To What Extent Does Relative Maturity Affect Test Scores Between Tracked and Untracked Education Systems? Evidence From TIMSS 2019

by

Qi Xuan Khoo

Robert Garlick, Faculty Advisor

Honors Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University

Duke University
Durham, North Carolina
2023

Acknowledgements

I would like to thank Professor Robert Garlick for his guidance and mentorship throughout this journey. His knowledge of and passion for development economics were invaluable during the research and writing of this thesis. I would also like to express gratitude to fellow undergraduate researchers of the Duke Economic Analytic Lab with whom I have discussed this project and from whom I have learnt a lot about the art of economics research. Lastly, I would like to thank my parents who are the only reason I am who I am today. Soli Deo gloria.

Abstract

Most education systems enforce a cutoff birth date for school entry, and some group students based on their perceived ability—a practice known as tracking. While the former policy leads to maturity gaps among early learners, the concomitant performance gaps may or may not be exacerbated by the latter. Analyzing the Trends in International Mathematics and Science Study (TIMSS) 2019 dataset to study how relative maturity affects test scores with tracking, this paper finds that older students outperform their younger peers. This relative maturity test score premium is accentuated by tracking, and these effects are found to be more significant in mathematics than in science.

JEL classification: I2, I24, I28

Keywords: Education and Inequality, Education and Research Institutions, Government Policy

1 Introduction

While there are features of the modern education system that vary distinctively by country, two policies remain somewhat common: a single annual school entry cutoff birth date, which generates at least an eleven-month age range within each grade cohort, and ability-based sorting into curriculum groups—a practice commonly known as tracking. The school entry cutoff birth date determines the relative maturity of students in the same cohort; students born right before the cutoff date may be eleven months younger than their peers born just after the cutoff. On a similar note, tracking may lead to performance gaps in certain subject areas as a result of different specialization and peer group effects.

The relative age effect was perhaps first popularized in public discourse by Malcolm Gladwell who demonstrated the prevalence of birthdays in the first quarter of the year among elite Canadian hockey players (Gladwell, 2008). Since a decade ago, redshirting, the delaying of kindergarten entry by a year, has become increasingly popular in the U.S. (Reeves, 2022), especially among white male students from a higher socioeconomic background (Bassok and Reardon, 2013). If relative age effects are significant in early primary grades but diminish with age, this phenomenon may not be of particular importance for the economy. However, if they persist through the human capital accumulation process to higher levels of education, they may have important implications for future adult outcomes and productivity. Motivated by a similar concern for optimizing the rate of human capital accumulation, the main rationale behind tracking has been the potential gains from specialization; in theory, by teaching a more practical and vocational curriculum in the lower track and a theory-orientated academic curriculum in the higher track, tracking increases students' specialization in different skill sets. Yet, this may translate into gaps in learning outcomes for universal subjects such as science and mathematics between students of different tracks. Hanushek and Woessmann (2007) discovered that relative inequality in test scores increases with tracking and decreases without tracking across 18 countries by analyzing data from the Trends in International Mathematics and Science Study (TIMSS) 1995 - 2003 and the Program for International Student Assessment (PISA) 2003. However, current literature has yet to arrive at a consensus on the definitive effects of tracking on performance inequality among students, as discussed in the next section. While the current literature on relative age and tracking have shed light on their respective effects on student outcomes, the combined effects of relative age and tracking on student outcome along with their relationship remain unclear.

The challenge of identifying the causal impact of relative maturity and tracking on later outcomes lies in the fact that age enters into educational decisions in at least four important ways. Firstly, relative age is only determined by school cutoff dates if the rules are strictly followed; this means that students in a sample are not necessarily a random draw. To distinguish observed relative age from the relative age at which a student should be observed based on their birth date relative to the school cutoff date, the latter measure is referred to as assigned relative age. Secondly, students who are young are more likely to repeat a grade, resulting in omitted variable bias that might lead to downward bias in reduced-form coefficient estimates; it is likely that students who are relatively younger perform better during the retention year as they catch up on their studies. Observing their academic performance alongside other students who did not go through grade retention in the same cohort may attenuate the true performance gap attributable to relative maturity. Thirdly, relative maturity may at least partially determine academic program placement during primary school. Finally, relatively older students will be more mature at young ages and hence score higher on achievement tests, independent of program placement.

While observed relative age is rather endogenous, a student's birth, and hence birth month, is arguably random and hence exogenous. The impact of assigned relative age on test scores reflects both differential school entry and grade retention or failure across the assigned relative age distribution, as well as differences in program placement and skill acquisition. Given that both observed age and assigned age are known, this paper estimates the causal impact of relative age using assigned relative age as an instrument for observed age while controlling for school fixed effects for within-grade samples. Overall, this paper found evidence for substantial relative maturity effects on test scores among grade 4 students, but arguably small age-based performance differential among grade 8 students: the oldest grade 4 students are expected to score 13 and 9 percentiles higher than the youngest members of the cohort in mathematics and science respectively, whereas the oldest grade 8 students are expected to score 5 percentiles higher than their youngest peers in both subjects.

As only a few countries participating in TIMSS employ "clean" education systems where essentially all children enter on time and pass from one grade to the next on schedule, as evidenced by high first-stage coefficient estimate, results from the clean countries are compared with the rest of the countries to arrive at a more robust estimate of the relative age effect. The estimated relative age effects on test scores are greater in these "clean" countries: the oldest grade 4 students here are expected to score 16 and 14 percentiles higher than their youngest peers in mathematics and science, whereas the percentile premiums are similar as before among grade 8 students.

Using information on the tracking policy of each country from the PISA-OECD database, the interaction effect of tracking and relative maturity on test scores is captured by an interaction term, with country- and school-level differences controlled for through school fixed effects. In addition, since no country tracks students at grade 4, a falsification test against placebo tracking effects in the grade 8 sample using the grade 4 sample can be attempted. Overall, the oldest tracked student in the grade 8 cohort is expected to score around 4-5 percentiles higher than their youngest untracked peers. While the relative maturity estimates are largely consistent with current literature, the relative maturity and the interaction effect estimates between tracking and age on test scores are perhaps a novel contribution of this paper. Using data from 22 countries allows this paper to ascertain the pervasiveness of relative age effects between tracked and untracked education systems.

The remainder of the paper is as follows: Section 2 discusses existing results from current literature. Section 3 describes the econometric framework. Section 4 discusses the data used in the analysis and investigates birth date targeting. Section 5 reports and explores the relative age and tracking estimates for grade 4 and grade 8 samples. Section 6 discusses the limitations of this study and concludes.

2 Literature Review

2.1 Relative Age Effects

Recent studies have consistently found evidence for relative maturity effects on test scores across different settings where older members outperform their younger peers in the same grade cohort: Bedard Dhuey (2006) analyzed data from TIMSS 1995, 1999 and found that the youngest members of each cohort score 4-12 percentiles lower than the oldest members in grade 4 and 2-9 percentiles lower in grade 8. Moreover, studies using large-scale international assessment data and intrastate inter-school data both find that relatively younger male students are more likely to score significantly lower than female schoolchildren of the same relative age (Diris, 2017; Hemelt and Rosen, 2016), and students who delay their enrolment in a school year score significantly higher than schoolchildren enrolled in their corresponding cohort (Dhuey et al., 2019). Using data from PISA, Givord (2020) analyzed the impact of a student's month of birth on cognitive and noncognitive outcomes, showing that the youngest members of a cohort are more likely to perform poorer and experience lower self-confidence. However, longitudinal studies have found that these differences at the onset of primary education are compensated for as students progress through primary school or when they enter secondary school, as relatively younger students have higher levels of concentration (Nam, 2014); at age 12, there are still differences in favor of relatively older students; these differences disappear by age 15 (Pehkonen et al., 2015).

2.2 Relative Maturity and Tracking

Studies on the effects of tracking on student outcomes have largely focused on intrastate experiments: a study in Kenya found evidence that tracking primary-school students by prior achievement increases the test scores of students in high-achievement and low-achievement classes, because homogeneous classrooms allow teachers to focus their teaching at the right level(Duflo et al., 2011). Investigating the opposite direction of a possible relationship between relative maturity and tracking, a study in Austria where students are tracked in grade 5 and subsequently in grade 9 found a strong positive relative-age effect on track choice in grades 5-8 and the persistence of the relative age effect beyond grade 8 for students from lower socio-economic background (Schneeweis and Zweimüller, 2014).

Instead of focusing on the effects of tracking and relative maturity within an institution (school, state or country), this paper attempts to investigate the extent to which relative maturity affects academic performance in the presence of tracking, and the persistence of such effects by comparing two cohorts which are four years apart while controlling for important background factors identified by current literature such as birth month seasonality, parents' education level, home study support etc.

3 Empirical Strategy

The three main variables of interest in this study are test score (left-hand-side variable), age and tracking (right-hand-side variables). Test score is a proxy for learning attainment or education outcome, age is a proxy of relative maturity whereas tracking can be represented by a dummy variable indicating whether the country / education system. While the causal effects of relative age on test score can be estimated through OLS regression, the causal impact of tracking on the effects of relative age on test score can be estimated through a difference-in-differences regression, as explained below.

3.1 Assigned Relative Age as an Instrument

One of the key issues with regards to a regression of test scores on observed age and tracking is that the causal effects of observed age might be confounded by unobservable factors such as nonrandom grade retention or intentionally delayed school entry. There could also be simultaneous causality at play as it is possible that students' poor test scores might cause some parents to hold them out of school for a year to help them catch up with school and increase test scores. If these were true, the OLS estimate of the causal impact of age may be downward biased (as discussed in the results section), since older students who are retained will likely perform poorer than other students in the same cohort, suggesting a plausibly negative parameter coefficient of age. Thus, an instrumental variable approach is proposed to address these issues.

Assigned relative age, defined as the difference in months between a student's birth date and the cutoff date for school entry, is proposed as an instrumental variable for observed age. For assigned relative age to be a consistent estimate, three conditions must be satisfied:

- Relevance: Assigned relative age determines birth month, since it is the difference in months between the school entry cutoff and the student's birth month.
- Exogeneity: Since birth month is random and there is no relationship between the national school entry cutoff date and a student's birth month, assigned relative age is not determined by either test score or birth month, as it is reasonable to assume no relationship between a national school entry cutoff date and any individual student's birth date. A possible violation of this condition might be unobserved effects of birth month seasons, which will need to be explored further in exploratory data analysis.

- Exclusion: Since the school entry cutoff date is an exogenous institutional variable, assigned relative age only affects test score through observed age.

The measure of assigned relative age (R) is defined as the relative difference in months between the school entry cutoff date and the birth month of a student i.e. $R = 0$ for students born in the last eligible month and R = 11 for students born the first eligible month. For instance, if the cutoff date is February 1, January babies are the youngest ($R = 0$) and March babies are the oldest ($R = 11$). Actual age in months (A) is constructed using the test date and birth date, both reported in months.

3.2 Estimating Equations

A simple regression model of the relationship between test scores and observed age can be assumed as such:

$$S_{cgi} = \gamma_0 + \gamma_1 A_{cgi} + \xi \mathbf{X}_{cgi} + \epsilon \tag{1}$$

where S_{cgi} denotes the test score (student outcome) for student i in country c, grade g, A is the observed age (test date - student birth date in months) and ϵ is the error term. $\mathbf{X}_{cgi}$ is a vector of control variables such as the level of home study support, parents' education level and school fixed effects. γ_1 captures the estimated causal impact of observed age on test score.

As outlined above, equation (1) is likely to suffer from endogeneity issues and γ_1 might be downward biased. Hence, an instrumental variables regression using Two-Stage Least Squares is proposed:

$$A_{cgi} = \pi_0 + \pi_1 R_{cgi} + \pi_2 \mathbf{X}_{cgi} + u \tag{2}$$

$$S_{cgi} = \theta_0 + \theta_1 \hat{A}_{cgi} + \theta_2 \mathbf{X}_{cgi} + v \tag{3}$$

where equation (2) is the first stage equation with π_1 capturing an estimate of the strength of R's relevance (R is assigned relative age), and equation (3) is the second stage equation with θ_1 capturing the TSLS estimate of age on test scores. In addition, the reduced form equation on the relationship between test score and assigned relative age may be of interest for key policy implications, with κ_1 capturing the relative maturity effect:

$$S_{cgi} = \kappa_0 + \kappa_1 \hat{A}_{cgi} + \kappa_2 \mathbf{X}_{cgi} + e \tag{4}$$

From equation (3), a difference-in-differences TSLS regression is used to estimate the impact of tracking on the relationship between test score and relative age:

$$S_{cgi} = \beta_0 + \beta_1 \hat{A}_{cgi} + \beta_2 T_c + \beta_3 T_c \hat{A}_{cgi} + \beta_4 \mathbf{X}_{cgi} + \epsilon \tag{5}$$

where T_c is a dummy variable that encodes whether students have undergone tracking before grade g in country c when they sat the test. The estimated difference between tracked and untracked students is captured by β_2 whereas the estimated difference in relative age effects on test score between tracked and untracked students is captured by β_3. Note that since the schools are identified in the sample, country fixed effects are captured by school fixed effects, and a variant of equation (4) with school fixed effects is:

$$\boxed{S_{cgi} = \alpha_c + \tau_1 \hat{A}_{cgi} + \tau_2 T_c \hat{A}_{cgi} + \tau_3 \mathbf{X}_{cgi} + \epsilon} \tag{6}$$

Equation (6) will be the main equation estimated in this study to investigate the relationship between relative maturity and test scores in tracked and untracked education systems. The reduced-form equation (4) is estimated via OLS quantile regression in a further analysis of the breakdown of this relationship for students who scored in different test score percentiles.

4 Data

The data used in this study are obtained from two large-scale international assessments—the 2019 Trends in International Mathematics and Science Study (hereafter TIMSS) and the 2018 Program for International Student Assessment (hereafter PISA). TIMSS—the main dataset—provides nationally representative mathematics and science achievement results for students in grades four and eight, in addition to students' age, sex and background control variables. Given that TIMSS did not collect information on country-level tracking and schooling policies, PISA is used as a complementary source for data on country-level tracking, first grade at tracking and school entry cutoff birth date.

Given the focus of this study on relative maturity, tracking and test scores, only countries that enforce an official school entry cutoff birth date are considered. Out of the 64 participating and 8 benchmarking countries surveyed by TIMSS, only 38 matched this criterion with complete data on school entry cutoff birth date and first grade at tracking from the PISA dataset. Thus, the final sample consists of 22 and 34 countries for grade 8 and grade 4 samples respectively.

Subsetting the student sample based on the selected countries, the total number of grade 8 and grade 4 students who participated in TIMSS are 115,917 and 172,941 respectively. Among grade 8 students, 64 did not report their birth date and 37 did not report their sex whereas for grade 4 students, the numbers are 142 and 91 respectively. Considering only students with complete birth date and sex data, 67 8th graders and 145 4th graders were omitted, and the only grade 4 students from the grade 8 sample countries are included. Thus, the final sample consists of **115,850** grade 8 students and **102,553** grade 4 students. The following sections discuss how the variables of interest are measured and constructed.

Table 1: Summary Statistics - grade 8

Statistic	N	Mean	St. Dev.	Min	Max
Observed Age	115,850	170.193	7.800	109	228
Assigned Relative Age	115,850	5.502	3.462	0	11
Birth Month	115,850	6.481	3.436	1	12
Sex (F=1, M=0)	115,850	0.506	0.500	0	1
1st Math PV	115,850	500.169	103.766	90.682	905.724
1st Science PV	115,850	497.736	107.624	18.567	863.029
Average Math PV	115,850	500.494	101.982	152.013	856.873
Average Science PV	115,850	497.920	103.959	107.047	829.914

Initial Sample Size = 115,917

Table 2: Summary Statistics - grade 4

Statistic	N	Mean	St. Dev.	Min	Max
Observed Age	102,553	122.100	6.877	76	180
Relative Age	102,553	5.491	3.480	0	11
Birth Month	102,553	6.565	3.433	1	12
Sex (F=1, M=0)	102,553	0.507	0.500	0	1
1st Math PV	102,553	523.330	99.308	89.430	851.946
1st Science PV	102,553	510.86	97.367	16.610	861.779
Average Math PV	102,553	510.445	96.273	181.600	811.367
Average Science PV	102,553	509.900	94.084	57.860	802.862

Initial Sample Size = 172,941

4.1 Key Variables

4.1.1 Test Scores

The test scores in TIMSS are reported as a result of imputation, not raw scores; each student is associated with five plausible values for their performance in mathematics and another five for science. Students were randomly given only two out of eight assessment booklets, one for each mathematics and science, and the questions they did not answer are considered Missing Completely At Random due to the random assignment of the booklets. Multiple imputation is then applied to obtain the five plausible values (hereafter PVs) for each subject—the final score that the student would have obtained had they been tested on all questions. Hence, the PVs are random draws from a posterior distribution and are computed using information on students' performance and their characteristics, including the school attended and the average scores obtained by the other pupils.

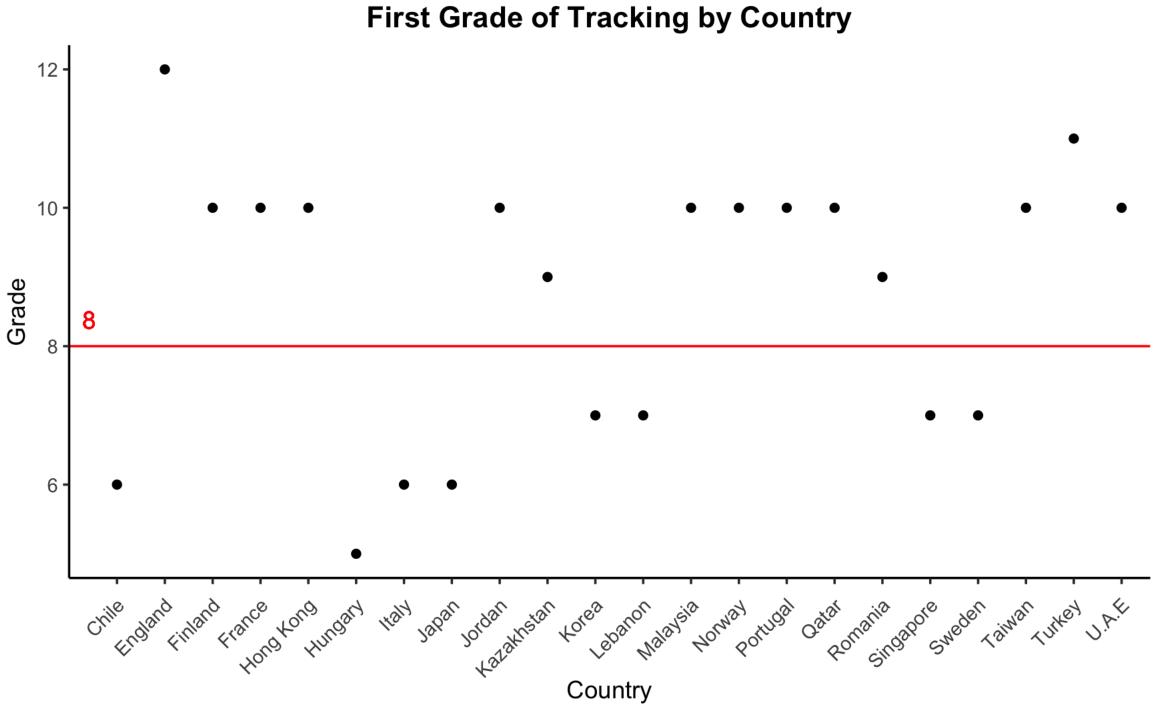
The regression models specified in the previous section are estimated for each of the 5 plausible values for both subjects, and the coefficient estimates reported in this study are pooled from these results. The standard errors of these pooled estimates are computed to also include the within-imputation and between-imputation variances with the weights provided in the TIMSS dataset, based on the Rubin's Rule for Multiple Imputation.

4.1.2 Age

For ease of measuring relative maturity among students, age is measured in months through two key statistics—observed age and assigned relative age. A student's observed age is computed by taking the difference in months between the test date and their birth date. To isolate the relationship between a student's age and test scores that are unconfounded by unobserved variables, assigned relative age is introduced as an instrument that measures a student's relative age through the difference in months between their birth month and their country's school entry cutoff birth date. In the case of a student who was born in January, their assigned relative age would be 11 in France, where the cutoff birth date is Dec. 31, and 0 in Malaysia, where the cutoff birth date is Jan. 2.

4.1.3 Tracking

Country-level tracking data from PISA are encoded as dummy variables, and the details. Since none of the countries tracked students before grade 4, the *has_tracking* dummy encodes whether a country introduces tracking at any grade throughout the public education system for the grade 4 sample. Estimating the grade 4 coefficient estimate for tracking provides a falsification test against possible omitted variable bias that might confound tracking coefficient estimates at grade 8, since any evidence found for the effects of tracking on test scores at grade 4 level are merely "placebo effects". Among all countries that participated in grade 8 TIMSS, 14 track students after grade 8 and 8 before, as shown in figure below.

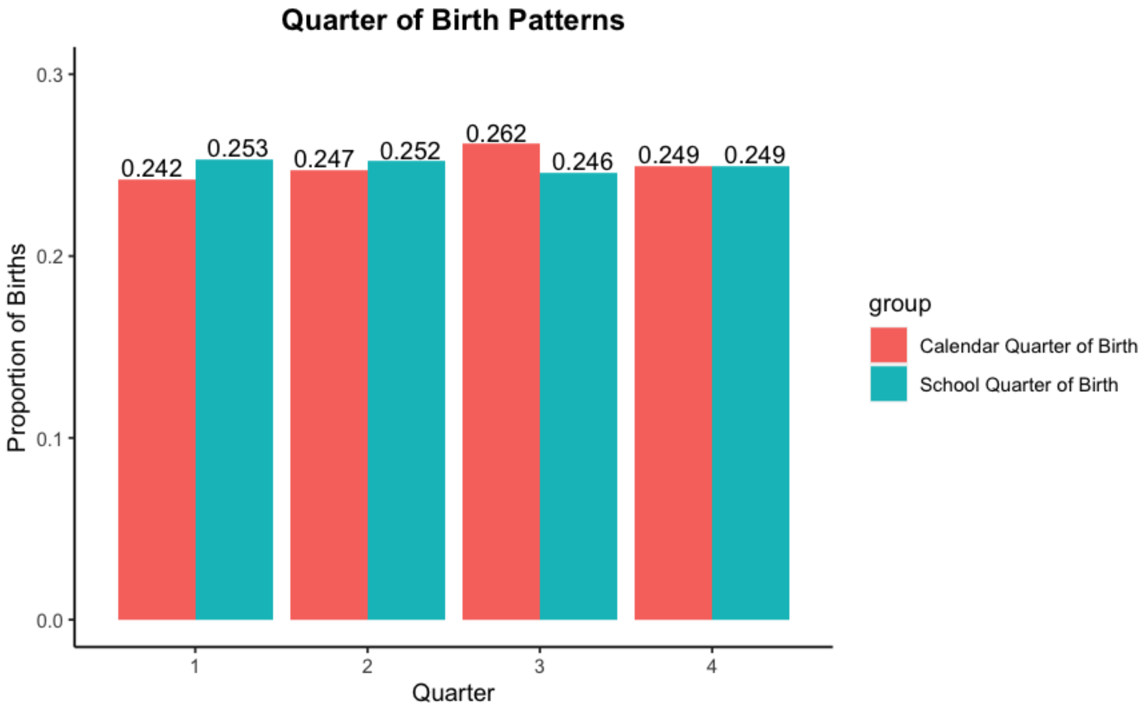


4.1.4 Control Variables

Sex, parents' highest education level and the amount of home study support are control variables included in the analysis of both grade 4 and grade 8 samples. Sex is a dummy variable for male or female, and parents' highest education level is encoded as a categorical variable—university or higher, post-secondary but not university, upper secondary, lower secondary, some primary, lower secondary or no school. The amount of home study supports is encoded as a categorical variable that indicates different levels of sufficiency. Both parents' highest education level and home study support data were collected in the TIMSS student home context questionnaire.

In addition to the common control variables, the analysis of the grade 4 sample also included a categorical variable for the amount of home resources and a dummy for preschool education; the former indicates whether a student has many, some or few resources whereas the latter indicates whether a student has attended preschool education. This data point was collected in the home context questionnaire.

4.2 Assigned Relative Age—Random or Planned?



Before delving into the results of this study, it would be instructive to examine the potential endogeneity of relative age due to birth date targeting by parents aimed at ensuring that their child is among the oldest in class. The figure above illustrates the fraction of all students and across Grades 4 and 8 born in each calendar and school quarter (i.e. January - March is Quarter 1 and October - December is Quarter 4).While most births occur in the third calendar quarter (June - Aug), births seem to be rather evenly distributed over the school quarters, with the first two being the slightly more popular season. Strictly speaking, one would expect parents who target birth date to have a somewhat higher fourth school quarter birth rate than parents who do not target age at school entry. While the observations here seem to provide some supporting evidence for the randomness of relative age within the sample, seasonality makes detecting birth date targeting difficult.

5 Results

5.1 First Stage Results

Table 3: First Stage Coefficient Estimates

Country	grade 8	grade 4[1]
Pooled	**0.460***** (0.115)	**0.449***** (0.032)
Norway	0.980*** (0.019)	0.981*** (0.004)
England	0.953*** (0.017)	0.956*** (0.013)
Korea	0.935*** (0.007)	0.980*** (0.005)
Sweden	0.929*** (0.014)	0.918*** (0.011)
Finland	0.913*** (0.012)	0.929*** (0.011)
France	0.835*** (0.019)	0.913*** (0.013)
Hong Kong	0.784*** (0.035)	0.641*** (0.029)
Italy	0.615*** (0.023)	0.594*** (0.017)
Chinese Taipei	0.574*** (0.013)	0.543*** (0.014)
Singapore	0.554*** (0.019)	0.533*** (0.013)
Malaysia	0.534*** (0.013)	N/A
Japan	0.526*** (0.013)	0.556*** (0.013)
Jordan	0.439*** (0.020)	N/A
Hungary	0.225*** (0.024)	0.080*** (0.025)
Lebanon	0.209** (0.045)	N/A
Qatar	0.151*** (0.038)	0.206*** (0.025)
Chile	0.113*** (0.038)	0.277*** (0.028)
Kazakhstan	0.087*** (0.028)	0.065*** (0.025)
Turkey	0.074** (0.022)	−0.127*** (0.011)
Portugal	−0.006 (0.042)	0.077*** (0.024)
U.A.E	−0.085*** (0.014)	−0.111*** (0.024)
Romania	−0.106*** (0.029)	N/A
Significance Levels	*p<0.1; **p<0.05; ***p<0.01	

Apart from birth month targeting by parents, the strength and exogeneity of assigned relative age may be undermined by unobserved grade retention or delayed entry. If grade retention or delayed entry were non-existent, one would expect the first-stage coefficient estimate to be close to 1. Table 3 reports the cross-country (pooled) and country-specific first-stage coefficient estimates of π_1 from Equation (2). Assigned relative age appears to be a strong instrument for observed age, with statistically significant coefficient estimates of 0.460 and 0.449 for grade 8 and grade 4 samples

[1] grade 4 coefficient estimates are only displayed for grade 8 countries. Pooled grade 4 estimate includes all grade 4 countries

respectively. However, the fact that the estimates are much less than 1 indicates that the overall mapping of assigned relative age to observed age is not exact, showing evidence of grade retention or delayed entry in some countries.

On one hand, the breakdown of the first-stage coefficient estimates by country in Table 3 shows that grade retention or delayed entry might in fact be prevalent in countries with small coefficient estimates. Rather unexpectedly, the grade 8 estimates for Portugal, U.A.E. and Romania are small and negative. As shown in Table 4, a cross-tabulation of the observed and relative ages of grade 8 students from Romania, the country with the most negative estimate, shows two countervailing trends at play: most students in the first twelve-month cohort, from 165 to 176 months old, have $R > 6$ whereas most students in the second twelve-month cohort, from 177 to 188 months old, have $R < 6$. This suggests that relatively younger grade 8 students might be more likely to experience grade retention than older members of the same cohort in this sample, thus leading to the confounding of the positive correlation between observed age and assigned relative age—possibly negative first-stage estimate.

Table 4: Cross Tabulation of Observed Age vs. Relative Age

Romania (grade 8)[2]

Assigned Relative Age (months)

Observed Age (months)	0	1	2	3	4	5	6	7	8	9	10	11
120	0	0	0	1	0	0	0	0	0	0	0	0
						…						
164	2	0	0	0	0	0	0	0	0	0	0	17
165	30	1	0	0	0	0	0	0	0	0	0	0
166	0	32	3	0	0	0	0	0	0	0	0	0
167	0	0	49	4	0	0	0	0	0	0	0	0
168	0	0	0	50	8	0	0	0	0	0	0	0
169	0	0	0	1	81	10	0	0	0	0	0	0
170	0	0	0	0	1	100	9	0	0	0	0	0
171	0	0	0	0	0	2	116	9	0	0	0	0
172	0	0	0	0	0	0	5	149	17	0	0	0
173	0	0	0	0	0	0	0	0	210	21	0	0
174	0	0	0	0	0	0	0	0	5	283	23	0
175	0	0	0	0	0	0	0	0	0	2	299	19
176	18	0	0	0	0	0	0	0	0	0	2	319
177	310	24	0	0	0	0	0	0	0	0	0	3
178	0	348	22	0	0	0	0	0	0	0	0	0
179	0	7	292	13	0	0	0	0	0	0	0	0
180	0	0	3	264	20	0	0	0	0	0	0	0
181	0	0	0	4	252	22	0	0	0	0	0	0
182	0	0	0	0	2	244	12	0	0	0	0	0
183	0	0	0	0	0	3	170	9	0	0	0	0
184	0	0	0	0	0	0	3	182	4	0	0	0
185	0	0	0	0	0	0	0	0	81	1	0	0
186	0	0	0	0	0	0	0	0	0	52	2	0
187	0	0	0	0	0	0	0	0	0	0	29	3
188	2	0	0	0	0	0	0	0	0	0	0	23
						…						
220	0	0	0	0	0	0	0	2	0	0	0	0

On the other hand, countries with strong and positive first-stage coefficient estimates show evidence of a "clean" education system where virtually all students move on from one grade to

[2] Similar results are found from the cross-tabulation analysis of samples from U.A.E and Portugal.

another on schedule. A cross tabulation of countries grade 8 students' observed age and assigned relative age from countries with negative coefficient estimates shows that the mapping from observed age to assigned relative age is near exact. In particular, Norway, England and Korea appear to have the "cleanest" system with first-stage coefficient estimates of greater than 0.9—a result consistent with current literature (Bedard and Dhuey, 2006;), suggesting that the age rules have been consistent for the past two decades in these countries. In general, the sign of the coefficient estimates for all countries are consistent for both grade 8 and grade 4 samples, indicating a similar unobserved grade retention / delayed entry pattern at both grade levels.

5.2 Estimated Effects of Relative Maturity and Tracking on Test Scores

Since strong first-stage estimates show little evidence of unobserved grade retention / delayed entry, the analysis is replicated with a smaller sample that includes countries with first-stage estimates greater than 0.8—Norway, England, Korea, Sweden, Finland and France. In theory, the estimated effects of relative maturity would be more robust for this sample (hereafter strong sample) as they are much less confounded by unobserved grade retention or delayed entry. Furthermore, given that the strong sample consists of richer developed countries in the full sample, the differences in the estimates for the two samples may also be attributable to inter-group differences.

Therefore, based on the first-stage results, the regression analysis is performed on the full sample and subsequently the strong sample for robustness.

5.2.1 Grade 4

The grade 4 results for the full and strong samples are reported in Table 5. Looking first at relative maturity effects, there is substantial evidence for a performance gap which is more pronounced in mathematics than science: the oldest members of the grade 4 cohort is expected to score 32.912 (11 months × 2.992) and 25.570 (11 months × 1.870) in mathematics and science respectively. Note that the age coefficient estimates for the strong sample are larger, verifying the theoretical prediction that any violation of school entry age rules such as grade retention may exert a downward omitted variable bias on the estimates. As each plausible value is an imputation of the unobservable latent achievement for each student (Aparicio et al., 2021), it is instructive to contextualize these relative age test score premiums in terms of the relative percentile ranking of the youngest and oldest students in the same cohort[3]. Given that the mean math test score (first plausible value) of students with $R = 0$ corresponds to the 44th and 48th percentiles in the full and strong samples respectively, the age coefficient estimates from both samples translate into a 13- and 16-percentile test score premium enjoyed by the oldest students of the cohort. As for science, an 11-month age gap is expected to translate into an 9- and 14-percentile test score premiums for the full and strong sample respectively.

[3] Quantile plots for grade 4 and grade 8 samples are included in the appendix.

Table 5: OLS Coefficient Estimates (grade 4)

	Math	Science	Math	Science
	(1)	(2)	(3)	(4)
Age	2.992***	1.870***	3.185***	2.674***
	(0.168)	(0.179)	(0.155)	(0.146)
Age×$Has\_Tracking$	0.086***	0.765**	-	-
	(0.029)	(0.029)		
Sex (female)	-0.302	-3.374	8.984***	1.142
	(4.225)	(3.921)	(1.087)	(1.002)
Sex (female)×$Has\_Tracking$	5.461	3.694	-	-
	(4.406)	(1.527)		
Intercept	113.825***	165.992***	121.728***	230.547***
	(58.416)	(54.737)	(45.715)	(43.327)
Sample	all	all	strong countries	strong countries
Observations	102,553	102,553	24,113	24,113
Significance Levels			*p<0.1; **p<0.05; ***p<0.01	

Although no country tracks students before grade 4, the dummy variable $Has\_Tracking$—1 if a country has a tracking policy in place—and interaction term $Age \times Has\_Tracking$ are included as a robustness check in the form of falsification; if the coefficient estimates of tracking are statistically significant in grade 4 samples where all students are untracked, it suggests that any evidence found for the effects of tracking at grade 8 are likely to be confounded by country-level omitted variables. Unfortunately, the only country that does not track students throughout the education system is Azerbaijan in the full sample, and all countries in the strong sample practice educational tracking. This limitation meant that the $Has\_Tracking$ coefficient estimates essentially captures marginal differences between grade 4 students in Azerbaijan and other countries in the full sample and fall short of providing evidence for or against any placebo effects of tracking on test scores in the grade 8 sample. The coefficient estimates for sex are consistent with current literature on the stylized fact that male students tend to outperform their female peers in less developed economies but the opposite is true in more developed countries which, in this case, constitute the strong sample.

5.2.2 Grade 8

This subsection analyzes the estimated effects of relative maturity and tracking on test scores based on the main sample—grade 8 students. The first major observation here is that while there is evidence for the effects age on test scores, the magnitude of the relative maturity effects are much less than those in the grade 4 sample; with a statistically significant estimate of 0.887 and 0.797

Table 6: OLS Coefficient Estimates (grade 8)

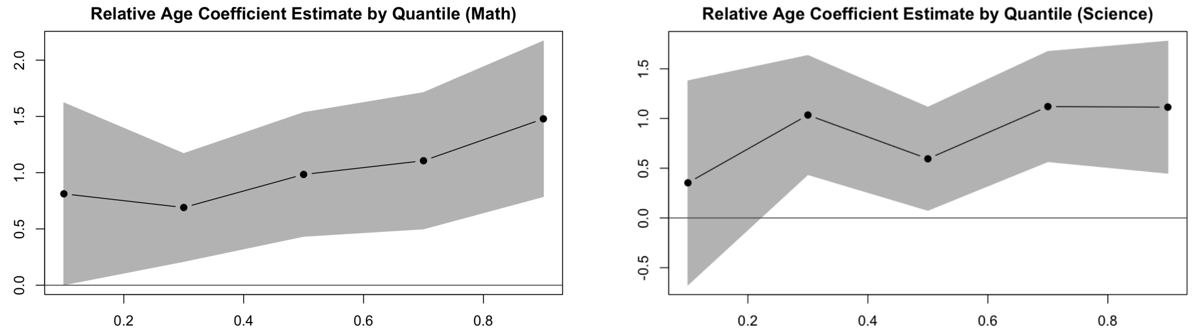
	Math	Science	Math	Science
	(1)	(2)	(3)	(4)
Age	0.887***	0.882***	0.797***	0.901***
	(0.191)	(0.184)	(0.198)	(0.181)
Age×*Tracking*	0.233***	0.112**	0.432***	0.194***
	(0.358)	(0.007)	(0.017)	(0.017)
Sex (Female)	-2.381***	-4.411***	6.456***	1.504
	(0.782)	(0.891)	(0.752)	(0.810)
Sex (Female)×*Tracking*	5.505***	10.588***	0.793	7.219***
	(1.584)	(1.527)	(2.742)	(2.836)
Intercept	345.433***	357.783***	298.009***	269.703***
	(132.696)	(126.734)	(41.201)	(43.150)
Sample	all	all	strong countries	strong countries
Observations	115,850	115,850	24,530	24,530
Significance Levels			*p<0.1; **p<0.05; ***p<0.01	

in the full and strong samples, the oldest grade 8 students are expected to score 9.757 (11 months × 0.887) and 8.767 (11 months × 0.797) points more than their youngest peers in mathematics respectively; the same estimates are 9.702 and 9.911 for science. While the mean math test score (first plausible value) of students with $R = 0$ corresponds to the 50th and 40th percentiles in the full and strong samples respectively, the age coefficient estimates from both samples both translate into a 4-percentile test score premium enjoyed by the oldest students of the cohort, and this percentile premium is the same for science. The performance gaps in both subjects are remarkably smaller than those among grade 4 students. It is also noteworthy that the positive age coefficient estimates which are larger than the reduced form estimates (see Table 7 in Appendix) imply that the omitted variable bias from grade retention or delayed school entry is indeed downward bias, since a student's assigned relative age masks any information about red-shirting or grade retention.

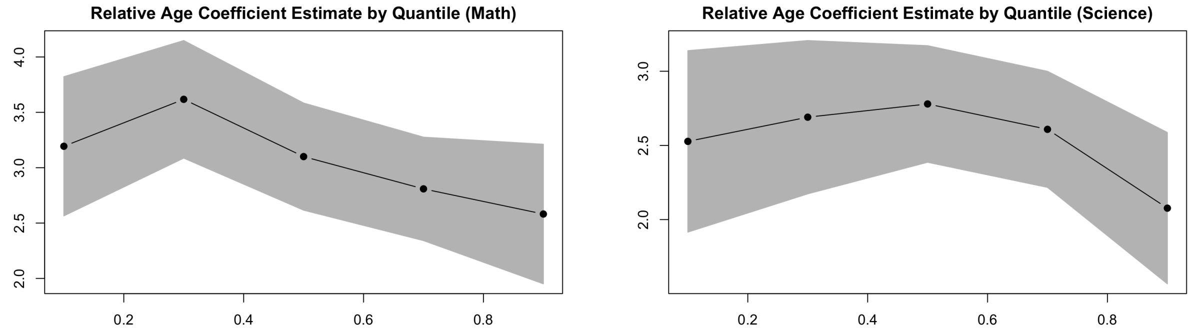
Similar to the estimated age effects on test scores, while there is statistically significant evidence for the interaction effects of tracking and age on test scores, the estimates are negligible: the oldest tracked students are expected to perform 2.563 and 2.178 higher in mathematics than the oldest untracked students in the full and strong samples respectively, and 1.232 and 1.991 respectively in science. To put things in perspective, the oldest tracked members of the grade 8 cohort are expected to score 4 and 5 percentiles higher in mathematics, and 4 percentiles higher in science than the youngest untracked members based on results from the full and strong sample respectively.

Having found substantial evidence of relative maturity effects on test scores in the strong samples, it would be instructive to analyze an age premium breakdown for students in different test score percentiles in the strong sample. Equation (5) is estimated with OLS conditional quantile regressions where the observed age predictor is replaced with assigned relative age, covering the 10th to the 90th quantiles. The coefficient estimates conditional on different quantiles are summarized in the quantile progress graphs below (shaded areas represent 99% confidence intervals).

Quantile Progress Graph for grade 8 (strong sample)



Quantile Progress Graph for grade 4 (strong sample)



To begin with, the relative age coefficient estimates are statistically significant across all quantiles for both grade 4 and grade 8 strong samples, with no evidence for the age-tracking interaction effects across the quantiles. For mathematics, there appears to be a divergence between the quantile relative maturity premium between grade 8 and grade 4: eighth graders who score in the higher quantiles can expect higher relative age test score percentile premium whereas the reverse is true for fourth graders. To put things into perspective, the eleven-month relative age test score premium among eighth graders who score in the 90th percentile is expected to be 7.337 points more than the same premium among eighth graders who score in the 10th percentile. However, the eleven-month relative age test score premium among fourth graders who score in the 90th percentile is expected to be 6.732 points less than the same premium among fourth graders who score in the 10th percentile.

While there seems to be no consistent pattern in the science test score age coefficient quantiles, the same divergence in the trend of the eleven-month test score premium magnitude between the highest and the lowest quantiles is observed between grade 4 and grade 8 strong samples. In general, the effects of relative maturity on test scores are greater for grade 8 students who perform at the top of their cohort, but smaller for high-performing grade 4 students for both mathematics and science.

6 Conclusion

In answering the question of the extent to which relative maturity affects test scores between tracked and untracked education systems, this study found substantial evidence for relative maturity premiums in mathematics and science test scores among grade 4 students, small yet statistically significant relative maturity effects among grade 8 students with positive interaction effects in the presence of tracking.

In terms of relative maturity, the expected difference in test scores between the oldest and the youngest within the grade 4 cohort is greater than that within the grade 8 cohort. The most robust relative age estimates reveal that the oldest members of the grade 4 cohort are expected to gain a 16- and 14- percentile test score premiums in mathematics and science over their youngest peers. The same is true for grade 8 students, except that the eleven-month relative age test score premiums for mathematics and science are about 4 percentiles. In addition, the effects of relative age on test scores are greater among high-performing grade 8 students compared to other peer groups, and smaller among high-performing grade 4 students as compared to students who scored in the lower quantiles. Overall, the relative age effects on test score persist but shrink in magnitude at higher grades, based on the the analysis of grade 4 and grade 8 TIMSS 2019 samples—consistent with the consensus of current literature. Furthermore, the expected performance differential attributable to tracking is statistically significant yet negligible: the expected performance differential between the tracked and untracked oldest grade 8 students is close to 1 percentile. Taking into account both relative maturity and tracking, the oldest tracked grade 8 student is expected to score at least 4 percentiles higher in both mathematics and science than the youngest untracked student.

Key findings aside, this study faced several major limitations that hindered a more robust investigation of the research question. Firstly, the coefficient estimates of relative age and tracking likely suffer from omitted variable bias. Although assigned relative age turns out to be a strong instrument for observed age, the incidence of grade retention in both samples, as revealed by the first-stage results, confounds the regression estimates, as evident in the large standard errors in the full sample analyses for both grades. This bias is, indeed, downward bias, as corroborated by the increase in robust coefficient estimates when only countries with strong first-stage estimates are included in the analysis.

Secondly, the estimated interaction effects of tracking and relative maturity are likely confounded by omitted variable bias, as mentioned above. Given that the tracking dummy essentially estimates the average difference in test scores between countries that track students before grade 8 and those that track after grade 8, any unobserved factors shared by either group of countries will confound the estimates. For instance, if countries that track students before grade 8 in the sample coincidentally have better science and mathematics curricula and collectively spend more on education expenditure than countries that track students after grade 8, the tracking estimates will suffer from a downward bias. The bias will be upward for the vice versa case. Given the positive tracking-age coefficient estimates found in this study, the former is likely to be true, and future studies should include stronger controls or a more robust measure of tracking to better estimate the effects of interest.

Lastly, it is likely that measurement errors exist in the variables of interest measured in this study. Given that the tracking information is measured at the country level, it is possible that not all schools surveyed in TIMSS 2019 adhere to the tracking policy, causing possible mislabeling of untracked students as tracked and vice versa. In addition, as shown in the results between the full and strong samples, measurement errors in relative maturity, in the form of grade retention and delayed entry, introduce a downward bias in the estimates. Although these errors have been addressed by restricting the sample to only countries with the least amount of grade retention, the estimates would have been more robust and accurate had all students were included in the analysis.

References

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. Biometrika, 86, 948-955

Bassok, D., & Reardon, S. (2013). Academic redshirting in kindergarten: Prevalence, patterns, and implications. Educational Evaluation and Policy Analysis, 35, 283–297.

Campbell, T. (2014). Stratified at seven: in-class ability grouping and the relative age effect. British Educational Research Journal, 40(5), 749–771. http://www.jstor.org/stable/43297615

Cook, Philip J. & Kang, Songman, 2020. "Girls to the front: How redshirting and test-score gaps are affected by a change in the school-entry cut date," Economics of Education Review, Elsevier, vol. 76(C).

Crawford, C., Dearden, L., & Greaves, E. (2014). The drivers of month-of-birth differences in children's cognitive and non-cognitive skills. Journal of the Royal Statistical Society: Series A (Statistics in Society), 177(4), 829–860

Dhuey, Elizabeth & Bedard, Kelly. (2006). The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects. The Quarterly Journal of Economics. 121. 1437-1472. 10.1162/qjec.121.4.1437

Dhuey, E., Figlio, D., Karbownik, K. & Roth, J. (2019), School Starting Age and Cognitive Development. J. Pol. Anal. Manage., 38: 538-578. https://doi.org/10.1002/pam.22135

Duflo, E., Dupas, P., Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. The American Economic Review, 101(5), 1739–1774. http://www.jstor.org/stable/23045621

Elizabeth U. Cascio & Diane Whitmore Schanzenbach. (2016). First in the Class? Age and the Education Production Function. Education Finance and Policy; 11 (3): 225–250. doi: https://doi.org/10.1162/EDFP a_00191

Givord, P. (2020). How a student's month of birth is linked to performance at school: new evidence from PISA. OECD Education Working Paper No. 221

Gladwell, Malcolm. (2008). Outliers : the story of success. New York :Little, Brown and Company

Hanushek, E. & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. The Economic Journal (Royal Economic Society), 116, C63 - C76

Hemelt, S. & Rosen, R. (2016). School Entry, Compulsory Schooling, and Human Capital Accumulation: Evidence from Michigan . The B.E. Journal of Economic Analysis Policy, 16(4), 20150219. https://doi.org/10.1515/bejeap-2015-0219

Aparicio, Juan, Jose M. Cordero, and Lidia Ortiz. 2021. ”Efficiency Analysis with Educational Data: How to Deal with Plausible Values from International Large-Scale Assessments” Mathematics 9, no. 13: 1579. https://doi.org/10.3390/math9131579

Meier, Volker & Schütz, Gabriela. (2007). The economics of tracking and non-tracking. ifo Working Paper Series No. 50, ifo Institute - Leibniz Institute for Economic Research at the University of Munich

Nam, K. (2014). Until when does the effect of age on academic achievement persist? Evidence from Korean data. Econ. Educ. Rev., 40, 106–122.

Page, L., Sarkar, D. & Silva-Goncalves, J. (2019). Long-lasting effects of relative age at school. Journal of Economic Behavior Organization, Vol. 168, 166-195

Pehkonen, J,; Viinikainen, J., Böckerman, P., Pulkki-Råback, L., Keltikangas-Järvinen, L. & Raitakari, O. (2015). Relative age at school entry, school performance and long-term labour market outcomes. Applied Economics Letters, 22, 1345–1348.

Reeves, Richard. (2022). Who Redshirts?. Brookings Institution Publisher

Ron Diris. (2017). Don’t Hold Back? The Effect of Grade Retention on Student Achievement. Education Finance and Policy; 12 (3): 312–341. doi: https://doi.org/10.1162/EDFP_a_00203

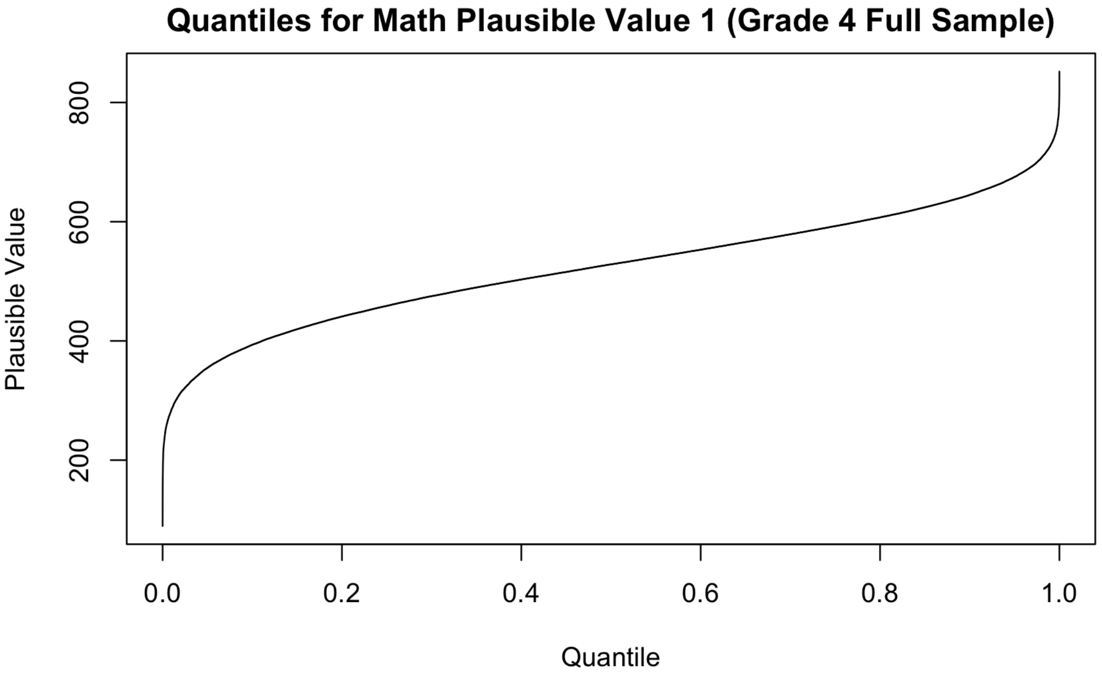
Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Hoboken, NJ: Wiley

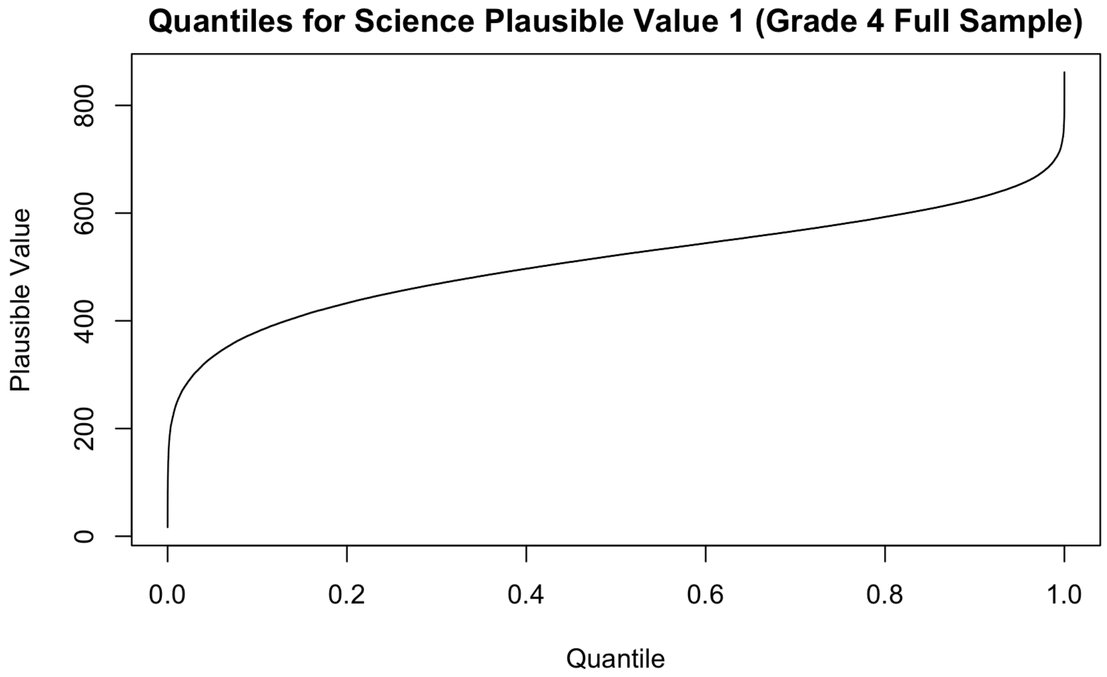
Schneeweis, N., & Zweimüller, M. (2014). Early Tracking and the Misfortune of Being Young. The Scandinavian Journal of Economics, 116(2), 394–428. http://www.jstor.org/stable/43673672

7 Appendix

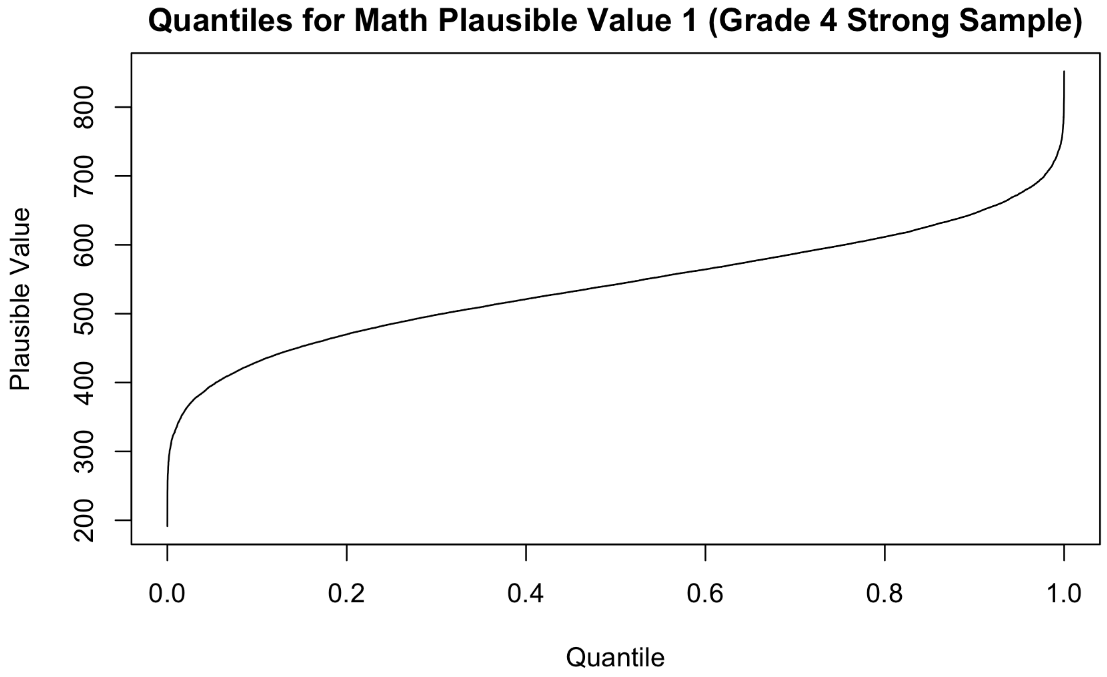
Table 7: OLS Reduced-form Coefficient Estimates (grade 8)

	Math	Science	Math	Science
	(1)	(2)	(3)	(4)
Age	-1.110***	-0.338***	-1.168***	-0.110***
	(0.117)	(0.102)	(0.197)	(0.184)
Age×*Tracking*	4.291***	2.178**	7.926***	3.865***
	(0.153)	(0.120)	(0.253)	(0.254)
Sex (Female)	-9.005***	-7.548***	-1.627***	-2.378
	(0.749)	(0.847)	(1.422)	(1.425)
Sex (Female)×*Tracking*	22.956***	19.186***	34.533***	23.655***
	(1.584)	(1.169)	(2.200)	(2.089)
Intercept	504.840***	508.010***	441.473***	430.523***
	(130.033)	(122.505)	(35.819)	(36.483)
Sample	all	all	strong countries	strong countries
Observations	115,850	115,850	24,530	24,530
Significance Levels				*p<0.1; **p<0.05; ***p<0.01

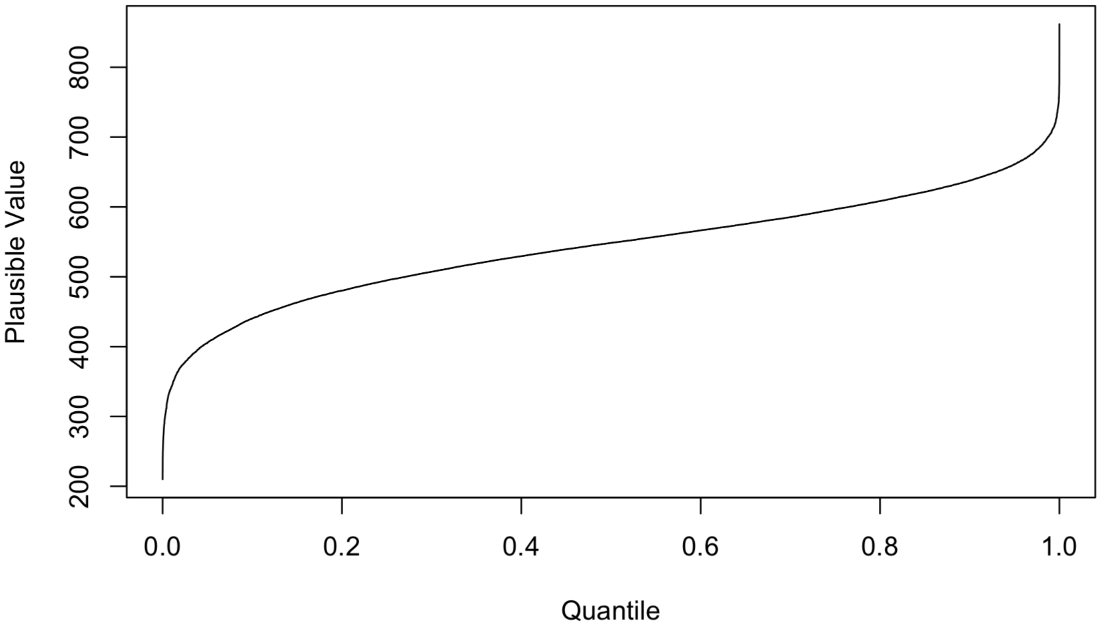




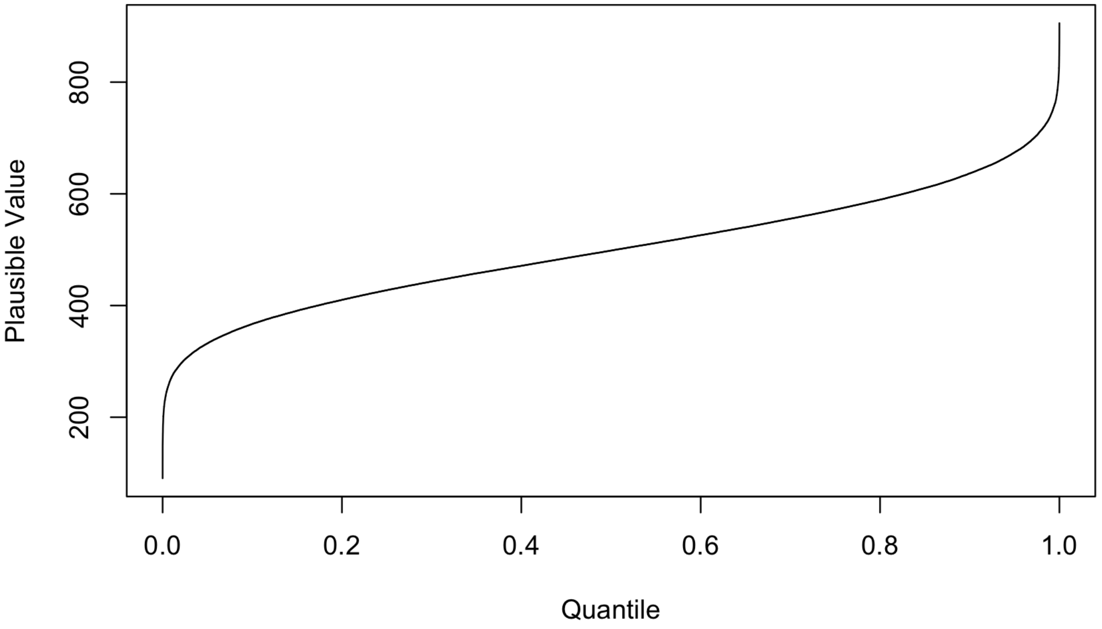
Quantiles for Science Plausible Value 1 (Grade 4 Full Sample)



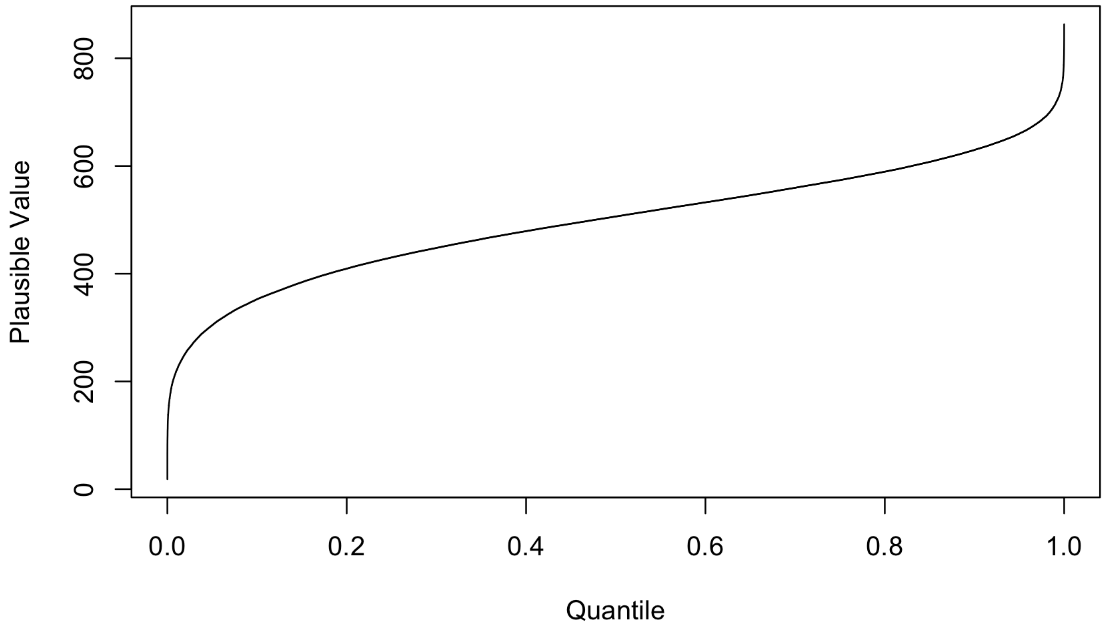
Quantiles for Math Plausible Value 1 (Grade 4 Strong Sample)

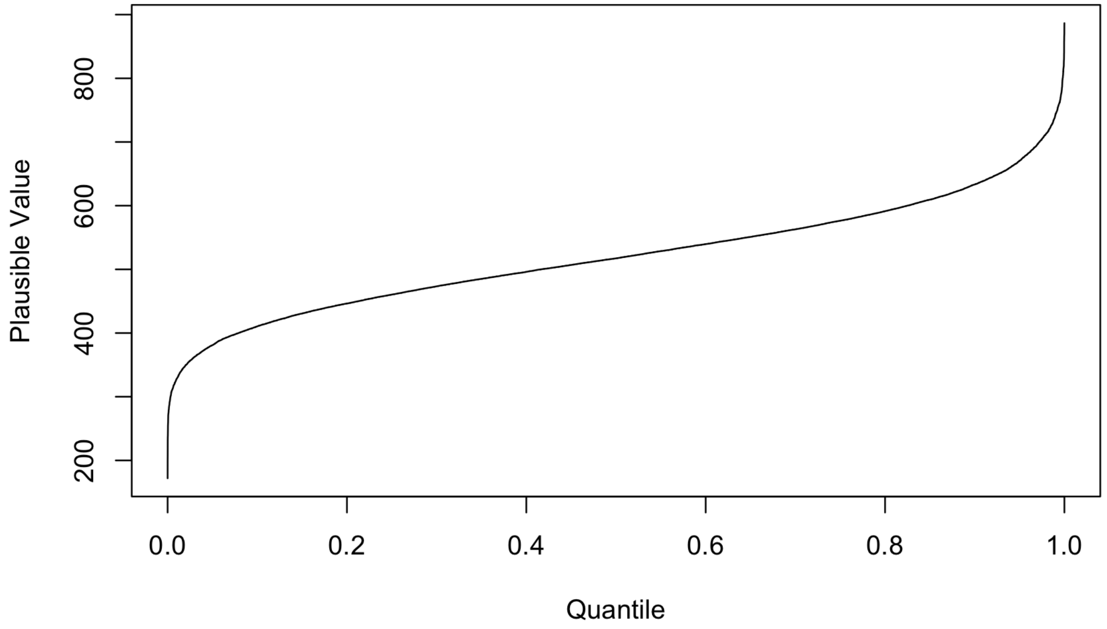


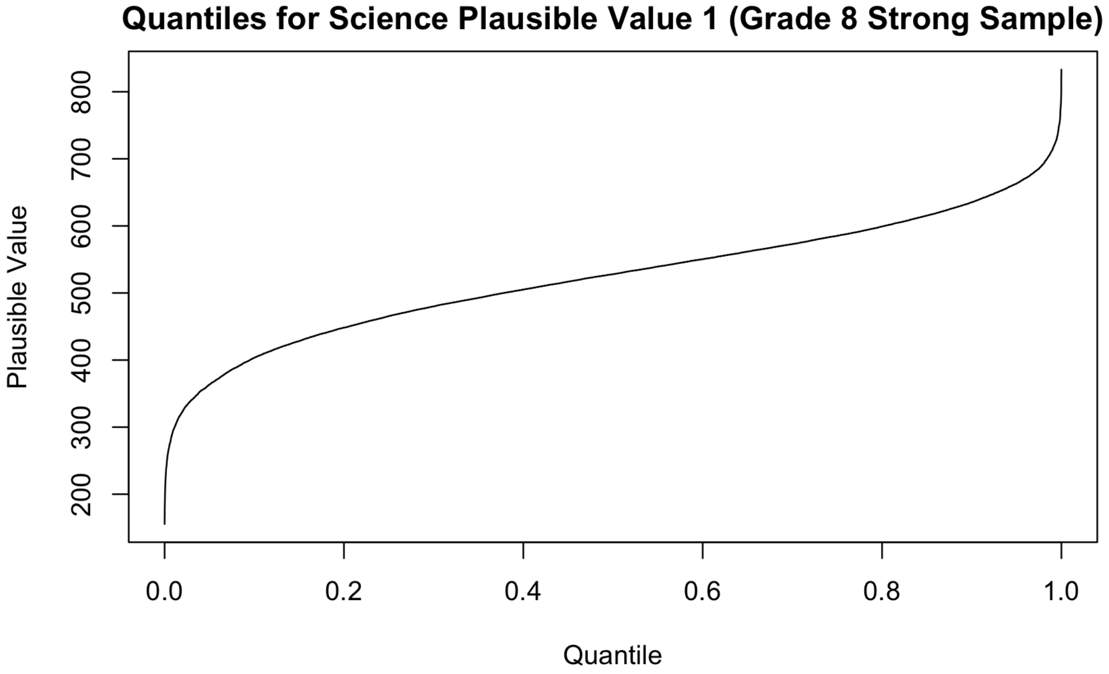
Quantiles for Science Plausible Value 1 (Grade 4 Strong Sample)



Quantiles for Math Plausible Value 1 (Grade 8 Full Sample)


Quantiles for Science Plausible Value 1 (Grade 8 Full Sample)


Quantiles for Math Plausible Value 1 (Grade 8 Strong Sample)


Quantiles for Science Plausible Value 1 (Grade 8 Strong Sample)
Plausible Value
200
300
400
500
600
700
800
Quantile
0.0
0.2
0.4
0.6
0.8
1.0