows for comparing the grade effect to the effect of relative age, and investigate how the effect of relative age changes over grade years (Cahan & Cohen, 1989; Cliffordson, 2010; Cliffordson & Gustafsson, 2008; Kyriakides & Luyten, 2009; Luyten, 2006; Luyten, Merrell, & Tymms, 2017; Gerritsen & Webbink, 2013). National tests are indeed conducted on two adjacent grade years, meaning we can compare the effect students' amount of schooling and age on tests for reading and numeracy in grade 8 and 9. We cannot investigate the grade effect for grade 5, since there are no adjacent grade years for comparison in the national test format.

Furthermore, national tests has been lacking a procedure that links tests for different grade years onto the same baseline scale. Recently, Ræder, Tokle and Olsen (2019) provided a report which proposes a vertical linking design for numeracy scales on national assessments. The vertical linking of scale scores is the most unique contribution this paper will bring to the existing body of literature on RAE on school performances. Unfortunately, vertical linking for national reading and English tests are currently not available.

Another important contribution from this study is that, in addition to providing descriptive statistics on the differences in performance, we provide a robust estimate of the effect of relative age and the grade effect on performance in national tests. RAE on national tests in Norway have previously been reported in the form of comparing mean scores across birth quartiles (Aune, Ingvaldsen, Vestheim, Bjerkeset, & Dalen, 2018; Vestheim, Husby, Aune, Bjerkeset, & Dalen, 2019), while in this study RAE is modelled as a linear effect of birth month. In addition, this study contributes to and adds to this literature in the following ways:

- A robust methodological approach by applying a regression discontinuity design (RD-design). RD-designs reflects almost the same causal force as those from a randomized trial, when standards and assumptions are sufficiently met (Shadish, Cook, & Campbell, 2002; Schochet, et al., 2010).
- As compared to the previous studies, this analysis makes use of more recent data, and includes comparisons across all the three domains of testing. This provides a more holistic picture of the impact of RAE and the grade effect on national tests.
- The present study applies item response theory ('IRT') – calibrated scale scores instead of raw scores from the national tests. Furthermore, by using the vertically linked scales this allows the scale scores to be placed on to the same baseline scale. In other words, this allows for a direct comparison of grade 8/9 scale scores with grade 5 scale scores in

numeracy, when applying the vertical linking technique developed by Ræder, Tokle and Olsen (2019). This gives the present study the opportunity to verify the use of vertically linked scale scores in numeracy.

The study investigates the following research questions:

- RQ 1: What is the linear effect of relative age on students' national test performance, across the various subjects?

  Sub-questions related to RQ 1:

  A) What are the differences in the effect of students' relative age on national test performance across the various subjects and grade years?

  B) What are the gender differences in the effect of students' relative age on national test performance across the various subjects and grade years?

- RQ 2: What is the ratio of RAE over the grade effect in grade 8 and 9 in numeracy and reading?

  Sub-question related to RQ 2:

  A) What is the ratio of RAE over the grade effect in grade 8 and 9 in numeracy and reading, for each gender separately?

- RQ 3: How does RAE change across grade 5 to grade 8 and 9 in numeracy?

Methods

The method section starts with a description of the dataset that is used for the present study, including descriptive statistics that is relevant for describing samples (i.e sample size and distribution of students for all birth months). Next, we provide a description of the variables that are used in the various analyses conducted in this paper as well as justifications for each method used in the study.

Sample and data

This study utilized data on national test results from Norwegian students in grade 5, 8 and 9 in the school year 2018/2019. The dataset was provided and prepared by the Norwegian Directory of Education and Training ('DET'). We separated the dataset into sub-groups, containing one set for each grade year and subject (e.g., 5th grade reading and 8th grade numeracy etc). A total of 38 observations had invalid birth months in the raw data we received from DET (i.e larger birth month values than 12). These observations were removed from the dataset that was ultimately used for all analyses. Table 1 provides an overview of the sample sizes in each grade year and subject. We have no code that links test results to individual students. This means we have no indication of which students that are absent on one test but present on the others (e.g present on the numeracy test but not on the reading and English tests).

Table 1. Sample sizes tabulated by subject and grade year.

Subject/grade	**Numeracy**	**Reading**	**English**
Grade 5	N= 60,665	N= 59,995	N= 59,998
Grade 8	N= 59,171	N= 59,043	N= 58,873
Grade 9	N= 58,880	N= 58,802	

Figure 1 provides graphical insight into the distribution of students per birth month for all grade years included in the study. Interestingly, the proportion of children born in November and Decmeber in all grade years are smaller relative to the other months apart from February. We would expect students born in February to consist of the smallest amount of students, because February is the shortest month of the year. However, according to Statistics Norway, there are indeed fewer children born in November and December compared to the rest of the year (Statistics Norway, 2019). There is a satisfactory large overlap between the number of students participating in national tests per birth month and the number of children born in the respective years. We can therefore assume that the sample size used in this study reflects the student populations for these respective years. See appendix 3 for the actual distribution of children born per month**.**

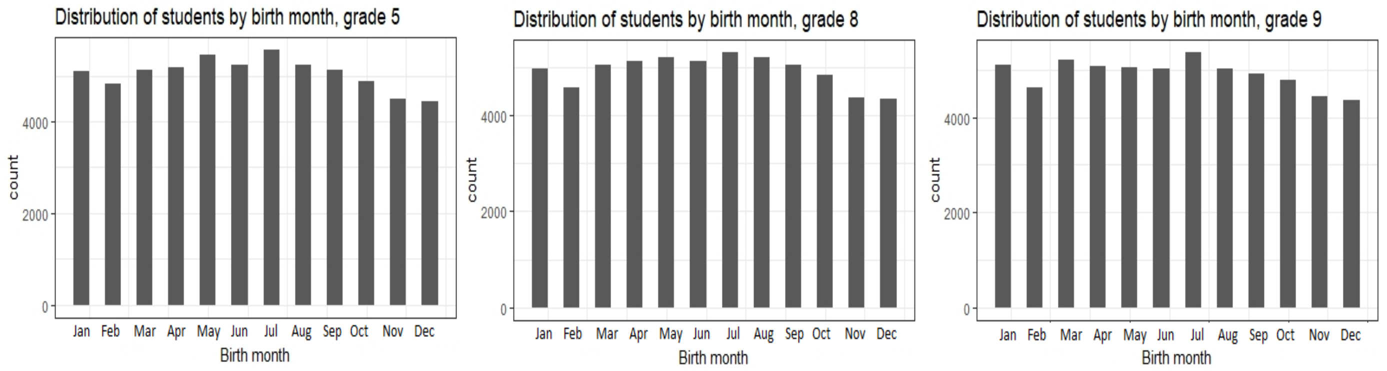


Figure 1. Histograms for distribution of students in each birth month for each year. The results reflects the distribution for numeracy tests, similar results was found for reading and English but were not included here.

Since the data set contains no variables that can directly identify any individuals in the data set, no declaration according to the regulations from GDPR (General Data Protection Regulation) was needed to ethically approve this study. The application form and confirmation of ethical approval from the Norwegian centre for research data can be found in appendix 1. All data management and analyses were conducted in the statistical software R, and the coding-script can be found in appendix 2.

Measures/variables

There are three independent variables in this study. The independent variables are birth month, gender and grade. In addition, we included two interaction terms – One for gender and birth month and the other for grade and birth month. The interaction terms are not main effects, but indicates whether RAE differs between genders and grade years. In the present study, birth month were used as an independent variable representing relative age. Birth months are reverse coded, meaning that December is coded as 0.5 and January 11.5. There are two reasons that motivates this decision:

- By using half-intervals, the birth months are now representing the average birth date within each month.
- The interpretation of the intercept in all regression analyses now becomes more meaningful, since the intercept now represents the scores for the youngest students born at the cutoff-date.

The dependent variable for all analyses was the students' scale score for each subject in the data set. The national tests uses scale points which is based on standardized scores with a mean score of 50 and standard deviation of 10 points. Furthermore, the national test uses a calibration procedure to measure changes in student cohorts' proficiency over time, which is based on models used in item response theory (IRT). This calibration procedure has been administrated since 2014 for national numeracy and English tests (Björnsson, 2018), and reading tests in 2016 (Björnsson, 2016). The achievement scores on national tests are used to assign the individual students' scores to different mastery levels. The mastery levels were normatively distributed in the respective calibration years (i.e., 2014 for Numeracy and English, and 2016 for reading). In the following years, the original normative distribution is used as a criterion for characterizing the scores in various mastery levels. In practice, this means that in grade 5, 25% of the students are allocated in mastery level 1, 50% of the students are allocated in mastery level 2 and 25% of the students are allocated in mastery level 3. For grade 8 and 9 - 10% of the students are allocated to level 1, 20% to level 2, 40 % to level 3, 20 % to level 4 and 10% to level 5 (Björnsson, 2016). See table 1.3 in appendix 3 for a description of which mastery level a given achievement score is characterized as on the national tests.

Figure 2 uses stacked barcharts to visualize the distribution of all mastery levels across grade years and birth months in the present data set. As expected, we can see a clear tendency which shows that the largest proportion of students which achieved the highest mastery levels are born between January and March. The largest proportion of students which achieved the lowest mastery levels are born between October and December. In addtion, we can see that the percentage of students which achieves the lowest mastery levels decreases over grade years. This in turn results in larger groups of students achieving higher mastery levels.

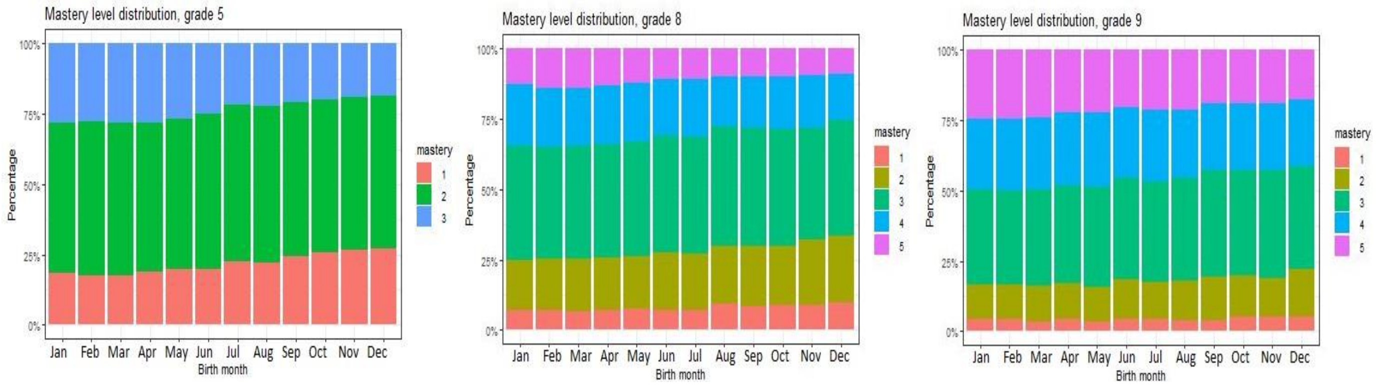


Figure 2. Stacked barcharts for the distribution of mastery levels across birth months for all grade years on national tests. These figures are based on the distribution for numeracy tests, similar results was found for reading and English but were not included here.

Statistical methods

This section provides descriptions of the statistical methods used to answer the research questions that has previously been stated in section "The present study". Further, we provide descriptions of assumptions and standards for the regression analyses that needs to be met to ensure satisfactory internal validity. Lastly we provide explanations for why these particular methods were chosen.

Ordinary least squared regression

In terms of the statistical methods used, the study applied ordinary least squared regression (OLS) as the statistical method for answering RQ 1, including its sub-questions. OLS regression linearly models the relationship between a dependent variable and a set of independent variables. This allows us to test for how well the independent variables predicts the dependent variable, and how much the independent variables accounts for the variance (i.e. R-squared estimate) in the dependent variable (Bruce & Bruce, 2017). More specifically, the OLS regression analyses allows the present study to investigate the impact of age on performance for national tests within a full year age cohort. In addition to using birth month as an independent variable representing relative age, we also include gender and an interaction term for gender and birth month as predictor variables. This allows for testing whether RAE differs between genders on national test performances. It is important to note, as mentioned earlier, that this paper is not concerned with the main effect of gender differences on national test results. The gender variable is only included to investigate the interaction effect of relative age and gender. The equation for the linear regression analyses is then modelled as

$$Yi = \beta0 + \beta1 \cdot \text{ Birth month} + \beta2 \cdot \text{ Gender} + \beta3 \cdot \text{Birth month x Gender} + \epsilon i, \quad (1)$$

where Yi is the national test score for student i, β1 represents the regression coefficient for age in birth months. β2 represents the regression coefficient for gender which is a dichotomous variable where boys is used as the reference group. β3 represents the regression coefficient for the interaction term birth month x gender. The interaction term is used for investigating the effects of birth month on both genders when applied as a predictor for the score on the national assessment. ϵi represents the coefficient for the random error component (i.e. residual) for student i.
OLS regression analyses was conducted on each subject for each grade year separately.

OLS regression models follows a set of assumptions regarding the independent and dependent variables, including the relationship among them. The assumptions must be sufficiently met in order to claim that the regression models can make any predictions between the set of independent variables and dependent variable (Bruce & Bruce, 2017). Descriptions of how these assumptions are met for all the OLS regression models will be included in the result section. The relevant figures, parameter estimates and more detailed descriptions of the results regarding the assumption tests can be found in appendix 3**.** These assumptions includes:

1. Normality of relationship between variables Y|X (and hence also of ϵ) (i.e. the relationship between the dependent and independent variables including residuals should be normally distributed)
2. Homoscedasticity over X for Y|X (including ϵ) (i.e. constant variance)
3. Linear relationship between Y and X (i.e. The relationship between the dependent variable and independent variables should be linear in its form)
4. Mutual independence among residuals (i.e. variance between residuals should be equal to zero (σ ϵi, ϵi’ = 0))
5. Independence of residual errors and predictors (i.e. Absence of influential outliers and extreme values in the independent variables, (σ ϵi, X = 0))

Regression discontinuity design

Regression discontinuity design (‘RD-design’) refers to a quasi-experimental pretest-posttest design which allows for assignments of a treatment and control condition. In RD-design it is common to apply a substantially meaningful continuous predictor with a threshold value, which defines a criteria for assigning study units to different groups (i.e control and treatment groups) in a population. Apart from assigning study units to different groups there is nothing else that differentiates the study units in each group. Furthermore, RD-design enables the possibility to evaluate the causal effects of the given conditions. In addition, RD-designs measures the effects of individuals close to the cutoff point of the assignments of conditions (Shadish, Cook, & Campbell, 2002). In practice, the effect of the intervention that is measured in the RD-design is estimated by the sudden leap at the discontinuity (i.e the cutoff-point). In order to investigate the ratio of RAE on the grade effect on performance in reading and numeracy, we applied a regression discontinuity design to estimate the overall difference across two neighboring year cohorts with birth month as a continuous independent variable (i.e. forcing variable).

When studying education-related interventions, RD-designs are increasingly used to obtain unbiased estimates. RD-designs are applicable when a continuous scoring rule is applied to assign the intervention to study units. In this paper, the continuous scoring rule concerns the assignment of grade 8 students to the control group (values below the pre-set cutoff value) and grade 9 students to the treatment group (values above the pre-set cutoff value). The cutoff point is then set at the time point between January for grade 8 and December for grade 9. Stated differently, December to January for grade 8 is coded as -11.5 (December) to -0.5 (January) and 0.5 (December) to 11.5 (January) for grade 9. The cutoff point is set to 0. Hence, the intercept of the RD design should be interpreted as the average score in achievement for the oldest student in grade 8.

In RD-design, an effect occurs if there is a discontinuity in the two regression lines (or curves) at the cutoff. In practice, it is not a large difference in age between 9th grade students born late in December and 8th grade students born early in January. The smallest possible difference in age between the oldest grade 8 student and the youngest grade 9 student can be a matter of seconds. However, by the time of testing according to the school starting age policy in Norway, the youngest 9th grade students born late in December would have spent one year extra in school compared to the oldest 8th grade students born early in January. Hence, it would be of interest to not only investigate RAE within 8th and 9th grade, but also to investigate the impact of having spent one more year in school when looking at their performance on national tests (i.e. the grade effect). Therefore, a RD-design is an appropriate method to use as it is able to utilize exogenous influence (i.e. school starting age policy) on how Norwegian students are assigned into different classes. RD-designs also allows to investigate whether it is relative age or grade that has the strongest impact on students' performance in school. Further, in this RD-design, the grade effect is logically estimated by the difference in achievement between the youngest student in grade 9 and the oldest student in the grade 8. In practice, this is done by entering the grade as a dummy variable into the regression equation.

RD will generate unbiased estimates if (1) the relationship between the outcome variable and forcing variable can be modelled correctly, and (2) the forcing variable (birth month) was not manipulated to influence the treatment assignments. In addition to this, the forcing variable in RD-designs are recommended to be at least be ordinal in its nature. It must also include at least four unique values above and below the cutoff point. In order to apply RD-designs correctly, the study must sufficiently satisfy the following set of standards (Schochet, et al., 2010):

- **Standard 1: Integrity of the forcing variable**
  No systematic changes in units from their true values (i.e manipulation) to influence treatment assignments.
- **Standard 2: Attrition**
  Attrition rates must be low. RD-studies have to report the number of study units (e.g number of students) that were assigned to the treatment and control group.
- **Standard 3: Continuity of the relationship between the outcome and forcing variable**
  When there is absence of an intervention, there would be presence of a smooth relationship between outcome and forcing variable at the cutoff point.
- **Standard 4: Functional form and bandwidth**
  Involves controlling for the forcing variable when estimating the treatment effect, including choice of appropriate functional forms and bandwidth of the forcing variable.

In the present study, the equation for the RD-analyses are modelled as

$$Yi = \beta0 + \beta1\cdot \text{ Birth month} + \beta2\cdot \text{ Grade year} + \epsilon i, (2)$$

where Yi is the national test score for student i. β1 represents the regression coefficient for age in birth months. β2 represents the regression coefficient for grade year, which is a dichotomous variable indicating grade year where 8th grade being the control group (i.e. coded as ‘0’), and 9th grade is then the treatment group (i.e. coded as ‘1’). ϵi represents the coefficient for the random error component for student i.

A description of how the aforementioned standards for RD-designs were met will be presented in the results section. To test for standard 1 (Schochet, et al., 2010), the RD-model is extended with an interaction effect of age and grade year. If standard 1 is satisfied we have evidence to suggest that RAE has the same functional relationship with the outcome variable at both sides of the cutoff-point. Thus, the RD-equation (2) would be modified to model the relationship between birth month across grade years and achievement scores as

$$Yi = \beta0 + \beta1\cdot \text{ Birth month} + \beta2\cdot \text{ Grade year } + \beta3\cdot\text{Birth month x Grade year} + \epsilon i, (3)$$

All the regression coefficients in equation 3 are interpreted as the coefficients in equation 2. In addition, β3 represents the regression coefficient for the interaction term between age in birth months and grade year which allows us to test if RAE significantly changes across grade years. Furthermore, to test if the effect relative age and grade could be related to other exogenous variables, we conducted separate analyses for boys and girls. More detailed descriptions of how the present study meets the standards for RD can be found in appendix 3.

Vertical linking of numeracy scores

For the analyses on the numeracy scale regarding research question 3, this study applied the results from a vertical linking technique which encompasses results from national assessments in 5th grade with equivalent results from national assessments on 8th grade. Vertical linking entails that tests with comparable constructs for different target populations with different ability levels gets linked together (Kolen & Brennan, 2014). Recently, a vertical linking design for national numeracy tests has been developed (Ræder, Tokle & Olsen 2019). The vertical linking design for national numeracy tests was developed by constructing linking tests in numeracy for grade 6 and 7 which consisted of a substantial amount of overlapping anchor items from the national numeracy tests for grade 5 and 8. Vertical linking of national numeracy test scales is therefore possible due to the following reasons - By utilizing IRT models for the various items it is possible to use national test items for scaling students' scores on a given national test, and place two or more tests on the same scale. This allows for direct comparisons of scores in grade 5 with grade 8/9. Second, the constructs measured in national numeracy test for grade 5 and 8 has recently shown to be measuring a common construct which does not differ across the scales. This allows results from the two national numeracy tests to be placed and compared on the same baseline scale (Ræder & Olsen, 2020).

It has to be noted that the vertical linking design in numeracy is limited by the following aspects – The total sample of schools that participated in the linking tests cannot be considered representative for the whole Norwegian student population in grade years 6 and 7 (71 schools participated out of the 226 schools that were invited). However, based on aggregated information about the participating schools, there are no reasons to suspect large discrepancies between the sample and the population. This claim is further supported by the item parameters for the linking tests, which suggests that the vertical linking is not less stable because of a small sample used for the linking tests (Ræder & Olsen, 2020). Therefore, we can assume that the vertical linking of numeracy scale scores allowed the present study to successfully investigate how RAE changes over grade years.

Results

Results from OLS regression analyses

With regards to the assumptions for OLS regression, we ran diagnostic tests for each regression model which examines the residual distribution for each model separately. These diagnostic tests includes testing:

1) Normality of the relationship between variables Y|X, and the error term
(i.e. Inspecting Q-Q plots for residual variance)
2) Homoscedasticity over X for Y|X (including the error term)
(i.e. Checking the extent to which standardized residuals has a constant variance across the fitted values)
3) Linear relationship between Y and X (i.e. Checking for normal distribution by comparing fitted values to the residuals)
4) Mutual independence among residuals (i.e. Checking whether there are any observations with unusual high leverage on the regression model)
5) Independence of residual errors and predictors (i.e. Checking for Cook's distance in each residual value to determine whether there are any influential outliers and extreme values in the independent variables.)

In general, all regression models applied in this study showed satisfactory degrees of linearity. We conclude, from the residual diagnostic tests, that the residual errors were also sufficiently homoscedastic, normally distributed and did not influence the linearity of the models. Therefore, we can claim that the models in this present study met the assumptions of OLS regression. A more detailed explanation for these tests can be found in appendix 3.

Table 2. Ordinary least square regression coefficients, tabulated by the various subjects of national tests and grade year. Significant results are bolded.

Variables	Numeracy	Reading	English
__(5th grade)__			
__Intercept__	__49.509 (0.112)***__	__47.348 (0.109)***__	__49.096 (0.114)***__
__Birth month__	__0.282 (0.016)***__	__0.290 (0.016)***__	__0.308 (0.016)***__
__Gender (female)__	__-2.855 (0.158)***__	__1.527 (0.154)***__	__-1.368 (0.162)***__
__Birth month x gender__	__0.050 (0.022)*__	0.029 (0.022)	-0.014 (0.023)
R-squared	__0.029***__	__0.020***__	__0.017***__
__(8th grade)__			
__Intercept__	__49.269 (0.114)***__	__47.520 (0.112)***__	__49.686 (0.115)***__
__Birth month__	__0.220 (0.016)***__	__0.235 (0.016)***__	__0.168 (0.016)***__
__Gender (female)__	__-1.228 (0.163)***__	__2.144 (0.159)***__	__-1.070 (0.164)***__
__Birth month x gender__	-0.001 (0.023)	0.001 (0.022)	0.012 (0.023)
R-squared	__0.009__***	__0.019***__	__0.006***__
__(9th grade)__			
__Intercept__	__53.177 (0.118)***__	__51.204 (0.118)***__	
__Birth month__	__0.184 (0.017)***__	__0.189 (0.017)***__	
__Gender (female)__	__-1.464 (0.168)***__	__2.029 (0.168)***__	
__Birth month x gender__	0.021 (0.024)	0.030 (0.024)	
R-squared	__0.008***__	__0.017***__	

*Significance codes: ***= p<.001,**= p.<.01 , * = p<.05 (1)Birth month is reverse-coded for all grade years and subjects. December is recoded to 0.5. (2) Gender is coded as a dummy variable, where males is the reference group. (3) Birth month:gender refers to an interaction term between the independent variables. Standard errors are reported in the parentheses.*

Table 2 shows the output from the OLS regression analyses. Overall, the results from the various OLS regression analyses show a similar trend. We found that RAE is statistically significant in all grade years and subjects at hand. In grade 5, we found that the effect of one full year difference in age is 3.38 points in numeracy (i.e RAE-estimate multiplied by 12), 3.48 points in reading and 3.69 points in English. Figure 3 shows the linear relationship between birth month and achievement score in English for 5th grade. This figure confirms expected results regarding theory on RAE and school performance. Similar findings are found in the other subjects and grades, hence figure 3 also serves the purpose of working as an example figure for what the other OLS regression analyses would look like.

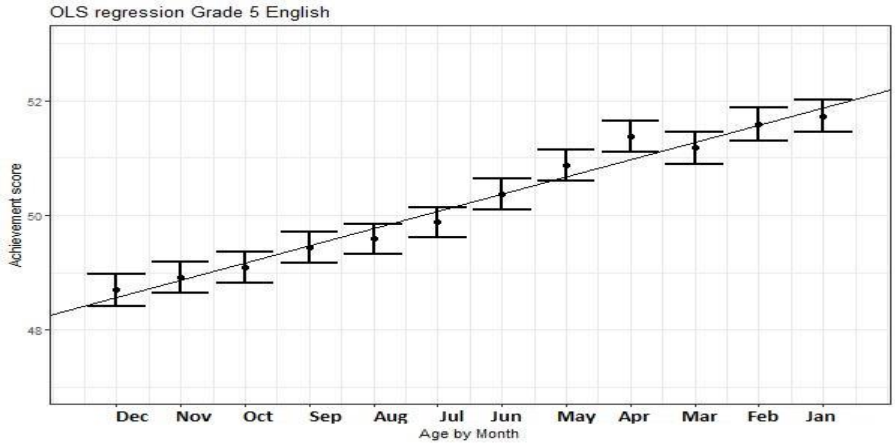


Figure 3. *OLS regression estimate of achievement scores in English for 5th grade students in error bars which is based on 95% confidence intervals of average achievement score per birth month. The straight line represents the regression line for average achivement score per birth month.*

In grade 8, we found that the effect of one full year difference in age is 2.64 points in numeracy, 2.82 points in reading and 2.01 points in English. In grade 9, we find that the effect of one full year difference in age is 2.20 in numeracy and 2.26 in reading.

For grade 5, the r-squared statistic showed statistically significant results for numeracy meaning that this regression model only accounts for 2.9% of the variation in numeracy achievement for 5th grade students on national tests. The r-squared statistic showed statistically significant results for reading, meaning that this regression model accounts for 2.0% of the variation in reading achievement for 5th grade students on national tests. The r-squared statistic also showed statistically significant results for English, meaning that this regression model accounts for 1.6% of the variation in English achievement for 5th grade students on national tests. For grade 8, the r-squared coefficient showed statistically significant results for numeracy, meaning that this regression model accounts for 0.9% of the variation in numeracy achievement for 8th grade students on national tests. The r-squared statistic showed statistically significant results for reading, meaning that this regression model accounts for 1.9% of the variation in reading achievement for 8th grade students on national tests.. The r-squared statistic also showed statistically significant results for English, meaning that this regression model accounts for 0.6% of the variation in English achievement for 8th grade students on national tests. For grade 9, the r-squared statistic showed statistically significant results for numeracy, meaning that this regression model accounts for 0.8% of the variation in numeracy achievement for 9th grade students on national tests.. The r-squared statistic also showed statistically significant results for reading, meaning that this regression model explains 1.7% of the variation in reading achievement for 9th grade students on national tests.

Although we could not compare the coefficients for RAE directly with each other, it can be noted that the R-squared values systematically decrease over grade years. This finding can be interpreted as RAE having a decreasing effect on national test achievement scores as students proceeds through the grade years in the Norwegian compulsory school system. These findings were expected as literature suggests that the impact of RAE diminishes as students gets older (Martin, Mullis, & Foy, 2011; Olsen & Björnsson, 2018). Interestingly, we found that the R-squared estimates are more consistent in reading, compared to the other tests. The statistically significant results for RAE can further be interpreted as; the earlier students are born in the year - the better they perform on average. We found no evidence for gender differences in RAE, with an exception for a small interaction effect between RAE and gender in numeracy grade 5, which shows that RAE has a somewhat larger impact on girls than boys. Further, we found that boys have larger standard deviations than girls in all subjects and grade years, but the larger spread among boys cannot be attributed to a larger RAE for boys than girls. The reason for this is because there is an absence of significant interaction effects of gender and RAE in almost all test formats. See table 2.3 in appendix 3 for more details on descriptive statistics for gender-specific subsets of the data.

Results from regression discontinuity analyses

We conducted several tests to check for compliance with the four standards of RD-analyses (Schochet, et al., 2010). These standards were presented in section about RD-design in the method section. Table 3.3 in appendix 3 provides an overview of the results which consists of different varieties of the RD-model. These different varieties were used to test whether the present study meets the standards of RD-analyses. Based on these preliminary analyses we can claim that the present study meets the standards for RD-analyses, set forward by Schochet and colleagues (2010). Therefore we conclude that it is reasonable to apply a RD-analysis to model the effect of attending school for one additional year on achievement scores.

Table 3. Regression discontinuity coefficients for reading, tabulated by test scores and grade year. Significant results are bolded.

Variable	**Reading (8/9th grade)**	**Numeracy (8/9th grade)**
(Intercept)	**51.317 (0.063)*** **	**51.218(0.064)*** **
Birth month	**0.219 (0.008)*** **	**0.206 (0.008)*** **
Grade	**0.793 (0.115)*** **	**1.163 (0.116)*** **
R-squared	0.035	0.038
Bandwidth	[-11.5 : 11.5]	[-11.5 : 11.5]
N	117,845	118,051

*Significance codes: ***= p<.001,**= p.<.01 , * = p<.05*

Table 3 shows the output from the main regression discontinuity analyses. In numeracy, we found that the effect of a whole year difference in age, on average, is 2.47 points on achievement scores. The grade effect showed that the difference between the youngest student in grade 9 and the oldest student in grade 8 is about 1.16 points. In reading, we found that the effect of a whole year difference in age, on average is 2.62 points on achievement scores. The grade effect showed that the difference between the youngest student in grade 9 and the oldest student in grade 8 is about 0.793 points.

The r-squared statistic showed statistically significant results for numeracy, meaning that this regression discontinuity model accounts for 3.5% of the variation in reading achievement for 8th and 9th grade students on national tests. The r-squared statistic also showed statistically significant results in reading, meaning that this regression discontinuity model accounts for 3.8% of the variation in reading achievement for 8th and 9th grade students on national tests.

An inspection of figure 4 and 5 reveals that we found that the effect of a whole year difference in age has a larger effect than the grade effect on national test scores in numeracy and reading**.** Although we found clear evidence for RAE in the RD-results it is important to note that the R-squared values are low. Hence, we need to acknowledge that relative age and grade is indeed explaining some variance in reading and numeracy achievement, but its overall impact is not considerably strong.

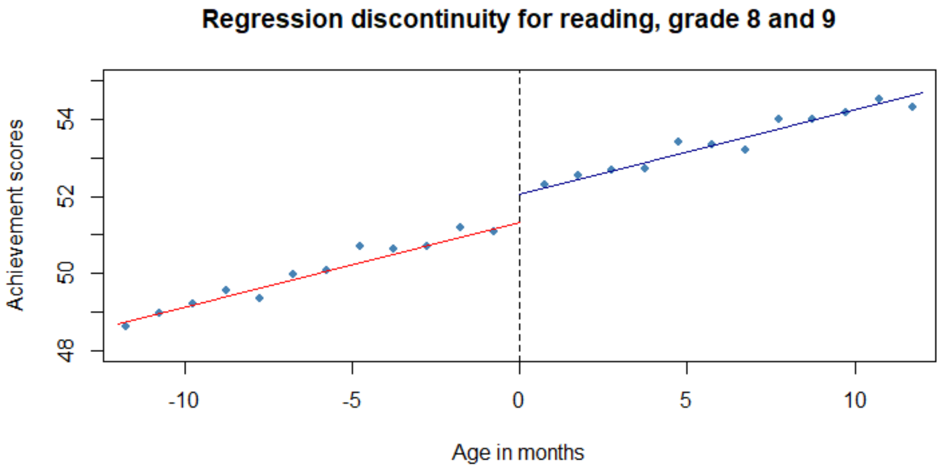


Figure 4. *Regression discontinuity estimate of achievement scores in reading for 8th and 9th grade students, per birth month. Students are separated by their birth month on the x-axis. Months are reverse scored. The red line represents the fitted values for each birth month in grade 8, the blue line represents the fitted values for each birth month in grade 9. The dots represents the observed scores. The dashed line represents the cutoff-point between January grade 8 and December grade 9.*

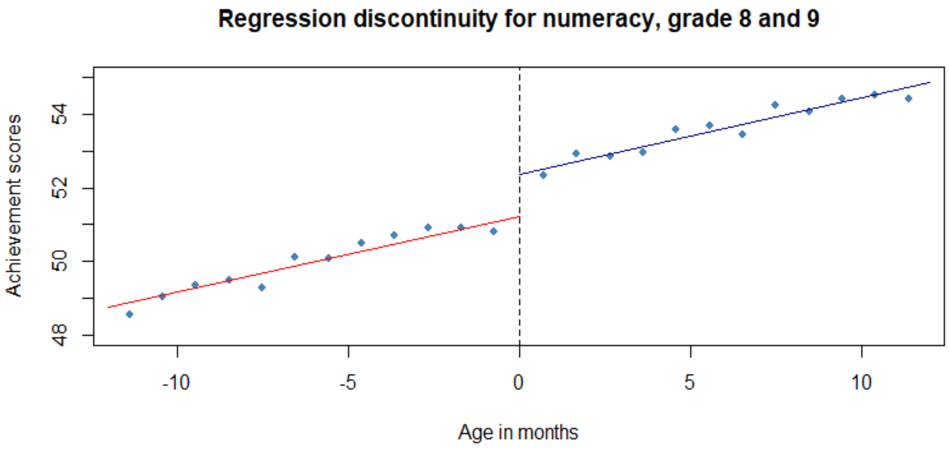


Figure 5. *Regression discontinuity estimate of achievement scores in numeracy for 8th and 9th grade students, per birth month. Students are separated by their birth month on the x-axis. Birth months are reversed. The red line represents the fitted values for each birth month in grade 8, the blue line represents the fitted values for each birth month in grade 9. The dots represents the observed scores. The dashed line represents the cutoff-point between January grade 8 and December grade 9.*

Change in RAE over grade years (numeracy)

Before we could investigate the change in RAE over grade years in numeracy, we linearly transformed the numeracy scores according to the vertical linking technique by Ræder, Tokle and Olsen (2019). The end-result is that the scores from grade 8 and 9 could be placed on the 5 grade numeracy scales. The mean scores and standard deviations for each of the grades are presented in table 4.

Table 4. Mean scores and standard deviations off vertical linked numeracy scores on grade 5 numeracy scale.

Grade	**Mean (standard deviation)**
Grade 5	49.97 (9.60)
Grade 8	61.79 (11.80)
Grade 9	66.24 (12.18)

Further, we conducted separate OLS regression analyses of the vertically linked numeracy scores, in order to test if the regression slopes for RAE are significantly different from each other. These models are needed to obtain the regression slopes of interest for comparison. We compared the regression slopes using independent samples t-tests. This is how we investigated how RAE changes over grade years. The results for OLS regression analyses of vertically linked scales scores are found in table 5. Further, the results of the independent t-tests are found in table 6.

Table 5. OLS regression estimates of RAE on vertically linked numeracy scores. Scores for grade 8 and 9 are now placed on the same scale as the national numeracy test for grade 5.

Coefficients	**Grade 5**	**Grade 8**	**Grade 9**
(Intercept)	**49.089(0.80)*****	**60.165(0.100)*****	**64.797(0.103)*****
Birth month	**0.309(0.011)*****	**0.267(0.014)*****	**0.238 (0.015)*****
R-squared	**0.012*****	**0.006*****	**0.004*****

Significance codes: ***= p<.001,**= p.<.01 , * = p<.05

Table 6 shows the results of three independent t-tests for two samples of regression slopes of RAE in each analysis, by using the results from table 5. Overall, the results of this analysis suggests that RAE changes significantly from grade 5 to grade 8 and 9. Furthermore, we found that there is no significant difference in RAE from grade 8 to 9, respectively. We found in table 4 that the standard deviations increase over grade years. This means that the spread in numeracy scores in grade 8 and 9 is larger than in grade 5. However, when looking at the R-squared estimates in table 5, we found that the amount variance accounted for in these models decline across grade years. This means that although the spread is larger in the higher grade years, this variation has less to do with the effect of relative age differences.

Figure 6 shows a graphical representation of how the linear effect of relative age changes over grade years, using vertically linked numeracy scale scores. Further, a graphical inspection of this figure suggests that RAE has a considerably linear effect across grade years on national numeracy tests.

Table 6. Two-sample independent t-tests for regression slopes and its standard errors of RAE in numeracy, grade 5,8 and 9

	Model 1	**Model 2**	**Model 3**
Mean difference	0.042 (0.02)	0.071 (0.019)	0.029 (0.021)
95% Confidence interval [Lower bound:Upper bound]	[0.007:0.076]	[0.031:0.107]	[-0.011:0.069]
T-value	2.365	3.833	1.413
Degrees of freedom	119,834	119,543	118,049

Model 1 = Grade 5 and grade 8; Model 2 = Grade 5 and grade 9; Model 3 = Grade 8 and grade 9

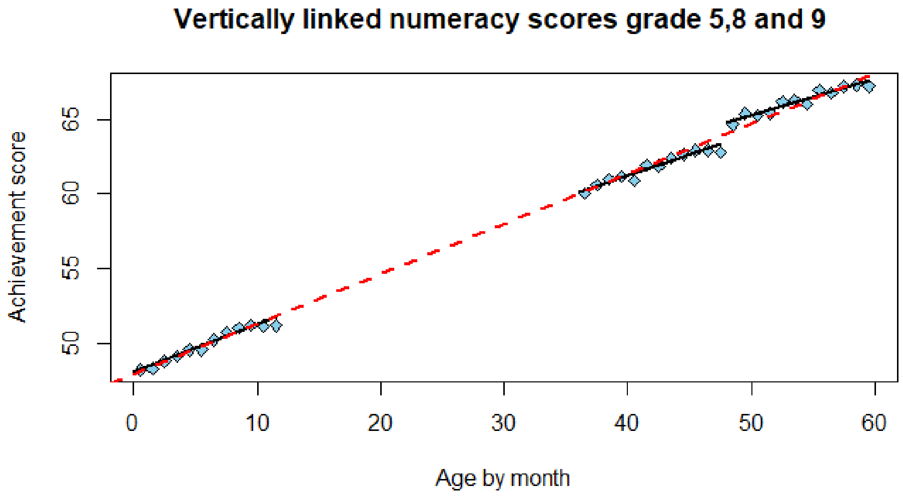


Figure 6. *Vertically linked scores of mean achievement scores per birth month across grade 5,8 and 9. Regression lines represent the various regression slopes for the analyses conducted in table 6. The dashed regression lines represents the predicted values for an overall regression model which includes all grade years. Students are separated by their birth month on the x-axis. Months are reverse scored (i.e from youngest to oldest).*

Discussion

This study presents findings regarding the effects of relative age and grade on Norwegian grade 5,8 and 9 students on national tests in reading, numeracy and English. When using results from these tests, we confirm findings from previous literature that RAE impacts students' school outcomes (Aune, Ingvaldsen, Vestheim, Bjerkeset, & Dalen, 2018; Martin, Mullis, & Foy, 2011; Olsen & Björnsson, 2018; Vestheim, Husby, Aune, Bjerkeset, & Dalen, 2019). In the discussion section proceeds through each research question, that has been studied in the present paper, to highlight the main findings and limitations of the present study. The main findings are followed by implications, recommendations for further studies and conclusions.

Research question 1

"What is the linear effect of relative age on students' national test performance, across the various subjects?"

For the first research question we expected to find a linear effect of relative age on students national test performance, across the various subjects. Our analyses confirmed that there is a positive linear effect of relative age on national test performance. In general, the findings suggests that the older the students are at the time of testing the better they perform, on average.

With regards to sub-question 1 *("What are the differences in the effect of students' relative age on national test performance across the various subjects and grade years?")*, we found that the estimates of RAE is largest in grade 5, and that RAE has a decreasing impact on performance in higher grade years. Specifically, we found that the average difference between the youngest and oldest student in grade 5 is approximately 3.5 points across subjects. Considering grade 9, we found that the average difference between the youngest and oldest is approximately 2.2 points across subjects. These findings are also in line with the literature which suggests that the effect of relative age impacts younger students more strongly than older students (Bedard & Duhey, 2006). We found no substantial differences in RAE across subjects, which suggests that RAE is a stable phenomenon across the various subjects in national tests. It seems reasonable to assume that the effect of students relative age affects their performance in a quite consistent manner, regardless of subject. However, it is worth mentioning that the estimates seems to be more stable in reading than in English and numeracy. This finding is also in line with more consistent r-squared estimates for the former test format, which could suggest that relative age has a more consistent impact on tests related specifically to reading skills.

With regards to sub-question 2 (*"What are the gender differences in the effect of students' relative age on national test performance across the various subjects and grade years?"*), we found marginal evidence for gender differences in RAE in numeracy grade 5. This indicates that girls have a marginally larger RAE in numeracy than boys. Apart from that finding, we found no further evidence to suggest that RAE is affected by gender. This adheres to the absence of significant interaction effects of RAE and gender on Norwegian PISA and TIMSS results over the last twenty years (Olsen & Björnsson, 2018). Gender differences in RAE could have been expected because boys have larger intra-sex differences in cognitive and biological mechanisms associated with maturity (Lehre, Lehre, Laake, & Danbolt, 2009). Recently a group named 'Stoltenbergutvalget' was selected by the Norwegian Ministry of Education and Research to investigate gender differences on school performances in Norwegian schools. In their report, they suggest that a larger proportion of the most immature children (particularly the youngest boys) would benefit from flexible school start, meaning they (especially the youngest boys) could take advantage from starting school later (Stoltenbergutvalget, 2019). We have not studied the possible effect of a more flexible school start for immature children. Accordingly, we do not draw any conclusions regarding school starting age policies. However, our results demonstrates that in grade 5 there is no observable difference in the effect of relative age between boys and girls apart from in numeracy which is small, but nevertheless statistically significant. If the hypothesis regarding maturation differences should hold, this should be evident as significant gender differences in RAE for the youngest students on school performances (Olsen & Björnsson, 2018). However, our findings suggests that we cannot support this hypothesis.

Research question 2

"What is the ratio of RAE over the grade effect in grade 8 and 9 in numeracy and reading?"

For the second research question, we aimed to investigate the ratio of RAE over the grade effect on grade 8 and 9 students in numeracy and reading achievement. When we compare the mean achievement scores of two adjacent grade years such as grade 8 and 9, we need to be mindful of how to interpret the difference between these mean scores. The mean difference between grade 8 and 9 represents the difference between students that are one year older, and has attended school for one additional year. By applying a RD-design we decompose the effect of relative age from the grade effect. Therefore we can estimate the extent to which the mean difference between grade 8 and 9 can be attributed to having attended school for one additional year, and to being one year older.

In the present study, we found evidence of a larger within-grade variation (i.e estimates of RAE x 12) than between-grade variation (i.e estimates of the grade effect) in reading and numeracy achievement. This finding is interesting as previous studies usually suggest that the opposite is the case. Grade usually has a stronger impact than age on school outcomes and intelligence tests (Black, Devereux, & Salvanes, 2008; Cliffordson & Gustafsson, 2008; Kyriakides & Luyten, 2009; Olsen & Björnsson, 2018). However, there are studies which support our findings. Using data from English primary schools, Luyten, Merrell and Tymms (2017) found that the grade effect declines in cognitive tests and school achievement as students progress from grade 1 to 6. Similar results have been found in Norwegian PIRLS data (Martin, Mullis, & Foy, 2011), where the effect of a whole year difference in age is larger than the grade effect between grade 4 and 5 students.

With regards to sub-question 1 (*"What is the ratio of RAE over the grade effect in grade 8 and 9 in numeracy and reading, for each gender separately?"*), we investigated whether the ratio of RAE over the grade effect differs among genders. At the same time, this sub-question investigated the extent to which the present study meets the one of the four standards for RD-design. We tested for gender differences in the RD-design to ensure the integrity of the birth month variable (i.e forcing variable) (Schochet et al., 2010). The effect of a full year difference in age was larger than the grade effect for both genders. Although we found that boys have larger standard deviations in all subjects and grade years than girls, this does not mean that RAE is larger for boys than girls. The spread is due to other factors that is beyond the scope of this study to investigate. Ultimately, we found no gender differences in RAE and the grade effect across national reading and numeracy tests in grade 8/9. We can conclude that the present study was successfully able to utilize a RD-design for estimating the effects of age and added grade years on national test achievement.

Research question 3

"How does RAE change across grade 5 to grade 8 and 9 in numeracy?"

One limitation in the results of the first research question is that although we find that RAE declines in higher grade years, this finding is based on analyses of scales that do not have a common baseline for direct comparisons. This study investigated RAE across three different subjects where each subject has two different scales – one for grade 5 and one for grade 8/9. In order to directly test if the impact of RAE declines in higher grade years, then two or more grade years needs to be compared on a mutual scale. This limitation was the motivation to address the third research question.

With regards to research question 3 we applied the vertical linking design developed by Ræder and Olsen (2020) for the national numeracy scales, to test how RAE changes from grade 5 to 8/9. Our results confirmed that RAE changes significantly from grade 5 to 8/9 in numeracy. Further we found that RAE does not change significantly from grade 8 to 9, suggesting that the effect of relative age has a similar impact on achievement scores in these grade years. The application of vertical linking is one of the more unique contributions to the existing literature on RAE and grade effects on school achievement. However, this approach to studying RAE has previously been applied by Luyten, Merrell and Tymms (2017), who used RD-design with multiple cut-off points and vertically equated scale scores to investigate the learning gains for English students in grade 1 to 6. However, this is the first study to apply vertical linking to investigate RAE on school achivement, in a Norwegian school context.

Another interesting remark on the results of the vertical linking is that the standard deviations are larger for grade 8 and 9. This means that the spread in numeracy scores in grade 8/9 are larger than in grade 5. On the one hand the increasing standard deviations in higher grade years suggest that the difference between high and low-performers in numeracy are increasing in higher grade years. On the other hand, this increasing difference in numeracy achievement scores has less to do with the impact of RAE due to decreasing estimates. The latter statement is further supported by decreasing r-squared estimates of the regression models in grade 8/9. Therefore, we found that RAE accounts for decreasing amounts of variation on numeracy achivement later in school.

Limitations of the present study

The limitations concerning the present study is firstly the issues regarding the small r-squared coefficients for all regression models that were applied. From the OLS regression analyses, the amount of variance explained for the OLS regression models ranges from 2.9 to 0.6 percent. The r-squared estimates decrease with grade. These findings are consistent with previous studies suggesting that RAE declines as individuals get older (Bedard & Duhey, 2006; González-Vallinas, Librero, Peiró, & San Fabián, 2019). Although the impact of RAE is smaller by the end of compulsory school in Norway, the effect is still not negligible. This might support an assumption that the oldest students still have a slight advantage when they enter upper secondary school.

Considering the RD-analyses, the amount of variance explained was 3.8 % in numeracy and 3.5% in reading. An explanation for why the amount variance explained is larger in the RD-analyses is that these analyses included grade as a predictor variable. The grade effect, has in some cases, shown to be a stronger predictor of school achievement than age (Cahan & Cohen, 1989; Cliffordsson, 2010; Cliffordsson & Gustafsson, 2008). When comparing the amount of variation explained in the present study to similar studies, we found various results. For example, Olsen and Björnsson (2018) studied RAE and the grade effect on Norwegian TIMSS data from the last 20 years. The amount of variance explained in the RD-models in their studies accounted for 4% in grade 8/9, these findings is consistent with the estimated r-squares in the present study. On the other hand, Kyriakides and Luyten (2009) investigated the effect of schooling on cognitive development and curriculum-based tests in language and mathematics on a sample of Cypriot students, using a RD-design. The amount of variance explained in their study ranged from 13.9 to 33.0 %. Evidently, the amount of variance increases if other variables such as various cognitive functions are included in the RD-models. The study from Cyprus is therefore not directly comparable, because we have no data on cognitive functions in the present study. Nevertheless the finding demonstrates that age and grade effects does not account for a large amount of variation in students achievement scores. Ultimately, none of the models in the present study are providing substantial explanations to why students' achievement scores differ on national tests. However, given the sample sizes in the present study, the standard errors are low for all RAE-estimates. This means that we are finding small but robust estimates of the relationship between birth month and achievement scores. Considering the robustness of the estimates in the present study, we can not claim that RAE, gender differences in RAE or grade effects are among the strongest predictors of achievement scores in national tests.

The present study is limited by a data set that contains few variables which would account for larger amounts of variance in students achievement scores. There are many other factors that could, in accordance with findings on RAE, explain why students differ on achievement scores in school. It has previously been shown that factors such as SES and other family background variables (e.g., parents' education level) explains larger amounts of variation in school performance. Interestingly, it is possible that relatively young students with high SES-background might be less affected by RAE than relatively young students with low SES-background. Parents with higher education tends to intervene faster when their children's performance in school drops. Hence, it is possible that students with high SES can compensate for the implications of being born late in the year if they receive additional educational support in the home (Buckles & Hungerman, 2013; Crawford, Dearden, & Greaves, 2011; Currie, 2009; Solli, 2017).

Implications

RAE has implications for students' school performance. Therefore, it is important that teachers and school leaders are aware of the differences in students' maturation rate. This in turn is, to some extent, related to students' relative age. There are implications regarding the findings on RAE from the sports literature which is applicable to the literature on RAE and school performance. Considering these findings, it seems evident that individuals born close to the annual cut-off point in annual age-grouped sports are disadvantaged, simply because of the time in the year they are born. A lot of potential talent might be overlooked. The relatively younger athletes needs more time to realize their potential and to catch up with the performances of the older peers in their cohort.

Transferred to the school context. The findings from the sports literature suggests that if school systems applied mechanisms where students are selected into different tracks based on performance at an early age, this would lead to systematic inequity. The reason is that children's birth date would then partly determine their future educational pathway. Even if systematic selections into different educational tracks does not exist, such as in Norway, RAE may work indirectly through the expectations of teachers (May, Kundert, & Brent, 1995; Rubie-Davies, 2006; Rosenthal & Jacobson, 1968; Sykes, Bell, & Roderio, 2009). In sports, it has been found that coaches perceive the oldest athletes in the youth levels as stronger, faster and more capable of high performance than the youngest athletes, due to advanced physical maturation (Cobley, Baker, Wattie, & McKenna, 2009). Similarities might occur in the classroom where the youngest students are usually less mature and is, on average, performing relatively poorer on curriculum-based tests.

Further studies

As mentioned in the limitations of the present study, the results of this study is limited in the sense that it lacks other predictor variables which would make the regression models account for more variance in achievment scores. In addition to the aformentioned proposal for including indicators of SES, it would be relevant to consider inclusion of a variable which identifies students with specific and diagnosed learning disorders. One example of a disorder that causes strong learning difficulties is ADHD. In countries where ADHD and learning disorders are frequently being identified in screening assessments of students, a disproportionate amount of these students are born late in the year (Chen, et al., 2016; Disanto, et al., 2012; Fuller, Rawlings, Ennis, Merrill, & Flores, 1996; Morrow, et al., 2012; Rihmer, et al., 2011; Tochigi, Okazaki, Kato, & Sasaki, 2004). It would be of interest for further studies to investigate the probabilities of students born in the 4th quarter of the calendar year in mastery level 1 with being diagnosed with learning disorders. This would be an interesting new approach to study the impact of RAE, in the most severe settings. In addition, it would provide school leaders and teachers with new insights on how to interpret the potential implications of classifying students performance on national tests at mastery level 1. Furthermore, it would also be interesting to investigate whether RAE would contribute more substantially to test formats with higher stakes (e.g. exam scores) and tests which are more curriculum-specific than the national tests. The measures used in national tests are not directly related to curriculum-specific skills, rather they represent skills that are developed over time. These types of studies would also benefit from proper longitudinal studies which would allow for measuring RAE on the same students over time, and across different scales that are possible to link together.

Conclusion

It is difficult to propose clear actions for addressing the impact of RAE on the youngest students in classrooms. RAE may not be the strongest long-term predictor of educational success, but there is enough evidence present to suggest that it is indeed a contributing factor. In this study, we find clear evidence of RAE and the grade effect on national tests. Therefore school leaders and teachers needs to be sensitive to larger impact of RAE in the lower grade years of formal schooling. The youngest students are not less capable, but they need more time to mature in order to perform at the same level as their older peers.

References

Aune, T. K., Ingvaldsen, R. P., Vestheim, O. P., Bjerkeset, O., & Dalen, T. (2018). Relative age effects and gender differences in the national test of numeracy: a population study of Norwegian children. *Frontiers in psychology, article: 1091*, ss. 1-8. doi:10.3389/fpsyg.2018.01091

Baker, J., & Horton, S. (2004). A review of primary and secondary influences on sport expertise. *High ability studies, 15*, ss. 211-228. doi:10.1080/1359813042000314781

Balke-Aurell, G. (1982). *Changes in ability as related to educational and occupational experience.* Gothenburg: Acta Universitatis Gothoburgensis. Obtained from https://core.ac.uk/download/pdf/16318045.pdf

Barnsley, R. H., & Thompson, A. H. (1988). Birthdate and success in minor hockey: The key to the NHL. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 20*, ss. 167-176. Obtained from https://www.researchgate.net/profile/Roger_Barnsley/publication/232490968_Birthdate_and_success_in_minor_hockey_The_key_to_the_NHL/links/5806fcbd08aeb85ac85f5cb5.pdf

Bedard, K., & Dhuey, E. (2012). School-entry policies and skill accumulation across directly and indirectly affected individuals. *Journal of Human Resources, 47*, ss. 643-683. Obtained https://www.researchgate.net/profile/Kelly_Bedard/publication/241768569_School-Entry_Policies_and_Skill_Accumulation_Across_Directly_and_Indirectly_Affected_Individuals/links/577289f308aeef01a0b6577a/School-Entry-Policies-and-Skill-Accumulation-Across-Dir

Bedard, K., & Duhey, E. (2006). The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects. *The Quarterly Journal of Economics*, *121*, ss. 1437–1472. doi:10.1093/qje/121.4.1437

Björnsson, J. K. (2016). *Metodegrunnlag for nasjonale prøver.* Oslo: Utdanningsdirektoratet. Obtained from https://www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-for-nasjonale-prover-august-2018.pdf

Björnsson, J. K. (2018). Om lenkefeil og ekvivaleringsmetoder på nasjonale prøver: Evaluering av endring over tid. *Acta Didactica Norge*, ss. 1-24. doi:10.5617/adno.6273

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2008). Too young to leave the nest? The effects of school starting age. *The Review of Economics and Statistics, 93*, ss. 455-467. doi:10.1162/REST_a_00081

Bruce, P., & Bruce, A. (2017). *Practical statistics for data scientists: 50 essential concepts.* Sebastopol: O'Reilly Media, Inc.

Buckles, K. S., & Hungerman, D. M. (2013). Season of Birth and Later Outcomes: Old Questions, New Answers. *Review of Economics and Statistics, 95*, ss. 711-724. doi:10.1162/REST_a_00314

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

Carlsson, M., Dahl, G. B., Öckert, B., & Rooth, D.-O. (2014). The Effect of Schooling on Cognitive Skills. *The Review of Economics and Statistics, 97*, ss. 533-547. doi:10.1162/REST_a_00501

Chen, M.-H., Lan, W.-H., Bai, Y.-M., Huang, K.-L., Su, T.-P., Tsai, S.-J., . . . Hsu, J.-W. (2016). Influence of Relative Age on Diagnosis and Treatment of Attention-Deficit Hyperactivity Disorder in Taiwanese Children. *The Journal of Pediatrics, 172*, ss. 162-167. doi:10.1016/j.jpeds.2016.02.012

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

Cliffordson, C., & 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*, ss. 143-152. doi:10.1016/j.intell.2007.03.006

Cobley, S., Baker, J., Wattie, N., & McKenna, J. (2009). Annual age-grouping and athlete development. *Sports medicine, 39*, ss. 235-256. doi:10.2165/00007256-200939030-00005

Cools, S., Schøne, P., & Strøm, M. (2017). Forskyvninger i skolestart: Hvilken rolle spiller kjønnog sosial bakgrunn? *Søkelys på arbeidslivet,34*, ss. 273-289. Obtained from https://www.idunn.no/spa/2017/04/forskyvninger_i_skolestart_hvilken_rolle_spiller_kjoenn_og_

Côté, J., Baker, J., & Abernethy, B. (2007). Practice and play in the development of sport expertise. I R. Eklund, & G. Tenenbaum, *Handbook of sport psychology, 3* (ss. 184-202). Hoboken : Wiley.

Crawford, C., Dearden, L., & Greaves, E. (2011). *Does when you are born matter? The impact of month of birth on children's cognitive and non-cognitive skills in England.* Obtained from https://www.ifs.org.uk/bns/bn122.pdf

Crawford, C., Dearden, L., & Greaves, E. (2013). *When you are born matters: evidence for England.* London: IFS report R80. Obtained from https://dera.ioe.ac.uk/33086/1/r80.pdf

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *JOURNAL OF ECONOMIC LITERATURE, 47* , ss. 87-122. doi:10.1257/jel.47.1.87

Dee, T. S., & Sievertsen, H. H. (2018). The gift of time? School starting age and mental health. *Health economics, 27,* ss. 781-802. doi:10.1002/hec.3638

Dhuey, E., Figlio, D., Karbownik, K., & Roth, J. (2019). School Starting Age and Cognitive Development. *Journal of Policy Analysis and Management, 38,* ss. 538-578. doi:10.1002/pam.22135

Directory for Education and Training. (2017). Rammeverk for nasjonale prøver [The framework for national tests]. Oslo. Obtained from https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover/

Directory of Education. (2020). Skoleporten [The School Portal]. Oslo. Obtained from https://skoleporten.udir.no/rapportvisning/grunnskole/laeringsresultater/nasjonale-proever-fritatt-og-ikke-deltatt/nasjonalt?periode=2018-2019&orgaggr=a&kjonn=a&sammenstilling=1&fordeling=4

Disanto, G., Morahan, J. M., Lacey, M. V., DeLuca, G. C., Giovannoni, G., Ebers, G. C., & Ramagopalan, S. V. (2012). Seasonal Distribution of Psychiatric Births in England. *PLoS ONE, 7*. doi:10.1371/journal.pone.0034866

Dobkin, C., & Ferreira, F. (2010). Do school entry laws affect educational attainment and labor market outcomes? *Economics of education review, 29*, ss. 40-54. doi:10.1016/j.econedurev.2009.04.003

Elstad, E. (2009). Schools which are named, shamed and blamed by the media: school accountability in Norway. *Educational Assessment, Evaluation and Accountability, 21,* ss. 173-189. doi:10.1007/s11092-009-9076-0

Fuller, T. E., Rawlings, R. R., Ennis, J. M., Merrill, D. D., & Flores, D. S. (1996). Birth seasonality in bipolar disorder, schizophrenia, schizoaffective disorder and stillbirths. *Schizophrenia Research, 21*, ss. 141-149. doi:10.1016/0920-9964(96)00022-9

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

González-Vallinas, P., Librero, J., Peiró, S., & San Fabián, J. L. (2019). Relative age impact on language and mathematics school achievement in primary education in Asturias county. *Revista de Educación, 386*, ss. 165-186. Obtained from https://www.researchgate.net/profile/Jl_Maroto/publication/336346230_Relative_age_impact_on_language_and_mathematics_school_achievement_in_primary_education_in_Asturias_county_Paula_Gonzalez-Vallinas/links/5d9ca098299bf1c36301dd7f/Relative-age-impact-on-l

Helsen, W. F., van Winckel, J., & Williams, A. M. (2005). The relative age effect in youth soccer across Europe. *Journal of Sports Sciences, 23*, ss. 629-636. doi:10.1080/02640410400021310

Hovdhaugen, E. (2016). National Tests in Norway: An Undeclared Standard in Education? Practical and Political Implications of Norm-Referenced Standards. I S. Blömeke, & J.-E. Gustafsson, *Standard Setting in Education: The Nordic countries in an international perspective* (ss. 161-179). Cham: Springer.

Jeronimus, B. F., Stavrakakis, N., Veenstra, R., & Oldehinkel, A. J. (2015). Relative Age Effects in Dutch Adolescents: Concurrent and Prospective Analyses. *PloS one, 10*, s. e0128856. doi:10.1371/journal.pone.0128856

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3. utg.). New York: Springer-Verlag. doi:10.1007/978-1-4939-0317-7

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

Larsen, E. R., & Solli, I. F. (2016). Born to run behind? Persisting birth month effects on earnings. *Labour Economics,*, ss. 200-210. doi:10.1016/j.labeco.2016.10.005

Lehre, A.-C., Lehre, K.-P., Laake, P., & Danbolt, N. C. (2009). Greater intrasex phenotype variability in males than in females is a fundamental aspect of the gender differences in humans. *Developmental Psychobiology, 51*, ss. 198–206. doi:10.1002/dev.20358

Lund, T., & Thrane, V. C. (1983). Schooling and intelligence: A Methodological and logitudinal study. *Scandinavian Journal of Psychology, 24*, ss. 161-173. doi:10.1111/j.1467-9450.1983.tb00489.x

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

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*, ss. 374-405. doi:10.1080/09243453.2017.1297312

Martin, M. O., Mullis, I. V., & Foy, P. (2011). Age distribution and reading achievement configurations among fourth-grade students in PIRLS 2006. *ERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4*, ss. 9-33.

May, D. C., Kundert, D. K., & Brent, D. (1995). Does Delayed School Entry Reduce Later Grade Retentions and Use of Special Education Services? *Remedial and Special Education, 16*, ss. 288-294. doi:10.1177/074193259501600505

Morrow, R. L., Garland, E. J., Wright, J. M., Maclure, M., Taylor, S., & Dormuth, C. R. (2012). Influence of relative age on diagnosis and treatment of attention-deficit/hyperactivity disorder in children. *CMAJ*, ss. 755-762. doi:10.1503/cmaj.111619

Musch, J., & Grondin, S. (2001). Unequal competition as an impediment to personal development: A review of the relative age effect in sport. *Developmental review, 21*, ss. 147-167. doi:10.1006/drev.2000.0516

OECD. (2018). *OECD - Library*. Obtained from https://www.oecd-ilibrary.org/education/education-at-a-glance-2018/starting-and-ending-age-for-students-in-compulsory-education-and-starting-age-for-students-in-primary-education-2016_eag-2018-table221-en

Olsen, R. V., & Björnsson, J. (2018). Fødselsmåned og skoleprestasjoner. . I &. R. J. Björnsson, *Tjue år med PISA og TIMSS i Norge: Trender og nye analyser* (ss. 76-93). Oslo: Universitetsforlaget. doi:10.18261/9788215030067-2018-05

Pedraja-Chaparro, F., Santín, D., & Simancas, R. (2015). Determinants of grade retention in France and Spain: Does birth month matter? *Journal of Policy Modeling, 37,* ss. 820-834. doi:10.1016/j.jpolmod.2015.04.004

Ponzo, M., & Scoppa, V. (2014). The long-lasting effects of school entry age: Evidence from Italian students. *Journal of Policy Modeling, 36,* ss. 578-599. doi:10.1016/j.jpolmod.2014.04.001

Puhani, P., & Weber, A. (2008). Does the early bird catch the worm? I F. B. Dustmann C., *The Economics of Education and Training. Studies in Empirical Economics* (ss. 105-132). Heidelberg: Physica-Verlag HD. doi:10.1007/978-3-7908-2022-5_6

Rihmer, Z., Erdos, P., Ormos, M., Fountoukalis, K., Vazquez, G., Pompili, M., & Gonda, X. (2011). Association between affective temperaments and season of birth in a general student population. *Journal of Affective Disorders, 132,* ss. 64-70. doi:10.1016/j.jad.2011.01.015

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual attainment* . New York: Holt, Rinehart & Winston.

Rubie-Davies, C. M. (2006). Teacher expectations and student self-perceptions: Exploring relationships. *Psychology in the Schools, 43,* ss. 537-552. doi:10.1002/pits.20169

Rubie-Davies, C. M., Flint, A., & McDonald, L. G. (2012). Teacher beliefs, teacher characteristics, and school contextual factors: What are the relationships? *British journal of educational psychology, 82,* ss. 270-288. doi:10.1111/j.2044-8279.2011.02025.x

Ræder, H. G., & Olsen, R. V. (2020). *Utvikling av Nasjonale Prøver Rapport 2b.* Oslo: Centre for Educational Measurement.

Ræder, H. G., Olsen, R. V., & Blömeke, S. (2020). Large-Scale Assessments in the Norwegian Context. I H. Harju-Luukkainen, N. McElvany, & J. Stang, *Monitoring Student Achievement in the 21st Century* (ss. 195-206). Cham: Springer.

Ræder, H. G., Tokle, O. D., & Olsen, R. V. (2019). *Utvikling av nasjonale prøver – rapport 1 og 2a Rapportering av potensielle underdimensjoner og implementering av vertikalt lenkedesign for de nasjonale prøvene i regning.* Oslo: Centre for Educational Measurement (CEMO).

Salib, E., & Cortina-Borja, M. (2006). Effect of month of birth on the risk of suicide. *The British Journal of Psychiatry, 188,* ss. 416-422. doi:10.1192/bjp.bp.105.009118

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J., Porter, J., & Smith, J. (2010). Standards for Regression Discontinuity Designs. . *What Works Clearinghouse*, ss. 1-8. Obtained from http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf

Schorer, J., Baker, J., Lotz, S., & Büsch, D. (2010). Influence of early environmental constraints on achievement motivation in talented young handball players. *International journal of sport psychology, 41*, ss. 42-57. Obtained from https://www.researchgate.net/profile/Bradley_Young2/publication/236124336_Young_BW_Salmela_JH_2010_Examination_of_practice_activities_related_to_the_acquisition_of_elite_performance_in_Canadian_middle_distance_running_International_Journal_of_Sport_Psycho

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston : Houghton Miffin.

Sharp, C. (1995). What's age got to do with it? A study of patterns of school entry and the impact of season of birth on school attainments. *Educational research, 36*, ss. 251-265. doi:10.1080/0013188950370304

Sherar, L. B., Baxter-Jones, A. D., Faulkner, R. A., & Russell, K. W. (2007). Do physical maturity and birth date predict talent in male youth ice hockey players? *Journal of Sports Sciences, 25*, ss. 879-886. doi:10.1080/02640410600908001

Solli, I. (2017). Left behind by birth month. *Education Economics, 25(4)*, ss. 323-346. doi:10.1080/09645292.2017.1287881

Statistics Norway. (2017, September 26). Guttene havner bakpå [Boys are trailing behind]. Obtained from https://www.ssb.no/utdanning/artikler-og-publikasjoner/guttene-havner-bakpa

Statistics Norway. (2018). STATBANK - Marks, lower secondary school: 07496: Overall achievement marks, by subject, sex and parents' educational attainment level (C) 2009 - 2019. Obtained from https://www.ssb.no/en/statbank/table/07496/

Statistics Norway. (2019). 05531: Live births, by month 1966 - 2019. Oslo. Obtained from https://www.ssb.no/en/statbank/table/05531/

Stoltenbergutvalget. (2019). *Nye sjanser - Bedre læring: Kjønnsforskjeller i skoleprestasjoner og utdanningsløp.* . Oslo: Kunnskapsdepartementet. Obtained from https://nettsteder.regjeringen.no/stoltenbergutvalget/files/2019/02/nou201920190003000ddd pdfs

Sykes, E. D., Bell, J. F., & Roderio, C. V. (2009). *Birthdate effects a review of the literature from 1990-on - school starting age. Cambridge Assessment.* Cambridge Assessment. Obtained from statesassembly.gov.je: ttps://statesassembly.gov.je/scrutinyreviewresearches/2016/res

Thompson, A. H., Barnsley, R. H., & Dyck, R. J. (1999). A New Factor in Youth Suicide: The Relative Age Effect. *The Canadian Journal of Psychiatry, 44*, ss. 82-85. doi:10.1177/070674379904400111

Thompson, A. H., Barnsley, R. H., & Stebelsky, G. (1991). "Born to Play Ball" The Relative Age Effect and Major League Baseball. *Sociology of Sport Journal*, ss. 146-151. doi:10.1123/ssj.8.2.146

Tochigi, M., Okazaki, Y., Kato, N., & Sasaki, T. (2004). What causes seasonality of birth in schizophrenia? *Neuroscience Research, 48*, ss. 1-11. doi:10.1016/j.neures.2003.09.002

Tveit, S. (2014). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice, 21*, ss. 221-237. doi:10.1080/0969594X.2013.830079

Vestheim, O. P., Husby, M., Aune, T. K., Bjerkeset, O., & Dalen, T. (2019). A Population Study of Relative Age Effects on National Tests in Reading Literacy. *Frontiers in psychology, article: 1761*, ss. 1-7. doi:10.3389/fpsyg.2019.01761

Ward, P., & Williams, A. M. (2003). Perceptual and cognitive skill development in soccer: The multidimensional nature of expert performance. *Journal of sport and exercise psychology, 25*, ss. 93-111. doi:10.1123/jsep.25.1.93

Wattie, N., Schorer, J., & Baker, J. (2015). The Relative Age Effect in Sport: A Developmental Systems Model. *Sports Medicine, 45*, ss. 83-94. doi:10.1007/s40279-014-0248-9

Weinstein, R. S., Marshall, H. H., Sharp, L., & Botkin, M. (1987). Pygmalion and the student: Age and classroom differences in children's awareness of teacher expectations. *Child development, 58*, ss. 1079-1093. doi:10.2307/1130548

Wilson, G. (2000). The effects of season of birth, sex and cognitive abilities on the assessment of special educational needs. *Educational Psychology, 20*, ss. 153-166. doi:10.1080/713663714

Appendix 1

NSD Documents and ethical approval

NOTIFICATION FORM – NSD

Oscar Skovdahl Jørstad

Which personal data will be processed?

Response: Gender and birth month (without date or year, only month)

Name
Comment from NSD: First name and surname
Response: No

National ID number or other personal identification number
Comment from NSD: 11-digit personal identifier, D number, or other national identification number
Response: No

Date of birth
Response: No

Address or telephone number
Response: No

Email address, IP address or other online identifier
Comment from NSD: An email address is a unique address that is assigned to the user of an electronic mail service. An IP address is a unique address that is assigned to a device (e.g. a computer) in a computer network like the Internet. Dynamic IP addresses may also be considered personal data in certain cases. Cookies are an example of an online identifier. NB! If you are going use an online survey, and the service provider (data processor) will have access to email addresses or IP addresses, you must indicate this here.
Response: No

Photographs or video recordings of persons
Comment from NSD: Photographs and video recordings of faces are usually considered to be personal data
Response: No

Audio recordings of persons
Comment from NSD: Audio recordings where personal data are recorded and/or where there exists a scrambling key that links the audio recordings to individual persons on the recordings. The voice of the person speaking may be considered personal data in combination with other background information.
Response: No

GPS data or other geolocation data

Comment from NSD: Data which indicate the geographical location of a person

Response: No

Demographic data that can identify a natural person

Comment from NSD: E.g. a combination of information such as municipality of residence, workplace, position, age, gender etc.

Response: Yes (Gender)

Genetic data

Comment from NSD: Personal data relating to the inherited or acquired genetic characteristics of a natural person, which give unique information about the physiology or health of that person.

Response: No

Biometric data

Comment from NSD: E.g. fingerprint, handprint, facial form, retina and iris scan, voice recognition, DNA.

Response: No

Other data that can identify a natural person

If you think that you will be processing personal data but cannot find a suitable alternative above, indicate this here.

Response: No

Will special categories of personal data or personal data relating to criminal convictions and offences be processed?

Racial or ethnic origin

Comment from NSD: This includes belonging to an ethnic group, population, cultural sphere or society that has common characteristics. For example, information that a person is Sami is not considered to say anything about race but it says something about ethnicity.

Response: No

Political opinions

Comment from NSD: That a person is a member of a political party and/or what a person voted in an election, including political opinions and beliefs. However, this does not include information that a person is a conservative, radical or labour party supporter.

Response: No

Religious beliefs

Comment from NSD: That a person is a member of a religious organisation/congregation. This does not include information that a person has a subscription to a religious newspaper.

Response: No

Philosophical beliefs

Comment from NSD: That a person is a member of a philosophical association, or that a person believes that knowledge is acquired through logical speculation and observation.

Response: No

Trade Union Membership
Comment from NSD: That a person is a member of a trade union that organises employees within the same industry/subject area, e.g. LO, NTL, NAR etc.
Response: No

Health data
Comment from NSD: Personal data concerning a natural person's physical or mental health, including use of healthcare services.
Response: No

Sex life or sexual orientation
Comment from NSD: A person's sexual orientation (homosexual, lesbian, bisexual etc.) and/or sexual behaviour (e.g. that a personal has been unfaithful, indecent exposure, offensive gestures/language)
Response: No

Criminal convictions and offences
Comment from NSD: Personal data concerning convictions and offences, or related to security measures.
Response: No

Project Information

Edit project Register new project Chose existing project
under 'Register new project':

Title

Response: Relative age effect on outcomes in national assessments in Norway

Project description

Response: The purpose of the project is to investigate whether students birth month are able to predict the outcomes of their performance on the national assessments in Norway across 5th, 8th and 9th grade in the subjects Norwegian (reading), English (reading) and Mathematics (numeracy). We also want to investigate potential gender differences across the grade years and subjects. The idea is to investigate how large the effect on the outcome variable is influenced by students that are born later in the year compared to students born earlier in the year. The project will use a data set provided by the Norwegian Directory of Education and Training ("Utdanningsdirektoratet") which consists of the following variables:

1. Test year
2. Type of test and grade year (i.e., Mathematics – 5th grade, Reading, 9th grade)
3. Gender
4. Scale scores
5. Birth month

Subject area

- Other subject areas

Will the collected personal data be used for other purposes, in addition to the purpose of this project?

Response: The collected personal data will not be used for other purposes.

Comment from NSD: Personal data should only be processed for specified, explicit and legitimate purposes. This means that each purpose for processing personal data must be identified and described clearly and accurately. In order for a purpose to be considered legitimate, it must also be in accordance with ethical and legal norms.

Explain why it is necessary to process personal data.

Response: National assessments is one of main measures for quality assurance of the Norwegian curriculum and also serves as a foundation for contemporary evaluation of the current status of Norwegian students performance in the given subjects (Reading (Norwegian and English) and numeracy). Furthermore, Norway has a very strict policy for re-doing years in elementary school where less than 2 per cent of the population usually does so from year to year. This makes grade year and age almost perfectly coinciding which very ideal in the research field of relative age effects, because the variation in age is less than 12 months in each grade year. This is also the reason for using information about students birth month as a measure of differences within the age for each grade year. Gender differences is also of interest to investigate as the Ministry of Education in Norway has been looking into gender differences in terms of performance in school, especially in recent years.

Comment from NSD: Explain why the personal data are adequate, relevant and limited to what is necessary for the purposes for which they are being processed. This includes limiting the amount of collected data to that which is necessary to realise the purposes of data collection.

External funding

Response: This project has no external funding.

Type of project

- Student project, Master's thesis

Responsibility for data processing

Data controller

University of Oslo/CEMO – Centre for Educational Measurement

Project leader (research assistant/ supervisor or research fellow/phD candidate)

Name: Rolf Vegar Olsen
Position: Professor
Email address: r.v.olsen@cemo.uio.no
Telephone number: 22844510

Will the responsibility for processing personal data be shared with other institutions (joint data controllers)?

Response: Yes

Comment from NSD: If two or more institutions together decide the purposes for which personal data are processed, they are joint data controllers.

Joint data controllers

Institution: Norwegian Directory of Education and Training
Country: Norway
Postal address: Schweigaards gate 15 B
Email address: post@udir.no
Telephone number: 23301200

Whose personal data will be processed?

You must describe each group of people whose personal data you will be processing. Add and describe each sample individually.

Sample 1

Describe the sample

Response: All Norwegian students in 5th, 8th or 9th grade, i.e approximately 117,000 students.

Recruitment or selection of the sample

Comment from NSD: Describe how the sample will be recruited and how initial contact with the sample will be made. For example, whether you will make initial contact during field-work or via your own network, or whether a school, hospital or organisation will contact its pupils, patients or members on your behalf. If the sample will not be recruited but will be selected from a registry or an administrative system etc., describe how the selection will be carried out and what the selection criteria will be.

Response: Recruitment of the sample is done through the students that participated in the national assessments in the school year 2018-2019.

Age

Response: 9-15 years of age

Will you include adults (18 år +) who do not have the capacity to consent?

Comment from NSD: i.e. the person has reduced capacity or lacks capacity to consent. For example, the person may have mental/cognitive impairment, significant physical/emotional ailments, or may be unconscious, conditions which make it difficult or impossible for the person to gain sufficient understanding in order to give valid consent. The central aspect is whether the person is capable of understanding the purpose of the processing/project in question, and of understanding potential positive and negative consequences (immediate and long-term).

Response: No

Types of personal data - sample 1

- Other data that can identify a natural person

Methods /data sources - sample 1

Other

Description:

Response: The data set consists of scale scores from the various national assessments which is arranged by the

Norwegian Directory of Education and Training. Hence, the given application describes a collaboration project between UiO/CEMO and Norwegian Directory of Education and Training. It is the latter organization that is responsible for data collection. In contrast to the original data set gathered by Norwegian Directory of Education and Training, this data set does not contain any information about which school, birth month, name or municipality that the student is related to, which ensures anonymity in the provided data set I will be working with.

Legal basis for general categories of personal data

Response: A task in the public interest or in the exercise of official authority (art. 6 nr .1 e)

Explain your choice of legal basis

Response: This project is motivated by the purpose of the national assessments which is to improve the quality of the Norwegian curriculum and circumstances surrounding the performance on the tests. This study provides research on inter-individual differences in students' performance across the different subjects and grade years.

Information - sample 1

Will you inform the sample about processing their personal data?
Response: Yes

How?
Response: Written information (on paper or electronically)

Information letter:
Response: https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover/

Third persons

Comment from NSD: Will you be processing personal data about third persons? This includes data about persons who are not included in the sample/are not participating in the project; information provided by a data subject that relates to another identified or identifiable natural person. Examples of this are when a data subject is asked about their mother's and father's education or country of origin, or when pupils are asked about their teacher's teaching methods.
Response: No

Documentation

Total number of data subjects in the project
(Data subjects: persons whose personal data you will be processing)
Response: 100.000+

How can data subjects get access to their personal data or how they can have their personal data corrected or deleted?
Comment from NSD: Rights of data subjects (participants) include the right to access one's own personal data and to receive a copy of one's data if asked for. A data subject can request that their personal data are corrected if they feel that the information is wrong or lacking, and the data subject

can withdraw consent and request that their personal data are deleted. Give a short description of the procedure for how a data subject can get access to their personal data, and how they can have their personal data corrected or deleted.

Response: Information about the results of the national assessments are publicly available at udir.no but are anonymized in the sense that only school leaders and schools can access information at individual level. As mentioned earlier, in the provided data set I will be working with, the students are not able to be identified as the only information regarding each student is limited to gender and birth month.

Other approvals

Will you obtain any other approvals or permits for the project?
Response: No

Processing

Where will the personal data be processed?
"Processing" includes any collecting, registering, storing, collating, transferring etc. of data. You must indicate all processing of personal data that will take place in the project.

Response: Computer belonging to the institution responsible for the project
Comment from NSD: Computer owned/operated by the data controller. For example, processing data in a private or communal user area on the institution's server.
Response: Private device
Comment from NSD: Data collection or storage on private devices such as your own computer or mobile phone etc. is not recommended and must be clarified with the institution responsible for the project.
Data collection, storing or archiving on private devices such as your own computer, mobile phone, memory stick etc. is not recommended and must be clarified with the institution responsible for the project.

Who will be processing/have access to the collected personal data?

Response:

- Project leader
- Student (student project)
- Internal co-workers

Employees of the data controller
Which others will have access to the collected personal data?
Response: None.

Will the collected personal data be made available to a third party or international organisation outside the EEA?
Comment from NSD: This includes when personal data are sent to and stored in a country outside the EEA, or when persons outside this area are given access to personal data stored within the EEA. This means that you cannot use a service provider or outsourced supplier outside the EEA, unless there is a valid basis for the transfer of personal data.
Response: No

Information Security

Will directly identifiable personal data be stored separately from the rest of the collected data (in a scrambling key)?

Comment from NSD: It is common practice to remove directly identifiable data (name, national ID number, contact details etc.) from the collected data and give each data subject a code/number. A scrambling key is the file/list of names and codes that makes it possible to directly identify data subjects in the collected data. It should be stored separately from the rest of the collected data. In practice, this means that the scrambling key cannot be stored in the same network as the rest of the data, unless the scrambling key is encrypted.
Response: Yes

Which technical and practical measures will be used to secure the personal data?

Response: Personal data will be anonymised as soon as no longer needed
Comment from NSD: Anonymisation involves processing the data in such a way that no individual persons can be identified in the data that you're left with, i.e. the data can no longer be linked to individual persons in any way.
Anonymisation usually involves: *deleting directly identifiable personal data (including scrambling key/list of names) *deleting or rewriting indirectly identifiable personal data (e.g. deleting or categorizing variables such as age, place of residence, school etc.) *deleting or editing audio recordings, photographs and video recordings

Duration of project

Project period
Response: 1.6.2019 – 30.6.2020

Will personal data be stored beyond the end of project period?
Comment from NSD: Personal data should not be further processed a way that is inconsistent with the initial purpose(s) for which the data were collected. Anonymous/anonymised data may be stored indefinitely, so long as nothing else has been agreed to by the data subjects.
Response: No, the collected data will be stored in anonymous form

Which anonymization measures will be taken?
Response: Personally identifiable information will be removed, re-written or categorized

Additional information

Will the data subjects be identifiable (directly or indirectly) in the thesis/publications for the project?
Comment from NSD: If personal data are to be published, there should be a scientific purpose for this. Data is usually published in anonymous form.
Response: No

Other attachments

Response : None.

Ethical approval

NSD sin vurdering

PROSJEKTTITTEL

Relative age effect in school outcomes

REFERANSENUMMER

504915

REGISTRERT

16.09.2019 av Oscar Skovdahl Jørstad - oscarsj@student.uio.no

BEHANDLINGSANSVARLIG INSTITUSJON

Universitetet i Oslo / Det utdanningsvitenskapelige fakultet / CEMO - Centre for Educational Measurement

PROSJEKTANSVARLIG (VITENSKAPELIG ANSATT/VEILEDER ELLER STIPENDIAT)

Rolf Vegar Olsen , r.v.olsen@cemo.uio.no, tlf: 22844510

FELLES BEHANDLINGSANSVARLIGE INSTITUSJONER

TYPE PROSJEKT

Studentprosjekt, masterstudium

KONTAKTINFORMASJON, STUDENT

Oscar Skovdahl Jørstad, oscarskovdahl@gmail.com, tlf: 92825166

PROSJEKTPERIODE

01.06.2019 - 30.06.2020

STATUS

26.09.2019 - Vurdert anonym

NSD Personvern
26.09.2019 13:53

Det innsendte meldeskjemaet med referansekode 504915 er nå vurdert av NSD.

Følgende vurdering er gitt: Basert på meldeskjemaet forstår vi det slik at utelukkende er kjønn, fødselsmåned, prøveresultater, type prøve for det gitte årstrinn (Norsk, Engelsk og Matematikk for 5., 8. og 9. trinn) og hvilket år prøvene ble administrert som skal innhentes fra Nasjonale prøver. Kombinasjonen av disse opplysningene er ikke nok til å identifisere enkeltperson og datafilen er derfor anonym. Det er vår vurdering at det ikke skal behandles direkte eller indirekte opplysninger som kan identifisere enkeltpersoner i dette prosjektet, så fremt den gjennomføres i tråd med det som er dokumentert i meldeskjemaet 26.09.2019 med vedlegg. Prosjektet trenger derfor ikke en vurdering fra NSD.

HVA MÅ DU GJØRE DERSOM DU LIKEVEL SKAL BEHANDLE PERSONOPPLYSNINGER? Dersom prosjektopplegget endres og det likevel blir aktuelt å behandle personopplysninger må du melde dette til NSD ved å oppdatere meldeskjemaet. Vent på svar før du setter i gang med behandlingen av personopplysninger.

VI AVSLUTTER OPPFØLGING AV PROSJEKTET Siden prosjektet ikke behandler personopplysninger avslutter vi all videre oppfølging.

Lykke til med prosjektet!

Kontaktperson hos NSD: Belinda Gloppen Helle Tlf. Personverntjenester: 55 58 21 17 (tast 1)

Appendix 2

Data management and analysis code

Software used: R studio

```
### Master thesis code ###
### Installing necessary packages for all analyses
#install.packages("ggplot2")
#install.packages("foreign")
#install.packages("dplyr")
#install.packages("moments")
#install.packages("tidyverse")
#install.packages("memisc")
#install.packages("scales")
#install.packages("rddtools")
#install.packages("psych")
#install.packages("reshape2")
library(reshape2)
library(ggplot2)
library(foreign)
library(dplyr)
library(moments)
library(tidyverse)
library(memisc)
library(scales)
library(rddtools)
library(psych)
################################################################################
### Loading data
Test <- read.table('NP2018 skalapoeng til CEMO.txt', header = TRUE)
### Ordering data in descending order
Test1<- Test[order(Test$Test, # Column which contains test format
                Test$Grade, # Column which contains grade year
                Test$Year, # Column which contains test year
                Test$Test.1, # Column which contains test format and grade year
                Test$Month, # Column which contains birth month
                Test$Gender, # Column which contains gender
                Test$Score),] # Column which contains test score
### Data management
Test1<- as.data.frame(Test1)
Test1$Score<- gsub(",",".",Test1$Score)
Test1$Score<- as.numeric(as.character(Test1$Score))
Test1$new_month<- as.numeric(Test1$new_month)
options(digits = 5)
summary(Test1)

### Subsetting data by grade year
g5_data <- Test1 %>% filter(Grade==5)
g8_data <- Test1 %>% filter(Grade==8)
g9_data <- Test1 %>% filter(Grade==9)

## Removal of invalid birth months
g5_data <- subset(g5_data, g5_data$Month<= 12)
summary(g5_data)
```

```
g8_data <- subset(g8_data, g8_data$Month<= 12)
summary(g8_data)
g9_data <- subset(g9_data, g9_data$Month<= 12)
summary(g9_data)

### Reverses birthmonths, meaning that December (12) is now 1
### and january (1) is now 12
g5_data_new <- g5_data %>% mutate(new_month = (Month-13)*(-1))
g8_data_new <- g8_data %>% mutate(new_month = (Month-13)*(-1))
g9_data_new <- g9_data %>% mutate(new_month = (Month-13)*(-1))

########## Subset preparation for further analyses ################
### Grade 5 subjects
g5Maths<- subset(g5_data_new, g5_data_new$Test.1=="NPREG05") # Numeracy
g5Eng<- subset(g5_data_new,g5_data_new$Test.1== "NPENG05") # English
g5Read<- subset(g5_data_new,g5_data_new$Test.1== "NPLES05") # Reading
summary(g5Maths)
summary(g5Eng)
summary(g5Read)

### Grade 8 subjects
g8Eng<- subset(g8_data_new,
          g8_data_new$Test.1== "NPENG08") # English
g8Read<- subset(g8_data_new,
           g8_data_new$Test.1== "NPLES08") # Reading
g8Maths<- subset(g8_data_new,
            g8_data_new$Test.1=="NPREG08") # Numeracy
summary(g8Eng)
summary(g8Read)
summary(g8Maths)

### Grade 9 subjects
g9Read<- subset(g9_data_new,g9_data_new$Test.1== "NPLES09") # Reading
g9Maths<- subset(g9_data_new, g9_data_new$Test.1=="NPREG09") # Numeracy
summary(g9Read)
summary(g9Maths)

########## Descriptive statistics ##########
### Histogram for student distribution per birth month
### Grade 5 sample distribution
ggplot(data=g5Maths, # Using numeracy as example due to largest sample size
     aes(g5Maths$Month)) + # Visualizing sample size per birth month
      geom_histogram(bins = 23)+ # Histogram-command, bins used to separate columns for
readability
      xlab("Birth month")+ # Adding new label for X-axis
      theme_bw()+ # Adding black-white background theme
      ggtitle("Distribution of students by birth month, grade 5") # Adding title for figure
### Comparing the figure with a frequency table
table(g5Maths$Month) #N = 60665

### Grade 8 sample distribution
ggplot(data=g8Maths,
     aes(g8Maths$Month)) +
     geom_histogram(bins=23)+
     xlab("Birth month")+
     theme_bw()+
     ggtitle("Distribution of students by birth month, grade 8")
```

```
### Comparing the figure with a frequency table
table(g8Maths$Month) # N = 59171

### Grade 9 sample distribution
ggplot(data=g9Maths,
     aes(g9Maths$Month)) +
     geom_histogram(bins=23)+
     xlab("Birth month")+
     theme_bw()+
     ggtitle("Distribution of students by birth month, grade 9")
### Comparing the figure with a frequency table
table(g9Maths$Month) # N = 58880

### Descriptive statistics (Mastery levels) ###
### Adding mastery levels to each subject and grade
### Grade 5 ###
g5Maths$mastery <- ifelse(g5Maths$Score <= 42, "1", # Level 1 if score <=42
                  ifelse(g5Maths$Score >= 57, "3", "2")) # Level 3 if score >= 57, if else level 2
g5Eng$mastery <- ifelse(g5Eng$Score <= 42, "1",
                ifelse(g5Eng$Score >= 57, "3", "2"))
g5Read$mastery <- ifelse(g5Read$Score <= 42, "1",
                 ifelse(g5Read$Score >= 58, "3", "2")) # Level 3 if score >= 58, if else level 2
### Descriptive statistics of each mastery level by score
describeBy(g5Maths[c("Score", "mastery")], g5Maths$mastery, fast =T)
describeBy(g5Eng[c("Score", "mastery")], g5Eng$mastery, fast =T)
describeBy(g5Read[c("Score", "mastery")], g5Read$mastery, fast =T)

### Grade 8 ###
g8Eng$mastery<- ifelse(g8Eng$Score<=37, "1", # Level 1
                ifelse(g8Eng$Score>38 & g8Eng$Score<=43, "2", # Level 2
                     ifelse(g8Eng$Score>44 & g8Eng$Score<=54, "3", # Level 3
                          ifelse(g8Eng$Score>55 & g8Eng$Score<=62, "4", # Level 4
                               ifelse(g8Eng$Score>=63, "5","0") # Level 5
                          )
                     )
                )
)
g8Read$mastery<- ifelse(g8Read$Score<=37, "1",
                 ifelse(g8Read$Score>38 & g8Read$Score<=43, "2",
                      ifelse(g8Read$Score>44 & g8Read$Score<=54, "3",
                           ifelse(g8Read$Score>55 & g8Read$Score<=62, "4",
                                ifelse(g8Read$Score>=63, "5", "0")
                           )
                      )
                 )
)
g8Maths$mastery<- ifelse(g8Maths$Score<=36, "1",
                  ifelse(g8Maths$Score>36 & g8Maths$Score<=44, "2",
                       ifelse(g8Maths$Score>44 & g8Maths$Score<=55, "3",
                            ifelse(g8Maths$Score>55 & g8Maths$Score<=62, "4",
                                 ifelse(g8Maths$Score>=62, "5", "0")
                            )
                       )
                  )
)
### Descriptive statistics of each mastery level by score
describeBy(g8Eng[c("Score", "mastery")], g8Eng$mastery, fast =T)
describeBy(g8Read[c("Score", "mastery")], g8Read$mastery, fast =T)
describeBy(g8Maths[c("Score", "mastery")], g8Maths$mastery, fast =T)
### Grade 9 ###
g9Read$mastery<- ifelse(g9Read$Score<=37, "1",
                        ifelse(g9Read$Score>38 & g9Read$Score<=43, "2",
                               ifelse(g9Read$Score>44 & g9Read$Score<=54, "3",
                                      ifelse(g9Read$Score>55 & g9Read$Score<=62, "4",
                                             ifelse(g9Read$Score>=63, "5", "0")
                                      )
                               )
                        )
)
g9Maths$mastery<- ifelse(g9Maths$Score<=36, "1",
                         ifelse(g9Maths$Score>36 & g9Maths$Score<=44, "2",
                                ifelse(g9Maths$Score>44 & g9Maths$Score<=55, "3",
                                       ifelse(g9Maths$Score>55 & g9Maths$Score<=62, "4",
                                              ifelse(g9Maths$Score>=62, "5", "0")
                                       )
                                )
                         )
)
### Descriptive statistics of each mastery level by score
describeBy(g9Read[c("Score", "mastery")], g9Read$mastery, fast =T)
describeBy(g9Maths[c("Score", "mastery")], g9Maths$mastery, fast =T)

### Preparing data for 100% stacked bar charts of mastery levels by birth month
### Grade 5# Level 1 in the first column and level 3 in the third column
table(g5Maths$Month,g5Maths$mastery) # Frequency table for mastery level by birth month
Num5mastery<- read.table(text = " 1    2    3  # Creating data frame number of students in each mastery level by birth month
  1   936 2725 1428 # Level 1 in the first column and level 3 in the third column
  2   840 2639 1342
  3   906 2772 1453
  4   975 2740 1470
  5  1077 2912 1462
  6  1039 2901 1298
  7  1245 3096 1217
  8  1169 2907 1171
  9  1247 2815 1056
  10 1258 2649  981
  11 1200 2450  846
  12 1199 2421  823", sep = "", header = TRUE)
datnum5 <- Num5mastery %>% # Preparing data frame for stacked bar chart figure
    mutate(month = factor(row_number())) %>%
    gather(mastery, value, -month) # Creates column with the count of number of mastery level
### Stacked bar chart
stacked5<- ggplot(data=datnum5,
                  aes(x= month, # Birth months on X-axis
                      y= value, # Count of mastery level achievements on Y-axis
                      fill=mastery)) + # colors the different parts of the columns based on frequency of each mastery level
                  geom_bar(position = position_fill(reverse = TRUE), # stacks data from level 1 on bottom and level 3 on top of the columns
                  stat = "identity") +
                  scale_y_continuous(labels = scales::percent_format())+ # Converts values on Y-axis into percentages
                  ggtitle("Mastery level distribution, grade 5")+
            ylab("Percentage")+
            xlab("Birth month")+
            theme_bw()
### Storing figure as jpeg file
jpeg("stacked5.jpeg", width = 500, height = 300) ## Adjust width and height to desired sizes
item_plot5 <- stacked5
print(item_plot5)
dev.off()

### Grade 8
table(g8Maths$Month, g8Maths$mastery)
Num8mastery<- read.table(text = "1    2    3    4    5
  1   343  895 2005 1087  633
  2   328  824 1824  978  630
  3   330  944 2039 1043  701
  4   372  963 2064 1071  669
  5   381  987 2106 1109  627
  6   369 1050 2133 1022  555
  7   379 1069 2192 1090  574
  8   473 1076 2201  934  511
  9   414 1090 2110  935  492
  10  418 1029 2012  901  473
  11  387 1015 1752  809  417
  12  429 1016 1783  717  391", sep = "", header = TRUE)
datnum8 <- Num8mastery %>%
      mutate(month = factor(row_number())) %>%
      gather(mastery, value, -month)
### Stacked bar chart
stacked8<-ggplot(data=datnum8, aes(x=month, y= value,fill=mastery)) +
      geom_bar(position = position_fill(reverse = TRUE),stat = "identity") +
      scale_y_continuous(labels = scales::percent_format())+
      ggtitle("Mastery level distribution, grade 8")+
      ylab("Percentage")+
      xlab("Birth month")+
      theme_bw()
### Storing figure as jpeg file
jpeg("stacked8.jpeg", width = 500, height = 300) ## Adjust width and height to desired size]
item_plot8 <- stacked8
print(item_plot8)
dev.off()
### Grade 9
table(g9Maths$Month, g9Maths$mastery)
num9mastery<- read.table(text = " 1    2    3    4    5
  1   210  640 1719 1278 1239
  2   191  577 1543 1176 1132
  3   185  649 1791 1340 1242
  4   208  663 1746 1333 1118
  5   166  627 1801 1318 1130
  6   211  709 1823 1251 1020
  7   239  698 1921 1357 1141
  8   201  698 1843 1220 1059
  9   199  750 1857 1172  924
  10  243  697 1794 1142  896
  11  231  615 1689 1052  841
  12  233  723 1607 1032  770", sep = "", header = TRUE)
dat9 <- num9mastery %>%
      mutate(month = factor(row_number())) %>%
     gather(mastery, value, -month)
 ### Stacked bar chart
 stacked9<-ggplot(data=dat9, aes(x=month, y= value,fill=mastery)) +
     geom_bar(position = position_fill(reverse = TRUE),stat = "identity") +
     scale_y_continuous(labels = scales::percent_format())+
     ggtitle("Mastery level distribution, grade 9")+
     ylab("Percentage")+
     xlab("Birth month")+
     theme_bw()
 ### Storing figure as jpeg file
 jpeg("stacked9.jpeg", width = 500, height = 300) ## Adjust width and height to desired sizes
 item_plot <- stacked9
 print(item_plot)
 dev.off()
 ## Note that numeracy was chosen for all these figures due to having the largest sample sizes
 #####################################################
 ########### OLS REGRESSION ANALYSES #################
 ## Grade 5
 ## Intercept recoding (December is coded as 0.5 to make intercept meaningful)
 g5Eng$new_month<- g5Eng$new_month-0.5
 g5Read$new_month<- g5Read$new_month-0.5
 g5Maths$new_month<- g5Maths$new_month-0.5
 ## OLS regression analyses
 Eng5LM<- lm(Score~new_month*Gender, data = g5Eng)
 Read5LM<- lm(Score~new_month*Gender, data = g5Read)
 Num5LM<- lm(Score~new_month*Gender, data = g5Maths)
 ## Summary of OLS regression grade 5
 summary(Eng5LM)
 summary(Read5LM)
 summary(Num5LM)

 ## Grade 8
 ## Intercept recoding (December is coded as 0.5 to make intercept meaningful)
 g8Eng$new_month<- g8Eng$new_month-0.5
 g8Maths$new_month<- g8Maths$new_month-0.5
 g8Read$new_month<- g8Read$new_month-0.5
 ## OLS regression analyses
 Eng8LM<- lm(Score~new_month*Gender, data = g8Eng)
 Num8LM<- lm(Score~new_month*Gender, data= g8Maths)
 Read8LM<- lm(Score~new_month*Gender, data= g8Read)
 ## Summary of OLS regression analyses
 summary(Eng8LM)
 summary(Num8LM)
 summary(Read8LM)

 ## Grade 9
 ## Intercept recoding (December is coded as 0.5 to make intercept meaningful)
 g9Maths$new_month<- g9Maths$new_month-0.5
 g9Read$new_month<- g9Read$new_month-0.5
 ## OLS regression analyses
 Num9LM<- lm(Score~new_month+Gender+new_month*Gender, data=g9Maths)
 Read9LM<- lm(Score~new_month+Gender+new_month*Gender, data=g9Read)
 ## Summary of OLS regression analyses
 summary(Num9LM)
 summary(Read9LM)
 ## Summary table of all OLS regression analyses
 mtable(Eng5LM,Read5LM,Num5LM,Eng8LM,Read8LM,Num8LM,Read9LM, Num9LM)
```

```
        ####### Data visualization of Grade 5 English OLS regression #########################
        g5_means<- g5Eng %>% # Making data frame with mean score per birth month
          group_by(new_month)%>% # Group scores by birth month
          summarise(mean = mean(Score), # Adds mean scores
                sd = sd(Score), # Adds standard deviations
                sem = sd(Score)/sqrt(n()), # Adds standard error of measurement
                n = n(), # Adds sample sizes
                upper = mean(Score) + 2*sd(Score)/sqrt(n()), # Adds upper bound of 95 % confidence
interval
                lower = mean(Score) - 2*sd(Score)/sqrt(n())) # Adds lower bound of 95 % confidence
interval
        # Plotting results wtih error bars
        Engplot<- ggplot(g5_means, aes(x = new_month, # Adding birth month to X-axis
                          y = mean, # Adding mean scores to Y-xis
                          ymin = lower, # Sets the lower bound of error bars to the lower bound of
confidence interval
                          ymax= upper)) + # Sets the upper bound of the error bars to the upper bound of
confidence interval
                        scale_x_continuous(breaks =
as.numeric(as.character(levels(factor(g5_means$new_month))))) + # Formats the X-axis
                        geom_errorbar(size = 1)+ # Adding error bars
                        ylim(c(47,53))+ # Adjusting the limits of the Y-axis values
                        geom_point(size =2)+ # Adds points to error bars representing the mean score per
month
                        geom_abline(intercept = 48.41887, # Addding regression line to error bars
                        slope = 0.30100)+
                        theme_bw()+ # Adding black and white background theme for the figure
                        ggtitle("OLS regression Grade 5 English")+ # Adding title to the figure
                        ylab("Achievement score")+ # Changing the label on the Y-axis
                        xlab("Age by Month") # Changing the label on the X-axis
        ## Storing figure as jpeg file
        jpeg("engplotfinal.jpeg",
           width = 600,
           height = 400) ## Adjust width and height to desired sizes
        eng_plot <- Engplot
        print(eng_plot)
        dev.off()
        ################################################################
        ### Regression discontinuity for numeracy grade 8 and 9 ###########
        ## Step.1: Preparing datasets
        # Numeracy
        g8Maths$new_month<- g8Maths$new_month-12 # Making grade 8 birth months negative values
        RDDMath<- full_join(g8Maths,g9Maths) # Merging grade 8 and 9 numeracy data
        RDDMath$Grade<- RDDMath$Grade-8 # Recoding grade to "0" for grade 8 and "1" for grade 9
        RDDMath$Gender<- gsub("G","Boys", RDDMath$Gender) # Recoding "G" to boys (reference
group)
        RDDMath$Gender<- gsub("J","Girls", RDDMath$Gender) # Recoding "J" to girls

        # Reading
        g8Read$new_month<- g8Read$new_month-12 # Making grade 8 birth months negative values
        RDDRead<- full_join(g8Read,g9Read) # Merging grade 8 and 9 numeracy data
        RDDRead$Grade<- RDDRead$Grade-8 # Recoding grade to "0" for grade 8 and "1" for grade 9
        RDDRead$Gender<- gsub("G","Boys", RDDRead$Gender) # Recoding "G" to boys (reference
group)
        RDDRead$Gender<- gsub("J","Girls", RDDRead$Gender) # Recoding "J" to girls

        ### Step.2: Controlling for gender proportions
```

```
    ## installing necessary package
    # install.packages("plyr")
    library(plyr)
    # Numeracy
    RDDMath$group[RDDMath$Grade %in% 0] = 'Grade8' # Creating new column containing
strings,
    RDDMath$group[RDDMath$Grade %in% 1] = 'Grade9' # which indicates the grade each
observation belongs to.
    group.counts = as.data.frame(with(RDDMath, table(group, Gender, new_month))) # Creating a
new data frame with the grade strings, gender and birth month
    group.counts = dcast(melt(group.counts), group * new_month ~ Gender) # Splits gender into two
variables, boys and girls respectively, to count the amount of each gender per birth month and grade
    group.counts$proportion.boys = group.counts$Boys / (group.counts$Girls + group.counts$Boys)
# Adding column which contains the proportion of boys per birth month and grade year
    group.counts$group = factor(group.counts$group, levels = c("Grade8","Grade9")) # Ordering data
per birth month
    group.counts$new_month = as.numeric(as.character(group.counts$new_month)) # Formatting
birth months into string
    group.counts<- group.counts[complete.cases(group.counts),] # Formatting observed cases
    propci = function(r) prop.test(matrix(c(r$Boys, r$Girls), nrow=1))$conf.int # Creating function
that calculates 95% confidence intervals for each gender proportion
    group.counts = adply(group.counts, 1, propci) # Applying confidence interval function to the data
frame
    ## Making gender proportion figure with error bars
    GenNUM<-  ggplot(group.counts, aes(x = new_month, # Adding month on X-axis
                              y = proportion.boys, # Adding the proportion of boys on Y-axis
                              ymin = V1, # Setting the lower limit on the error bars as the lower bound of
the confidence interval
                              ymax = V2)) + # Setting the upper limit on the error bars as the upper bound
of the confidence interval
                           scale_x_continuous(breaks =
as.numeric(as.character(levels(factor(group.counts$new_month))))) + # Formatting the X-axis so error
bars occur for each specific birth month
                           geom_errorbar(aes(color = group), # Adding error bars, grade 8 in red color and
grade 9 in blue color
                                  size=1) +
                           geom_hline(yintercept = 0.5070741) + # Adding horizontal line indicating
average proportion of boys in the dataset
                           geom_point(aes(color = group), # Adding points to error bars which reflects the
mean proportion of boys
                                 size = 4)+
                           scale_color_discrete(name = "Grade")+ # Adding legend
                           ylim(c(0.45,0.55)) + # Adjusting the limits for the values on the Y-axis
                           theme(axis.text.x = element_text(angle = 45, hjust = 1), # Formatting text in
figure on X-axis
                              text = element_text(size=20)) +
                           ggtitle("Gender proportion, numeracy")+ # Adding title to figure
                           xlab("Age by month")+ # Adding X-axis label
                           theme_bw() # Adding a black and white background theme to the figure
    ### Storing figure as jpeg file
    jpeg("Gender_Proportion_NUM.jpeg",
      width = 900,
      height = 400) # Adjust width and height to desired sizes
    Prop_plot1 <- GenNUM
    print(Prop_plot1)
    dev.off()
    # Reading (same procedure as for numeracy)
```

```
RDDRead$group[RDDRead$Grade %in% 0] = 'Grade8' # Creating new column containing strings,
RDDRead$group[RDDRead$Grade %in% 1] = 'Grade9' # which indicates the grade each observation belongs to.
group.countsRead = as.data.frame(with(RDDRead, table(group, Gender, new_month))) # Creating a new data frame with the grade strings, gender and birth month
group.countsRead = dcast(melt(group.countsRead), group * new_month ~ Gender) # Splits gender into two variables, boys and girls respectively, to count the amount of each gender per birth month and grade
group.countsRead$proportion.boys = group.countsRead$Boys / (group.countsRead$Girls + group.countsRead$Boys) # Adding column which contains the proportion of boys per birth month and grade year
group.countsRead$group = factor(group.countsRead$group, levels = c("Grade8","Grade9")) # Ordering data per birth month
group.countsRead$new_month = as.numeric(as.character(group.countsRead$new_month)) # Formatting birth months into string
group.countsRead<- group.countsRead[complete.cases(group.countsRead),] # Formatting observed cases
group.countsRead = adply(group.countsRead, 1, propci) # Applying confidence interval function to the data frame
## Making gender proportion figure with error bars
GenREAD<- ggplot(group.countsRead, aes(x = new_month, # Adding month on X-axis
                         y = proportion.boys, # Adding the proportion of boys on Y-axis
                         ymin = V1, # Setting the lower limit on the error bars as the lower bound of the confidence interval
                         ymax = V2)) + # Setting the upper limit on the error bars as the upper bound of the confidence interval
                       scale_x_continuous(breaks = as.numeric(as.character(levels(factor(group.countsRead$new_month))))) + # Formatting the X-axis so error bars occur for each specific birth month
                       geom_errorbar(aes(color = group),
                              size=1) + # Adding error bars, grade 8 in red color and grade 9 in blue color
                       geom_hline(yintercept = 0.505809) + # Adding horizontal line indicating average proportion of boys in the dataset
                       geom_point(aes(color = group), # Adding points to error bars which reflects the mean proportion of boys
                            size = 4)+
                       scale_color_discrete(name = "Grade")+ # Adding legend
                       ylim(c(0.45,0.55)) +
                       theme(axis.text.x = element_text(angle = 45, hjust = 1), # Formatting text in figure on X-axis
                         text = element_text(size=20)) + # Adjusting the limits for the values on the Y-axis
                       ggtitle("Gender proportion, reading")+ # Adding title to figure
                       xlab("Age by month")+ # Adding X-axis label
                       theme_bw() # Adding a black and white background theme to the figure
## Storing figure as jpeg file
jpeg("Gender_Proportion_READ.jpeg",
  width = 900,
  height = 400) ## Adjust width and height to desired sizes
Prop_plot2 <- GenREAD
print(Prop_plot2)
dev.off()

##### Step.3: Controlling for standards of regression discontinuity (Schochet et.al., 2010) #####
## Step 3.1: Regression discontinuity (RD) for numeracy with interaction
```

```
RDnum1<- lm(Score~new_month*Grade, data = RDDMath) # Numeracy
RDread1<- lm(Score~new_month*Grade, data = RDDRead) # Reading
## Summary of RD analyses with interaction
summary(RDread1) # Model 1 (table 3)
summary(RDnum1) # Model 2 (table 3)

## Step 3.2: Gender specific RD analyses
## Preparing subsets for boys
ReadRDboys<- subset(RDDRead, RDDRead$Gender=="Boys") # Reading
NUMRDboys<- subset(RDDMath, RDDMath$Gender=="Boys") # Numeracy
## RD analyses for boys subset
RDDBoys1<- lm(Score~new_month+Grade, ReadRDboys)
RDDBoys2<- lm(Score~new_month+Grade, NUMRDboys)
## Summary of RD analyses for boys subset
summary(RDDBoys1) # Model 3 (table 3)
summary(RDDBoys2) # Model 5 (table 3)
## Preparing subsets for girls
ReadRDgirls<- subset(RDDRead, RDDRead$Gender=="Girls") # Reading
NUMRDgirls<- subset(RDDMath, RDDMath$Gender=="Girls") # Numeracy
## RD analyses for girls subset
RDDGirls1<- lm(Score~new_month+Grade, ReadRDgirls) # Reading
RDDGirls2<- lm(Score~new_month+Grade, NUMRDgirls) # Numeracy
## Summary of RD analyses for girls subset
summary(RDDGirls1) # Model 4 (table 3)
summary(RDDGirls2) # Model 6 (table 3)

## Step 3.3: Testing alternative bandwidth of RD-analysis
## Preparing data sets for alternative bandwidth
## with six months on each side of the cutoff-point.
g8NumAlt<- subset(g8Maths, g8Maths$new_month>= -5.5) # Numeracy grade 8
g9NumAlt<- subset(g9Maths, g9Maths$new_month<= 5.5) # Numeracy grade 9
g8ReadAlt<- subset(g8Read, g8Read$new_month>= -5.5) # Reading grade 8
g9ReadAlt<- subset(g9Read, g9Read$new_month<= 5.5) # Reading grade 9
## Merging separate data sets for grade 8 and 9
RDDMathAlt<- full_join(g8NumAlt,g9NumAlt) # Numeracy
RDDReadAlt<- full_join(g8ReadAlt,g9ReadAlt) # Reading
## RD analyses with alternative bandwidth
RDReadAltBan<- lm(Score~new_month+Grade, data = RDDReadAlt) # Reading
RDNumAltBan<- lm(Score~new_month+Grade, data = RDDMathAlt) # Numeracy
## Summary of RD analyses with alternative bandwidth
summary(RDReadAltBan)  # Model 7 (table 3)
summary(RDNumAltBan) # Model 8 (table 3)

## Regression discontinuity analyses (table 4)
RDnum<- lm(Score~new_month+Grade, data = RDDMath) # RD analysis numeracy
RDread<- lm(Score~new_month+Grade,data = RDDRead) # Rd analysis reading
## Summary of RD analyses (table 4)
summary(RDread)
summary(RDnum)
### Plotting RD
## Preparing datasets for visualizing RD-results
# Numeracy
RDnumdata<- rdd_data(y= Score, # Adding scale score as outcome variable
                    x=new_month, # Adding birth month as predictor variable
                    data=RDDMath, # Applying data frame for RD analyses of numeracy
                    cutpoint = 0) # Applying the value 0 as cutpoint between grade 8 and 9
# Reading
```

```
        RDreaddata<- rdd_data(y= Score,# Adding scale score as outcome variable
                     x=new_month, # Adding birth month as predictor variable
                     data=RDDRead,# Applying data frame for RD analyses of numeracy
                     cutpoint = 0) # Applying the value 0 as cutpoint between grade 8 and 9

        ## Applying RD datasets for visualizing RD-results
        ## using functions from the 'rddtools'-package
        # Numeracy model
        num_mod <- rdd_reg_lm(rdd_object = RDnumdata, # Uses RD data as object RD figure
                     slope = "same") # Specifies that the regression slope is equal at both sides of the
cutpoint
        # Reading model
        read_mod <- rdd_reg_lm(rdd_object = RDreaddata, # Uses RD data as object RD figure
                      slope = "same") # Specifies that the regression slope is equal at both sides of the
cutpoint
        ## Plotting RD results
        # Numeracy
        plot(RDnumdata, # Plotting data
           cex = 0.90,  # Adjusting size of mean score points
           col = "steelblue",  # Color of mean points
           xlab = "Age in months", # Label on X-axis
           ylab = "Achievement scores", # Label on Y-axis
           ylim = c(48,55), # Adjusting the value range on the Y-axis
           main = "Regression discontinuity for numeracy, grade 8 and 9") # Adding title to figure
        summary(num_mod) ## Extracting intercept and regression slope
        lines(c(-12,0),c(48.73506,51.21834), col = "red", lwd =3) ## Plotting regression lines for 8th
grade, Start point = (RAE-estimate * 12) - Intercept, End point = Intercept estimate
        lines(c(0,12),c(52.38212,54.8654), col= "darkblue", lwd = 3) ## Plotting regression lines for 9th
grade, Start point = Intercept + Grade, End point = Intercept + Grade + (RAE-estimate * 12)

        # Reading
        plot(RDreaddata, # Plotting data
           cex = 1, # Adjusting size of mean score points
           col = "darkblue", # Color of mean points
           xlab = "Age in months", # Label on X-axis
           ylab = "Achievement scores", # Label on Y-axis
           ylim = c(48,55), # Adjusting the value range on the y-axis
           main = "Regression discontinuity for reading, grade 8 and 9") # Adding title to figure
        summary(read_mod) ## Extracting intercept and regression slope
        lines(c(-12,0),c(48.67972,51.31756), col = "red", lwd=2) ## Plotting regression lines for 8th
grade, Start point = (RAE-estimate * 12) - Intercept, End point = Intercept estimate
        lines(c(0,12),c(52.05146, 54.6893), col = "steelblue", lwd=2) ## Plotting regression lines for 9th
grade, Start point = Intercept + Grade, End point = Intercept + Grade + (RAE-estimate * 12)
        ##############################################################################
########
        ################################################################
        ## RQ3: Investigation of how RAE changes over time from the various ages
        ## Preparing vertical linking of mean scores in numeracy (grade 5,8 and 9)
        g8Maths$new_month<- g8Maths$new_month + 48 ## Adjusting birth month values for grade 8
        g9Maths$new_month<- g9Maths$new_month + 48 ## Adjusting birth month values for grade 9
        ## Vertical linking of grade 8 and 9 scores to grade 5
        g8_data_g5Scale<- g8Maths %>% mutate(new_score = ((Score-50)/10)*12.18+61.58) # Linear
transformation of numeracy scale scores grade 8 (Ræder and Olsen, 2019)
        g9_data_g5Scale<- g9Maths %>% mutate(new_score = ((Score-50)/10)*12.18+61.58) # Linear
transformation of numeracy scale scores grade 9 (Ræder and Olsen, 2019)
        ## Making data frames with mean scores per birth month
```

```
g5_month_mean<- g5Maths %>% group_by(new_month) %>% dplyr::summarise(month_mean = mean(Score, na.rm = T)) # Grade 5
g8_month_mean<- g8_data_g5Scale %>% group_by(new_month) %>% dplyr::summarise(month_mean = mean(new_score, na.rm = T)) # Grade 8, new_score = vertically linked score
g9_month_mean<- g9_data_g5Scale %>% group_by(new_month) %>% dplyr::summarise(month_mean = mean(new_score, na.rm = T)) # Grade 9, new_score = vertically linked score
## Binding mean scores together in one dataset
NumScaled<- bind_rows(g5_month_mean, g8_month_mean, g9_month_mean)
## Running OLS regression vertically linked scale scores
mathscal<-lm(month_mean~new_month,data = NumScaled)
summary(mathscal) # Using these results to calculate parallel regression slopes for each grade year
## Plotting the results
plot(x=NumScaled$new_month, # Adding birth month which now ranges from 0.5 to 59.5
    y= NumScaled$month_mean, # Adding mean numeracy scores per birth month and grade
    xlab = "Age by month", # Adding label on X-axis
    ylab = "Achievement score", # Adding label on Y-axis
    main = "Vertically linked numeracy scores grade 5,8 and 9 ", # Adding title on figure
    cex = 0.95, # Adjusting size on mean points
    pch = 23, # Shape of points
    bg = "skyblue" # Color of points
)
## Adding parallel regression lines
lines(c(0,11.5),c(47.940206,51.89997), lwd =2.5) # Parallel regression line for grade 5
lines(c(36,47.5),c(59.81949,63.77925), lwd = 2.5) # Parallel regression line for grade 8
lines(c(48,59.5),c(63.77925,67.73901), lwd = 2.5) # Parallel regression line for grade 9
abline(mathscal, lty=2, lwd = 2.5, col = "red") # Regression line for all scores
#######################################################
## Code used for making APPENDIX 3
## Assumption testing of regression models

# Making a function to extract all parameters of interest
# for testing the assumptions of OLS regression models
diagnostic <- function(data, model){
  data_diag <- data.frame(
    yhat = round(fitted.values(model), 2), # Fitted values
    raw_resid = round(residuals(model), 2), # Raw residuals
    leverage = round(hatvalues(model), 2), # Leverage
    std_resid = round(rstandard(model), 2), # Standardized residuals
    influential = round(cooks.distance(model),2)) # Cook's distance
  data_diag <- cbind(data, data_diag)
  return(data_diag) # Returns a table with parameter estimates included in this function for the given OLS regression model
}
## Diagnostic statistics
# Grade 5
Eng5stats <- diagnostic(data = g5Eng, model = Eng5LM)
Num5stats <- diagnostic(data = g5Maths, model =  Num5LM)
Read5stats <- diagnostic(data = g5Read, model =   Read5LM)
summary(Eng5stats) # Diagnostics English
summary(Num5stats) # Diagnostics Numeracy
summary(Read5stats) # Diagnostics Reading
```

```
# Grade 8
Eng8stats <- diagnostic(data = g8Eng, model = Eng8LM)
Num8stats <- diagnostic(data = g8Maths, model =  Num8LM)
Read8stats <- diagnostic(data = g8Read, model =   Read8LM)
summary(Eng8stats) # Diagnostics English
summary(Num8stats) # Diagnostics Numeracy
summary(Read8stats) # Diagnostics Reading

# Grade 9
Num9stats <- diagnostic(data = g9Maths, model =  Num9LM)
Read9stats <- diagnostic(data = g9Read, model =   Read9LM)
summary(Num9stats) # Diagnostics Numeracy
summary(Read9stats) # Diagnostics Reading


## Diagnostic plots
# Grade 5
par(mfrow=c(2,2)) # Arranging plots into 2 columns and 2 rows
plot(Eng5LM) # Plotting diagnostics of OLS model English
par(mfrow=c(2,2))
plot(Num5LM) # Plotting diagnostics of OLS model numeracy
par(mfrow=c(2,2))
plot(Read5LM) # Plotting diagnostics of OLS model reading

# Grade 8
par(mfrow=c(2,2))
plot(Eng8LM)
par(mfrow=c(2,2))
plot(Num8LM)
par(mfrow=c(2,2))
plot(Read8LM)

# Grade 9
par(mfrow=c(2,2))
plot(Num9LM)
par(mfrow=c(2,2))
plot(Read9LM)

### Descriptive statistics, tabulated by gender and grade year

## Making subsets of the data frames
g5_data_boys<- subset(g5_data_new, g5_data_new$Gender=="G") # Grade 5 boys
g5_data_girls<- subset(g5_data_new, g5_data_new$Gender=="J") # Grade 5 girls
g8_data_boys<- subset(g8_data_new, g8_data_new$Gender=="G") # Grade 8 boys
g8_data_girls<- subset(g8_data_new, g8_data_new$Gender=="J") # Grade 8 girls
g9_data_boys<- subset(g9_data_new, g9_data_new$Gender=="G") # Grade 9 boys
g9_data_girls<- subset(g9_data_new, g9_data_new$Gender=="J") # Grade 9 girls

## Grade 5 subjects
# Numeracy
NumBoys<- subset(g5_data_boys, g5_data_boys$Test.1=="NPREG05") # Boys
mean(NumBoys$Score) # Extracting mean score for boys
sd(NumBoys$Score) # Extracting std.deviation for boys' scores
NumGirls<- subset(g5_data_girls, g5_data_girls$Test.1=="NPREG05") # Girls
mean(NumGirls$Score) # Extracting mean score for girls
sd(NumGirls$Score) # Extracting std.deviation for girls' scores
# Reading
```

```
ReadBoys<- subset(g5_data_boys, g5_data_boys$Test.1=="NPLES05") # Boys
mean(ReadBoys$Score) # Extracting mean score for boys
sd(ReadBoys$Score) # Extracting std.deviation for boys' scores
ReadGirls<- subset(g5_data_girls, g5_data_girls$Test.1=="NPLES05") # Girls
mean(ReadGirls$Score) # Extracting mean score for girls
sd(ReadGirls$Score) # Extracting std.deviation for girls' scores
# English
EngBoys<- subset(g5_data_boys, g5_data_boys$Test.1=="NPENG05") # Boys
mean(EngBoys$Score) # Extracting mean score for boys
sd(EngBoys$Score) # Extracting std.deviation for boys' scores
EngGirls<- subset(g5_data_girls, g5_data_girls$Test.1=="NPENG05") # Girls
mean(EngGirls$Score) # Extracting mean score for girls
sd(EngGirls$Score) # Extracting std.deviation for girls' scores



## Grade 8 subjects
#Numeracy
Num8Boys<- subset(g8_data_boys, g8_data_boys$Test.1=="NPREG08") # Boys
mean(Num8Boys$Score) # Extracting mean score for boys
sd(Num8Boys$Score) # Extracting std.deviation for boys' scores
Num8Girls<- subset(g8_data_girls, g8_data_girls$Test.1=="NPREG08") # Girls
mean(Num8Girls$Score) # Extracting mean score for girls
sd(Num8Girls$Score) # Extracting std.devation for girls' scores
# Reading
Read8Boys<- subset(g8_data_boys, g8_data_boys$Test.1=="NPLES08") # Boys
mean(Read8Boys$Score) # Extracting mean scores for boys
sd(Read8Boys$Score) # Extracting std.deviation for boys' scores
Read8Girls<- subset(g8_data_girls, g8_data_girls$Test.1=="NPLES08") # Girls
mean(Read8Girls$Score) # Extracting mean scores for girls
sd(Read8Girls$Score) # Extracting std.deviation for girls' scores
# English
Eng8Boys<- subset(g8_data_boys, g8_data_boys$Test.1=="NPENG08") # Boys
mean(Eng8Boys$Score) # Extracting mean scores for boys
sd(Eng8Boys$Score) # Extracting std.deviation for boys' scores
Eng8Girls<- subset(g8_data_girls, g8_data_girls$Test.1=="NPENG08") # Girls
mean(Eng8Girls$Score) # Extracting mean scores for girls
sd(Eng8Girls$Score) # Extracting std.deviation for girls' scores

## Grade 9 subjects
# Numeracy
Num9Boys<- subset(g9_data_boys, g9_data_boys$Test.1=="NPREG09") # Boys
mean(Num9Boys$Score) # Extracting mean scores for boys
sd(Num9Boys$Score) # Extracting std.deviation for boys' scores
Num9Girls<- subset(g9_data_girls, g9_data_girls$Test.1=="NPREG09") # Girls
mean(Num9Girls$Score) # Extracting mean scores for girls
sd(Num9Girls$Score) # Extracting std.deviation for girls' scores
# Reading
Read9Boys<- subset(g9_data_boys, g9_data_boys$Test.1=="NPLES09") # Boys
mean(Read9Boys$Score) # Extracting mean scores for boys
sd(Read9Boys$Score) # Extracting std.deviation for boys' scores
Read9Girls<- subset(g9_data_girls, g9_data_girls$Test.1=="NPLES09") # Girls
mean(Read9Girls$Score) # Extracting mean scores for girls
sd(Read9Girls$Score) # Extracting std.deviations for girls' scores
```

Appendix 3

Supplemental material:

Diagnostics of OLS regression models

In order to test whether the OLS regression models meets the assumptions it is useful to first define all parameters of interest. Following the definitions of all parameters used to test the assumptions of OLS regression, we will provide rules of thumb from the literature as a guide to interpret the quality of the parameter estimates.

Fitted values: Refers to predicted values (ŷ) on the outcome variable (y-values) that is expected for the range of values on the x-axis. Fitted values are determined by how the regression model are specified.

Raw residuals: Refers to the unstandardized residual estimates of the y-values. Basically, if the OLS regression model holds the residuals ϵi = yi – ŷi should be random noise. In practice, random noise is evident if there is no systematic pattern when the raw residuals are plotted against the fitted values ŷ. This also holds if the residuals are normally distributed.

Leverage: Refers to how unusual a data point is in terms of its values on the predictors. Leverage hii reflects the distance between xi and the center of X. Further, leverage is used to test if residuals are mutually independent. The rule of thumb for leverage suggests to flag observations with an estimate hii>2k/n, where hii refers to leverage, k refers to the sum of leverage and n refers to sample size, which in this case means that all observations with hii > 0.00013 should be flagged.

Standardized residuals: Used to measure discrepancies among y-values. While raw residuals indicate the atypicality of yi given the model expected values, "standardized" residuals compares expected values in terms of its standard deviation (i.e RMSE). The rule of thumb for standardized residuals suggests that estimates larger than +/- 3 is to be considered outliers because the observed value is then larger than the expected value.

Cook's distance: Cook's distance is used to estimate influential values in the data set. Influential values can be understood as statistical outliers which influence the relationship between X and Y. Cook's distance estimates how influential a outlier is by calculating how much all predicted values change when the i^{th} observation is removed. The rule of thumb suggests that estimates with a an estimate larger than 1 (D>1) is to be considered an influential outlier and may then be removed.

In addition to the parameters used to test assumptions, as presented above, graphical inspections are also important to consider when testing OLS regression assumptions. For the following diagnostics, for each OLS regression model a set of 4 figures are provided:

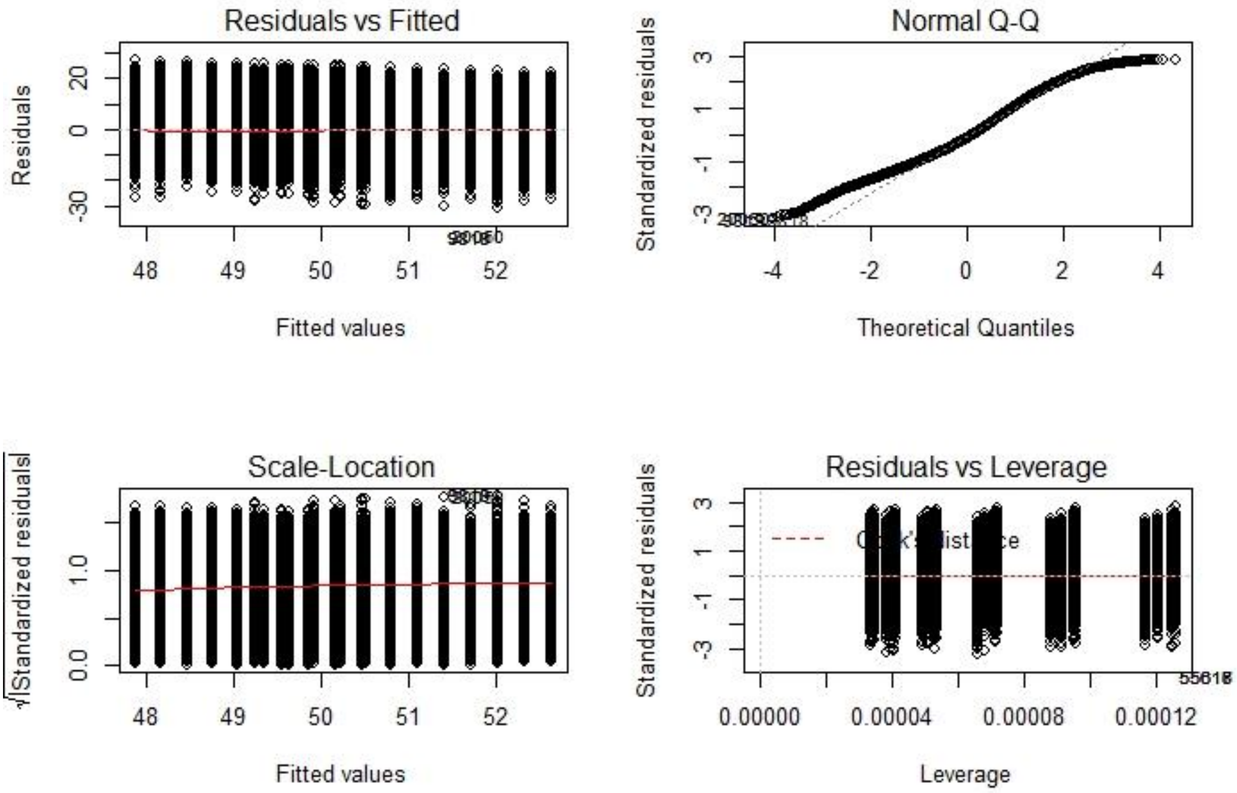
1. The first figure (i.e. “Residuals vs Fitted”) is used to check the assumption of linear relationships. A horizontal line without any distinct pattern indicates a linear relationship between residuals and fitted values.

2. The second figure (i.e. “Normal Q-Q”) is used to check how normally distributed the residuals are. If the residual points follow the dashed line, this indication of normally distributed residuals.

3. The third figure (i.e. “Scale-Location”) is used to check the assumption of homoscedasticity of residual variance. A horizontal line with an equal spread of data points suggests the residual variance is homoscedastic.

4. The fourth figure (i.e. “Residuals vs Leverage”) is used to check for influential outliers in the data set. The Y-axis represents the residuals on Y-value and the X-axis represents the leverage on the x-values. The horizontal line reflects Cook’s distance. A straight line indicates that no data points are influencing the regression output substantially.

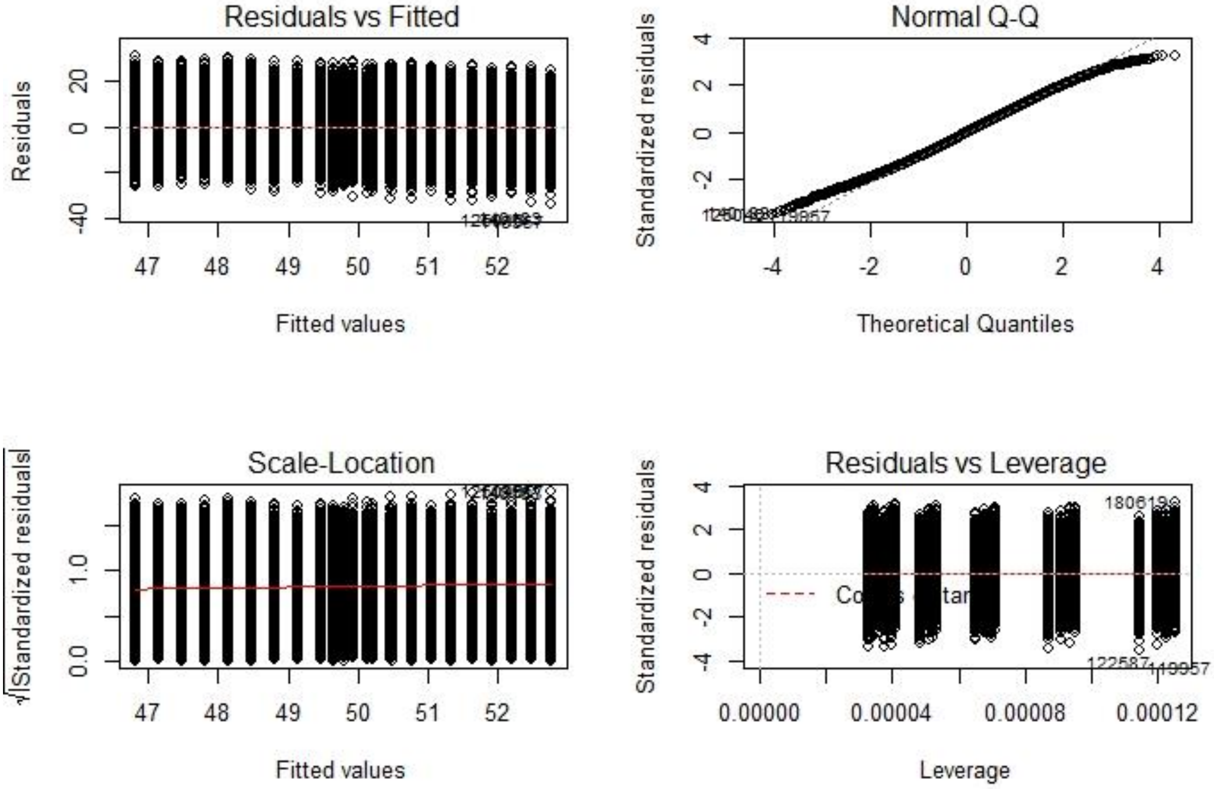
Grade 5 English diagnostics



	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	47.87	-30.86	0.000033	-3.22	0
1ST QUARTILE	49.34	-7.18	0.000039	-0.75	0.0000018
MEDIAN	50.22	-1.15	0.000053	-0.12	0.0000077
MEAN	50.22	0.00012	0.000066	0.000022	0.000016
3RD QUARTILE	51.10	6.64	0.000091	0.69	0.000021
MAXIMUM	52.64	26.99	0.00012	2.81	0.00026

The figures for **grade 5 English** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum value of the standardized residuals (Std. residual minimum = **-3.22**, Std. residuals maximum = **2.81**). This suggests that there is an outlier at the minimum of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 5 English**, these cannot be considered influential outliers on the regression results.

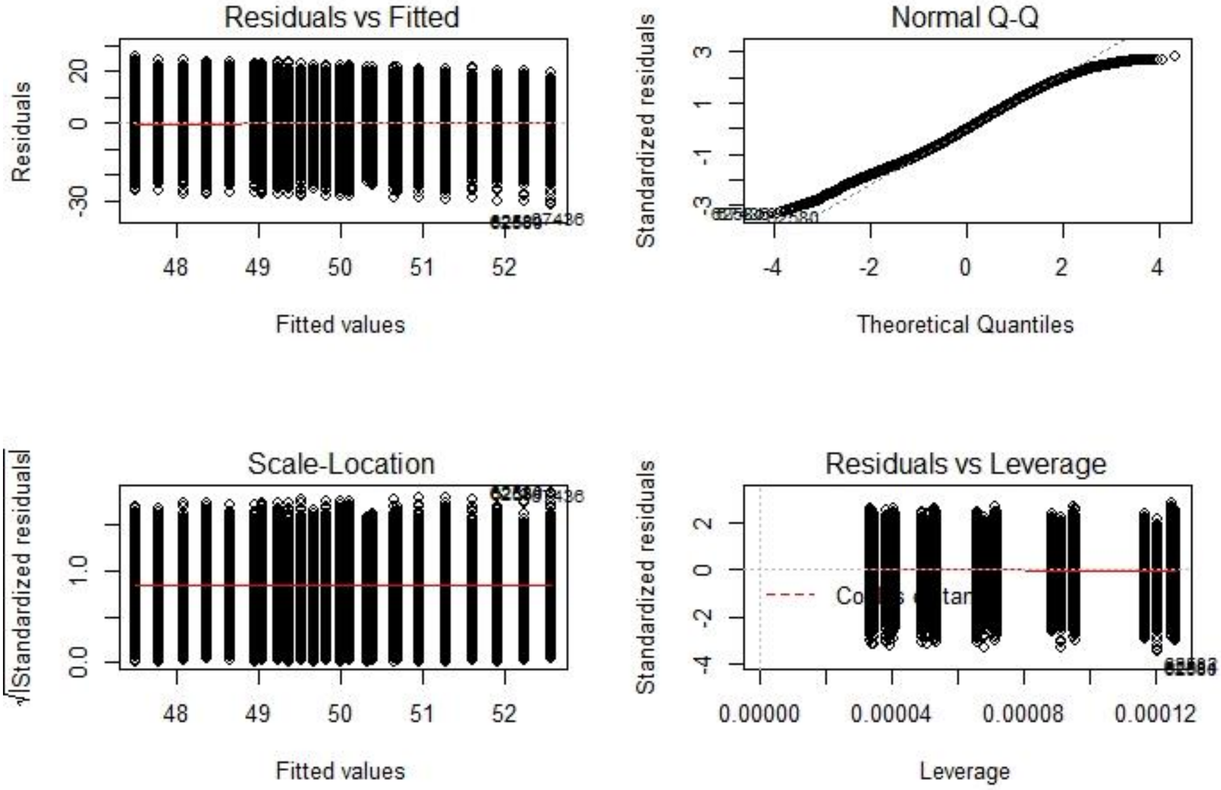
Grade 5 numeracy diagnostics



	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	46.82	-33.39	0.000032	-3.53	0
1ST QUARTILE	48.81	-6.73	0.000039	-0.71	0.0000015
MEDIAN	50.14	-0.12	0.000053	-0.01	0.0000071
MEAN	49.98	0.00003	0.000065	0.000009	0.000016
3RD QUARTILE	51.34	6.58	0.000091	0.70	0.000021
MAXIMUM	52.75	30.47	0.00012	3.22	0.00035

The figures for **grade 5 numeracy** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum and maximum value of the standardized residuals (Std. residual minimum = **-3.53**, Std. residuals maximum = **3.22**). This suggests that there is an outlier at the minimum and maximum of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 5 numeracy**, these cannot be considered influential outliers on the regression results.

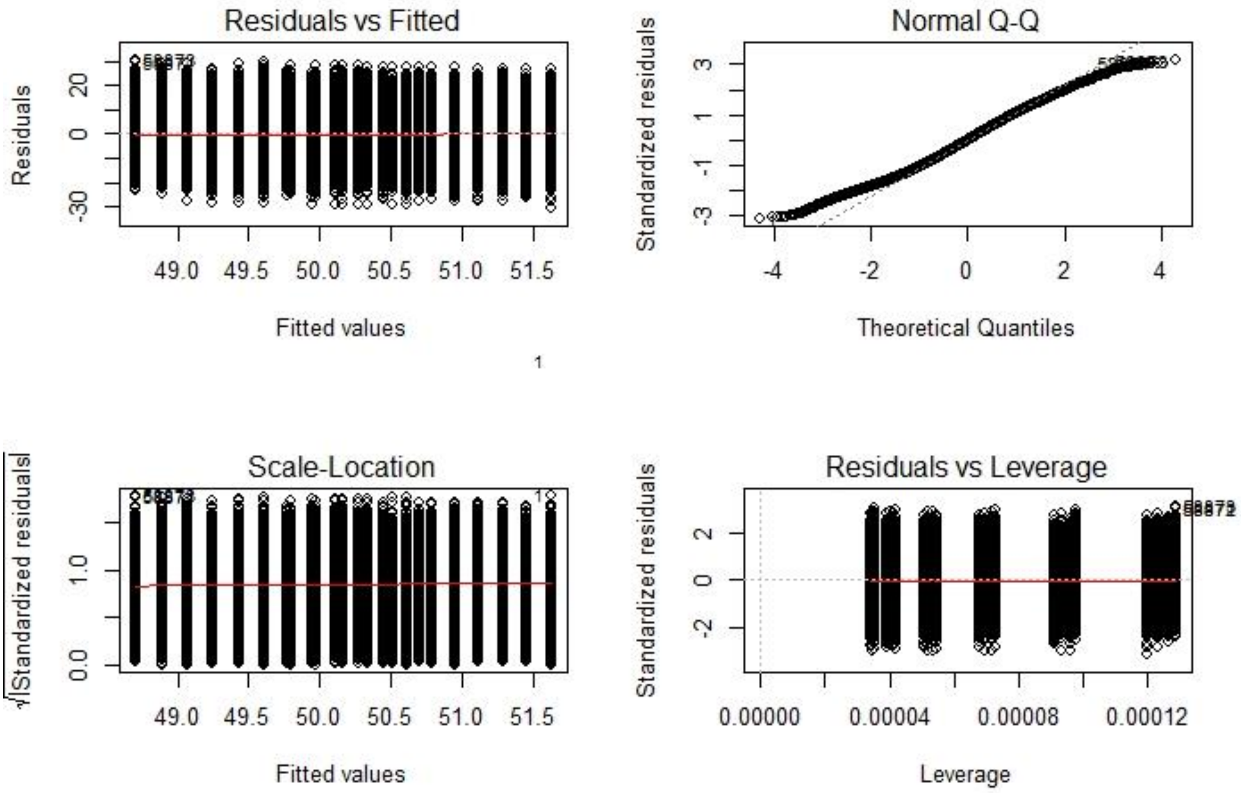
Grade 5 Reading diagnostics



	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	47.49	-31.50	0.000033	-3.44	0
1ST QUARTILE	49.04	-6.80	0.000039	-0.74	0.0000017
MEDIAN	49.99	-0.30	0.000053	-0.03	0.0000077
MEAN	49.96	-0.000022	0.000066	0.000009	0.000016
3RD QUARTILE	50.95	6.55	0.000091	0.71	0.000022
MAXIMUM	52.55	25.63	0.00012	2.80	0.00035

The figures for **grade 5 reading** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum value of the standardized residuals (Std. residual minimum = **-3.44**, Std. residuals maximum = **2.80**). This suggests that there is an outlier at the minimum of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 5 reading**, these cannot be considered influential outliers on the regression results.

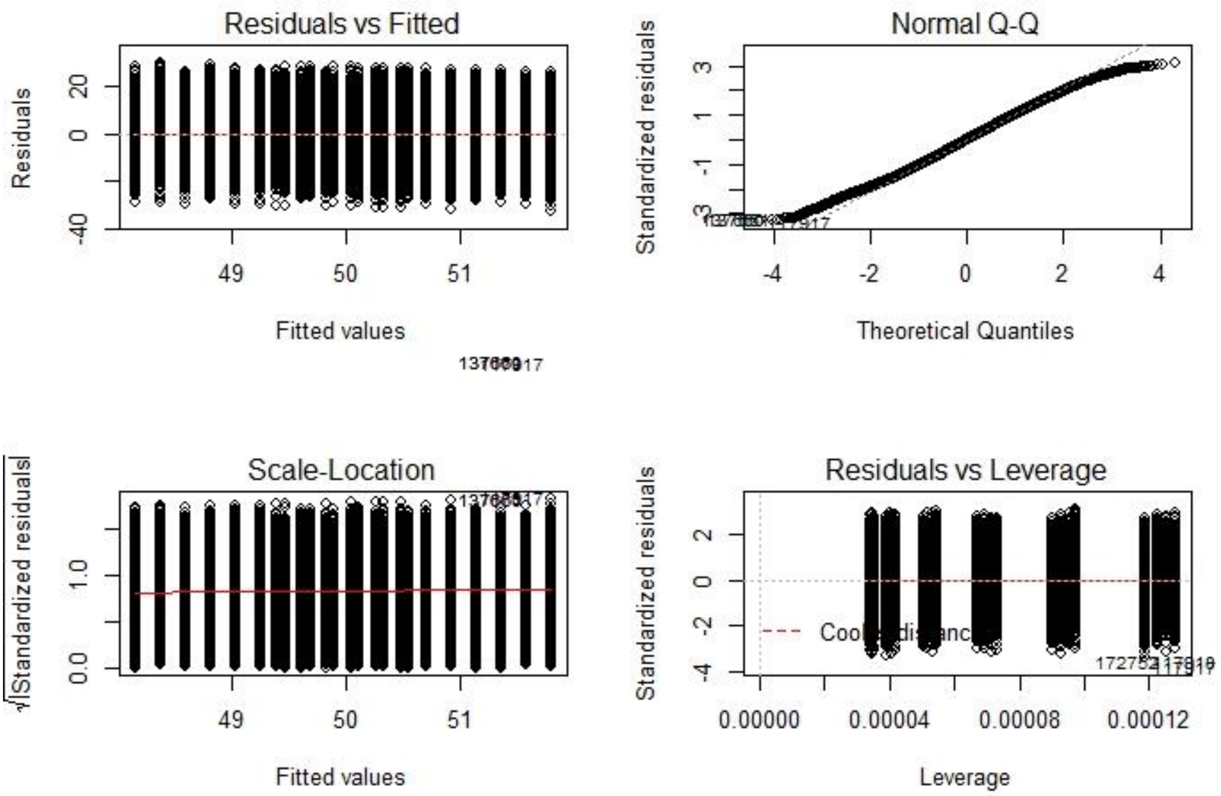
Grade 8 English diagnostics



	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	48.71	-30.51	0.000034	-3.15	0
1ST QUARTILE	49.77	-7.25	0.000040	-0.74	0.0000018
MEDIAN	50.27	-0.09	0.000054	-0.009	0.0000080
MEAN	50.22	-0.000026	0.000067	0.000007	0.000017
3RD QUARTILE	50.78	7.01	0.000093	0.72	0.000022
MAXIMUM	51.62	30.40	0.00012	3.14	0.00031

The figures for **grade 8 English** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum and maximum value of the standardized residuals (Std. residual minimum = **-3.15**, Std. residuals maximum = **3.14**). This suggests that there is an outlier at the minimum and maximum of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 8 English**, these cannot be considered influential outliers on the regression results.

Grade 8 numeracy diagnostics

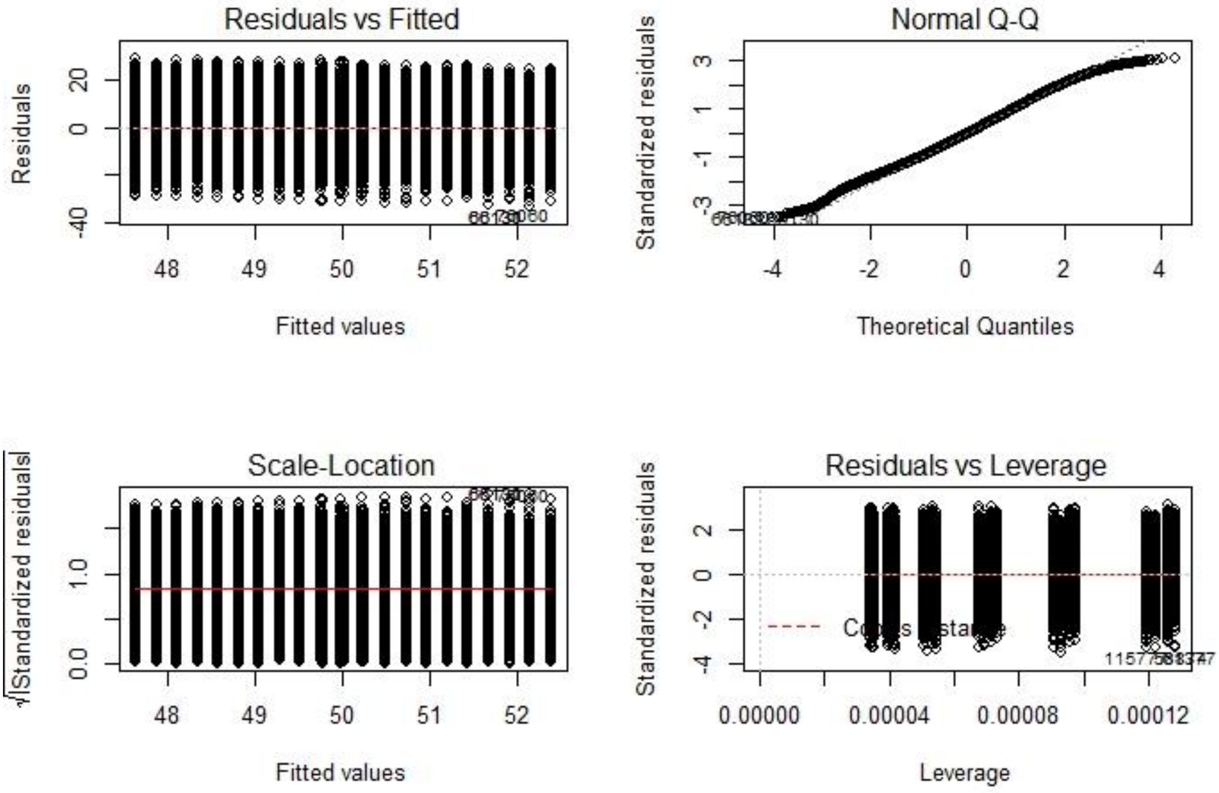


	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	48.15	-32.41	0.000033	-3.37	0
1ST QUARTILE	49.38	-6.86	0.000040	-0.71	0.0000016
MEDIAN	50.04	-0.01	0.000054	0.00	0.0000074
MEAN	50.00	0.00023	0.000067	0.000001	0.000016
3RD QUARTILE	50.70	6.79	0.000093	0.71	0.000021
MAXIMUM	51.79	29.75	0.00012	3.09	0.00033

The figures for **grade 8 numeracy** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum and maximum value of the standardized residuals

(Std. residual minimum = **-3.37**, Std. residuals maximum = **3.09**). This suggests that there is an outlier at the minimum and maximum of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 8 numeracy**, these cannot be considered influential outliers on the regression results.

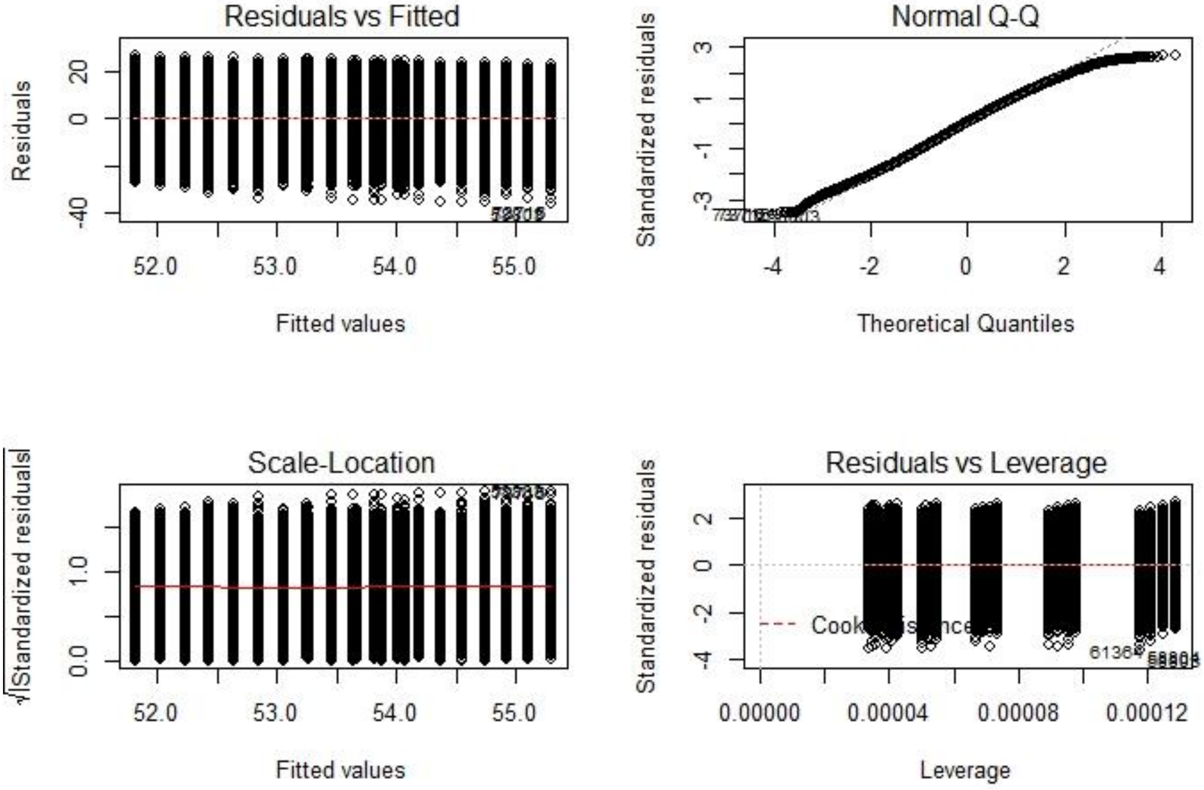
Grade 8 Reading diagnostics



	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	47.64	-33.25	0.000034	-3.54	0
1ST QUARTILE	49.05	-6.71	0.000040	-0.71	0.0000016
MEDIAN	50.02	-0.31	0.000054	-0.03	0.0000072
MEAN	50.02	0.00002	0.000067	0.000003	0.000016
3RD QUARTILE	51.20	6.41	0.000092	0.68	0.000021
MAXIMUM	52.39	28.90	0.00012	3.08	0.00034

The figures for **grade 8 reading** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum and maximum value of the standardized residuals (Std. residual minimum = **-3.54**, Std. residuals maximum = **3.08**). This suggests that there is an outlier at the minimum and maximum of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 8 reading**, these cannot be considered influential outliers on the regression results.

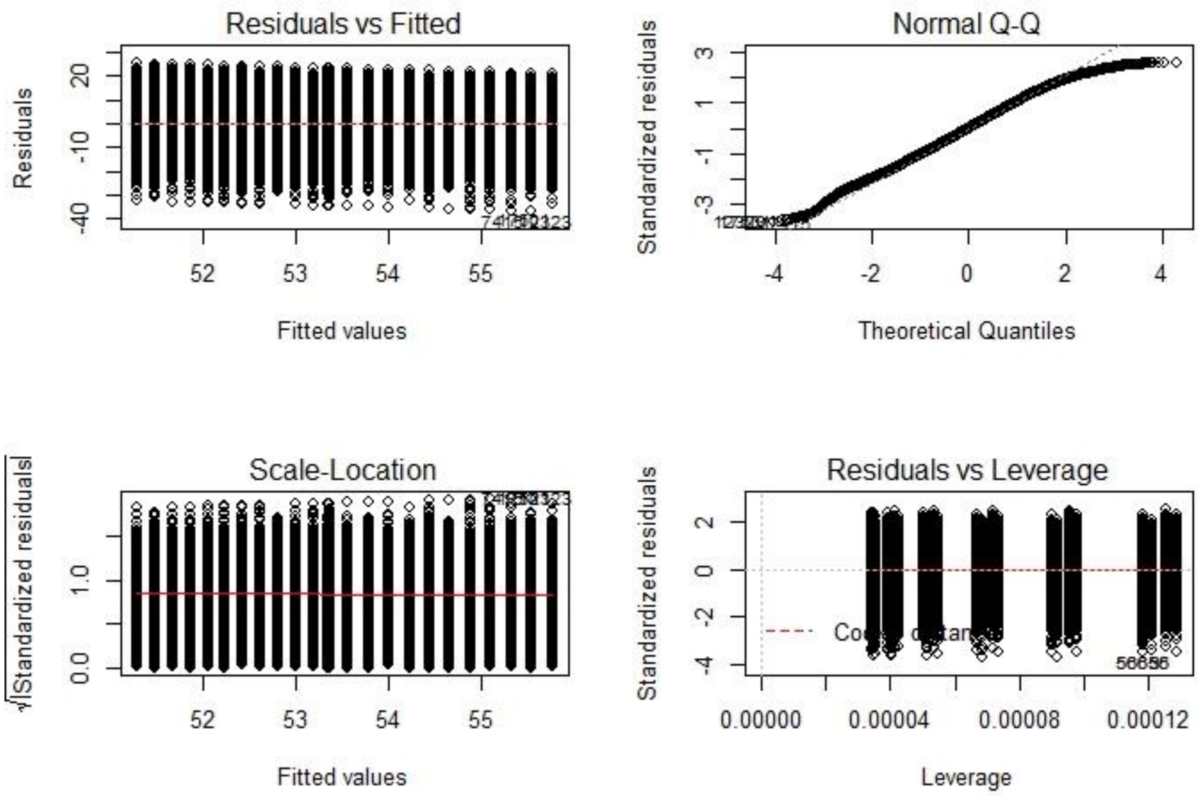
Grade 9 Numeracy diagnostics



	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	51.82	-35.92	0.000033	-3.61	0
1ST QUARTILE	53.05	-6.90	0.000040	-0.69	0.0000016
MEDIAN	53.67	0.38	0.000053	0.04	0.0000074
MEAN	53.65	0.00006	0.000054	-0.000001	0.000017
3RD QUARTILE	54.38	7.12	0.000092	0.72	0.000021
MAXIMUM	55.30	26.30	0.00012	2.65	0.00038

The figures for **grade 9 numeracy** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum value of the standardized residuals (Std. residual minimum = **-3.61**, Std. residuals maximum = **2.65**). This suggests that there is an outlier at the minimum value of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 9 numeracy**, these cannot be considered influential outliers on the regression results.

Grade 9 reading diagnostics



	FITTED VALUES	RAW RESIDUALS	LEVERAGE	STANDARDIZED RESIDUALS	COOK'S DISTANCE
MINIMUM	51.30	-36.64	0.000034	-3.70	0
1ST QUARTILE	52.43	-7.01	0.000040	-0.71	0.0000016
MEDIAN	53.38	0.05	0.000054	0.01	0.0000075
MEAN	53.45	-0.00008	0.000068	-0.000011	0.000017
3RD QUARTILE	54.66	7.11	0.000091	0.72	0.000022
MAXIMUM	55.76	25.24	0.00012	2.54	0.00038

The figures for **grade 9 reading** diagnostics suggests that there is a satisfactory linear relationship between the fitted values and the raw residuals. The Q-Q plot suggests that the standardized residuals are mostly normally distributed since the standardized residuals are smaller than +/- 3, but there is an exception in the minimum value of the standardized residuals (Std. residual minimum = **-3.70**, Std. residuals maximum = 2.54). This suggests that there is an outlier at the minimum value of the range of the residuals. The residuals was mutually independent (i.e., all estimates of leverage were smaller than the rule of thumb hii > 0.00013). There was no influential outliers in the regression models (i.e., all Cook's distance estimates were considerably smaller than the rule of thumb which suggests to flag observations with an estimate of larger than 1). Therefore, we can conclude that although there are outliers in the OLS regression model for **grade 8 reading**, these cannot be considered influential outliers on the regression results.

Standard-testing for RD-analyses (Schochet, et al., 2010):

In the following each of the four standards are evaluated, leaning heavily on the results for the RD-analyses presented in table 3.3:

- **Standard 1:** Manipulation of the forcing variable in this context is difficult as this is hard to do with birth months. The scoring rule and the cutoff-point for this study is based on the Norwegian school policy, which has strict rules for assignment of students into grade years. Deferred school start in Norway is extremely rare (less than 2% of students are deferred annually), hence students are assigned to the formally correct grade year in the data set. Hypothetically, we could think of a number of variables that may influence our interpretation at the cut-off in our RD design. However, with the limitations of the data available, we can only study how the gender variable potentially could influence the forcing variable at and around the cut-off. Figure 1.3 and 2.3 provides a graphical presentation of the gender distribution in both subjects and grades and suggests that the gender distribution is even, and that there is no tendency of an unequal distribution on each side of the cut-off. In addition, we applied separate gender analyses of the RD analysis to test whether RAE and grade effects on achievement scores works differently for boys and girls. In table 3.3, we find that model 3-6 confirms that both RAE and the grade effect has very similar effects for both genders when analyzed in separate data sets for boys and girls in each subject.

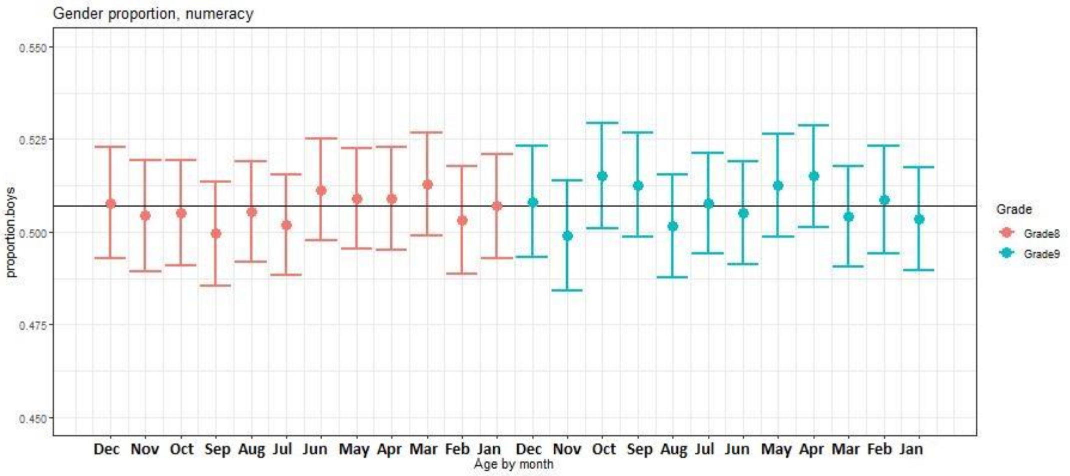


Figure 1.3. Gender distribution among national numeracy test takers in grade 8 and 9. The gender distribution is measured as the proportion of boys by birth month on the X-axis. The red error bars represent the proportion of boys per birth month in grade 8 and the blue error bars bars represent the proportion of boys per birth month in grade 9. The straight line represents the average proportion of boys for all students in grade 8 and 9.

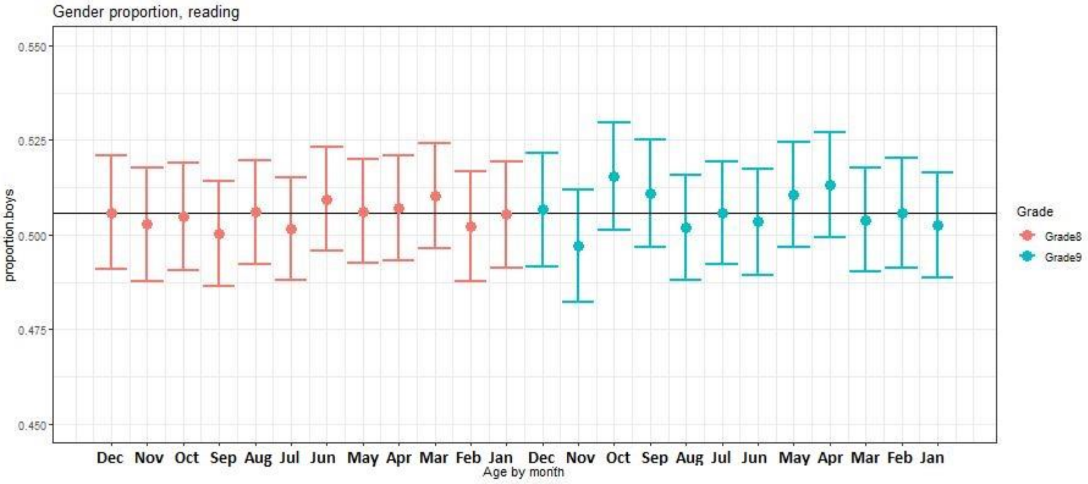


Figure 2.3. Gender distribution among national reading test takers in grade 8 and 9. The gender distribution is measured as the proportion of boys by birth month on the X-axis. The red error bars represent the proportion of boys per birth month in grade 8 and the blue error bars bars represent the proportion of boys per birth month in grade 9. The straight line represents the average proportion of boys for all students in grade 8 and 9.

- **Standard 2:** The present paper has no attrition by treatment status, meaning that the study units included in the RD-design are either in grade 8 or 9. A potential concern regarding whether the available data meets this standard is the relatively smaller amounts of students born in November and December. It is possible to hypothesize that this is systematically related to which students schools decide to exclude from participation on national tests. However, this is not a concern for the present study, according to Statistics Norway’s Statbank which shows that there is indeed born fewer students in November and December, compared to the number of children born in the relevant calendar years. See figures distribution of children born per birth month (Statistics Norway, 2019) in appendix 3.
- **Standard 3:** This standard can be checked with statistical tests and graphical inspections. Figure 4 and 5 (in the results-section regarding RD-design) provides graphical images of the RD-analyses which suggests that the relationship between the forcing variable and outcome variable are continuous on both sides of the cutoff point. In addition, there is no discontinuity within grades. We also included an initial RD-analysis with an interaction term between RAE and grade year, the results were non-significant suggesting there are no significant differences in RAE from grade 8 to 9. The results can be found in model 1 and model 2 in table 3. More specifically, model 1 (reading) and model 2 (numeracy) shows that the interaction term is not statistically significant in reading and numeracy. This means that the relationship between RAE and achievement scores is equal across grades. These findings serves the purpose of being an initial analysis of the RD analysis to test if RAE changes from grade 8 to 9.

- **Standard 4:** We control for the forcing variable when estimating the treatment effect by controlling for the distribution of students per birth month in each grade years. Figure 1 provides a descriptive statistic of this distribution which shows that students per birth month is more or less evenly distributed. Furthermore, we control the functional form by running gender specific RD-analyses which showed that both RAE and the grade effect affects both genders similarly, results are shown in table 3 (model 3-6). Lastly we compared the bandwidth of the forcing variable with an alternative bandwidth consisting of 6 unique units on each side of the cutoff-point. This means that we ran an RD analysis with students born in the first six months of the calendar year for grade 8 and students born in the last six months of the calendar year for grade 9. The results of the RD analysis with an alternative bandwidth of 6 months on each side of the cutoff-point provided similar results as the actual bandwidth used in the main analysis meaning that the functional form and bandwidth of the RD-analysis is appropriate to use. Results of this analysis is found in table 3 (model 7-8). Model 7 and 8 from table 3 confirms that the impact of RAE and the grade effect is still present in both subjects when we adjusted the bandwidth to six months for each grade year.

Supplementary tables

Table 1.3 Mastery level characteristics tabulated by subjects in the various grade years (udir.no).

Grade 5	**Level 1**	**Level 2**	**Level 3**		
Reading	≤ 42 points	43 – 57 points	≥ 58 points		
Numeracy	≤ 42 points	43 – 56 points	≥ 57 points		
English	≤ 42 points	43 – 56 points	≥ 57 points		
Grade 8/9	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Reading	≤ 37 points	38 – 43 points	44 – 54 points	55 – 62 points	≥ 63 points
Numeracy	≤ 37 points	38 – 43 points	44 – 54 points	55 – 62 points	≥ 63 points
English (grade 8 only)	≤ 36 points	37– 43 points	44 – 55 points	56 – 62 points	≥ 63 points

Table 2.3 with mean scores for gender specific subsets of each subject per grade. Standard deviations are reported in the parentheses.

	Numeracy	**Reading**	**English**
Grade 5			
Boys	**51.23 (9.80)**	**49.12 (9.33)**	**50.98 (9.80)**
Girls	**48.67 (9.21)**	**50.81 (9.09)**	**49.51 (9.48)**
Grade 8			
Boys	**50.60 (9.80)**	**48.95 (9.61)**	**50.71 (9.82)**
Girls	**49.36 (9.50)**	**51.10 (9.23)**	**49.71 (9.54)**
Grade 9			
Boys	**54.30 (10.04)**	**52.36 (10.21)**	
Girls	**52.96 (9.87)**	**54.57 (9.65)**	

Table 3.3 Regression discontinuity models conducted to test the standards of regression discontinuity analyses. Significant results are bolded.

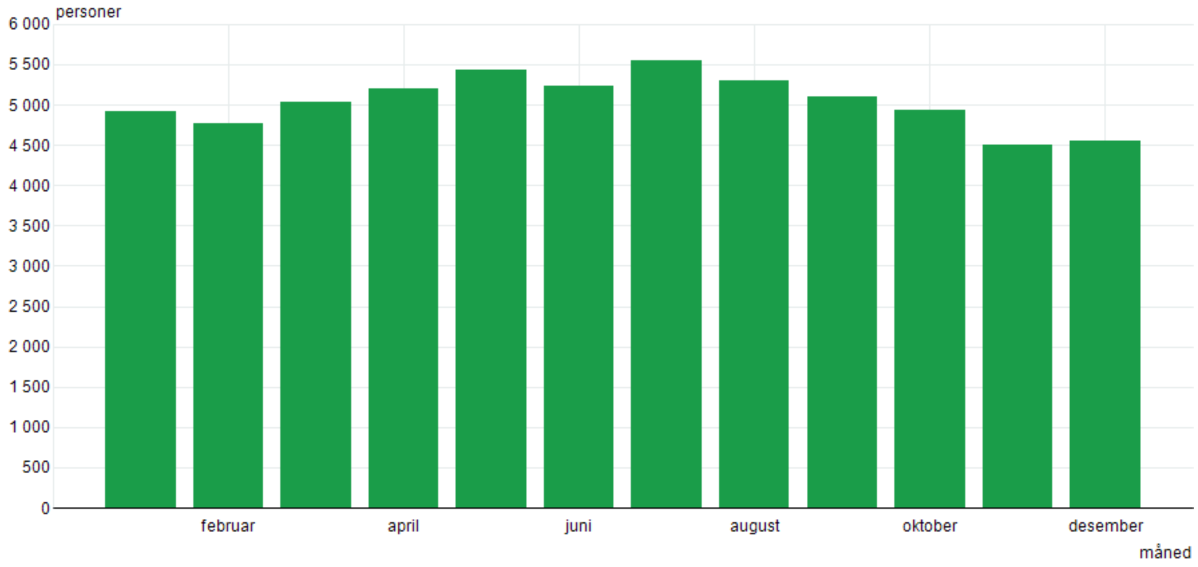
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
Intercept	**51.409(0.08)*** **	**51.291(0.08)*** **	**50.207(0.09)*** **	**52.454(0.09)*** **	**51.799(0.09)*** **	**50.619(0.09)*** **	**44.369(1.41)*** **	**41.584(1.42)*** **
Birth month	**0.235(0.01)*** **	**0.219(0.01)*** **	**0.212(0.01)*** **	**0.228(0.01)*** **	**0.201(0.01)*** **	**0.211(0.01)*** **	**0.205(0.02)*** **	**0.199(0.02)*** **
Grade	**0.795(0.11)*** **	**1.165(0.11)*** **	**0.857(0.16)*** **	**0.729(0.15)*** **	**1.271(0.16)*** **	**1.055(0.163)*** **	**0.871(0.16)*** **	**1.208(0.16)*** **
Birth month x Grade	-0.031(0.01)	-0.024(0.01)						
R-squared	**0.035*** **	**0.038*** **	**0.034*** **	**0.039*** **	**0.038*** **	**0.038*** **	**0.013*** **	**0.016*** **
Bandwidth	[-11.5 :11.5]	[-11.5 : 11.5]	[-11.5 : 11.5]	[-11.5 : 11.5]	[-11.5 : 11.5]	[-11.5 : 11.5]	[-5.5 : 5.5]	[-5.5 : 5.5]
N	117,845	118,051	59,612	58,233	59,864	58,187	58,834	58,926

*Model 1 represents the RD-model with an interaction term RAE x Grade in reading. Model 2 represents the RD-model with an interaction term RAE x Grade in numeracy. Model 3 represents the RD-model for boys only in reading, Model 4 represents the RD-model for girls only in reading, Model 5 represents the RD-model for boys only in numeracy, Model 6 represents the RD-model for girls only in numeracy, Model 7 represents the RD-model with alternative bandwidth in reading, Model 8 represents the RD-model with alternative bandwidth in numeracy. Significance codes: ***= p<.001,**= p.<.01 , * = p<.05*

Distribution of children born per month

Grade 5 (Born in 2008)

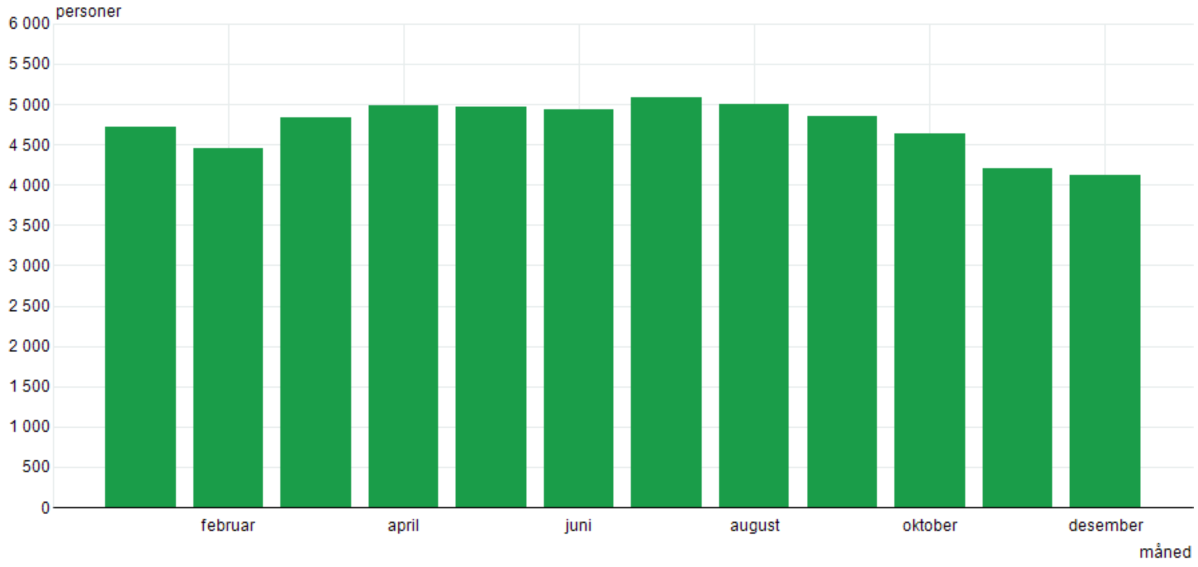
05531: Levendefødte, etter måned. Levendefødte, absolutte tall, 2008.



Kilde: Statistisk sentralbyrå

Grade 8 (Born in 2005)

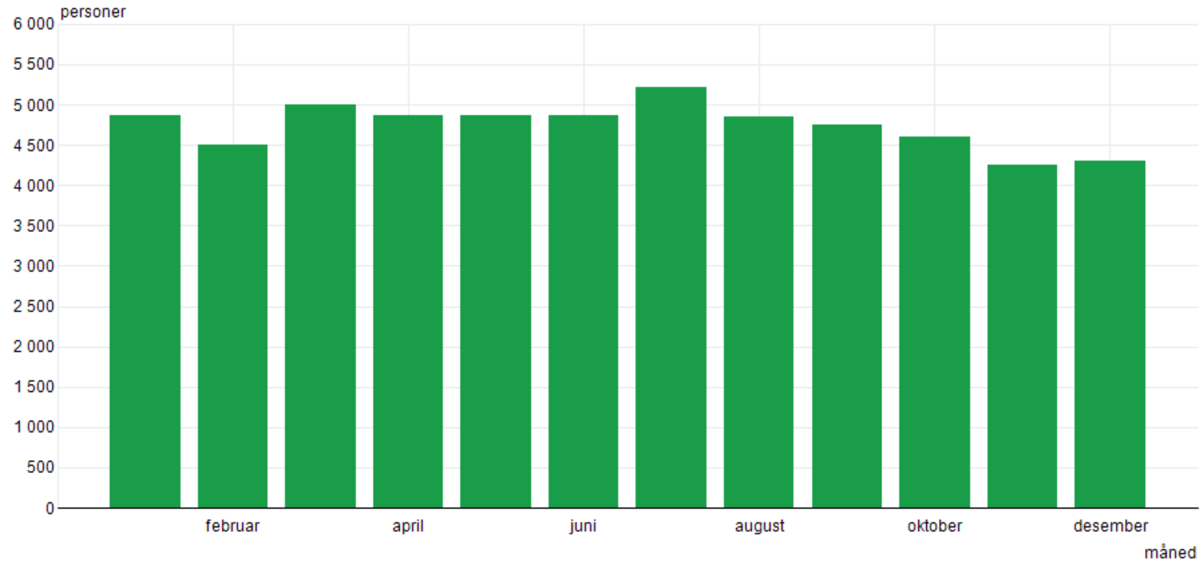
05531: Levendefødte, etter måned. Levendefødte, absolutte tall, 2005.



Kilde: Statistisk sentralbyrå

Grade 9 (Born in 2004)

05531: Levendefødte, etter måned. Levendefødte, absolutte tall, 2004.



Kilde: Statistisk sentralbyrå