Erasmus School of Economics

The Effects of Ordinal Class Ranks on Student Performance and Motivation

Benjamin Alessandro Dahmen
456352 — Erasmus School of Economics
July 2019

Bachelor Thesis for the Policy Economics Major
International Bachelor Economics and Business Economics

Supervisor: Dinand Webbink
Second Assessor: Matthijs Oosterveen

Abstract

In this paper I analyse the impact a student's ordinal ranking among her classmates has on her long-term school performance. In effect, I compare students with similar grades but different rankings in their respective classes. I use data from a state-wide experiment in the US that randomised student's allocation into classes and tracked their performances from kindergarten to the end of high school. My results generally confirm the positive effects of being ranked higher in one's class. However, the effect is dominant only within subjects and there is inconclusive evidence for precise mechanisms.

1 Introduction

Parents, as well as schools, invest a great amount of effort to choose the best learning environment for students. Next to other issues, one factor that stands out in this decision is: Who will be the student's peers? While the notion that it is good to be surrounded by well-performing classmates is well accepted, there are different concepts to be considered when thinking about peers. The concept that this paper will focus on deals with a student's rank. Here the idea is that if a student stands out relatively to his classmates (has a high rank in his class), this will benefit his later performances. Academically, the effects of students' peers and class environments have been a subject of analysis for a long time already. Much of the literature, however, deals with the classic idea that students can profit from well-performing peers. In his literature survey, Sacerdote (2011) summarises the findings to that date: students generally benefit from better-performing peers around them. However, many studies fail to grasp some of the heterogeneities embedded in this effect. It seems high-performing students benefit most from other high-achievers while students with lower achievements benefit from peers that are performing slightly better than them. Furthermore, these peer influences seem not to be restricted to academic performances only but can be observed in all other kinds of behaviours. In their randomised experiment, Booij, Leuven, and Oosterbeek (2017) confirm the general positive effect of high-performing peers. However, they also find that a higher standard deviation [SD] in ability within a group can decrease performances.

A Big Fish in a Little Pond

Also the concept of rank effects has been studied in previous literature: The "Big Fish in a Little Pond" effect was first proposed by Marsh (1987). He describes how a student's self-concept has an important impact on one's academic performances. Standing out in a particularly low-performing peer group - *being a big fish in a little pond* - leads students to overestimate their actual ability and develop higher expectations. This poses the question whether better peers do not always have a positive influence. Instead, the same student might perform worse in a better environment because he feels relatively inferior to his classmates. Put differently, the student's lower/higher rank causes him to lose/gain self-confidence and perform worse/better. This happens because outperforming weak peers gives a student an incorrect belief about his actual ability. He erroneously assumes

that belonging to the smartest students in a small group will be the same in a bigger population - effectively, he fails to understand the law of small numbers. Marsh's findings therefore lay the conceptual basis for my research question: **Does a student's ordinal rank in his childhood impact his long-term academic performance?** Goulas and Megalokonomou (2015) show that telling students about their performance relative to their peers further emphasises the pre-existing differences, good students do even better and bad ones do worse. Tincani (2017) formalises the rank idea into a theoretical model. Furthermore, she finds that a higher dispersion of ability among peers reduces incentives as it becomes harder to outperform others.

This paper however, follows most closely a research design that can be seen in Elsner and Isphording (2017). As proposed by the theory, they find that a student benefits if he is exposed to a relatively weaker cohort. This can be partly explained by students raising their expectations for rewards to further education. For other possible mechanisms no conclusive evidence is found. Furthermore, a student's rank does not only have an impact on his academic performance but also on out-of-class behaviours. A lower-ranked student is more likely to engage in a number of risky behaviours (such as drinking, violence or unsafe sex). Placing a lower value on one's future education seems to be playing a role again but also the fact that students select their group of friends according to their relative rankings plays a big role (Elsner & Isphording, 2018). Lastly, Murphy and Weinhardt (2018) use a slightly different research design and also find a positive effect of ordinal ranks on future academic performances within the same subject. Again the effect seems to work through a higher self-confidence for well-performing students. The Murphy and Elsner studies are contradicting each other in what gender heterogeneities they find. While Elsner reports the rankings affect mainly females, Murphy concludes the opposite.

Concluding, I can say that the literature confirms the existence of rank effects. Dominant mechanisms include the uncertainty every student faces regarding his actual ability. Students use their relative performance in the class environment as a heuristic to fill this uncertainty, which in turn can lead to changes in motivation and self-confidence as well as the expected returns to future education. Furthermore, there exist some theoretical mechanisms that have not yet been supported by the data. Institutions such as universities could be using ordinal ranks as a filter when choosing among applicants. Alternatively, environmental actors, such as parents or teachers, could be reacting stronger to a student's ordinal rank than to his actual ability (a teacher giving a lower grade to a student because

of well-performing classmates or a parent being disappointed in a student's good absolute but bad relative results).

This paper adds to the group of new literature discussing ranking effects additionally to the established classic peer effects (Sacerdote, 2011). My contribution is most comparable to the Elsner and Isphording (2017) literature but offers new perspectives at the same time: Most importantly, I use a data set where students' entry into classes is randomised. This allows me to compare students in different classes rather than different cohorts, which alleviates the concern of time trends biasing the results. Furthermore, I use public test results to rank students rather than unpublished results of cognitive ability tests. With public results, students are more likely to be aware of their relative rankings and react to them. Lastly, I have data on entire cohorts of participating schools, which frees me of the sampling problems existent in the Elsner paper. Next to the academic sphere, my findings will also help both parental actors as well as educational institutions in understanding the optimal learning environments for students.

In the analysis I exploit the idiosyncratic variation in ability composition that exists in every class due to the law of small numbers, to compare students with similar achievements but different ordinal ranks. I use an OLS regression analysis with school fixed effects that regresses long-term academic outcomes on ranking measures. My results show a sizeable positive rank effect within the mathematics subject but I fail to find trends for more general measures. An increased motivation and valuation of school itself seems to be the driving mechanism behind these results. Overall I find stronger results for females and classes with higher scores to begin with. Section 2 will introduce my empirical strategy, section 3 takes a closer look at the data, section 4 gives an overview of my results, section 5 examines robustness issues and section 6 concludes.

2 Empirical Strategy

The STAR Experiment

The STAR experiment was conducted in Tennessee starting in 1985. It was originally designed to test the influence of class size on student performances. Researchers randomised the allocation of students into classes within each of the 76 schools participating in the experiment. Overall 11,601 students were involved in the randomisation process. Chetty

et al. (2011) provide some evidence on the successful randomisation of students. The initial plan was to randomise student allocation in the 1985 kindergarten cohort and then closely evaluate the students during their development up until the end of third grade in 1989. However, due to issues with parent satisfaction and many students joining schools only after the kindergarten year, the classes were re-randomised at the beginning of first grade. Although the experiment finished after that, researchers continued to obtain a variety of achievement measures up until the student's high school graduation. Such measures include a student's graduation GPA, his performance in college entrance exams (ACT or SAT) and a dummy that indicates whether a student likely graduated.[1] Unfortunately, many of the post-experiment measures are only available for a fraction (often significantly less than half) of the students. Next to the graduation-related measures they include two participation studies conducted in fourth and eighth grade as well as a school identification study conducted in eighth grade. In the participation studies, teachers were asked to assign scores to each student among a range of questions related to their participation and behaviour in class. In the school identification study, students were directly asked about their attitude towards school. Moreover, students took an annual SAT test during every year of the experiment (G1 to G3). This test comprises multiple areas such as reading, writing and mathematics.

Research Concept and Econometric Design

In the following I explain how my research setup helps me to answer my research question: **Does a student's ordinal rank in his childhood impact his long-term academic performance?** First, I create class rankings based on students' performances in their early primary school years. Then I test what effect those rankings have on outcome measures during the students' graduation 12 years after they started primary school in first grade [G1]. For the ranking variable I calculate a student's ordinal percentile rank in her respective class. That is how many percent of a student's peers are ranked below her - the best student's rank being 1 and the worst one's 0. Using percentiles rather than absolute ranks removes the issue of different class sizes - being the 10^{th} best student in a class of of 15 is a different type of achievement than in a class of 25. After simply obtaining

[1] Originally researchers generated a categorical variable that indicated how likely a student had graduated. Later this variable was simplified into a dummy that coded "1" if a student graduated or likely graduated and "0" for the rest.

a student's absolute class rank the ordinal percentile rank [OPR] is calculated with the following equation:

$$PercentileRank_{ic} = \frac{AbsoluteRank_{ic} - 1}{ClassSize_c - 1} \tag{1}$$

Due to the problems mentioned earlier[2], I do not use any achievement data from the kindergarten year. Instead I calculate an OPR both for the first grade [G1] and the third grade [G3] to allow the analysis to appreciate that rankings from different years might have differently strong effects on a student.[3] As a measure on which to rank students, I use scores from the SAT tests that students in all participating schools had to take. I generate both a general rank (a weighted average of all SAT test scores) as well as a maths-specific rank.

I will use a set of different specifications to estimate a number of general effects, explore mechanisms and investigate the existence of heterogeneities. All my specifications rely on a simple OLS regression that uses school fixed effects and a number of control variables. Which specific controls are employed, is determined by the demands made by the discoveries of related literature as well as peculiarities to the STAR experiment. The most straightforward control variable is a student's own score that is also used in the ranking process. This enables me to compare students with *the same grades* but *differing OPRs*. Furthermore, as discussed by Sacerdote (2011) and Booij et al. (2017), I include both the class' mean SAT scores as well as their SD to control for the classic peer effects caused by these variables. Lastly, because of the fact that the STAR experiment exhibits classes of varying size, I employ also class size as a control variable.

After understanding the need to control for the aforementioned variables, it remains unclear in what functional relationship each one stands to the outcome variable.[4] Figure A1 shows scatter and residual plots for a student's graduation GPA and each of the control variables:

[2] As students were reshuffled after the end of kindergarten and many new students joined only after that, a ranking constructed from data in the kindergarten period would be unrepresentative as students were not exposed to the environment their rank indicates.

[3] The G1 ranking allows for a longer time between the ranking and outcome for the effect to take place, whereas the G3 ranking might be more important for a student as at an older age he might be more aware of his relative performance and its importance.

[4] As will be explained later, I use an array of different outcome variables throughout the specifications, however I will use the graduation GPA as the most representative one in these tests. The results are robust to using similar outcome measures.

class size, mean SAT scores and the SD of SAT scores. None of the scatter plots seem to suggest a non-linear relationship and also the residual plots show a mostly uniform distribution of residuals. Following these graphs I include each of the variable as a linear control. From this results the following generalised regression equation:

$$\text{AchievementVariable}_{ics} = \alpha_s + \beta \text{OPR}_{ics} + \gamma Z_{fcs} + \epsilon_{ics} \tag{2}$$

where AchievementVariable is the respective achievement (GPA etc.) of student i in class c and schools s in each specification; OPR is a student's ordinal percentile rank with β being the coefficient of interest; Z is a vector of f control variables (namely: SAT score, class size, mean of class SAT scores and SD of class SAT scores); and ϵ is the error term. It is important to note that my decision to use a simple OLS regression with school fixed effects rests entirely on the assumption that the error term is uncorrelated with the OPR variable ($E[\epsilon \mid \text{OPR}_{ics}] = 0$). I will motivate the intuition for this assumption in the following section.

Identification Strategy

Effectively, the ultimate aim of my econometric design is to compare students from the same school who have similar grades but differing OPRs. To achieve this, I use the idiosyncratic variation in the composition of classes within a school. As Hoxby (2000) explains, such a variation is created by the random cut-off beyond which children are placed into the next school year. If slightly more children are born before this cut-off, this creates a variation in how easily one can achieve a certain OPR in that cohort.

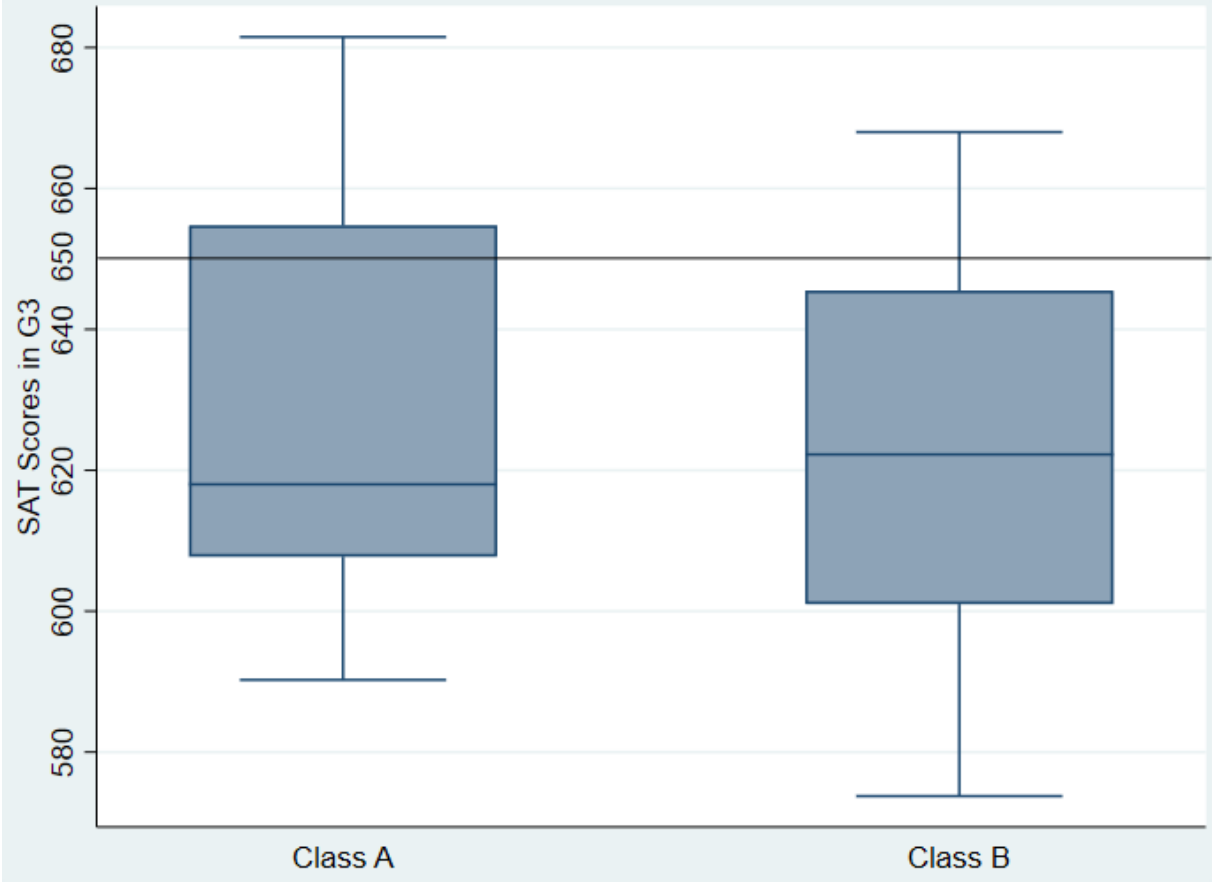


Figure 1: Boxplot comparing grade composition in two classes of a school

Figure 1 illustrates how different class compositions can lead to variation in OPRs. Class A and B are two randomly chosen classes of the same size and from the same school. It is easy to see that despite a similar SD, class A has a higher mean grade. Looking at the horizontal line for an SAT score of 650, one can see that a student with this score would have been between the 50^{th} and 75^{th} percentile in class A but surely higher than the 75^{th} percentile in class B. As in the STAR experiment, entry into classes is randomised, the given student's assigned OPR is random too. Furthermore, I use a set of school fixed effects to control for pre-existing differences that can be due to selection into certain schools and biases if some schools are placed in neighbourhoods dominated by different socio-economic backgrounds than others. This explains why, once we include school fixed effects we can expect that $E\left[\epsilon \mid \text{OPR}_{ics}\right] = 0$.

Specifications

I will use the introduced econometric design for an array of different specifications. First I examine a general rank effect. Here I use rankings from overall SAT scores in both G1 and G3 as independent regressors for three different achievement measures, namely a student's graduation GPA, his ACT scores[5] and a dummy that indicates whether a student likely graduated. Next to that, I examine whether the effect differs when I use measures that remain in the same subject. Therefore I use mathematics SAT ranks in G1 and G3 as independent variables on a students mathematics graduation GPA and his mathematics ACT scores.

Lastly, I explore a couple of mechanisms which could be potentially driving the rank effects. Also the mechanisms are split up into general effects as well as within-subject effects. For the former, I again use a student's G1/3 overall SAT Rank as explanatory variables and both the effort score in the G4 participation study and the valuation score in the G8 identification study. The effort score is assigned by a teacher depending on how much a student participates in class and the valuation score is given to the student based on his answer regarding how much he values school education. For the within-subject mechanism, I use G1/3 mathematics SAT ranks and measure their effect on the mathematics effort score in the G8 participation study. This score was assigned similarly to the G4 participation study, however only the mathematics teacher could influence it. Furthermore, I also test for heterogeneities. For those specifications that deliver significant results, I will include interaction terms between the OPR and each of the control variables to test whether they show any differential effects. Figure 2 gives an overview of all mentioned specifications.

[5]The ACT test is a college entrance exam that student can voluntarily take if they wish to continue on to college. Most students take the ACT test but some also the SAT test. For the latter students, the test scores were converted to an ACT level.

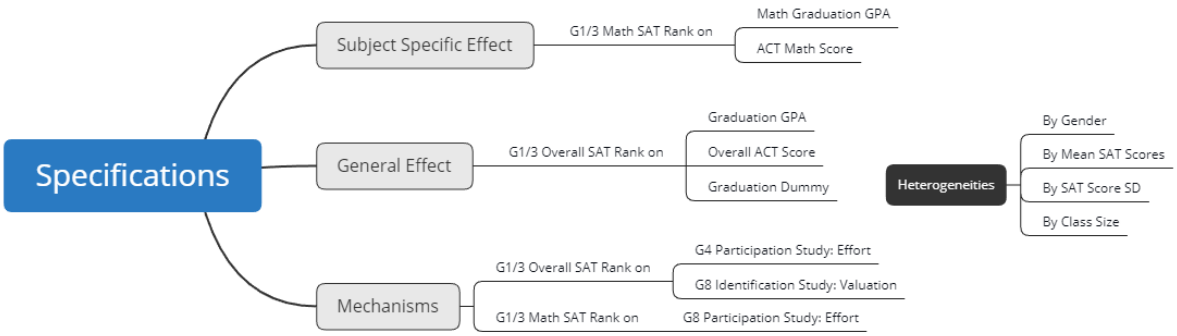
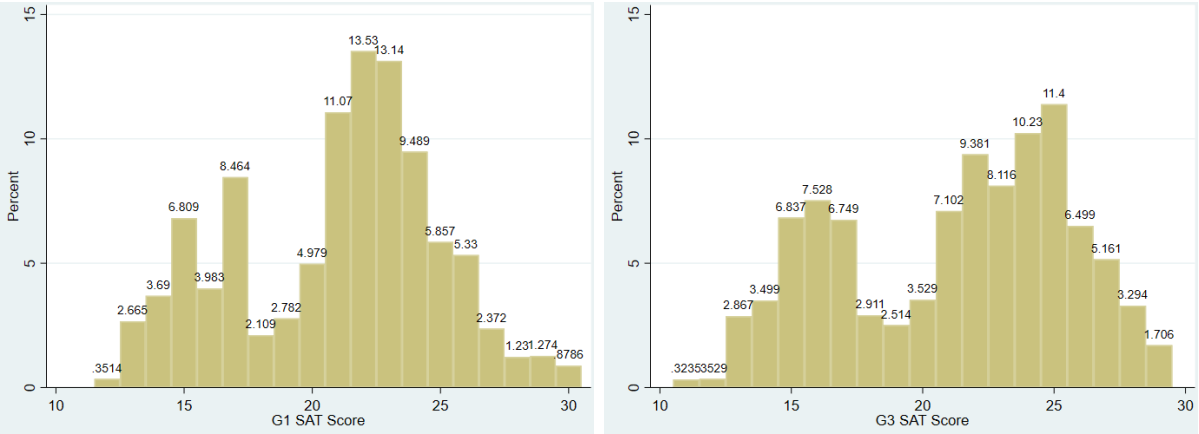


Figure 2: All different specifications

3 Data

Peculiarities in the Data and Summary Statistics

Before naively interpreting the results, one should consider that the experiment was originally designed to test the effect of different class sizes. As can be seen in Figure 3,



(a) in grade 1 (b) in grade 3

Figure 3: Histogram of distribution of class sizes

this resulted in the class sizes not being normally distributed around some average size but exhibiting two distinct spikes. One for the average small class and the average regular class respectively. Moreover, the fact that a number of students changed away from the initially assigned class becomes visible in the flattening out in the distribution in G3 compared

to G1 (compare panel (a) and (b)). This confirms my decision to include class size as a control variable.

As mentioned before, the entire experiment contains 11,601 students in 76 different classes. Table 1 gives an overview of the demographic composition of the study as well as heterogeneities in achievements across demographic groups. It stands out that the majority of the experiment (99.24%) is made up of black and white students while all other ethnicities only make up a negligible fraction of the sample. Furthermore, both the average Graduation GPA and the average ACT Scores show that black students are slightly under-performing in this sample while students of Asian descent far outperform the rest of the groups.

Besides the evidence presented by Chetty et al. (2011), I perform some additional tests

Table 1: Overview of student characteristics

Student Race/Ethnicity	Percentage	Graduation GPA	Average ACT Score
Male	52.88%	81.82	19.41
Female	47.12%	85.06	19.10
White	62.79%	83.99	20.30
Black	36.45%	81.89	16.55
Asian	0.28%	84.42	23.23
Hispanic	0.18%	81.53	19.20
Native Americans	0.12%	86.29	19.33
Others	0.17%	84.19	22.92

GPA is a percentage score measured between 0 and 100.
ACT test scores are measured between 0 and 36.

to evaluate how well the randomisation worked and specifically, how well it still held up in G3.

Table 2: Balancing test for distribution of students among classes

	(1) G1	(2) G3
Independent Variable:		
Gender	0.012	0.016
	(0.036)	(0.0.036)
Black	0.081	0.110
	(0.071)	(0.073)
Asian	0.199	0.488
	(0.323)	(0.357)
Hispanic	0.392	0.750
	(0.490)	(0.540)
Native American	-1.354***	
	(0.491)	
Other	0.282	-0.652
	(0.443)	0.539
Age (in days)	-0.000	-0.001**
	(0.000)	(0.000)
Free Lunch Status	0.002	0.057
	(0.044)	(0.044)
Teacher Experience	-0.004	0.006**
	(0.002)	(0.002)
Controls:		
School FE	YES	YES
Observations	6,595	6,439

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This table displays results from a balancing test that regresses different demographic characteristics on the student's class ID. An insignificant results indicates that this variable is balanced across classes while a signficant results indicates imbalances

Therefore I run a specification that regresses the class ID of a student's class on several demographic measures to test whether these are balanced across classes:

$$ClassID_i = \alpha + \beta X_{if} + \epsilon_i \tag{3}$$

where $ClassID$ is the ID of student i's class and X is a vector of f demographic control

variables for student i. The results are displayed in Table 2. There are two significant values in the G3 specification (Age and Teacher Experience) that might be due to students changing classes after G1, however, both values are very small when considering that age is measured in days and experience in years. In the G1 specification only the Native American dummy is significant. Although the value is quite high, this imbalance can be explained by the law of small numbers as Native Americans only make up a small fraction of the entire sample (compare Table 1) Even if it is not properly balanced, this marginal group is unlikely to have an impact on my results.

4 Results

In this section I will go through each set of specifications and explain their results. I will start with the general rank effects, then move on to the mechanisms and lastly discuss possible heterogeneities.

Rank Effects

Table 3: Effect of performance ranks on graduation measures

	(1) Graduation GPA	(2) ACT/SAT Scores	(3) Graduation Dummy
Independent Variable:			
G1 Rank	1.072	0.388	0.101**
	(0.810)	(0.454)	(0.0458)
G3 Rank	-0.753	0.505	0.137***
	(0.782)	(0.384)	(0.0478)
Controls:			
Class Mean GPA	YES	YES	YES
Class GPA SD	YES	YES	YES
Class Size	YES	YES	YES
GPA in Ranked Year	YES	YES	YES
School FE	YES	YES	YES
Observations in G1/G3	2,232/2,506	2,361/2,651	2,890/3,204

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This table shows results from regressing general student rankings in first and third grade on graduation measures. Each coefficient shows a distinct regression. Each column stands for a different outcome variable while different rows depict different independent variables.

I will start with the impact of the overall SAT score rank. Its results are displayed in Table 3. The Table shows results for both the rankings in first and third grade (rows) and all three outcome measures (columns). One can immediately see that most results are positive, however only the specifications using the graduation dummy as an outcome are significant. An increase of one in the OPR leads to a 10.1(13.7) percentage point increase in the probability to have likely graduated. The effect of the G3 Rank on graduation GPA is even negative, however extremely insignificant. A reason why I only find significant results for the third (3) specification might be that it is potentially easier to pick up an

effect on a "rougher" measure such as whether a student graduated or not. Furthermore, there is the possibility that the dummy is subject to a measurement error as the variable is coded as showing "1" for students who graduated or likely graduated. However, the relatively small amount of observations falling under the "likely graduated" category make it unlikely that this could be driving the result. Another observation one can make, is that generally the G3 rank seems to have a stronger effect than the G1 rank. This makes sense when considering two things: Firstly, after third grade less students will switch classes compared to after first grade. This means that the ranking of G3 will be the real ranking facing the student for a longer time period. Secondly, if one assumes that the rankings can vary from year to year if students improve or slack off, it would seem that the G3 ranking is more meaningful to the student as she is more likely to be aware of her relative performances at an older age. Overall, I can confirm generally positive effects but altogether inconclusive evidence of an effect for the overall OPR.

Moving on, I look at the within-subject rank specifications. Results for these tests can be

Table 4: Effect of performance ranks on graduation measures within a subject

	(1) Math Graduation GPA	(2) ACT Math Scores
Independent Variable:		
G1 Math Rank	2.639***	2.486***
	(0.860)	(0.389)
G3 Math Rank	3.243***	1.869***
	(0.904)	(0.375)
Controls:		
Class Mean GPA	YES	YES
Class GPA SD	YES	YES
Class Size	YES	YES
GPA in Ranked Year and Subject	YES	YES
School FE	YES	YES
Observations in G1/G3	2,621/2,576	2,594/2,594

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This table shows results from regressing math student rankings in first and third grade on math graduation measures. Each coefficient shows a distinct regression; Each column stands for a different outcome variable while different rows depict different independent variables.

obtained from Table 4. The structure of this table is comparable to Table 3, however, both the independent and dependent variable focus on measures entirely within the subject of mathematics. Compared to the overall effect, the within-subject OPR seems to have a much stronger effect. Tests from both G1 and G3 and for both outcome measures are highly significant and very positive. A one unit increase in the OPR leads to a 2.639(3.243) points increase in a student's maths graduation GPA and a 2.486(1.869) in the student's ACT maths scores. All the coefficients are significant at the one percent level. In these specifications however, there does not seem to be a clear difference between G1 and G3 ranks. For each outcome measure the respective other coefficient is stronger and if one looks at the standard errors they can be considered equal. Concluding, it can be said that within the mathematics subject the OPR has clearly a strong and positive effect on a student's longer term academic performance.

Another area I explored in my analysis is testing which mechanisms could explain the observed effects. Table 5 shows the results of an analysis of how exactly a student's OPR affects his motivation and opinion of school. Potential effects would be in line with the findings of Goulas and Megalokonomou (2015) and Marsh (1987) who both talk about how information about one's relative performances has a direct effect on one's valuation of future education and the importance of school in general. Also the mechanism analysis is split up into a part looking at overall ranks and another one focussing on mathematics-specific measures. Just as in the test for achievement measures, there is very little evidence on the overall rank effects. All coefficients are comparatively small and very insignificant. This makes sense in light of the results displayed in Table 3. The within-subject specifications show a slightly more interesting picture. While the coefficient for the G3 maths SAT rank is insignificant, the G1 ranks has a significant positive effect on a student's exerted effort in his final year. Increasing a student's OPR by one is associated with an increase in his G8 mathematics effort subscore of 0.878. This effect is significant at the five percent level. This observation suggests that a higher-ranked student achieves better results because outperforming others boosts his motivation and the value he puts on education. However, this only holds within a subject.

As a last measure, I examined the existence of certain heterogeneities, that is, whether the observed OPR effects differ between certain subgroups or along different values of a certain variable. As the only entirely conclusive evidence I found was for the within subject specifications, that is where I will look for heterogeneities. Table 6 shows the results of

Table 5: Effect of performance ranks on measures of motivation

	(1) G4 Effort Score	(2) G8 Valuation Score	(3) G8 Math Effort Score
Independent Variable:			
G1 Rank	-0.017	0.147	
	(1.244)	(0.403)	
G3 Rank	1.211	0.294	
	(1.162)	(0.409)	
G1 Math Rank			0.878**
			(0.421)
G3 Math Rank			0.258
			(0.434)
Controls:			
Class Mean GPA	YES	YES	YES
Class GPA SD	YES	YES	YES
Class Size	YES	YES	YES
GPA in Ranked Year	YES	YES	YES
School FE	YES	YES	YES
Observations in G1/G3	1,869/2,061	2,160/2,697	1,997/2,208

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This table shows results from regressing general and subject-specific student rankings in first and third grade on different measures of motivation. Each coefficient shows a distinct regression; Each column stands for a different outcome variable while different rows depict different independent variables.

this analysis. I tested whether the observed effect differs between males and females or for different levels of a class' GPA mean or SD or class size. Many of the displayed results are strong and insignificant, however they are partially very inconsistent among different specifications. Each of the entries related to a variable exploring heterogeneity represents a separate regression that includes an interaction effect between that variable and the OPR in addition to what was described in Equation 2. For gender heterogeneities, my results stand in opposition to the findings of Murphy and Weinhardt (2018), who report stronger effects for males. For the mathematics graduation GPA, the effect is a lot stronger for females and practically zero for males. For ACT maths scores, the evidence is more contradicting. For the G3 Rank, there is a small significant *positive* interaction effect for males while the G1 rank effect is insignificant. Also my results for heterogeneity among different levels of a

Table 6: Heterogeneities in the within subject rank effects

	(1) Math Graduation GPA	(2) ACT Math Scores
Independent Variable:		
G1 Math Rank	4.866***	2.486***
Male	-4.653***	1.117
Mean Class GPA	0.033	0.039***
Class GPA SD	0.572***	0.046
Class Size	0.027	0.052
G3 Math Rank	3.243***	1.869***
Male	-4.524***	1.353***
Mean Class GPA	0.032**	0.264***
Class GPA SD	0.051	0.010
Class Size	0.029	-0.008
Controls:		
Class Mean GPA	YES	YES
Class GPA SD	YES	YES
Class Size	YES	YES
GPA in Ranked Year and Subject	YES	YES
School FE	YES	YES

Standard errors omitted
*** p<0.01, ** p<0.05, * p<0.1

This Table displays an extension of the analysis displayed in Table 4. All independent variables that are not listed in bold font present the coefficient of an interaction term of that variable with the rank variable. Each row-column combination presents a separate regression.

class' mean SAT score contradict previous literature, in this case the findings of Elsner and Isphording (2017). Most measures pick up a small but significant positive interaction effect, which indicates that better-performing classes experience stronger rank effects. Elsner had found the opposite in his paper. For differential effects for different levels of a class' SAT score SD and class size, I would have expected negative interaction effects which would comply with both the theory and empirical findings of Tincani (2017). However, for both variables, the effects are not clear and mostly insignificant, such that I cannot conclude any heterogeneous effects. Concluding, I can say that I find some heterogeneity suggesting stronger effects for females and in better performing classes. However, many results are inconclusive and often contradict the findings of previous literature.

5 Robustness

Missing values

The reader should be made aware that a significant number of observations have to be dropped in each specification. Before calculating the OPR for every respective specification all observations that have missing values for the respective independent and dependent variables are temporarily deleted so that the rankings are not biased by the number of missing values. The control variables, however, are calculated before dropping observations, so that they reflect the reality (the actual SAT SD) that students experienced in class. Table 7 gives an overview of missing values among the dependent and independent variables. Throughout all variables there is a big amount of missing values. With the exception of the graduation dummy, especially the dependent variables have many missing values. Table 8 helps to examine how missing values affected each of the regression specifications. The first row of the table should be interpreted as follows: In the specification where I regress the graduation GPA on the G1 SAT rank, 9,351 (80.61%) values were missing. As for an observation to be considered in a regression, it cannot have missing values neither in the dependent nor in the independent variable, the missing values per specification are a lot higher than for the variables alone. When examining the last column, one can find hints towards a systematic pattern. Generally, the within subject specifications (all significant at one percent) seem to have less missing values than the general specifications. Within the general specifications, the regressions using the graduation dummy have again fewer missing values while also being the only significant ones in that group. This indicates an overall negative relationship between the amount of missing values and the significance of the respective regression. This variation in missing values will only have a biasing influence on my results if it is somehow systematic, e.g. especially smarter students have more missing values in a certain specification compared to other students. Table A1 displays the results from an analysis, testing whether values are missing in a systematic manner. Missing values in the independent variables (G1/G3) can be neglected as 92.24% of these observations come from students who simply joined the STAR schools after the experiment had finished. The results show very clearly that, across all specifications, having a missing value in an outcome variable is associated with a very significantly lower performance in G1 and G3. This makes intuitive sense, as badly performing students are much more likely to change schools or never graduate. Keeping track of those students'

Table 7: Missing values per variable

ID	Variable	# Missing	% Missing
	Independent Variables		
1	G1 SAT	5,900	50.9
2	G3 SAT	5,677	48.9
3	G1 Maths SAT	5,003	43.1
4	G3 Maths SAT	5,524	47.6
	Dependent Variables		
5	Graduation GPA	7,947	68.5
6	Graduation Dummy	6,609	57.0
7	ACT Score	7,722	66.6
8	Maths Graduation GPA	7,830	67.5
9	Maths ACT Score	7,846	67.6

performances is especially problematic. It is difficult to conclude a definite bias from these observations, however, the following explanation shows how the findings could have a biasing effect: Student A has a real OPR of 0.3 in his class. However, due to many students at the lower end with missing observations, his OPR in the regression is 0.1. This means that student A will be compared to students in other classes who are actually much worse (but have the same OPR). If student A then achieves graduation performances typical for a student with a 0.3 OPR rather than a 0.1 OPR, this will diminish the negative effects of having a low rank. If he had had a 0.3 OPR, his graduation performance would have been average, but for an 0.1 OPR it is exceptional. The existence of such scenarios would bias my results downwards and also imply higher (false) variation in the student's improvements, hence reducing the probability of obtaining significant results.

6 Discussion & Conclusion

Interpretation of Results

The main insight of this paper regards the difference between the overall OPR effect and the within subject OPR effect. Only for the latter one can I clearly confirm the findings of Elsner and Isphording (2017) as well as Murphy and Weinhardt (2018). It is not immediately apparent why the rank effect does not hold for overall student performance,

Table 8: Patterns of missing values per regression

ID Combination	Frequency	Percentage	Significance
1 or 5	9,351	80.61	
1 or 6	8,705	75.04	**
1 or 7	9,230	79.56	
2 or 5	9,082	78.29	
2 or 6	8,393	72.35	***
2 or 7	8,946	77.11	
3 or 8	8,924	76.92	***
3 or 9	8,964	77.27	***
4 or 8	8,972	77.39	***
4 or 9	8,966	77.29	***

ID's as defined in Table 7.
Each row presents one of the regression specifications discussed earlier. The last column indicates the significance level of the explanatory variable in the respective regression.
*** p<0.01, ** p<0.05, * p<0.1

however it is very intuitive that the within subject channel is at the very least stronger. That is because students' ranks and achievements might vary a lot between subjects, such that on the overall level the effects could cancel out. Furthermore, if a given student is ranked very differently for different subjects, then there is no clear resulting self-confidence effect. Within a given subject however, students know their relative performance exactly and extract more self-confidence and motivation out of their rank. What remains unclear from the results on the other hand, is whether the rankings in first or third grade have a higher influence. There are reasons for both to be stronger. The G1 ranking is followed by a longer time span for the rank to take effect, whereas during the G3 ranking students might be more aware of their rank. Also, relative performances might have stabilised while they were more volatile in G1 which could lead to rank reversals.

The mechanism analysis is in line with my general results. There is no clear evidence for mechanisms along overall school rankings but some evidence suggesting a motivational channel for the within subject rank. Lastly, I find heterogeneity among gender and average performance of peers. It seems that ranks have a stronger effect on females and in classes with better performing peers, though that is not necessarily in line with some of the literature. Tests for other heterogeneities failed to establish conclusive evidence.

Limitations

There are a couple of problems with my identification strategy as well as the structure of the STAR experiment itself that could lead to biased results and question the interpretation of my results. A big issue is the possibility of parents delaying their children's school entry because of the mere existence of the experiment (Bietenbeck, 2015). As it was well known that the STAR experiment would only affect one cohort, certain parents might have been worried about the fact that they could not influence whether their children were assigned to a class of small or regular size. These parents might have decided to delay the children's school entry until after the experiment. If this decision had ben somehow systematic - for example, if only highly educated parents felt this threat - then the sample would be unrepresentative and suffer from a sample selection bias.
Another potentially problematic topic is selective attrition. Naturally, badly performing students are more likely to drop out at some point along the way to high school graduation. These cases will show up as missing values in the outcome variables and therefore not be considered in the regressions. As explained earlier this could potentially bias my results downwards. The fact that specifications with fewer missing values[6] are the only ones that show significant results, encourages the existence of such a bias. However, this cannot be fixed by coding a "0" graduation GPA for students that dropped out, as it is incorrect to assume a student who dropped out would have achieved a bad graduation result.
Finally, there are a couple of factors that threaten the exogeneity of my identification strategy. Two potentially problematic variables are: parental pressure and a student's own motivation (determined prior to knowing his rank). Both these variables affecting long-term academic outcomes might also influence the OPR. This is especially the case if a student (or his parents) are more concerned about outperforming certain other students rather than the student's absolute performance.

Outlook

Two topics stand out on which successive research could be particularly insightful. Firstly, a lot of my findings about heterogeneities stand in contrast to what was found in previous literature. This calls for further investigation about how different subgroups are exactly affected by the rank effects, particularly the gender groups. Another open area is the

[6]Within subject specifications and the specifications using the graduation dummy

within-subject effects. Due to data constraints, I restricted myself to the mathematics subject. However, it would be interesting whether the observed results can be replicated also for other subjects or rather seem specific to maths.
Concluding, I can say that altogether I manage to confirm the existence of rank effects on student achievements and motivation. This finding adds to the increasing evidence that rank effects should be considered when artificially composing groups in order to maximise learning possibilities for all members of the group. Also many other groups can benefit from this knowledge, to improve their decision-making when it comes to our children.

References

Bietenbeck, J. (2015). The long-term impacts of low-achieving childhood peers: Evidence from project star. *Journal of the European Economic Association*.

Booij, A. S., Leuven, E., & Oosterbeek, H. (2017). Ability peer effects in university: Evidence from a randomized experiment. *The review of economic studies*, *84*(2), 547–578.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics*, *126*(4), 1593–1660.

Elsner, B., & Isphording, I. E. (2017). A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics*, *35*(3), 787–828.

Elsner, B., & Isphording, I. E. (2018). Rank, sex, drugs, and crime. *Journal of Human Resources*, *53*(2), 356–381.

Goulas, S., & Megalokonomou, R. (2015). Knowing who you are: The effect of feedback information on exam placement. *University of Warwick, mimeo*.

Hoxby, C. (2000). *Peer effects in the classroom: Learning from gender and race variation*. National Bureau of Economic Research.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of educational psychology*, *79*(3), 280.

Murphy, R., & Weinhardt, F. (2018). *Top of the class: The importance of ordinal rank*. National Bureau of Economic Research.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the economics of education* (Vol. 3, pp. 249–277). Elsevier.

Tincani, M. (2017). Heterogeneous peer effects and rank concerns: Theory and evidence.

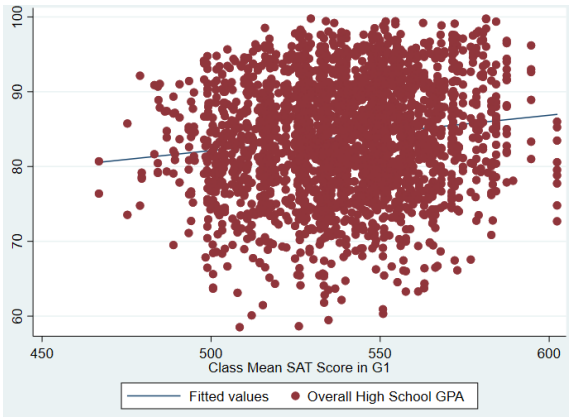
A Appendix

Table A1: Analysis of systematically missing values

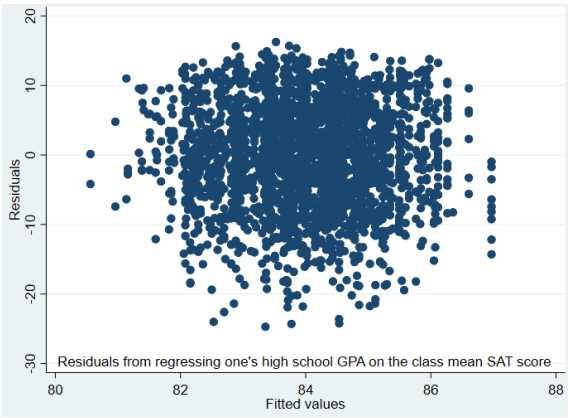
	Missing value in:		
	Graduation GPA	ACT Score	Graduation Dummy
General Rank			
G1 SAT Rank	-21.890***	-37.046***	-18.026***
	(1.059)	(0.987)	(1.064)
G3 SAT Rank	-15.971***	-28.142***	-12.440***
	(0.870)	(0.817)	(0.875)
	Missing value in:		
	Graduation Math GPA	Math SAT Score	
Subject Rank			
G1 Maths Rank	-20.951***	-32.650***	
	(1.050)	(1.007)	
G3 Maths Rank	-19.082***	-29.512***	
	(1.002)	(0.959)	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

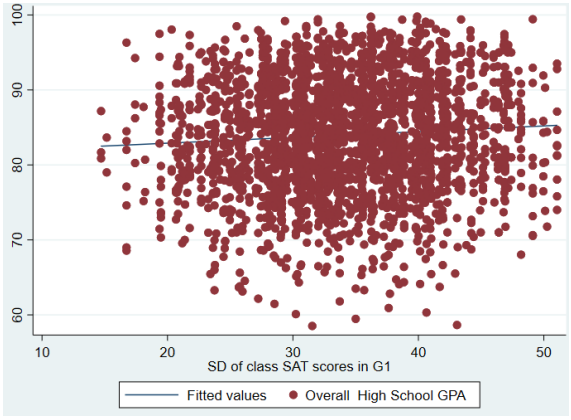
Each column-row combination depicts a separate regression. Columns show respective independent variables and rows show respective dependent variables. The table describes how having a missing value in one of the outcome variables is associated with G1 and G3 performances.



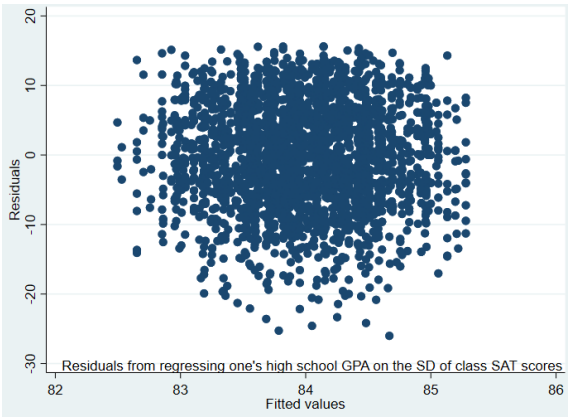
(a) Scatter plot: HS GPA and Class SAT score mean



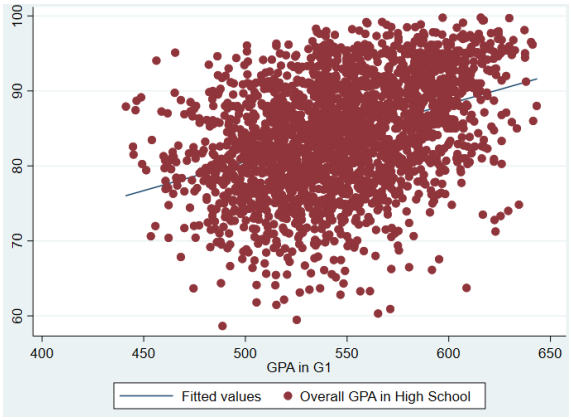
(b) Residuals: HS GPA and Class SAT score mean



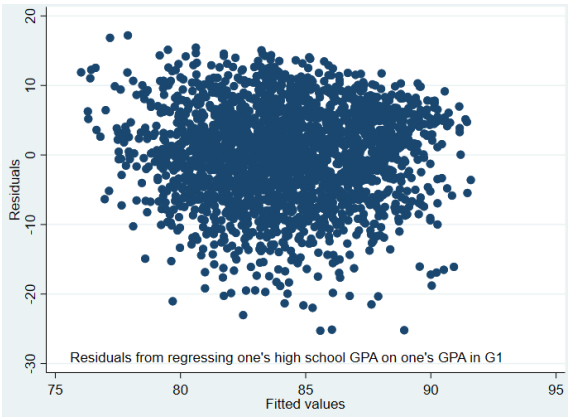
(c) Scatter plot: HS GPA and SD of class SAT score



(d) Residuals: HS GPA and SD of class SAT score



(e) Scatter plot: HS GPA and student's G1 GPA



(f) Residuals: HS GPA and student's G1 GPA

Figure A1: Scatter plots and residual plots for Control Variables