Journal of Economic Analysis

Homepage: https://www.anserpress.org/journal/jea

Does the 'Double Reduction' Policy Reduce 'Rat Race'? A Natural Experimental Study in China

Philip Arestis [a, *], Mianshan Lai [b], Jiale Yang [c], Yuexun Zhong [c]

[a] *Department of Land Economy, University of Cambridge, Cambridge, UK*

[b] *China Center for Special Economic Zone Research, Shenzhen University, Shenzhen, China*

[c] *College of Economics, Shenzhen University, Shenzhen, China*

ABSTRACT

This paper investigates the effects of an unprecedented education policy called 'double reduction' (DR) on students' performance, with evidence from China. One of the significant changes brought about by this policy is the transition of examination scores from a numerical system to a grading system. We found a junior middle school in a small town in Guangzhou that met the requirements for a natural experiment, and we used such unique data to examine three hypotheses. After the implementation of DR, the 'rat race' reflected on the students' grade is reduced. To examine heterogeneity, we apply quantile regression showing that the high-ability students exert less efforts on the 'rat race' than low-ability students. Meanwhile, we also provide possible explanations in terms of this phenomenon. Overall, this study not only does it assess the grading transformation that DR policy induced but also fills the gap in the field of DR from an economic perspective.

KEYWORDS

Rat Race; Double Reduction; Grading System; Quantile Analysis

* Corresponding author: Philip Arestis
E-mail address: pa267@cam.ac.uk

ISSN 2811-0943
doi: 10.58567/jea04030003

Received 21 January 2025; Accepted 16 April 2025; Available online 14 May 2025; Version of Record 15 September 2025

1. Introduction

The 'double reduction' policy, an education policy that aims to effectively reduce the heavy homework burden and extracurricular training load on students at the compulsory education stage, has been implemented and has had profound impacts in Chinese education since July 2021. Before the reduction policy, with the rapid economic growth leading to a more prosperous daily life in recent years, parents tended to arrange after-school classes for primary school students (Song, 2022) so that their children would not be fallen behind by other peers. Such an educational competition known as 'rat race' (the same as the Chinese slang term 'NeiJuan', which is widely used to reflect a fact that most of Chinese students must exert vast amounts of their leisure time to improve their study due to the fierce school entrance examination**,** which provides limited high-quality teaching resources. According to a recent report from UNICEF China, it is a common situation in China that the K-12 (primary and secondary education) students have excessive homework and extracurricular assignments. After the implementation of the DR policy, teachers need to help the students with their homework and solve the learning problem, making sure that the students finish most of the homework at school and no need is relevant to take it back home or ask the teachers in the institution to help them (Song, 2022). What a noteworthy effect the policy induces is that such an educational policy has greatly changed the K-12 students' learning mode and caused significant transformations in related industries. In recent years, there is a growing body of research on this policy, leading to various kinds of topics, such as off-campus training industry and innovations in pedagogy (Yin and Lai, 2021; Guo, 2022; Fu et al., 2023). And it is confirmed that the effects of extracurricular tutoring in different subjects on the academic performance of middle school students exhibit heterogeneity, and these effects on students of different ability levels also exhibit heterogeneity (Hu et al., 2021). However, these studies primarily focus on the fields of education and psychology, but rarely do they discuss DR policy from an economic perspective.

Most of the studies in terms of DR policy neglect a grading system transformation change, which only happened in several areas in China. During this period, some junior high schools transform their initial numerical and absolute score to letter and relative score. However, the incentives of the transformation that apply to students' learning remain unknown, especially in the case of China. To investigate such an issue, we obtained the academic performance of students over three cohorts from a middle school in Guangzhou as our research subject. These students enrolled in three different but consecutive years, entering school from 2018 to 2020, respectively. In their three-years study, DR policy coincidentally occurred during middle school education for the last two cohorts of students, thus ideally dividing our sample into an experimental group after policy induced and a control group before policy induced.

In our study, we have three main findings. Firstly, with a Difference-in- Differences (DID) strategy, the policy has a negative effect on students' mathematics score. Secondly, we employ quantile regression to conduct further research about student academic performance heterogeneity. Interestingly, all the students' scores of statistical significance fall to some extent, regardless of the score range they fall into. However, students originally in higher percentiles experienced a greater decline in scores. This further implies an alleviation of the 'rat race' phenomenon, which is quite widespread in Chinese education. Furthermore, using the label theory and former studies regarding grading system, we provide possible explanations for students' behavior. The transformation from absolute grading to relative grading (from numerical grading to letter one) implies that the initial 'label' of students has significantly been weakened as it becomes vague. Hence, students do not need to exert so much effort to achieve or maintain the 'status' among the classmates. On the other hand, from the perspective of rational actor, the course (letter) grading offers students less incentives to pay relentless effort to the diminishing marginal utility, which can be regarded as a reflection of academic scores, because there are more leeway for the students to remain the same appraisal unchanged. It is important to clarify that, in real-world educational settings, numerical or letter grading does not necessarily correspond to absolute or relative grading systems. However, under the context of the DR policy,

numerical grading functions as an absolute evaluation standard, whereas letter grading reflects a relative grading scheme.

This paper contributes in the following ways. First, our data set is unique. Due to remote geographical location and poor transportation accessibility of the selected school, the data set we use was not affected by private tutoring, our model substantially mitigates the effect of confounding factors. Second, we provide empirical evidence to the DR policy effect with a natural experiment, which fills the gap in the empirical study of DR policy. Third, we provide explanations for the student behavior to the transformation from absolute (numerical) grading to relative (letter) one and for labelling theory in Chinese education context. Besides, we substantiate that the traditional theoretical framework cannot exhaustively explain the reality based on different assumptions and situations - different national conditions and educational models, and it should be discussed case by case.

The remaining sections are structured as follows: Section 2 presents the literature review, Section 3 outlines the theoretical framework, Section 4 provides empirical evidence, and Section 5 concludes.

2. Literature Review

2.1. From Numerical- to Letter- Grading: A Transformation of Grading System

The DR policy in China has been widely implemented and has generated a profound effect in elementary education (Guo, 2022). The most conspicuous change is about the grading system, a transformation from numerical grading (like 100, 99, 98, …) to letter grading (like A, B, C), which is a more lenient and coarser one (Dubey and Gianakopulos, 2010). The debates concerning which grading system is better to motivate students has been extensively discussed since 2021. To be specific, studies have discussed in a theoretical aspect about which grading scheme is the better one to incentivize students, numerical grade, or letter grade. When the numerical grading is used, there is a more explicit 'anchor score', which helps students assess their intrinsic ability so that numerical grading informs students that their score closes to their anchor. In contrast, letter grading, a coarser scheme, may trigger demotivation because students may be degraded to an inferior range simply due to nuance. Therefore, when students' scores are highly likely to be close to their anchors, numerical grading always outperforms any arbitrary choice of uniform letter grade. But when the demotivational effect is smaller than the motivational effect, letter grading is better (Micha, Sekar, and Shah, n.d.). Similarly, another research also concludes that if students care mainly about their relative rank in class, the numerical examination score is a much better way to motivate students to work, instead of clumping them into letter grading (Dubey and Gianakopulos, 2010). However, another opposite view is that the letter categories in the form of ABCDF, for assigning grades, gained in popularity during the 20th century due to findings that more opportunities for errors were present with percentage scales (Huey, et al, 2022). Even though these studies provided sufficient theoretical evidence to support the grading system, they lack empirical analysis to explain reality. Based on different assumptions, different national conditions and educational models, scholars drew different conclusions in terms of the pros and cons of the grading system. Hence, the similar transformation of grading system in China, changed from numerical grading to letter one, should be examined case by case. To the best of our knowledge, academic community pay more attention to the DR policy's effect on the education industry or other aspects, such as public-private partnership school (Dai, 2023; Qian, Walker, and Xu, 2023). But there are few surveys about the concealing effect of scores that the policy developed, which means that using letter grading conveys vague information than the former numerical grading did. Using students' score data from a junior high school from Guangzhou and examining the heterogeneity of student performance stratification, our study offers an innovative and empirical perspective about this grading issue and tries to figure out the possible explanation and evaluates the DR effect in Chinese education.

2.2. Heterogeneity in Student Performance Across Different Score Ranges

Grading can evaluate the quality of student's work when not only is the student in school but also in the labor market (Betts, 2005; Walvoord and Anderson, 2011). Besides, it is a consensus that the grade is not just an output of the educational process but can also act as an input (Gray and Bunte, 2022), which could be a more important determinant of a student's progress even than teacher's level of education and experience (Betts, 2005). However, based on different assumptions and situations, the previous research drew diverse conclusions. For example, there are significant differences between accounting and business majors and non-business majors with different key aspects (Giacomino and Akers, 1998; Ridener, 1999). Theoretically, one general argument is that stricter grading motivates students to put more effort into their study (Adams and Torgerson, 1964; Johnson and Beck, 1988; McClure and Spector, 2005; Walvoord and Anderson, 2011). According to the model of educational achievement using education production functions, an increase in grading standard in academic achievement will also increase student effort, achievement, and wages for the majority of students who initially strictly preferred one letter grade over others, but for those who feel initially indifferent about two adjacent letter grades, they will suffer negative effect, strictly preferring the lower letter grade and exerting no effort (Betts, 2005). Empirically, the student-level data provided by the school board of Alachua county shows that high standards in grading system differentially affect students, with initially high-ability students experiencing the largest benefit (at least in reading) from high standards and that more interestingly, initially low-ability students benefit most from high standards when their classmates are of high-ability, while initially high-ability students benefit most from high standards when their classmates are of low-ability (Figlio and Lucas, 2004). Another study evaluated the effects of strict and lenient grading scales on students with high and low SAT scores, concluding that students with low SATs earned better test scores if they were graded on a strict rather than a lenient scale because they were grade-oriented and placed on an especially high value on grades (Johnson and Beck, 1988). Moreover, using the data from High School and Beyond survey, a study also concludes that higher standard raises test scores throughout the distribution of achievement, but that the increase is greatest toward the top of the test score distribution. In this way, more able students may increase effort to reach the new standard. Less able students must exert vast amounts of effort to increase their achievement. However, students at the lower end of the achievement distribution may be unaffected by the changes in grading standards (Betts and Grogger, 2003).

In addition to the strict or lenient grading scale, whether it is absolute or relative, can make a difference in students' behavior. Absolute grading system means that students will pass the class if they meet a given standard. While relative grading system means that the final assessment of one only, depends on his/her performance of students in the class. The author employed data of the students who studied in the University of Chile in twenty consecutive years. She found that low-ability students exert higher effort when the grading system changed from absolute to relative and that high-ability students reduced their effort after this change (Paredes, 2017). The grade system experienced the same transformation from absolute grading, numerical score, to relative grading based on each different score interval, letter criteria after the DR policy. In our study, we show some differences between Chile and China when it comes to the similar grading change. To be more specific, the transformation reduces the grades of students in all score ranges, with a more pronounced decline in the high-score range. Therefore, we try to sort out the reason why such a heterogeneity exists between Chile and China and to provide evidence in the empirical analysis about the grading change.

There are several views about the effect of grading system, including positive and negative comments. Boleslavsky and Cotton (2015) reckon that if high grades are assigned liberally, though a school would benefit, it would also cause 'grade inflation' that often imposes a welfare cost on employers and other evaluators. Sikora (2015) applied mathematical theory to proof that the overall best grading schemes are those which assign on average all possible passing grades with equal frequencies, while other grading schemes are sub-optimal. In our study, we

discuss the advantages and disadvantages of the grading system change that was induced from the DR policy. Our objective is not to give a universal answer about the grading system but to assess the applicability about the grading system in China and provide a new perspective for the world as a reference.

2.3. Labeling Theory

Labeling Theory, which focuses on how individuals are affected by social labels, leads to corresponding behavioral changes. When individuals are labeled (for example, as criminals), authority figures and peers begin to treat labeled people as though they possess an undesirable personality trait (Braithwaite, 1989; Lemert, 1951). Duxbury and Haynie (2020) focus on the relationship between school punishment and adolescent academic achievement. They propose two labeling mechanisms that change behavior: individuals alter their self-assessment, and there is the peer effect resulting from changes in the conventional social network. This labeling mechanism explains the association between the labels attached to students and the changes in student academic performance.

Boosting one's Academic Self-Concept (ASC) is linked to achieving a diverse range of educational goals. The more confident an individual is in her/his academic abilities (as indicated by a higher ASC), the more likely they are to excel academically. Conversely, a higher ASC is positively associated with academic proficiency (Marsh and Seaton, 2015).

Regarding the peer effects, Sacerdote (2011) points out that it has different mechanisms and degrees of influence on high-scoring and low-scoring students. Low-scoring students may experience pressure and frustration due to their high-scoring peers in their surroundings, which could lead to even poorer performance. As Hoxby (2000) noted, there is probably non-linear impact of peer effect, for instance, different genders or racial groups may vary. Furthermore, Hanushek, Kain, Markman, and Rivkin (2001) provide little evidence that the heterogeneity of peers in terms of variation in achievement levels affect growth in mathematics achievement, though students in the highest quartile do show a smaller peer effect. Burke and Sass (2011) confirmed that sizable effects observed in the non-linear models are obscured in the linear-in-means-models, after controlling both student fixed effects and teacher effects. Peer effects are somewhat smaller for students in the middle third of the distribution and smaller still for students in the highest third. Giannola, Busso and Berlinski (2022) indicate that students with lower grades can significantly influence the academic performance of high-achieving students.

In our research, a label can be considered as the impression of each student based on their past academic achievements, such as a high-score student or a low-score student. We identify a sample with blurred labels, allowing us to explore the impact of excluding certain labels on the academic performance of high school students—and this impact is heterogeneous. Our findings provide reverse confirmation of the conclusions reached by previous label theory, as both self-assessment and peer effect among students may be weakened potentially in our sample.

3. Theoretical Framework

Whether the students exert more effort or not greatly depends on the incentive they receive from the transformation of grading system. To better illustrate our findings, we introduce the labor market signaling model (Spence, 1978) and adapt it to explain the students' behavior. Because students do not exactly know what specific scores, they obtain from the examination but get a coarser grade, the information asymmetry is deteriorated. As Spence (1978) showed in his paper, we also suppose that a student's academic output is directly proportional to the level of educational effort or diligence and that the higher-ability students typically have higher initial talent or abilities than their lower-ability peers, as indicated in the Figure 1 below $-MP_H$ (the marginal production of more capable student) is higher but still parallels to MP_L (the marginal production of less capable student). The 'wage' used in Spence's model in y-axis represents the score in our paper. Initially, when the numerical grading is used, the

lower-ability student would spend a lower level of educational effort, represented as point e_L on the x-axis. At the same time, their indifference curve intersects with their marginal output curve at the equilibrium point E_L. On the contrary, the higher-ability student would spend a higher level of educational effort at e_H and reach equilibrium at E_H.

After the DR policy, the school changed the grading system from numerical score to letter one, resulting in a change of students' behavior. For the lower-ability students, in lower letter area, their marginal production curve no longer slopes upward. Because no matter how much effort students exert, as long as the grade locates in the same score interval, they still achieve same letter grade, leading to the same 'production level'. Therefore, the marginal production curve becomes a horizontal line parallel to the X-axis, represented as line $E_L'E_L$ on Figure 1**.** Within this range, regardless of the grade achieved, their output, as reflected in the letter grade, remains the same. Therefore, the lower-ability students would spend less effort than before, an educational effort on e_L'. The same situation also applies to higher-ability students, leading to a new marginal production curve $E_H'E_H$. Eventually, they would spend less effort and reach an equilibrium on E_H'. Hence, we can derive the first hypothesis from the policy implementation.

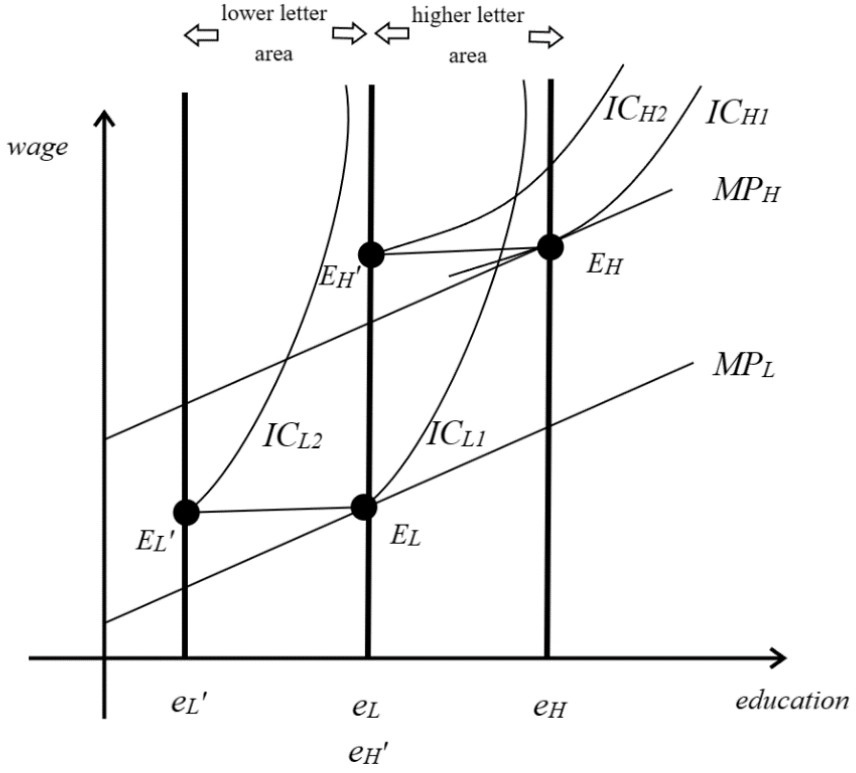


Figure 1. Effects of Grading Transformation on Students' Learning.

Source: Own construction.

Above, we have only discussed the extreme cases in which policies induce changes in student behavior. However, to what extent the student would be demotivated in one letter area remains unknown. Due to the law of diminishing marginal utility, high-ability students might reduce their efforts even further because of the excessive 'pain' they endure from investing extra effort in the 'rat race' for each additional point. Consequently, this could lead to a more significant decline in their scores. Therefore, for students of different abilities, we make the following heterogeneous assumptions.

Hypothesis 1: Compared to numerical grading, letter grading can lead to a demotivation of 'rat race' in students'

efforts in learning.

Hypothesis 2: The policy impact is heterogeneous and is expected to narrow the overall score difference (range/variance).

Hypothesis 3: The policy impact is heterogeneous. High-ability students' scores are expected to decrease more significantly compared to low-ability students.

4. Institutional setting and Data

4.1. The Compulsory Education System in China

Most schools in China are public institutions funded and staffed by the government's education authorities, granting the government significant control over schools. The Compulsory Education Law of the People's Republic of China mandates nine years of compulsory education, requiring all children aged six and above to attend school. In most regions of China, school-age children typically receive six years of primary education followed by three years of junior secondary education. Public primary and junior secondary schools generally admit children of school age based on their residence without entrance examinations.

After completing junior secondary education, students usually need to take the High School Entrance Examination (habitually called 'Zhongkao' in Chinese) to qualify for admission to senior high schools or vocational schools. The Ministry of Education determines the general curriculum structure and the proportion of instructional hours for compulsory education, but schools in different regions have some autonomy to make adjustments within these guidelines.

During the survey period for this study, in Guangzhou, the main subjects taught and assessed in primary schools included Chinese, Mathematics, and English. Other subjects, such as Music, Art, Physical Education, Natural Science, and Information Technology, were taught but not assessed. Chinese and Mathematics were introduced starting from the first grade, while English was typically introduced from the third grade. At the junior secondary level, the range of subjects taught and assessed expanded to include Politics, Geography, Biology, History, Physics, Chemistry, and Physical Education. Specifically, Politics, Geography, Biology, and History were introduced in the seventh grade (the first year of junior high school); Physics was added in the eighth grade; and Chemistry was introduced in the ninth grade. However, due to the focus of the Zhongkao in Guangzhou, which assesses only Chinese, Mathematics, English, Politics, Physics, and Chemistry, schools generally discontinue teaching Geography, Biology, and History in the ninth grade to allow students to concentrate on the subjects included in the examination.

4.2. Policy Background

In March 2021, six Chinese government departments, including the Ministry of Education, unexpectedly issued an order to primary and secondary schools: strict control of the frequency of examinations, and no disclosure of the examination results and rankings. Three months later, the Ministry of Education reiterated that students' examination scores, rankings, and other academic information should be made available to students and parents but should not be made public. In July 2021, the Chinese government officially introduced the highly anticipated DR policy.

By the end of August 2021, the Education Bureau of Guangzhou City responded to the Ministry of Education's instructions by issuing two notices: the first notice required subordinate district education bureaus to vigorously crack down on 'shadow education institutions', while the second notice provided six specific operational requirements for the primary and secondary schools under its jurisdiction. Among these requirements,[1] the first

[1] Other requirements of this policy are not related to our research.

and third are explicitly stated: (1) Strictly prohibit illegal supplementary classes. Compulsory education schools must not, under any circumstances or in any form, organize students for extra classes during holidays or rest times or organize or require students to attend training at external training institutions. (2) Strictly control the frequency of examinations. No more than 2 unified examinations per semester in junior high school. Holding special classes is strictly prohibited, as well as conducting any form of examinations or tests for the purpose of class allocation based on student performance. Student rankings based on examination results are not allowed, and examination scores should be presented using a level evaluation system. Public disclosure of student scores and rankings in any form is strictly prohibited. These two points can be considered as the primary impact of letter grading.

It is important to note that Zhongkao still employs a numerical scoring system. Candidates are required to submit their school preferences before the Zhongkao, and high schools rank and admit students based on their numerical Zhongkao scores. Therefore, the DR policy does not mandate the use of letter grading for assessing students in junior high and primary schools; instead, numerical scores are initially generated and later converted into letter grades, with only the latter being allowed to be communicated to students. However, the specific method of transitioning from numerical grading to letter grading is not explicitly stated in the above directive policy documents, which gives schools significant discretion. In practice, some schools directly convert the percentage scores into classic five-letter categories (ABCDF) with pluses and minuses, for example, 95-100 points are classified as A+. Some schools convert them into ABCD categories based on the ranking of scores. In our paper, we particularly examine the latter one, which is a relevant grading.

We found a junior middle school in a small town in Guangzhou that met the requirements for this natural experiment prompted by the DR policy. This is the only middle school in the town, with an enrollment range covering all eligible students in the town. Importantly, the town has never had any publicly listed middle school tutoring institutions, so students have had very limited opportunities to participate in external training institutions. Due to the high transportation and time opportunity cost, an overwhelming proportion of students would not participate in external courses before and after policy period. They have been largely unaffected by the education and training institution industry, which means that they can be considered unaffected by the policy that preceded the DR policy.

To sum up, for students who had already enrolled in the school before the fall semester of 2021, the substantial impact of the formal implementation of the DR policy was that they could no longer know their specific rankings after formal examinations and could only receive grade-based results (A, B, C, D). The four grades corresponded to the top 25%, 25%–50%, 50%–75%, and 75%–100% in terms of ranking. Since students had access to correct answers after the examinations (e.g., teacher explanations and online searches), they could generally infer their examination scores. Therefore, not disclosing the ranking essentially blurred the students' ranking information. Hiding their specific ranking information from students created asymmetrical information, making students the disadvantaged party in terms of information. How this information asymmetry will change students' motivation and ultimately affect their behavior (reflected in future performance) is worth exploring.

4.3. Data

We obtained the examination scores[2] of three cohorts of students (admitted in 2018, 2019, and 2020) from this junior high school in Guangzhou. Each cohort comprises approximately 330 students, divided into eight classes randomly. Except for students who were transferred during the school year, each student studied at the school for three years, with variations in the subjects studied in different school years. The school is allowed to organize two

[2] Some physical education scores are included in the dataset, but these scores is not affected by the 'DR' policy, because the results (such as running times) cannot be kept confidential technically.

formal examinations, mid-term and final, during each semester. The examination content is determined by the school's academic affairs department. Notably, there are no final examinations in the spring semester of the third year of junior high (the ninth grade); instead, the students take the Zhongkao. Therefore, a complete dataset would include the scores for eleven consecutive examinations taken by the students, spanning from the fall semester of 2018 to the spring semester of 2023. Particularly, we use the mathematics score first to substantiate our study for the reasons are not only that most of studies use mathematics score as a proxy to gauge students 'achievement, but also mathematics score is one of the essential components of students' academic performance and a significant reflection of their knowledge and abilities. Therefore, researching the impact of a student's mathematics grades and their peers on academic performance can provide a better understanding of how peers influence students' academic achievements, further emphasizing the importance of peer interactions in student learning. Additionally, mathematics grades offer an objective and quantifiable measure, making it easier to study and analyze the influence of peers. Furthermore, we use the Chinese and English score to execute our robustness test. Additionally, we have obtained a list of mathematics teachers, and the teachers do not change their assigned classes midway through the semester. We do not use actual calendar dates as time variables; instead, we use the sequence of examinations after students enter middle school. For instance, the first examination taken by students in the first semester of middle school is denoted as t = 1. This approach ensures that the scores of the three cohorts of students are comparable over time. For the students admitted in 2018 (graduating in the summer of 2021), they were not affected by the policy. For the students admitted in 2019, the policy took effect at t = 9, where examination scores were no longer publicly ranked, which means from t = 10 onwards, they would not know the scores of the previous examination due to the implementation of DR policy. Similarly, for the students admitted in 2020, the policy started at t = 5, and from t = 6 onwards, they would not know the scores of the previous exam. Table 1 below presents the descriptive statistics of the mathematics score and other variables of the 3 cohorts.

Table 1. Summary Description.

A. Descriptive statistics of mathematics score								
Cohort	N	Mean	Max	Min	Q25	Q50	Q75	Std
2018	3205	59.03	114	1	40	63	79	24.57
2019(pre-treat)	2638	60.21	120	0	41	63	80	25.55
2019(post-treat)	648	69.41	112	2	55	75	88	24.84
2020(pre-treat)	1760	67.65	120	0	49	73	89	27.27
2020(post-treat)	2075	65.42	116	0	44	72	88	27.80

B. Descriptive statistics of gender			
Cohort	Male	Female	Sum
2018	189 (57.8%)	138 (42.2%)	327
2019	177 (53.2%)	156 (46.8%)	333
2020	181 (50.7%)	176 (49.3%)	357

C. Descriptive statistics of mathematics teachers.		
Cohort	Name of Teacher	Number of Students
2018	Fu	77
	Lay	81
	Lei	82
	Liang	82
	Liu	41
	Wu	41
	Xu	41
	Zhong	80
	Zou	82
2019	Liang	82
	Tan	163

2020	Tang	83
	Xiang	83
	Xun	84
	Chen	87
	Cheung	88
	Guo	89
	Lee	89
	Leung	89
	Li	88
	Liu	44
	Tan	89
	Xu	45
	Zhang	88

Source: Own construction.

5. Empirical Strategy

5.1. The Standard Difference-in-Differences (DID) Model

We first construct the following multi-time period DID model to estimate the impact of the policy shock on overall student performance:

$$Math_{i,t} = \alpha + \beta_1 \cdot Conceal_{i,t} + \beta_2 \cdot Controls_{i,t} + \mu_i + \theta_t + \varepsilon_{i,t} \quad (1)$$

in which i and t represent students and examination periods, respectively. The explained variable, $Math_{i,t}$, which is the mathematics score achieved by student i in the t-th examination. The policy dummy variable $Conceal_{i,t}$ is taking value 1 when student i affected by the policy and the policy had been implemented before the t-th examination, otherwise, it takes the value 0. Individual fixed effects μ_i and time fixed effects θ_t are both controlled for, where the former captures individuals that do not change over time, and the latter captures common factors changing over time for individuals. $Controls$ are a set of control variables, including student i's gender and class, as well as the student i's mathematics teacher in the t-th examination period. $\varepsilon_{i,t}$ is the error term. It is important to note that, due to the absence of any inter-class transfers among all students, there exists a complete multicollinearity between individual fixed effects and either class fixed effects or cohort fixed effects.

To make full use of the dataset and improve the goodness of fit, we made a slight modification to Equation (1). The original individual fixed effects μ_i in Equation (1) were replaced by class fixed effects, as well as by the average ranking percentage $PGBHscore_i$ for student *i* in the four subjects of Politics, Geography, Biology, and History in their first examination after entering the middle school (i.e., t =1). The reason for selecting these 4 subjects is that students only study Chinese, Mathematics, and English in primary school, while Politics, Geography, Biology, and History are new subjects introduced in junior high. $PGBHscore_i$ is to some extent reflecting student i's academic foundation unrelated to a specific discipline. Table 2 reports the estimation results for Equations (1). To examine the robustness of the model, each column in Table 2 presents the estimated policy effect under different sets of control variables.

Table 2. Effect of Policy Shock Interpreted by Difference-in-Differences Model.

	(1)	(2)	(3)	(4)	(5)	(6)
Conceal	-3.245***	-3.368***	-3.249***	-3.248***	-3.366***	-3.365***
	(0.448)	(0.404)	(0.451)	(0.451)	(0.407)	(0.407)
gender			2.008**		1.993**	
			(0.857)		(0.856)	
PGBHscore			0.778***	0.774***	0.778***	0.774***

			(0.015)	(0.016)	(0.015)	(0.016)
Mathematics Teacher	Yes		Yes	Yes		
Class			Yes	Yes	Yes	Yes
Individual-fixed effect	Yes	Yes				
Time-fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10323	10323	10272	10272	10272	10272

*Note: Robust standard errors in parentheses are all clustered at the student level. * p<0.1, ** p<0.05, *** p<0.01. The first 2 columns control for individual fixed effects, while the subsequent 4 columns replace them with PGBHscore and control for class fixed effects. Column (1) presents the regression results of the baseline model represented by Equation (1); columns (2) and (5) do not control for mathematics teacher; columns (4) do not control for the gender; the estimates in columns (6) do not control for neither gender nor mathematics teacher. The gender variable is coded with females as the reference category, where females are assigned a value of 0, and males are assigned a value of 1. Source: Own construction.*

As seen in Table 2, the estimated coefficients of the policy shock dummy variable $Conceal_{i,t}$ that we are interested in are all statistically significant, and the estimation results that are numerically close also to some extent reflect the robustness of the model. This indicates that changing the grade disclosing policy has an overall impact on student scores, and the negative coefficient implies that the average effect of policy implementation on student scores is negative. This negative effect on scores is consistent with Hypothesis 1, which implies that transforming the numerical grading to letter grading results in a demotivation of 'rat race' in students' efforts in learning.

We provide the following explanation for this: for students, the time and effort they invested in their studies in the previous period can receive a positive incentive through the ranking of grades – better grades are the reward for diligent studying. However, after the policy implementation, this incentive is replaced by much coarser grade levels. The difficulty of moving up a grade level (e.g. from C to B) is far greater than gaining an extra point or improving by one place. Apart from a small group of students whose rankings are close to certain quartiles, the marginal cost of improving a grade level increases significantly for others. Consequently, students may be more inclined to maintain their current grade levels rather than seeking improvement. For one-quarter of students with an A grade level, there would not be any evidence of improvement on their report cards, something that was previously only seen for students who ranked first in their grade or scored full marks. Given that achieving a score of 99 and 85 both result in an A grade, but the former requires significantly more effort, considering the marginal cost, the optimal choice is to score just above the grade level. When most people make this choice, the average scores naturally decrease numerically.

To further observe the dynamic effects in each specific examination and simultaneously test the parallel trends assumption in the baseline DID model, we constructed the dynamic model as shown in Equation (2), using the relative time with reference to the examination when the policy started taking effect ($T - t_d = 0$). Here, t_d represents the first examination of each cohort for which numerical scores are not allowed to be disclosed, and $I(\cdot)$ is an indicator function, meaning that if the condition in parenthesis is satisfied, I = 1, otherwise, I = 0. $T - t_d$ indicates the timing of the t-th exam relative to the policy intervention, where negative values represent exams before the policy shock and positive values represent those after. For example, $T - t_d = 2$ refers to the second exam following the DR policy implementation. If the coefficients β_t^{precut} and β_t^{pre}s are not significantly different from 0, while the coefficients β_t^{post}s and $\beta_t^{postcut}$ are significantly different from 0, it indicates that the baseline DID model constructed in this study satisfies the parallel trends test.

$$Math_{i,t} = \alpha + \beta_t^{precut} \cdot I_i(T - t_d \le -5) + \sum_{t=-4}^{-1} \beta_t^{pre} \cdot I_i(T - t_d = t) + \sum_{t=1}^{4} \beta_t^{post} \cdot I_i(T - t_d = t) + \beta_t^{postcut} \cdot I_i(T - t_d \ge 5) + \beta \cdot Controls_{i,t} + \mu_i + \varepsilon_{i,t} \quad (2)$$

The estimated results of Equation (2) are presented in Table 3 and Figure 2. In Figure 2, the numbers on the horizontal axis correspond to the coefficient names in Table 3, representing the time relative to the policy

implementation point t_a. For example, 4 in the horizontal axis denotes the fourth examination after the policy took effect. It can be observed that coefficients for the β_{-2}^{pre}, the penultimate examination, or other earlier examinations before policy implementation are not statistically significant, with 95% confidence intervals covering zero line in Figure 2. In contrast, after policy took effect (points to the right of the dashed line), all coefficients are significantly negative, further confirming the robustness of the estimates in Table 2. However, as shown in Table 3, the coefficient for the examination just one period before policy implementation is significantly positive, suggesting the possible existence of some unaccounted-for destabilizing factors, probably attributed to student heterogeneity. In the placebo test in section 4.4.1, we artificially generated placebo variable, which moves the time of policy implementation one period ahead to further check this issue.

Table 3. Dynamic Effects in Each Examination.

△	-5	-4	-3	-2	-1	1	2	3	4	5
$\boldsymbol{\beta_t}$	-2.412	-0.407	0.985	-0.842	3.241***	-6.264***	-1.343*	-11.586***	-2.326*	-6.045***
	(1.28)	(0.91)	(0.73)	(0.60)	(0.45)	(0.49)	(0.63)	(0.75)	(1.08)	(1.27)

*Note: Robust standard errors in parentheses are all clustered at the student level. * p<0.05, ** p<0.01, *** p<0.001.*

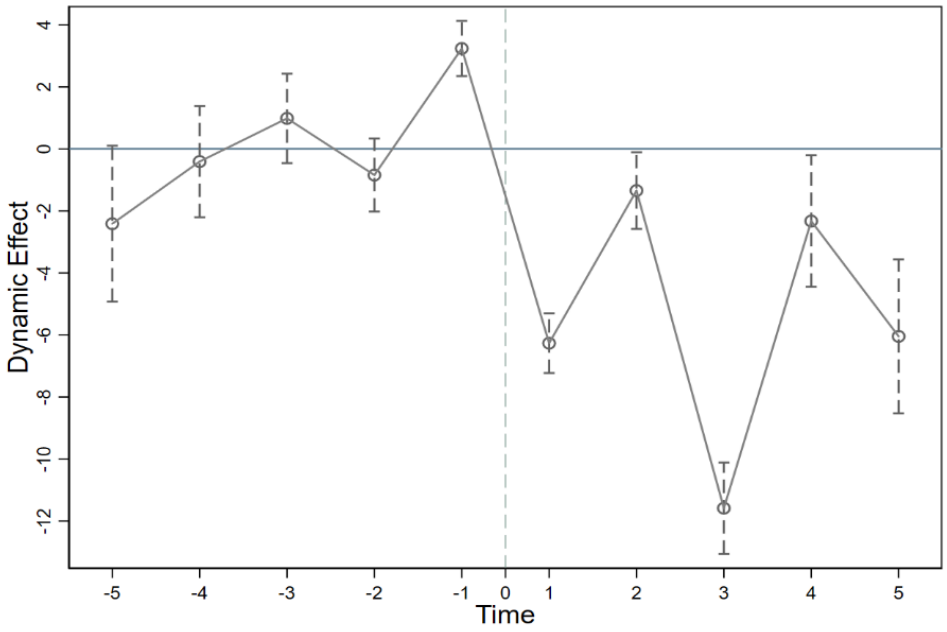


Figure 2. The Dynamic Effects in Each Examination.

Source: Own construction.

5.2. Heterogeneity program effects

We then conducted a Wilcoxon two-sample rank-sum test to non-parametrically examine whether the distribution of students' scores is different before and after the policy shock. The two samples involved in the test are the scores of the treatment group before and after the policy implementation.

From Table 4, it can be observed that the p-value is very small, allowing us to reject the null hypothesis, indicating that there is a significant difference in students' mathematics scores between before and after the treatment. This suggests the need to further investigate the heterogeneity of the treatment effect on top of the standard DID model.

Table 4. Wilcoxon Two-sample Rank-sum Test Result.

Cohort	N	Rank Sum	Expected
Pre-treat	4398	15181306	15661278
Post-treat	2723	10176575	9696603
Combined	7121	25357881	25357881
$\boldsymbol{Z = -5.694}$			Prob>$\lvert Z\rvert = 0.0000$

Note: $H0$: $Pre - treat = Post - treat$. *Source: Own construction.*

Hence, we construct a quantile DID model based on Equation (1). It seems inappropriate to reapply the dynamic effects model to test the parallel trends assumption in quantile regression. We draw inspiration from the innovative approach in the model proposed by Fang et al. (2020) and introduced the term $Pretrend_{i.t}$ into Equation (1), which takes the value 1 for students in the treatment group who have not yet affected the policy. The estimated coefficient γ, which is presented in Table 4, can be used to test the parallel trends assumption in the DID model.

$$Math_{i,t} = \alpha + \beta_1 \cdot Conceal_{i,t} + \gamma \cdot Pretrend_{i.t} + \beta_2 \cdot Controls_{i,t} + \beta_3 PGBHscore_i + \theta_t + \varepsilon_{i,t} \quad (3)$$

Subsequently, we performed simultaneous quantile regressions on each decile of Equation (3) to further explore the impact of the treatment effect on the shape of the student performance distribution. Standard errors were computed through bootstrapping for 400 repetitions. The results are presented in Table 5. It can be observed that the coefficients of $Pretrend_{i.t}$ are not significant at all quantiles, indicating that the parallel trends assumption is plausible. Regarding the estimated coefficients of the policy shock variable $Conceal_t$, at the lowest 20% of student scores (q10, q20), the t-test does not reject the null hypothesis, meaning that the policy did not have a significant impact on these students. However, from the 30th percentile (q30) up to the highest 90th percentile (q90), the coefficients are highly significant and appear to decrease progressively. In other words, compared to lower-scoring students, the policy has a more significant impact on higher-scoring students.

Table 5. The Policy Effect and Parallel Trend Test on Quantiles Interpreted by DID Model.

Quantile	q90	q80	q70	q60	q50	q40	q30	q20	q10
Conceal	-7.826***	-6.493***	-4.956***	-4.878***	-3.936***	-3.777**	-3.528**	-2.273	-2.508
	(1.443)	(1.153)	(1.156)	(1.284)	(1.364)	(1.349)	(1.298)	(1.459)	(1.769)
Pretrend	-2.167	-1.797	-2.721	-2.101	-1.967	-1.663	-1.192	-0.082	-2.558
	(1.668)	(1.403)	(1.445)	(1.477)	(1.405)	(1.550)	(1.415)	(1.763)	(1.993
Observations	10272								

*Note: Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001. Source: Own construction.*

To provide a more visual representation of how the coefficients change across quantiles, we have plotted Figure 3. As shown in the figure, the lines for each quantile coefficient slope downward. For students with lower scores, the policy's impact is approximately -3 points, but for the highest-performing students, the policy's effect can reach -8 points. Students in the middle range of performance experience impacts between these extremes, with negative treatment effects greater than those with lower scores and less than those with higher scores. This implies that the overall score differences among students have decreased as a policy consequence.

This fact is consistent with Hypothesis 2 and 3, which predict a smaller range in students' score and behavioral heterogeneity in different stratification, more demotivation for high-ability students and less demotivation for low-ability students. This result can be explained with Label Theory. Initially, students' self-assessment could be achieved as examination scores are released. However, with the DR policy, students can only obtain vague letter grades. The Academic Self-Concept of better-performing students loses its motivation, causing them to be more affected than their lower-performing counterparts. Meanwhile, peer effects among students have been weakened with the non-disclosure of numerical scores, reducing the academic psychological stress and frustration for lower-

performing students.

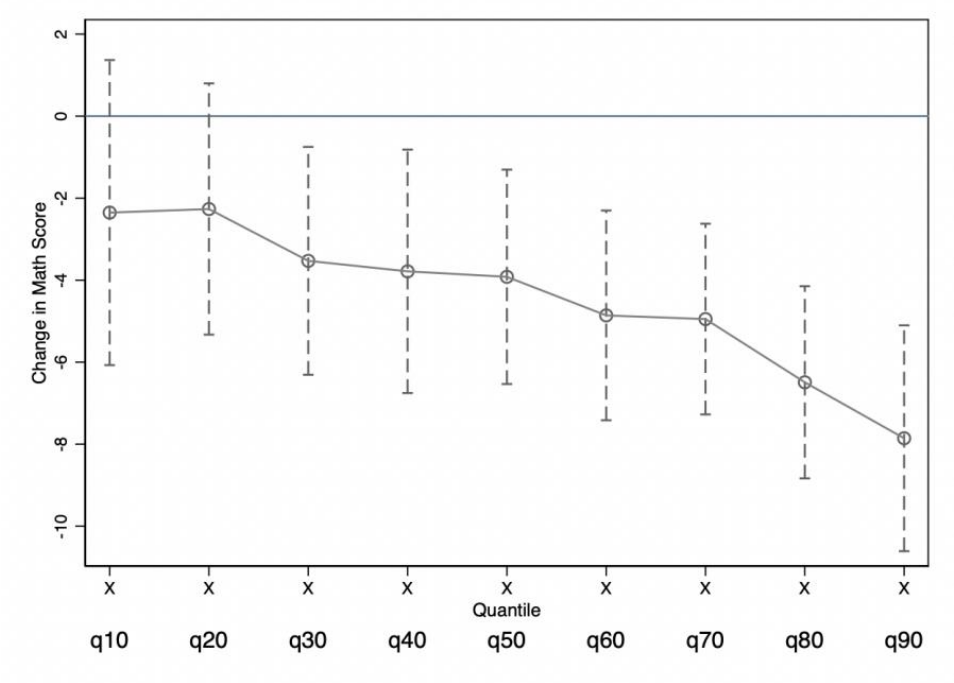


Figure 3. The Policy Effect on Different Quantile Interpreted by DID Model.

Source: Own construction.

Donald & Lang (2007) argue that in cases with few groups and specific group-level shocks affecting time trends, the traditional linear DID model may underestimate standard errors. Similarly, Bertrand, Duflo & Mullainathan (2005) point out that in multi-period DID models, the standard DID method can severely underestimate the standard deviation of estimates. Hence, we used an improved set of non-linear DID estimators introduced by Athey and Imbens (2005) - the Change-in-Change Estimator CIC - to re-estimate the treatment effect. In fact, even earlier, Meyer et al. (2007) and Poterba et al. (1995) applied DID estimates at specific quantiles, but Athey and Imbens' approach is applicable to the entire counterfactual distribution, making it more generalized than traditional DID. Additionally, CIC's conditions are less restrictive than standard DID and provide greater flexibility in its application. Athey and Imbens (2005) offer further details on the identification of the CIC model for both continuous and discrete outcomes, as well as its potential disadvantages. Lucas and Mbiti (2012), Borah, Burns, and Shah (2011) and Dai (2021) have empirically used the CIC model and pointed out some practical considerations.

Following the estimation framework from Athey and Imbens (2005), we employed the standard 2×2 model, using mathematical scores of students who admitted in 2018 as the control group and those who admitted in 2020 as the treatment group, with the same definition of the time window as before. The following assumptions are imperative to identifying the treatment effect of the DR policy using CIC estimator. At first the continuous variable Y denotes mathematics score, and it 'generated' by

$$Y^N = h(U, T) \tag{4}$$

where U represents unobservable individual characteristics, T is time indicator that takes the value 0 before policy and 1 after policy, and h is a nonlinear unknown function that does not vary across groups, and between-group differences are solely due to the distribution of U. Second, the 'production function' $h(u, t)$ is non-decreasing in u. Third, the individual characteristics U remain stationary over time within a given group:

$$U \perp T \mid G \tag{5}$$

Fourth, the support of the treatment group does not exceed the support of the control group:

$$supp[U|G=1] \subseteq supp[U|G=0] \tag{6}$$

Fifth, given a specific time T and outcome variable Y, individual characteristics U are independent of the group assignment G.

$$U \quad \perp \quad G|h(U,T),T \tag{7}$$

These assumptions are reasonable in our data context. The key assumption requests the distribution of the 'production function' U within groups remains constant over time because U's definition is based on individual student characteristics. Our choice of experimental and control groups depends on students' enrollment years (entry year), and there are no students who skipped grades or repeated grades in the sample. Meanwhile, we incorporated $PGBHscore$ as a covariate, which controls for individual characteristics of each student upon enrollment. There is ample reason to believe that these characteristics will not undergo significant changes over the relatively short observation period. Based on the above assumptions, the unobserved counterfactual CDF of the treatment group in the post-treatment period, $Y^N,11$ can be derived from the following formula:

$$F_{Y^N,11}(y) = F_{Y,10}\left(F_{Y,00}^{-1}\left(F_{Y,01}(y)\right)\right) \tag{8}$$

Compared to continuous variables, we consider mathematical numerical scores, which are recorded as integers, to be discrete variables, as these scores result from adding up the points earned for each correct answer on the test. Specifically, for discrete variable Y, the distribution can be obtained from the following equation:

$$F_{Y^N,11}(y) = \int_0^{F_{Y,01}(y)} f_{U,10}(u)du \tag{9}$$

where

$$f_{U,10}(u) = \sum_{k=1}^{K} 1\{F_{Y,00}(\lambda_{k-1}) < u \leq F_{Y,00}(\lambda_k)\} \cdot \frac{f_{Y,10}(\lambda_k)}{F_{Y,00}(\lambda_k) - F_{Y,00}(\lambda_{k-1})} \tag{10}$$

In that case, the Average Treatment Effect on the Treated (ATT) can be obtained from the following equations:

$$\tau^{DCIC} \equiv E[Y_{11}^I] - E[Y_{11}^N]$$

and

$$\tau_q^{DCIC} \equiv F_{Y^I,11}^{-1}(q) - F_{Y^N,11}^{-1}(q) \tag{11}$$

The results on each decile estimated by Equation (11) obtained in Table 6 and Figure 4 after repeating the bootstrap sampling 400 times are presented. The CIC estimates are slightly different numerically from our quantile regression estimates in Section 4.2, but as visually represented in Figure 4, the lines of the quantile coefficients still slope downward. Our previous conclusion remains valid and provide plausible evidence for our Hypothesis 2 and Hypothesis 3: the treatment effect of letter grading system has reduced overall score differences among students.

Table 6. The Policy Effect on Different Quantile Interpreted by CIC Model.

Quantile	q90	q80	q70	q60	q50	q40	q30	q20	q10
CIC	-8.164***	-7.633***	-6.482***	-5.521***	-5.067***	-4.160***	-2.665**	-1.836	-2.11
	(-1.975)	(-1.945)	(-1.305)	(-1.598)	(-1.217)	(-1.47)	(-1.335)	(-1.152)	(-1.36)
Observations	7024								

*Note: Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001. Source: Own construction.*

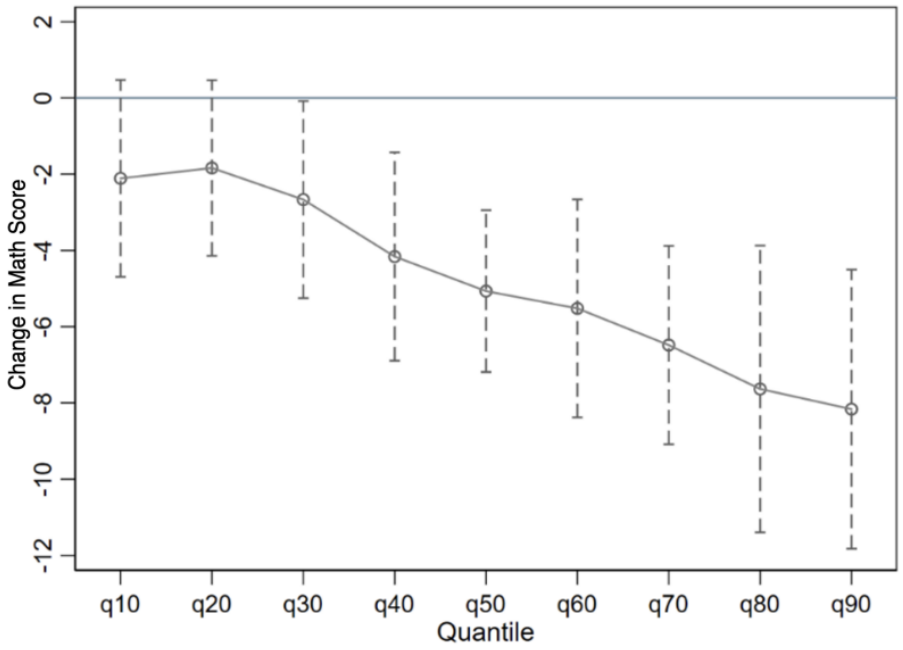


Figure 4. The Policy Effect on Different Quantile Interpreted by CIC Model.

Source: Own construction.

5.3. Robustness

5.3.1. Placebo Test

In order to explore the potential weaknesses of the previous evaluations, we conducted a placebo test to test their reliability. As seen in the dynamic effects model established by Equation (2), the conditions in the period prior to policy implementation appear somewhat questionable. Therefore, we attempted to move the time of policy implementation one period ahead to observe if it would yield a strong association that should not exist. Specifically, the core indicator variables $Conceal_{i,t}$ in Equations (1) and (3) were respectively replaced with $Placebo_{i,t} = Conceal_{i,t+1}$ (in particular, when t = 11 for students in the experimental group, $Placebo_{i,t}$ is set to 1). As shown in Table 7, the coefficients of $Placebo_{i,t}$ in the standard DID model are not significant, and most of the estimated results in the quantile models are also not significant. Compared to the significant effects observed in the previous section, this would suggest that the true policy intervention played a crucial role in the presence of causal effects, eliminating to some extent the temporal interfering factors.

Table 7. Placebo Test of Treatment Effect.

	Average	q90	q80	q70	q60	q50	q40	q30	q20	q10
Placebo	-0.802	-4.102**	-2.303	-0.586	-1.001	-0.673	-1.294	-0.778	-1.163	1.169
	(0.574)	(1.261)	(1.205)	(1.079)	(1.230)	(1.088)	(1.056)	(1.165)	(1.441)	(1.638)
Observations	10272	10272								

*Note: Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001. Source: Own construction.*

5.3.2. PSM (Propensity Score Matching)

To further examine the robustness of the conclusions drawn from the DID model, this study used propensity score matching combined with the double difference method, known as PSM-DID, to test the results. Because the changes in student performance in different score ranges may have inherent heterogeneity, i.e., with increasing age, high-scoring students may naturally have more pronounced changes in scores compared to low-scoring students.

This is determined by the characteristics of the student population itself, so the use of a simple DID model may not fully eliminate this type of sample selection bias, leading to biased policy effects. Therefore, this study, based on controlling for the average rank percentage of political, geographical, biological, and historical four subjects $PGBHscore_i$ and student gender variables sex_i, constructed a Logit model to determine whether students were affected by the score-concealing policy. This study then used radius matching with replacement to sample both the treatment and control groups for propensity score matching, and the matched sample was used for the baseline regression.

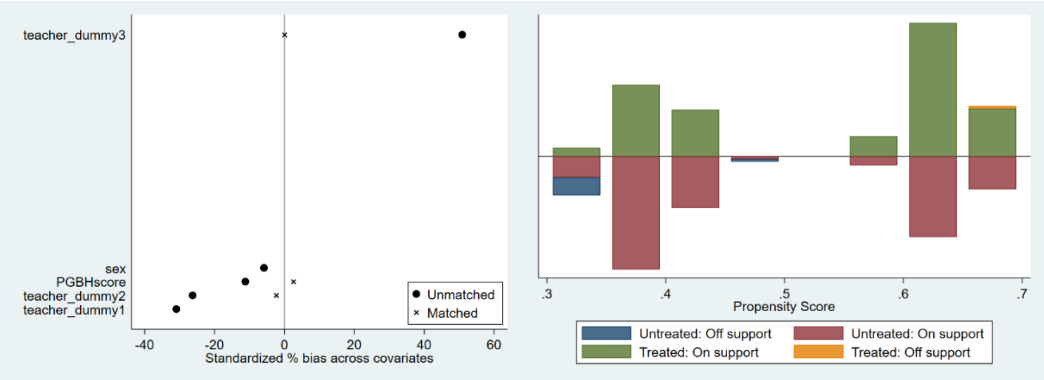


Figure 5. PSM Bias of Different Variable and PSM Propensity Score.

Source: Own construction

After propensity score matching, as shown on Figure 5, it was found that the matching bias of each control variable had decreased to less than 10%, and the matched sample indicated that the treatment and control group samples were more balanced, indicating a good match result.

The following Table 8 reports the regression results of the PSM-DID. The results show that the coefficient corresponding to $Conceal_{i,t}$ remains significantly negative, and numerically, the policy implementation is found to lead to a decrease of approximately 3.29 points in math scores. This result is consistent with the estimates of the baseline model before matching, indicating the reliability of the conclusions drawn from the original baseline model.

Table 8. Policy Effect After Using Propensity Score Match Strategy.

VARIABLES	Score
Conceal	-3.291***
	(-7.33)
Math Teacher	Yes
Sex	Yes
Class	-
PGBHscore	-
Time-fixed effect	Yes
Individual-fixed effect	Yes
Observations	10,264
R-squared	0.140

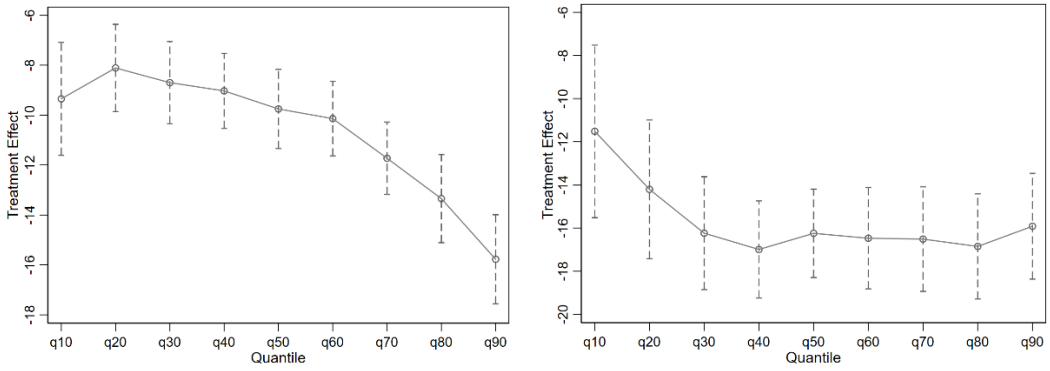
*Note: Robust t-statistics in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Source: Own construction.*

5.4. Other courses

Although we mentioned earlier that using Chinese and English scores as the dependent variables may have limitations, and it is possible that representing the treatment effect on students using regression coefficients, as in Equation (1) and (3), is likely unreliable, we are still interested in these results. As demonstrated in Table 9, the

regression results for Chinese and English exhibit significant differences in absolute values compared to mathematics, but the negative coefficients imply Hypothesis 1 can be true. Also, we compared the regression results of different quantiles in Figure 6, the trend provides probable evidence of Hypothesis 2 and 3.

Apparently, both Chinese and English grades at different quantiles decrease after the implementation of the DR policy, supporting Hypothesis 1. The differences in the absolute values of the estimated coefficients across various subjects reflect the heterogeneity of the impact of the DR policy on different subjects. Additionally, both Chinese and English grade decrease more in higher quantiles, but modestly in lower quantiles. It means that good students were affected more by DR policy than bad students. Such a fact supports Hypothesis 2 and Hypothesis 3.



(a) Chinese (b) English

Figure 6. The Policy Effect on Chinese and English Score.

Source: Own construction.

Table 9. The Policy Effect on Chinese and English Score.

	(a) Chinese	(b) English
Average	-9.778***	-17.43***
	(-0.477)	(-0.410)
q10	-9.348***	-11.514***
	(-1.154)	(-2.042)
q20	-8.111***	-14.206***
	(-0.892)	(-1.642)
q30	-8.700***	-16.238***
	(-0.84)	(-1.335)
q40	-9.030***	-16.993***
	(-0.767)	(-1.149)
q50	-9.754***	-16.245***
	(-0.808)	(-1.045_
q60	-10.141***	-16.472***
	(-0.763)	(-1.198)
q70	-11.728***	-16.512***
	(-0.739)	(-1.238)
q80	-13.345***	-16.853***
	(-0.901)	(-1.244)
q90	-15.776***	-15.917***
	(-0.911)	(-1.251)
Observations	8358	8360

*Note: Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001. Average estimates by function (2), and q* estimates by function (3). Source: Own construction.*

6. Summary and Conclusion

Our study delves into the impact of the DR policy in Chinese education. Since its implementation in July 2021, this policy has had profound effects on Chinese education. Prior to this policy, the competitive nature of education in China, often referred to as the 'rat race', led to extensive after-school classes and extracurricular activities for students, as parents sought to ensure their children remained competitive in the fiercely competitive school entrance examinations.

Our study presents three key findings. First, we use a difference-in-differences (DID) approach to estimate the policy's impact on student scores and find a negative effect on students' mathematics scores. Second, we reveal that students across all scores range experienced score declines, but students in higher percentiles experienced greater declines. This implies a reduction in the 'rat race' phenomenon in Chinese education. Additionally, we provide possible explanations for students' behavior based on labeling theory and grading system studies.

Our research contributes in several ways. Initially, we obtained a rare and valuable sample for studying students' academic performance, which allows us to exclude the influence of private tutoring institutions. It offers empirical evidence of the DR policy's impact from an innovative perspective and its applicability in the Chinese education context. Moreover, we provide insights into student behavior resulting from the transition of numerical grading to letter grading. We argue that traditional theoretical frameworks may not comprehensively explain the impact of such policy changes due to varying national conditions and educational models, requiring a case-by-case analysis.

This study uncovers the potential implications of information asymmetry on students' motivation and behavior, emphasizing the importance of addressing information disparities in education. The findings presented here offer valuable insights for the field of education in China and provide essential information for future research and policymaking related to the DR policy.

Funding Statement

This research was supported by the Youth Fund for Humanities and Social Sciences Research of the Ministry of Education of China, grant number 24YJC790084, and the Youth Fund of Philosophy and Social Sciences of Guangdong Province, grant number GD24YYJ20.

Acknowledgments

Acknowledgments to anonymous referees' comments and editor's effort.

Conflict of interest

All the authors claim that the manuscript is completely original. The authors also declare no conflict of interest.

References

Adams, G. S., and Torgerson, T. L. (1964). Measurement and Evaluation in Education, Psychology, and Guidance.

Athey, S., and Imbens, G. W. (2005). Identification and Inference in Nonlinear Difference-in-Differences Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.311920

Bertrand, M., Duflo, E., and Mullainathan, S. (2005). How Much Should We Trust Differences-in-Differences Estimates? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.288970

Betts, J. R. (2005). Do Grading Standards Affect the Incentive to Learn? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.76459

Betts, J. R., and Grogger, J. (2003). The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings. *Economics of Education Review*, 343–352.

https://doi.org/10.1016/s0272-7757(02)00059-6
Boleslavsky, R., and Cotton, C. (2015). Grading Standards and Education Quality. *American Economic Journal: Microeconomics*, 7, 248–79. https://doi.org/10.1257/mic.20130080
Borah, B. J., Burns, M. E., and Shah, N. D. (2011). Assessing the Impact of High Deductible Health Plans on Health-Care Utilization and Cost: A Changes-in-Changes Approach. *Health Economics*, 20, 1025–1042. https://doi.org/10.1002/hec.1757
Braithwaite, J. (1989). Crime, Shame and Reintegration. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804618
Burke, M. A., and Sass, T. R. (2011). Classroom Peer Effects and Student Achievement. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1260882
Dai, K. (2023). Double Reduction Policy in Education Industry and Firm Values: Evidence from China. *Finance Research Letters*, 54, 103696. https://doi.org/10.1016/j.frl.2023.103696
Dai, M. (2021). Enterprises' Environmental Performance and Environmental Protection Policy Evaluation. Jinan University.
Donald, S. G., and Lang, K. (2007). Inference with Difference-in-Differences and Other Panel Data. *Review of Economics and Statistics*, 221–233. https://doi.org/10.1162/rest.89.2.221
Dubey, P., and Gianakopulos, J. (2010). Grading Exams: 100, 99, 98, … or A, B, C? *Games and Economic Behaviour*, 69, 72–94. https://doi.org/10.1016/j.geb.2010.02.001
Duxbury, S. W., and Haynie, D. L. (2020). School Suspension and Social Selection: Labeling, Network Change, and Adolescent Academic Achievement. *Social Science Research*, 102365. https://doi.org/10.1016/j.ssresearch.2019.102365
Lemert, E. M. (1951). Social Pathology: A Systematic Approach to the Theory of Sociopathic Behavior. New York: McGraw-Hill Book Company, Inc. https://doi.org/10.2307/2571653
Fang, H., Wang, L., and Yang, Y. (2020). Human Mobility Restrictions and the Spread of the Novel Coronavirus (2019-nCoV) in China. *Journal of Public Economics*, 191, 104272. https://doi.org/10.1016/j.jpubeco.2020.104272
Figlio, D. N., and Lucas, M. E. (2004). Do High Grading Standards Affect Student Performance? *Journal of Public Economics*, 88, 1815–1834. https://doi.org/10.1016/S0047-2727(03)00039-2
Fu, C., Ou, H., Mo, T., and Liao, L. (2023). Effect Mechanism of Extracurricular Tuition and Implications on "Double Reduction" Policy: Extracurricular Tuition Intensity, Psychological Resilience, and Academic Performance. *Behavioral Sciences*, 13, 217. https://doi.org/10.3390/bs13030217
Giacomino, D. E., and Akers, M. D. (1998). An Examination of the Differences Between Personal Values and Value Types of Female and Male Accounting and Nonaccounting Majors. *Issues in Accounting Education*, 13, 565.
Giannola, M., Busso, M., and Berlinski, S. (2022). Helping Struggling Students and Benefiting All: Peer Effects in Primary Education. https://doi.org/10.1920/wp.ifs.2022.222
Gray, T., and Bunte, J. (2022). The Effect of Grades on Student Performance: Evidence from a Quasi-Experiment. *College Teaching*, 70, 15–28. https://doi.org/10.1080/87567555.2020.1865865
Guo, Y. (2022). The Current Impact of the Double Reduction Policy. In: Proceedings of the 2021 International Conference on Education, Language and Art (ICELA 2021). Atlantis Press, 147–152. https://doi.org/10.2991/assehr.k.220131.026
Hanushek, E. A., Kain, J. F., and Markman, J. M., Rivkin, S. G. (2001). Does Peer Ability Affect Student Achievement. *Journal of Applied Econometrics*.
Hoxby, C. M. (2000). Peer Effects in the Classroom: Learning from Gender and Race Variation. *National Bureau of Economic Research*.
Hu, Y., Yuan, J., and Wang, Y. (2021). The Heterogeneous Impact of Extracurricular Tutoring on Academic Performance in Different Subjects for Middle School Students. *Journal of Capital Normal University (Social Sciences Edition)*, 167–178.
Huey, M. E., Silvey, P. R., Vaughan, A. G., and Fisher, A. L. (2022). Assessing the Impact of Standards-Based Grading Policy Changes on Student Performance and Practice Work Completion in Secondary Mathematics. *Studies in Educational Evaluation*, 101211. https://doi.org/10.1016/j.stueduc.2022.101211
Spence, M. (1978). Job Market Signaling. In: Uncertainty in Economics. Elsevier, 281–306.
Johnson, B. G., and Beck, H. P. (1988). Strict and Lenient Grading Scales: How Do They Affect the Performance of College Students with High and Low SAT Scores? *Teaching of Psychology*, 15, 127–131. https://doi.org/10.1207/s15328023top1503_4
Lucas, A. M., and Mbiti, I. M. (2012). Access, Sorting, and Achievement: The Short-Run Effects of Free Primary Education in Kenya. *American Economic Journal: Applied Economics*, 4, 226–53. https://doi.org/10.1257/app.4.4.226

Marsh, H.W., Seaton, M. 2015. The Big-Fish–Little-Pond Effect, Competence Self-perceptions, and Relativity: Substantive Advances and Methodological Innovation. In: Advances in Motivation Science. Elsevier, 127–184. https://doi.org/10.1016/bs.adms.2015.05.002

McClure, J. E., and Spector, L. C. (2005). Plus/Minus Grading and Motivation: An Empirical Study of Student Choice and Performance. *Assessment & Evaluation in Higher Education*, 30, 571–579. https://doi.org/10.1080/02602930500260605

Meyer, B., Viscusi, W. K., and Durbin, D. (2007). Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *The American Economic Review*. https://doi.org/10.3386/w3494

Micha, E., Sekar, S., and Shah, N. (2023). What is Best for Students, Numerical Scores or Letter Grades?

Paredes, V. (2017). Grading System and Student Effort. *Education Finance and Policy*, 12, 107–128. https://doi.org/10.1162/EDFP_a_00195

Poterba, J. M., Venti, S. F., and Wise, D. A. (1995). Do 401(k) Contributions Crowd Out Other Personal Saving. *Journal of Public Economics*, 1–32. https://doi.org/10.1016/0047-2727(94)01462-w

Qian, H., Walker, A., and Xu, X. (2023). Running Schools on Two Legs: The Impact of Policy Oscillation on a Public-Private Partnership School in China. *International Journal of Educational Development*, 100, 102806. https://doi.org/10.1016/j.ijedudev.2023.102806

Ridener, L. R. (1999). Effects of College Major on Ecological Worldviews: A Comparison of Business, Science, and Other Students. *Journal of Education for Business*, 75, 15–21. https://doi.org/10.1080/08832329909598984

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. Elsevier, 249–277. https://doi.org/10.1016/B978-0-444-53429-3.00004-1

Sikora, A. S. (2015). Mathematical Theory of Student Assessment Through Grading.

Song, M. (2022). Under the Implementation of Double Reduction Policy. In: Proceedings of the 2021 International Conference on Education, Language and Art (ICELA 2021). Atlantis Press, 800–804. https://doi.org/10.2991/assehr.k.220131.146

Walvoord, B. E., and Anderson, V. J. (2011). Effective Grading: A Tool for Learning and Assessment in College. John Wiley & Sons.

Yin, Y., and Lai, Z. (2021). Research on the Transformation of Educational Institutions Under the Policy of Double Reduction. In: Proceedings of the 2021 4th International Conference on Humanities Education and Social Sciences (ICHESS 2021). Atlantis Press, 1530–1534. https://doi.org/10.2991/assehr.k.211220.258