International Journal of Educational Development 103 (2023) 102927

Contents lists available at ScienceDirect

International Journal of Educational Development

journal homepage: www.elsevier.com/locate/ijedudev

Association between ability tracking and student's academic and non-academic outcomes: Empirical evidence from junior high schools in rural China

Shriyam Gupta [a], Chengfang Liu [b,*], Shaoping Li [c], Fang Chang [d], Yaojiang Shi [d]

[a] Department of Sociology, University of Maryland, 20742, United States
[b] China Center for Agricultural Policy, School of Advanced Agricultural Sciences, Peking University, Haidian District, Beijing 100871, China
[c] School of Economics and Management, Huzhou University, Huzhou, Zhejiang Province 313000, China
[d] Center for Experimental Economics in Education, Shaanxi Normal University, Xi'an, Shaanxi Province 710119, China

ARTICLE INFO

Keywords:
Ability tracking
Academic outcomes
Non-academic outcomes
Rural China
Quasi-experiment
Peer effects

ABSTRACT

Ability tracking, a practice of grouping students into different classrooms based on their test scores, is prevalent in schools around the world. However, evidence on the effects of ability tracking on students' learning outcomes remains mixed. Drawing on a longitudinal dataset of 9170 students across 119 rural junior high schools from 23 counties in two provinces of China, this paper examines how ability tracking affects students' math score, math academic self-concept, and math anxiety. We find that ability tracking had no statistically significant effect on either students' academic or non-academic outcomes. Sub-group analyses by high and low-ability classrooms revealed similar results for math scores and math self-concept. However, ability tracking helped reduce the math anxiety of high-ability classroom students as they experienced a lower value added in their anxiety score by 0.103 SD ($p < 0.05$) relative to students in schools that do not practice ability tracking. Furthermore, heterogeneous analyses revealed that ability tracking is associated with a lower value added in math score of low-ability boarding students by 0.168 SD ($p < 0.05$) relative to non-boarding students in comparison schools.

1. Introduction

Classroom composition and peer effects are key concerns amongst both policymakers and researchers (Sacerdote, 2011), especially as a large body of literature has shown that peers can affect their classmates' learning outcomes (Coleman et al., 1966; Ding and Lehrer, 2007; Zhang et al., 2014). Particular attention has been paid to organizing classrooms based on academic 'ability'. Referred to as ability tracking/grouping, the process involves dividing students into different groups based on their 'ability'. The practice of ability tracking is common around the world and across all education levels including primary, secondary, and even college (Cheung and Rudowicz, 2003; Hanushek and W ößmann, 2006; Betts, 2011; Duflo et al., 2011; Carman and Zhang, 2012; Booij et al., 2016; Steenbergen-Hu et al., 2016; Li et al., 2018).

While common, ability tracking is still one of the most controversial educational policies around the world (Steenbergen-Hu et al., 2016). Some researchers argue that grouping students on similar skill levels could help with curriculum design and learning (Lou et al., 2000), while others argue that the policy leads to neglect of low-ability students, leading to greater learning disparity between students (Gamoran, 1992; Gamoran et al., 1995). Further, empirical evidence on the effect of ability tracking is still concentrated in the developed world with mixed results. Some show that ability tracking can be beneficial for high- and low-ability students (Collins and Gan, 2013), while others argue that it can have adverse effects, especially for low-ability students, and increase learning gaps amongst student groups (Argys et al., 1996; Fu and Mehta, 2018; Antonovics et al., 2022). However, most studies on ability tracking have come mainly from the developed world, reflecting data limitations and methodological challenges in developing contexts (Duflo et al., 2011; Betts, 2011; Sacerdote, 2011; Angrist, 2014).

Research over the last decade has extended the literature on ability tracking beyond the developed world (Duflo et al., 2011; Wang, 2015; Li et al., 2018). Nonetheless, prior findings on ability tracking from both developed and developing countries may not be applicable in the context of China for three reasons. First, peer effects are highly contextual (Paloyo, 2020), and findings may not translate across

* Corresponding author.
E-mail address: cfliu.ccap@pku.edu.cn (C. Liu).

https://doi.org/10.1016/j.ijedudev.2023.102927
Received 2 June 2021; Received in revised form 6 October 2023; Accepted 26 October 2023
Available online 7 November 2023
0738-0593/© 2023 Published by Elsevier Ltd.

education levels or geographic boundaries. Second, there exist significant institutional differences in the way ability tracking is practiced in schools (Zhang et al., 2014). In China, between-class ability tracking is not based on a single subject (like it is in the US) but is done on administrative class units where students are grouped across classrooms and 'key' classes (high-ability classrooms) are usually provided with the favoured educational resources, especially the best teachers, for all subjects. This form of ability tracking is usually formal and based on test scores (Zhang et al., 2014). It is unclear if different educational institutions may yield different effects of ability tracking on students. Last but not least, the high social value placed on academic achievement and competition may be particularly likely to create different social environments for high-ability and low-ability tracked students (Li et al., 2018). Research has shown different levels of competitiveness in cultures can lead to different effects of ability tracking (Thiemann, 2017). Thus, further studies are needed from new geographies to assess if ability tracking has a differential influence in different settings.

Using a longitudinal dataset where we followed the same set of students across two time periods from rural China, we take a quasi-experimental approach to assess the effects of ability tracking on students' academic and non-academic outcomes. We find that ability tracking had no effect on students' math test scores, math self-concept or math anxiety. Sub-group analyses reveal that within schools that practice ability tracking, students in high- and low-ability classrooms experience no statistically significant effect of ability tracking on their math score or self-concept when compared to students in the comparison group. However, high-ability classroom students in schools that practice ability tracking experience a statistically significant less value added in their anxiety by 0.103 SD ($p < 0.05$) relative to students in schools that do not practice ability tracking (the comparison group). Heterogeneous analyses revealed that ability tracking is associated with a lesser value added in math score of low-ability boarding students by 0.168 SD ($p < 0.05$) relative to non-boarding students in the comparison group.

The rest of the paper is organized as follows. Section 2 provides an overview of the literature and states the purpose of the study. Section 3 describes the data, followed by the research design and empirical strategy in Section 4. Section 5 presents the empirical results. Finally, Section 6 concludes with discussions and implications.

2. Literature review and purpose of the study

Ability tracking, or tracking or grouping, refers to the act of purposely sorting students based on their abilities (Betts, 2011; Belfi et al., 2012). 'Ability' here refers to academic ability and is often measured by test scores, although allotment can occur based on a combination of other factors such as IQ tests, or even teacher judgement (Hattie, 2002). Steenbergen-Hu et al. (2016) in their meta-analyses have identified four different kinds of ability grouping, namely, between-class ability tracking, within-class ability tracking, Joplin Plan, or ability grouping for special students. For the purpose of this paper, we focus on between-class ability tracking which refers to grouping students in the same grade into high- and low-ability classrooms based on their prior achievement or ability levels.

Proponents of ability tracking usually highlight the following three reasons. First, ability tracking makes it easier for instructors to adjust their curriculum by teaching a homogeneous group than a heterogeneous one (Lou et al., 2000). Second, ability tracking has economic benefits, as the school can direct, invest and match its resources to a given type of student (Oakes and Guiton, 1995; Betts, 2011). Third, ability tracking allows students to make progress proportional to their ability, and thus maintain interest and motivation. In other words, with ability tracking, students are less likely to be overshadowed by high-ability students and suffer negative self-concept, which is called the "big-fish-little-pond-effect" (Loyalka et al., 2018), or bogged down by slower low-ability counterparts, thus creating an ideal learning environment within the classroom (Hattie, 2002).

On the other hand, opponents of ability tracking also present their arguments. First, it is difficult to track students by their abilities. The use of standardized scores is unreliable to make student assignments (Betts, 2011), and ability tracking magnifies the initial learning gaps between low and high-ability students over time, causing the "Mathew effect" (Kerckhoff and Glennie, 1999). Second, the so-called "adjusted" curriculum for low-ability students may be less stimulating, directed at behavior management rather than learning, teaching them slowly with lesser content coverage and less analytical in instructional discourse (Hong et al., 2012). For these reasons, they are concerned that ability tracking is likely to demoralize low-ability students and make them prone to "delinquency, absenteeism, dropout, and other social problems'' (Slavin, 1990, pp. 473).

2.1. Prior studies on the effect of ability tracking

Empirical work on how ability tracking influences students' academic performance is mixed across different educational levels. With regards to elementary education, results from a randomized controlled trial found that primary school students (grade 1–3) in tracking schools in Kenya had higher scores than those in non-tracking schools, and the effects persisted even one year after the program (Duflo et al., 2011). Similar positive results are reported in their analysis of students in grade 3–5 by Collins and Gan (2013) in the United States. However, Fu and Mehta (2018) show that ability tracking may have a differential effect, with positive gains for high-ability but losses for low-ability students. At the secondary and high school level, results are more mixed. An earlier review of literature by Slavin (1993) in "middle schools" found essentially no effect of ability tracking on low, average or high school students, which is also reported by Betts and Shkolnik (2000) among students from 7th to 12th grade. However, analysis by Figlio and Page (2002) of students from 8th to 10th grade showed that there are no additional gains for high-ability students at the detriment of their low-ability counterparts, but with certain gains for low-ability students. In contrast, recent work by Antonovics et al. (2022) on students from grade 4th to 8th highlights that high-ability students benefit from ability tracking without any gains to low-ability students, leading to disparities in learning. Steenbergen-Hu et al. (2016) in their second-order meta--analyses (including 13 meta-studies) of ability tracking's impact on K-12 students found that between-class ability grouping was associated with no gains in overall student's academic achievement, though within-class ability tracking had a positive association with students' academic achievements. At the college level, Booij et al. (2016) found the opposite result, suggesting low and medium-ability students gained without any effect on high-ability students.

In addition to its association with academic outcomes, the effect of ability tracking on non-academic outcomes has gained the attention of scholars in recent years (Mulkey et al., 2005; Liu et al., 2005; Van Houtte, 2005; Wang, 2015). Research has shown that non-academic outcomes are directly correlated with student learning outcomes. For example, Marsh et al. (2008) and Marsh and Martin (2011) found that academic self-concept is positively correlated with subsequent academic achievement. Thus, it becomes essential to study the effect of ability tracking on non-academic outcomes separately. Some studies have evaluated the association of ability tracking on non-academic outcomes such as academic self-concept and anxiety, with mixed results at the secondary school level. Liu et al. (2005) showed that for secondary school students in the long term, while self-concept declined for both higher and lower-ability students, lower-ability students reported higher self-concept than their higher-ability counterparts. Evidence from middle schools in South Korea suggested that ability sorting decreased the likelihood of students feeling anxious about their grades (Wang, 2015). In contrast, Mulkey et al. (2005) found that in the long term, high-ability students who were tracked in middle schools are likely to suffer a decline in self-concept which subsequently negatively influences their achievements.

2.2. Ability tracking in China

Ability tracking was practiced in China in its early years, though with changes in later decades. At first, the so-called "Key School Policy" was a "fast lane to cultivate talented students who had limited resources" (Zhang et al., 2014, pp.81). However, concerns around inequity, combined with a demand for quality education (Wang, 2009) and unease around the rising concentration of high socio-economic profile students in key schools (Yang, 2005; Lai et al., 2015), facilitated many new changes. In response, measures were taken to cope with these challenges, including equalization of public expenditures and teacher salaries within municipalities, teachers from key schools being encouraged (or required) to teach at low-performing schools for a certain period of time, and most importantly, getting key schools to admit low-performing students (Zhang et al., 2014).

Whilst China formally prohibited ability tracking in junior-high schools in 2006 (Ministry of Education, 2006), the practice was still observed. Junior-high schools in China typically follow an S-shape allotment.[1] However, some junior high schools continued to track students. Schools especially those with large class sizes and high performance, assigned students with the high academic scores or with accolades to a few specially selected classrooms before assigning the remaining to other classes based on S-shape division rule (Lai, 2007; Carman and Zhang, 2012; Zhang et al., 2014). Why was ability tracking still practiced in some schools despite the prohibition? According to Li et al. (2018), the incentive system for junior-high-school principals and teachers, especially in rural schools, promotes the use of ability tracking. High school admission rates or the ability to gain admission into prestigious high schools and universities were treated as important indicators of teachers' and schools' reputations (Tsang, 2000; Wang et al., 2011; Carman and Zhang, 2012; Feng and Li, 2016). Securing such admissions, especially in poor rural counties, is an exceedingly difficult task (Liu et al., 2009; Wang et al., 2011; Loyalka et al., 2017). Given the slim chances of success, teachers and principals disproportionately invest their time and efforts in favor of the best students through ability tracking (Li et al., 2018).

Empirical work has assessed the effect of ability tracking on both academic and non-academic outcomes in China with mixed results. Zhang et al. (2014), in their study of ability tracking in high schools, find that high-ability classes ("key classes") do not benefit students in first-tier[2] schools (as compared to non-key classes in the same school). However, in second-tier schools, high-ability classes benefit due to ability tracking and the result is consistent across Math, English, and Chinese scores. With regard to non-academic outcomes, Li et. al (2018) show that fast-tracked students have higher confidence in all public institutions (schooling media, financial institutions, and government) than slow-tracked students. Cheung and Rudowicz (2003) reveal that ability-tracked classes had no influence on students' self-esteem, test anxiety, or academic self-concept. However, they do find that the students in higher banding schools had higher self-esteem and test anxiety.

2.3. Potential gaps in the literature and purpose of the study

A close examination of the literature reveals at least two potential gaps. First, most of the existing literature from China focused on peer effects (Ding and Lehrer, 2007; Carman and Zhang, 2012; Feng and Li, 2016; Lai, 2007), not explicitly on the role of ability tracking. Even when some studies have attempted to assess ability tracking, the comparison has been made between high-ability students and low-ability peers, with the latter being used as the comparison group (Li et al., 2018). This means the findings, therefore, reflect the difference between high-ability and low-ability groups, rather than ability tracking. Thus, the question of the effect of ability tracking remains unexplored.

Second, most of the literature on ability tracking from China comes from urban centers or municipalities (Zhang et al., 2014), while rural areas continue to be overlooked. With rising concerns about the inequity between urban-rural education in the country (Loyalka et al., 2017), and continued dropout amongst secondary schools in rural areas (Yi et al., 2012; Shi et al., 2015), greater attention should be paid to studying educational outcomes in rural China.

This paper attempts to fill in these potential gaps by assessing the effect of ability tracking on academic and non-academic outcomes in rural schools in China. Specifically, we have three objectives. First, we want to examine the effect of ability tracking on student's math score, math self-concept and math anxiety. Second, we examine the effect of ability tracking on high-ability and low-ability classroom student's math score, math self-concept and math anxiety. Third, we examine the heterogeneous effects of ability tracking by students' academic rank within the class, gender, boarding status and economic status. In doing so, the paper seeks to contribute to a growing body of literature on ability tracking in developing countries (McEwan, 2003; Duflo et al., 2011; Wang, 2015).

3. Data and sample construction

This paper draws on longitudinal data collected by the authors themselves in the Survey on the Quality of Middle Schools in Rural China. The data are longitudinal in that we followed the same set of students across three rounds: November (2015), January (2016) and June (2016). Given our focus on the effect of ability tracking on students, we use the data from the first and third rounds which were conducted in November 2015 (the baseline hereafter) and June 2016 (the follow-up or endline, hereafter), respectively. This allowed us to assess the influence of ability tracking on students over eight months.

The survey was conducted in 23 counties across 3 prefectures in two provinces of northwest China, Shaanxi and Gansu. These counties were nationally designated as "poverty" counties at the time of data collection. We took the following steps to collect the data. First, we got the list of all junior high schools in these counties and randomly chose 200 junior high schools as our sample schools. Second, in each sample school, we randomly chose the 7th-grade classroom as our sample class. Third, all students in the sampled 7th-grade classroom were surveyed. A more detailed description of the sampling strategy and data collection is available in Lu et al. (2017). Among the 200 sample schools, 81 schools (40.5%) had only one 7th-grade class, whereas the remaining schools had two or more 7th-grade classes. In total, 12704 seventh-grade students were included in this survey.

To collect the data, each round of the survey involved more than 100 enumerators. To ensure the quality of the survey, all enumerators attended a 2-day intensive training before they went to visit the sample schools to conduct the survey in a face-to-face, standardized manner following the same survey protocol. Each round of the survey collected

[1] S-shaped allotment is made to ensure students are equally divided by ability across all classrooms. Carman and Zhang (2012) describe the process as follows, "starting with the top three students, the 1st student is assigned to class one, the 2nd to class two, and the 3rd to class three. With the next three students, the order of class assignment is reversed: the 4th student is assigned to class three, the 5th to class two, and the 6th to class one. With the next three students, the order of class assignment starts with the second class and proceeds sequentially: the 7th student is assigned to class two, the 8th to class three, and the 9th to class one. With the next three students, the order is reversed: the 10th student is assigned to class one, the 11th to class three, and the 12th to class two." (pp. 225) The process continues until all students are allotted.

[2] For the purpose of their study, Zhang et al. (2014) describe the tiers of schools in their study as follows, "high-performing schools are labeled 1st tier schools, average-performing schools are labeled 2nd tier schools, and low-performing schools are labeled 3rd tier schools. The classification of these three categories is based on three sources of evidence: high school entrance exam (HSEE) admission line, historical reputation, and judgment of experts such as officers from the local educational authority and school administrators. Within the 1st and 2nd tier schools, key class settings are common." (pp. 83)

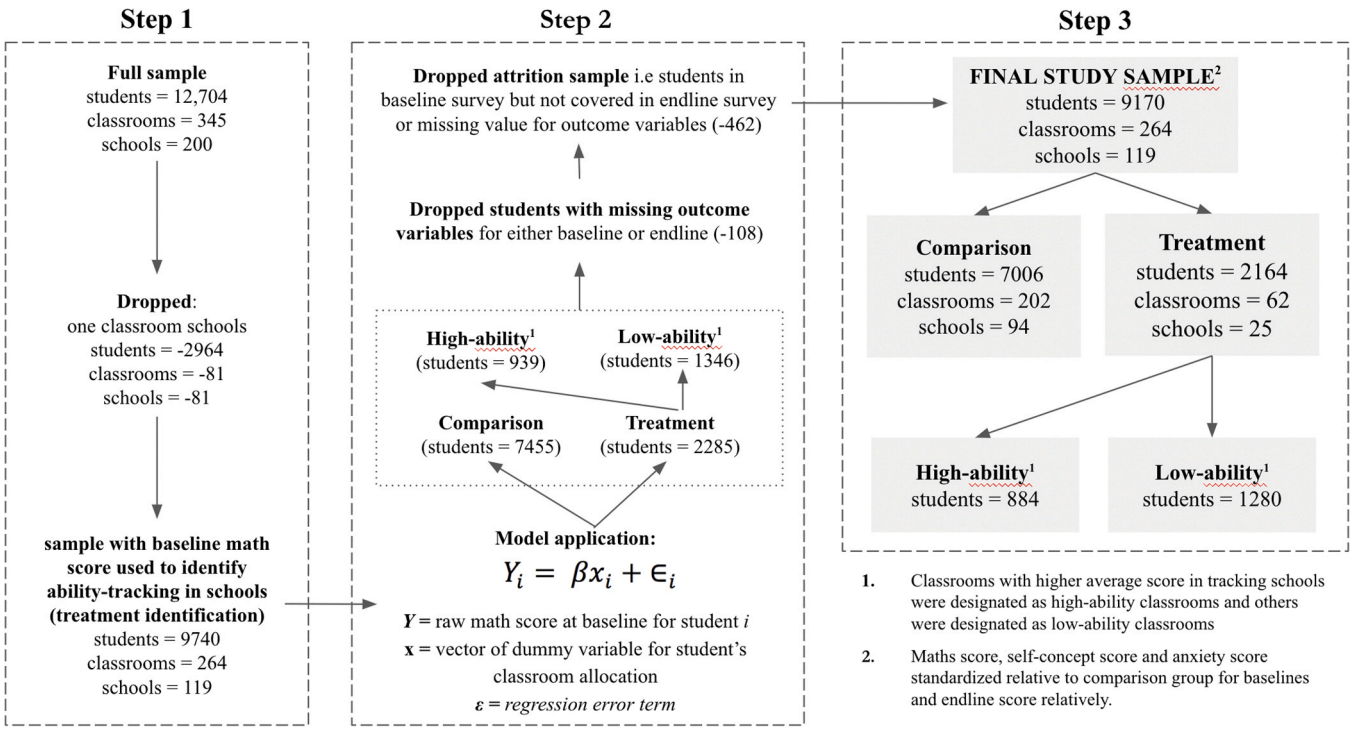

Fig. 1. Creating treatment and comparison groups.

rich information. For the purpose of this study, we draw on information from three blocks. In the first block, we asked each student to take a standardized math test on paper. In the second block, students were asked to fill out a survey instrument on paper that was designed to measure their math self-concept and math anxiety. In the third block, students were asked to provide information on basic demographic and family background characteristics, such as age, gender, boarding status, whether his/her father or mother completed at least junior high school, and family assets. Additionally, teacher and principals from the sampled school were surveyed to collect teacher and school attributes.

We focus on three outcome variables, namely math score, math self-concept and math anxiety. As described by Lu et al. (2017), the math score was computed from a 30-minute standardized mathematics test based on the Chinese National Curriculum Framework. The tests were strictly proctored and graded by the survey team. Consistent with the PISA 2012 survey, a set of five questions each were used to measure math self-concept and math anxiety score. Each items required a 4-point Likert-type response (OECD, 2012). Details of the questions used are available in Appendix Table 1. Higher values on the math self-concept index suggest that a student reported higher math self-concept, i.e., he/she has higher levels of confidence in his/her math ability. On the other hand, a positive math anxiety score indicates a higher level of math anxiety, i.e., a high math anxiety score means that a student suffers from higher levels of stress when doing math problems. For analysis, all three indicators were independently standardized into z-scores relative to the comparison group in the same survey round. This was done by subtracting the mean score and dividing it by the standard deviation (SD) of the comparison group at the relevant point of time.

4. Research design and empirical strategy

4.1. Research design

As discussed above, while ability tracking was prohibited in China's secondary education system in 2006, it is still practiced (Xinhua Daily, 2021). While schools are likely to be tracking students into higher ability and lower ability classes, they may hesitate to admit doing so publicly, which makes it unviable for researchers to establish whether ability tracking is being carried out within schools or not (Li et al., 2018). However, data can be exploited to make such an establishment. In fact, previous studies in China have attempted to identify ability tracking, including high and low track placement, to study the effects on student's outcomes by data exploitation (Li et al., 2018).

We rely on the available data to detect if ability tracking is practiced in the sample schools. Both international evidence and studies from China suggest that ability tracking is carried out based on pre-existing test scores of students (Slavin, 1993; Hattie, 2002; Zhang et al., 2014). In China, when students enter junior high school (grade 7), their new grade allocation is made based on prior test scores (Xinhua, 2010; Xinhua Daily, 2021). If there is a statistically significant disparity between average scores amongst classes in a school, it is highly likely that the school practiced ability tracking. Based on this assumption, we use raw baseline test score to test if ability tracking is practiced in a school. In the case of our study, the baseline survey was conducted after students were assigned to different classrooms in seventh grade, thus students' math test scores in the baseline capture their ability based on which they were placed in their respective classes. If the students were tracked during the seventh-grade class allocation, average class math test scores are likely to capture the effect. Therefore, schools practicing ability tracking would reflect a statistically significant difference in scores between its classes in the baseline survey.

We take a quasi-experimental approach to define the treatment and comparison groups. As explained by Gopalan et al. (2020), quasi-experimental approaches involve "nonexperimental (or non-researcher-induced) variation in the main independent variable of interest". In other words, treatment is assigned "not on a random basis" (pp. 44, emphasis our own). This is in contrast to randomized controlled trials (RCT) where the treatment is assigned (to designate the treatment group) and withheld (to designate the control group) randomly (Gopalan et al., 2020). As our treatment variable (ability tracking) was not assigned randomly, we use the term comparison group (as opposed to 'control' group which is the term primarily used in RCTs studies) to refer to our reference group.

In order to designate schools (and classrooms) into treatment group

Table 1
Distribution of sample across treatment and comparison groups.

No. of classrooms within each school	Total		Treatment		Comparison	
	Schools (1)	Students (2)	Schools (3)	Students (4)	Schools (5)	Students (6)
Panel A: Distribution of students across treatment and comparison groups (sample used to identify the treatment and comparison group)						
2	96	7215	15	1171	81	6044
3	20	2170	8	864	12	1306
4	3	355	2	250	1	105
Total	119	9740	25	2285	94	7455
Panel B: Final sample of students across treatment and comparison groups sued for analysis						
2	96	6773	15	1101	81	5672
3	20	2057	8	823	12	1234
4	3	340	2	240	1	100
Total	119	9170	25	2164	94	7006
Proportion	100%	100%	21%	23.60%	79%	76.40%

Note: From 9740 students we dropped students with missing outcome variables (math score, math self-concept and math anxiety either at baseline or endline (n = 108). Following this, we dropped the attrition students (n = 462). The remaining 9170 students was the final sample for the study.

(those that practice ability tracking) and comparison groups (those that do not practice ability tracking), we took the three steps to detect if schools practiced ability tracking (Fig. 1). To begin with, we dropped 81 schools where only one class was surveyed, this left us with 119 schools. Within the 119 schools with multiple classes, we dropped students with missing baseline data (n = 404), which left us with 9740 students. These 9740 students from 119 schools constitute the sample that we used to identify ability tracking schools (treatment group) and non-ability tracking schools (comparison group) for the rest of the study. As shown in Panel A of Table 1, the numbers of schools in the remaining sample with two, three, or four sample classes are 96 (80.67%), 20 (16.81%), and 3 (2.5%), respectively.

In the second step, we used linear regression to examine if there is any difference in baseline math scores between classes within schools to detect ability tracking. For each school, we ran the following model:

$$Y_i = \beta_o(class)_i + \in_i \quad (1)$$

where Y_i represents the raw math score at baseline survey for any student i. β_o represents the coefficients for a vector of class dummies. Finally, $\in$ is the error term. Additionally, for schools with three or four classes, we ran Model (1) by changing the base class to test for pair-wise differences in math scores across all possible combinations of the classes. This analysis produced mean differences in baseline math scores amongst classes in each school. If the coefficient for any class was statistically significant at 10% level ($p < 0.1$), the school was classified as a treatment school, i.e., the school practiced ability tracking. Otherwise, a school was classified as non-tracking and thus the comparison group. Regression results show that 25 schools (21%) practiced tracking, which is categorized as the treatment group. The rest 94 schools did not practice ability tracking (79%) and constitute the comparison group.

Within the treatment group (i.e., students in ability tracking schools), we further grouped the classes with the highest average score as the high-ability class, and the remaining classes as low-ability classes. Panel A of Appendix Table 2 details the sample distribution of treatment groups (schools with ability tracking) across high-ability and low-ability classes.

In the last step, after identifying treatment and comparison groups (including high-ability and low-ability classes), we dropped all students for whom there were missing values for outcome variables (math score, self-concept, and anxiety score, n = 108 in either of the two survey waves (baseline = 67 & endline = 41). Among the remaining students, 462 students participated in the baseline survey but were missing in the follow-up survey. In other words, the attrition rate is 4.8%. Appendix Table 3 compares the attrited (n = 462) and non-attrited students (n = 9170) across key characteristics.

There were two reasons to drop the students with missing information on outcome variables (for self-concept and math anxiety) and attrition students after using them to classify the treatment and comparison groups. First, having a greater sample of students can more precisely detect differences in test scores between classes and thus, provide more accurate evidence for whether ability tracking was practiced in a school. Second, whether a school practiced tracking only required students' entrance test scores (namely the baseline survey), but not their test scores in the follow-up survey. Further, detecting whether a school practiced tracking is unrelated to the overall effect of ability tracking on student learning outcomes. In other words, the former refers to the detection of ability tracking while the latter is concerned with its effects. After removing observations with missing outcome variables and attrition, we were left with 9170 students. Table 1, Panel B, shows the distribution of the sample across treatment and comparison groups, amongst schools with two, three, and four classrooms. Our data show that 21% of the schools (n = 119), and 24% of students (n = 9170) are in ability tracking schools. Further, within the treatment group, classrooms with higher average math test scores were designated as high-ability tracked classroom, whereas remaining were designated as low-ability tracked classrooms. 41% of the students were in high-ability tracked classes whereas the rest 59% in low-ability tracked classes (n = 2164) (Appendix Table 2, Panel B).

We use the 9170 students for all further analyses in the rest of the paper.[3] At this stage, all three dependent variables (math scores, self-concept, and anxiety scores) were independently standardized to z-scores relative to the comparison group in the same survey round. Fig. 2, Panel A plots the distribution of standardized math scores across the sample for treatment and comparison groups, while Panel B plots the difference between the highest-scoring and lowest-scoring classes within each school but across treatment and comparison groups.

Table 2 reports student, household, teacher, and school characteristics in the baseline survey across the entire study sample. Average standardized scores for math test score, math self-concept and math anxiety are 0.057, 0.010, and −0.017, respectively. Almost half of the sample students are female. About 60% of the students are boarding in the school. The average age of a student is about 13 years. Across the sample, only 25% of students' mothers have graduated from junior high school. Mothers (fathers) were not present in the household for the entire school year for 15% (41%) of the sample. The average household asset ownership score is 2.75 out of 7. The proportion of sample students who reported that they discussed homework at home at least once a week or more is 63%. About 36% of the teachers in the sample are

[3] Some observations, the control variables had missing values. These observations were dropped automatically when variable is added to the model. Missing value count for each variable is reported in Table 2. We also ran the analysis by dropping all observations for whom even a single control variable had a missing value, and our results still hold (not reported).

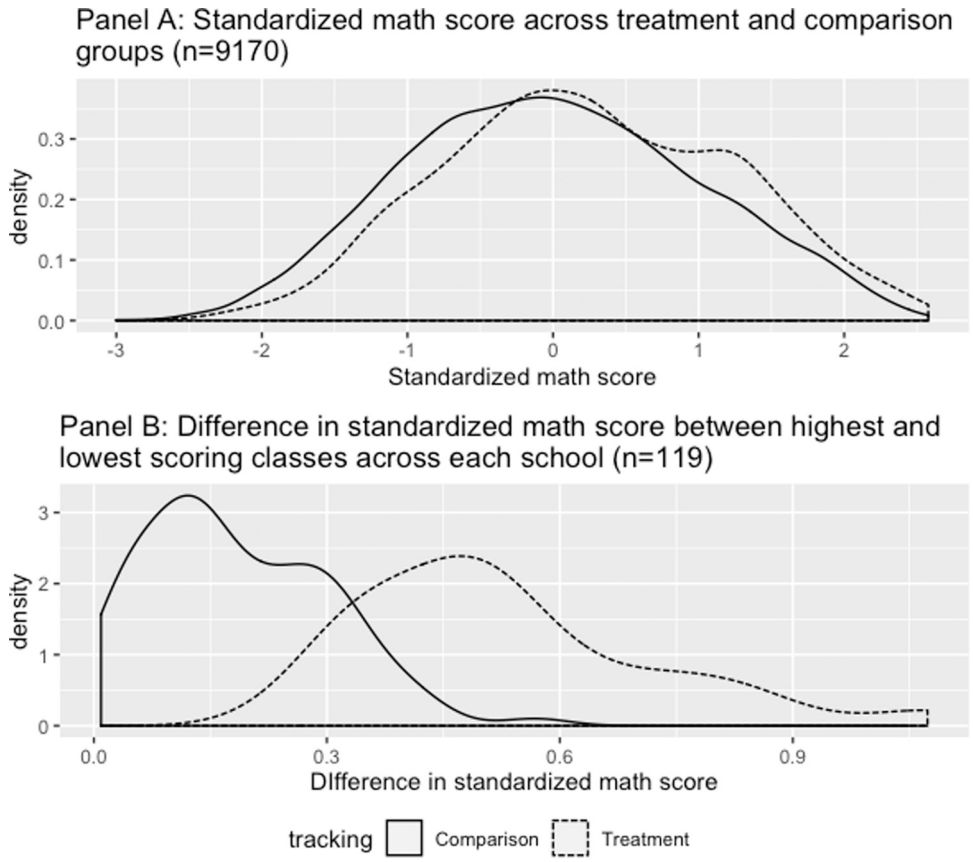

Fig. 2. Standardized math scores across treatment and comparison groups for students without missing values for outcome variables*. * *Note:* Panel A plots individual scores of students across treatment and comparison groups i.e., each point is one student. Panel B plots each school i.e., each point captures the difference in test scores between the highest and lowest-scoring classrooms within each school.

female. On average teachers have almost 10 years of teaching experience. About three-quarters of the teachers (72.3%) have a junior high school math teaching certificate. Lastly, sample schools have an average asset ownership score of 4.69 out of 5 and an average student-teacher ratio of 8.75.

When we compare the baseline characteristics by treatment status, results from t-tests show that there is a statistically significant difference at the baseline between treatment and comparison groups in terms of math, self-concept, and anxiety scores as well as student, household, and school characteristics. However, gender composition is equally distributed across treatment and comparison groups. These observed differences in the baseline imply the necessity to control for these characteristics in our multivariate analyses.

4.2. Empirical strategy

As explained in the above section, we took a quasi-experimental approach where we designate the treatment group as schools that practice ability tracking whereas the comparison group as schools that do not follow ability tracking. Having defined the treatment and comparison groups, following Koedel and Rockoff (2015), we take the linear value-added model (VAM) to assess the association between ability tracking and students' academic and non-academic outcomes. VAMs have been used to estimate value-added to student achievement for a variety of educational inputs. For example, VAMs have been widely used to estimate the effects of individual teachers on students (Hanushek and Rivkin, 2010; Chetty et al., 2014). To separate the effect of ability tracking on students, we follow Liu et al. (2010) and Guo et al. (2022) to specify our VAM model as follows:

$$\begin{aligned}\triangle \mathbf{Y_i} = {} & a + \beta_1(\text{tracking})_i + \beta_2(\text{baseline score})_i \\ & + \beta_3(\text{student characteristics})_i + \beta_4 \quad (\text{household characteristics})_i \\ & + \beta_5 \quad (\text{teacher characteristics})_i + \beta_6(\text{school characteristics})_i \\ & + \beta_7 \quad (\text{teacher incentive treatment})_i + \in_i \end{aligned} \tag{2}$$

where $\triangle \mathbf{Y}$ represents the value added in standardized math score (from baseline to endline surveys) for a given education-related outcome (i.e., math score, self-concept score, and anxiety score) for student i (Liu et al., 2010; Lei et al., 2018). Tracking is a dummy variable that takes the value of one if a student is from a school that practices ability tracking and zero if a student is from a school that does not practice ability tracking. Baseline score represents the standardized score for a given education-related outcome in the baseline survey, i.e., math score, math self-concept score, or math anxiety score.

We also control for student, household, teacher, and school characteristics in the baseline survey. Specifically, following the literature, we control for students' gender (1 = female) (Liu et al., 2010), their boarding status (1 = boarding in school) (Lei et al., 2018), and age in years (Booij et al., 2016). We include five variables for household characteristics: whether mother graduated from junior high school (1 =yes) (Lei et al., 2018), whether mother/father is absent from home for the school year (1 =yes) (Li et al., 2018; Lei et al., 2018), household economic status proxied by possession of durable assets (Guo et al., 2022; Lei et al., 2018), and whether students discuss homework at home (1 = more than once a week). We also control for three teacher characteristics: gender (1 = female) (Lei et al., 2018), years of teaching experience (Liu et al., 2010), and qualification (teacher has middle school math teaching certificate, 1 =yes) (Chu et al., 2015). For school

Table 2
Baseline characteristics by treatment status.

	Overall (n = 9170) (1)	Treatment (n = 2164) (2)	Comparison (n = 7006) (3)	p-value H0: (2)= (3) (4)
OUTCOME VARIABLES				
Standardized math score	0.057 (1.00)	0.242 (0.990)	0.00 (1.00)	< 0.01***
Standardized self-concept score	0.010 (1.00)	0.043 (1.01)	0.00 (1.00)	< 0.1*
Standardized anxiety score	-0.017 (1.01)	-0.072 (1.03)	0.00 (1.00)	< 0.01***
STUDENT CHARACTERISTICS				
Female (1 =yes)	0.499 (0.500)	0.496 (0.500)	0.500 (0.500)	0.735
Boarding at school (1 =yes)[1]	0.597 (0.491)	0.654 (0.476)	0.579 (0.494)	< 0.01***
Age (in years)	13.0 (0.940)	13.0 (0.976)	13.0 (0.929)	< 0.1*
HOUSEHOLD CHARACTERISTICS				
Mother graduated junior high school (1 =yes)[2]	0.251 (0.434)	0.276 (0.447)	0.243 (0.429)	< 0.01***
Mother absent for both semesters (1 =yes)[3]	0.149 (0.356)	0.132 (0.338)	0.155 (0.362)	< 0.01***
Father absent for both semesters (1 =yes)[4]	0.413 (0.492)	0.446 (0.497)	0.403 (0.491)	< 0.01***
Household assets (score 0–7)[5]	2.75 (1.70)	2.75 (1.66)	2.75 (1.71)	0.999
Students discuss homework with parents - more than once a week (1 =yes)[6]	0.633 (0.482)	0.633 (0.482)	0.633 (0.482)	0.997
MATH TEACHER CHARACTERISTICS				
Female (1 =yes)[7]	0.358 (0.479)	0.341 (0.474)	0.363 (0.481)	< 0.1*
Years of teaching experiences[8]	9.61 (7.07)	9.19 (6.65)	9.74 (7.19)	< 0.01***
Has middle school math teaching certificate (1 =yes)[9]	0.723 (0.447)	0.821 (0.384)	0.693 (0.461)	< 0.01***
SCHOOL CHARACTERISTICS				
School assets (score 0–5)	4.69 (0.620)	4.66 (0.563)	4.69 (0.636)	< 0.1*
Student/teacher ratio	8.75 (2.87)	7.83 (2.39)	9.04 (2.95)	< 0.01***

Note: N = number of observations; *p < 0.1; **p < 0.05; ***p < 0.01
[1] Missing =2 (Treatment = 1; Comparison = 1); [2] Missing =18 (Treatment = 7; Comparison = 11); [3] Missing =13 (Treatment = 6; Comparison = 7); [4] Missing =2 (Treatment = 1; Comparison = 1); [5] Missing =3 (Treatment = 0; Comparison = 3); [6] Missing =4 (Treatment = 0; Comparison = 4); [7] Missing =37 (Treatment = 0; Comparison = 37); [8] Missing =37 (Treatment = 0; Comparison = 37); [9] Missing =68 (Treatment = 0; Comparison = 68)

characteristics, we include two variables: school assets (scored out of 5) (Liu et al., 2010), and student-teacher ratio (Liu et al., 2010). Lastly, as the data were sourced from a randomized controlled trial on teacher incentive, we add a teacher incentive treatment school dummy variable which indicates if the student was from a school that was part of the treatment for the teacher incentive study. And $\in$ is the regression error term. We cluster the standard errors at the school level.

To further estimate the effect of ability tracking on students in high-ability and low-ability classes in schools that practice ability tracking, we modify the tracking variable in Model (2) to get an empirical specification as follows:

$$\begin{aligned}\triangle \mathbf{Y_i} = {} & a + \beta_1(\text{ability} - \text{tracking})_i + \beta_2(\text{baseline score})_i \\ & + \beta_3(\text{student characteristics})_i + \beta_4 \quad (\text{household characteristics})_i \\ & + \beta_5 \quad (\text{teacher characteristics})_i + \beta_6(\text{school characteristics})_i \\ & + \beta_7 \quad (\text{teacher incentive treatment})_i + \in_i \end{aligned} \tag{3}$$

where ability-tracking is a vector of two dummy variables with all students in schools that do not practice ability tracking serve as the reference. One dummy variable takes the value of one if a student is in a high-ability class in tracking schools and zero otherwise. The other dummy variable takes the value of one if a student is in a low ability classroom in schools that practice ability tracking and zero otherwise.[4] Other variables in Model (3) remain the same as in Model (2).

To estimate the heterogeneous effect of ability tracking on student ability, gender, boarding status, and economic status, we added their interaction terms with *tracking* and *ability-tracking* variables in Models (2) and (3), respectively. Four interaction variables include (a) student-ability, a vector of dummy variables that expresses a student's academic ranking within his/her class (Top 1/3rd, Middle 1/3rd, and Bottom 1/3rd in class) at the baseline survey, (b) student gender, (c) boarding status, and (d) economic status. We test for heterogeneity effects on boarding students as prior work has shown that boarding students are especially vulnerable groups within the school (Luo et al., 2009; Wang et al., 2016). Economic status is a dummy variable that indicates if the student is in the top-half economic level of the sample.

5. Results

Our design has two groups, namely schools with ability tracking (treatment group), and schools without ability tracking (comparison group). Within schools that practice ability tracking (treatment group), we have two sub-groups: (1) high-ability classroom students; and (2) low-ability classroom students. We focus on the effect of ability tracking on students' math score, math self-concept, and math anxiety.

5.1. Effect of ability tracking on students in schools that practice ability tracking

We compare students in schools that practice ability tracking (treatment group) to students in schools that do not practice ability tracking (comparison group). With regards to the effect of ability tracking, regression results from Model 2 show that ability tracking has no statistically significant effect on student's math score, math self-concept, or math anxiety (Table 3). However, when we look at the estimated coefficients on baseline scores, our results show that those students with higher baseline score tend to experience less value added from baseline to endline surveys in math score, math self-concept and math anxiety. Specifically, after controlling baseline characteristics and the treatment variable, results show that as baseline score of math score increases, the value added in math score from the baseline to the endline is significantly less by 0.408 SD ($p < 0.01$). Similar results are also found when we look at the value added in math self-concept (0.429 SD, $p < 0.01$) and math anxiety (0.513 SD, $p < 0.01$).

[4] In the context of our sample classroom allocation do not change through the school year i.e., students remain in the same classroom across the entire school year. Our data confirm that the classroom allocation did not change for 99.2% of the students from the baseline to the endline. Only 74 students reported having different classroom at the endline compared to the baseline (though their schools were still the same). We kept them in our analysis as we do not know when the classroom allocation changed across the entire year. Further, we also ran our analysis by dropping these 74 students, and our results still hold (not reported).

Table 3
Effect of ability tracking on students in school that practiced ability tracking on value-added math score, self-concept and anxiety.

	Math score (1)	Math self-concept (2)	Math anxiety (3)
Tracking	-0.075	0.039	-0.058
	(0.066)	(0.049)	(0.038)
Baseline math score (in SD)	-0.408***	-0.429***	-0.513***
	(0.013)	(0.010)	(0.012)
STUDENT CHARACTERISTICS			
Female (1 =yes)	-0.013	-0.143***	0.135***
	(0.018)	(0.020)	(0.018)
Boarding in school (1 =yes)	0.063**	0.034	-0.053**
	(0.030)	(0.025)	(0.025)
Age (in years)	-0.133***	-0.062***	0.067***
	(0.013)	(0.013)	(0.012)
HOUSEHOLD CHARACTERISTICS			
Mother graduated junior high school (1 =yes)	0.009	-0.019	-0.020
	(0.023)	(0.022)	(0.028)
Mother absent for both semesters (1 =yes)	-0.025	0.011	0.005
	(0.026)	(0.024)	(0.025)
Father absent for both semesters (1 =yes)	-0.003	-0.010	0.018
	(0.019)	(0.020)	(0.022)
Household assets (score 0–7)	0.013*	0.005	-0.016**
	(0.007)	(0.007)	(0.007)
Students discuss homework with parents - more than once a week (1 =yes)	0.008	0.040*	-0.025
	(0.021)	(0.021)	(0.020)
TEACHER CHARACTERISTICS			
Female (1 =yes)	0.010	0.042	-0.063*
	(0.053)	(0.044)	(0.035)
Years of teaching experience	-0.005	-0.004	0.005*
	(0.005)	(0.003)	(0.003)
Teacher has junior high school math teaching certificate (1 =yes)	-0.066	-0.015	0.009
	(0.049)	(0.038)	(0.035)
SCHOOL CHARACTERISTICS			
School assets (score 0–5)	0.064*	0.050**	-0.023
	(0.036)	(0.025)	(0.033)
Student/teacher ratio	-0.008	0.013**	-0.002
	(0.010)	(0.006)	(0.005)
Teacher incentive treatment (1 =yes)	0.009	-0.015	0.023
	(0.052)	(0.039)	(0.034)
Constant	1.517***	0.515**	-0.769***
	(0.288)	(0.231)	(0.260)
Observations	9061	9061	9061
R2	0.215	0.215	0.264

Note:
1SEs clustered at school level
2 $p < 0.1$; **$p < 0.05$; ***$p < 0.01$

5.2. Effect of ability tracking on students in high-ability and low-ability classrooms

We conducted two comparisons to assess the effect of ability tracking on students in high-ability and low-ability classrooms (within schools that practice ability tracking), relative to comparison schools. One is to compare students in the high-ability classrooms in schools that practice ability tracking relative to students in schools that do not practice ability tracking (i.e., the comparison group). The other is to compare students in the low-ability classrooms in schools that practice ability tracking relative to students in schools that do not practice ability tracking (i.e., the comparison group).

Regression results from Model (3) show that there is no statistically significant effect of ability tracking on math scores of high- or low-ability class students relative to students in schools that do not practice ability tracking (i.e., the comparison group). Similarly, ability tracking does not have any statistically significant effect on the math self-concept of high- or low-ability classroom students (Column 6, Panels A and B, Table 4) relative to students in the comparison group.

However, unlike math score and math self-concept, regression results show that ability tracking has a statistically significant effect on reducing the math anxiety of high-ability classroom students who started with higher math anxiety. Controlling for student, household, teacher, and school characteristics, compared with the comparison group (students in schools that do not practice ability tracking), high-ability classroom students in ability-tracking schools experience a statistically significant less value added in math anxiety from the baseline to the endline by 0.103 SD ($p < 0.05$). However, no such effect is observed on the math anxiety of low-ability class students (Column 6, Panel C, Table 4).

5.3. Heterogenous effect of ability tracking on students

Is there any heterogenous effect among ability tracking on students? To answer this question, we further investigated the heterogeneity in ability tracking by making three comparisons with a focus on four dimensions: students' academic rank, gender, boarding status, and economic status. First, we compare treatment against comparison. Then we compare high-ability classroom students (in treatment schools i.e., those that practice ability tracking) against comparison (all students in schools that do not practice ability tracking). And finally, we compare low-ability classroom students (in treatment school) against comparison.

Results from heterogeneity analyses exhibit no heterogeneity in the effect of ability tracking on students, regardless of the comparisons or dimensions (Tables 5 and 6). The only exception is that boarding students in low-ability classes experience statistically significant less value added from baseline to endline surveys in math test score (0.168 SD, $p < 0.05$) as compared to non-boarding students in comparison schools due to ability tracking (Column 3, Panel A, Table 6).

6. Discussions and implications

Using longitudinal data from 9170 students from 119 junior schools in rural China, the paper evaluated the association between ability tracking and students' academic and non-academic outcomes. Results from our analyses revealed that ability tracking did not affect math score, math self-concept, or math anxiety of students. Sub-group analyses revealed similar results for high-ability and low-ability classroom students (within schools that practice ability tracking) for math score and math self-concept. Our data did show that ability tracking, in about eight months of our study period, had a statistically significant effect in reducing the math anxiety of high-ability classroom students as indicated by 0.103 SD less value added from baseline to endline surveys ($p < 0.05$) relative to students in schools that do not practice ability tracking (i.e., comparison group). We also found that ability tracking has a heterogeneous effect on the math scores of low-ability boarding students.

Our findings are consistent with previous studies on both academic (Betts and Shkolnik, 2000) and non-academic outcomes (Ireson and Hallam, 2009).[5] In their meta-analysis Steenbergen-Hu et al. (2016) found that between class ability tracking had "negligible" effect on student's overall academic achievement, alongside previous works of Slavin (1993) who had also found no achievement effect from grouping in secondary schools. Studies in Asia have also found mixed effects of ability tracking on test scores (Zhang et al., 2014; Cheung and Rudowicz, 2003) and self-concept (Cheung and Rudowicz, 2003).

Our results showed that ability tracking has heterogeneous effects on the math score of low-ability boarding students, who experience less value added from baseline to endline surveys in their score (0.168 SD,

[5] Ireson and Hallam (2009) found that the policy did not influence student's subject-specific self-concepts in math, science, and English (though it did find a negative effect on students' general academic self-concept).

Table 4
Association between ability tracking and high-ability and low-ability classroom student's math self-concept and anxiety score.

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	**Value added: Math score**					
Tracking (High-ability)	-0.300***	-0.120	-0.104	-0.104	-0.097	-0.100
	(0.074)	(0.076)	(0.074)	(0.074)	(0.073)	(0.075)
Tracking (Low-ability)	-0.082	-0.053	-0.055	-0.052	-0.051	-0.057
	(0.059)	(0.070)	(0.069)	(0.066)	(0.066)	(0.065)
Baseline math score (in SD)		-0.361***	-0.394***	-0.399***	-0.400***	-0.407***
		(0.015)	(0.014)	(0.013)	(0.013)	(0.013)
Constant	-0.007	0.003	1.694***	1.602***	1.714***	1.514***
	(0.033)	(0.038)	(0.169)	(0.165)	(0.183)	(0.287)
Observations	9170	9170	9168	9128	9061	9061
R2	0.011	0.186	0.208	0.210	0.212	0.215
Panel B:	**Value added: Math self-concept score**					
Tracking (High-ability)	-0.022	0.005	0.004	0.005	0.011	0.025
	(0.057)	(0.054)	(0.053)	(0.053)	(0.051)	(0.050)
Tracking (Low-ability)	0.029	0.042	0.040	0.036	0.031	0.048
	(0.054)	(0.060)	(0.061)	(0.061)	(0.059)	(0.055)
Baseline self-concept score (in SD)		-0.404***	-0.422***	-0.424***	-0.428***	-0.429***
		(0.011)	(0.011)	(0.011)	(0.011)	(0.010)
Constant	0.017	0.021	0.807***	0.762***	0.858***	0.514**
	(0.033)	(0.034)	(0.182)	(0.185)	(0.190)	(0.231)
Observations	9170	9170	9168	9128	9061	9061
R2	0.001	0.198	0.207	0.208	0.213	0.215
Panel C:	**Value added: Math anxiety score**					
Tracking (High-ability)	-0.040	-0.095*	-0.094*	-0.094*	-0.100**	-0.103**
	(0.050)	(0.055)	(0.052)	(0.052)	(0.050)	(0.051)
Tracking (Low-ability)	-0.014	-0.038	-0.035	-0.031	-0.023	-0.027
	(0.044)	(0.049)	(0.049)	(0.049)	(0.047)	(0.046)
Baseline anxiety score (in SD)		-0.494***	-0.507***	-0.508***	-0.512***	-0.513***
		(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Constant	-0.020	-0.024	-0.949***	-0.817***	-0.893***	-0.772***
	(0.026)	(0.027)	(0.162)	(0.165)	(0.180)	(0.260)
Observations	9170	9170	9168	9128	9061	9061
R2	0.0005	0.249	0.258	0.259	0.264	0.264
Baseline score		YES	YES	YES	YES	YES
Student characteristics			YES	YES	YES	YES
Household characteristics				YES	YES	YES
Teacher characteristics					YES	YES
School characteristics						YES
Teacher incentive	YES	YES	YES	YES	YES	YES

Notes:
1Student characteristics: gender (1=female); boarding status (1=yes); and age (in years).
2Household characteristics: mother graduated junior-high school (1=yes); mother absent for both semesters (1=yes); father absent for both semesters (1=yes); household assets (out of 7); and students discuss homework with parents - more than once a week (1=yes).
3Teacher characteristics: gender (1=female); years of teaching experience; and teacher has junior high school math teaching certificate (1=yes).
4School characteristics: school asset (out of 5); and student-teacher ratio.
5Teacher incentive (1=yes)
6SEs clustered at school level
7*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

$p < 0.05$) relative to their non-boarding counterparts in comparison schools. Although the underlying reasons may be unknown, the interplay of two factors could potentially explain the results. First, studies have shown that boarding students perform much worse than their non-boarding counterparts across both mental health and academic outcome indicators (Luo et al., 2009; Wang et al., 2016). This combined with the potential adverse effect of having low-ability classmates could explain the negative association of ability tracking on the value added in math score of low-ability boarding students.

However, our results are inconsistent with some earlier studies on ability tracking. With regards to test scores, we find ability tracking had no effect on student test scores. This finding is inconsistent with those of Figlio and Page (2002), Duflo et al. (2011), Fuligni et al. (1995), Collins and Gan (2013), Antonovics et al. (2022), and Fu and Mehta (2018), where they all find the positive associations became visible after one year, or longer than the period of our study. This may imply that ability tracking takes longer to add value amongst junior high school students. Further, with regard to self-concept, the results are contradictory to those of Liu et al. (2005), and Mulkey et al. (2005). Liu et al. (2005) found that while low-ability tracked students initially reported lower academic self-concept, three years later they had higher self-concept than high-ability classroom students. The positive self-concept, as also discussed in the paper, may be attributed to Singapore's streaming policy that provides provision for additional years to low-ability stream students to finish secondary school. Prior research has shown that varying levels of flexibility in tracking could lead to differential effects (Gamoran, 1992). Mulkey et al. (2005) found effects two and four years post-tracking, which was longer than our study period of less than one year. This implies that our study period may have been too short to capture associations between ability tracking and self-concept.

The key finding of the study is that ability tracking had a statistically significant effect in reducing the math anxiety of high-ability classroom students (0.103 SD less value added, $p < 0.05$) relative to the students in schools that did not practice ability tracking. Previous studies have assessed the effect of ability tracking on grade anxiety and test anxiety (Cheung and Rudowicz, 2003; Wang, 2015), though they have not studied math anxiety.

In China, high-ability students may experience less anxiety because of three reasons. First, students in high-ability classrooms are often allotted better resources and teachers. Studies have shown that teachers

Table 5
Heterogeneous effect of ability tracking across students in schools that practice ability tracking.

	(1)	(2)	(3)	(4)
PANEL A:	***Value added: Math score***			
Tracking	-0.080	-0.067	-0.006	-0.095
	(0.061)	(0.071)	(0.074)	(0.068)
Baseline math score (in SD)	-0.312***	-0.408***	-0.408***	-0.408***
	(0.026)	(0.013)	(0.013)	(0.013)
Class rank (Middle 1/3)	-0.169***			
	(0.037)			
Class rank (Top 1/3)	-0.278***			
	(0.060)			
Top half economic status				-0.024
				(0.031)
Female	-0.015	-0.009	-0.012	-0.013
	(0.018)	(0.021)	(0.018)	(0.018)
Boarding student	0.050*	0.064**	0.087***	0.064**
	(0.029)	(0.030)	(0.034)	(0.030)
Tracking*class rank (Middle 1/3)	-0.048			
	(0.045)			
Tracking*class rank (Top1/3)	-0.005			
	(0.059)			
Tracking*Female		-0.015		
		(0.041)		
Tracking*Boarding student			-0.108	
			(0.069)	
Tracking*Top half economic status				0.040
				(0.065)
Constant	1.768***	1.516***	1.493***	1.525***
	(0.280)	(0.288)	(0.289)	(0.289)
Observations	9061	9061	9061	9061
R^2	0.222	0.215	0.215	0.215
PANEL B:	***Value added: Math self-concept score***			
Tracking	0.050	0.008	0.065	0.012
	(0.046)	(0.060)	(0.058)	(0.051)
Baseline self-concept score (in SD)	-0.462***	-0.429***	-0.429***	-0.429***
	(0.010)	(0.010)	(0.010)	(0.010)
Class rank (Middle 1/3)	0.126***			
	(0.022)			
Class rank (Top 1/3)	0.343***			
	(0.028)			
Top half economic status				-0.029
				(0.030)
Female	-0.128***	-0.158***	-0.142***	-0.143***
	(0.020)	(0.023)	(0.020)	(0.020)
Boarding student	0.037	0.033	0.043	0.035
	(0.025)	(0.025)	(0.028)	(0.025)
Tracking*class rank (Middle 1/3)	-0.019			
	(0.050)			
Tracking*class rank (Top 1/3)	-0.014			
	(0.060)			
Tracking*Female		0.062		
		(0.049)		
Tracking*Boarding student			-0.042	
			(0.060)	
Tracking*Top half economic status				0.052
				(0.054)
Constant	0.003	0.522**	0.506**	0.526**
	(0.238)	(0.232)	(0.235)	(0.230)
Observations	9061	9061	9061	9061
R^2	0.236	0.215	0.215	0.215
PANEL C:	***Value added: Math anxiety score***			
Tracking	-0.061	-0.062	-0.008	-0.091**
	(0.039)	(0.042)	(0.049)	(0.036)
Baseline anxiety score (in SD)	-0.532***	-0.513***	-0.512***	-0.513***
	(0.011)	(0.012)	(0.012)	(0.012)
Class rank (Middle 1/3)	-0.116***			
	(0.024)			
Class rank (Top 1/3)	-0.280***			
	(0.027)			
Top half economic status				-0.027
				(0.029)
Female	0.116***	0.132***	0.135***	0.134***
	(0.018)	(0.021)	(0.018)	(0.018)
Boarding student	-0.056**	-0.053**	-0.036	-0.052**
	(0.025)	(0.025)	(0.029)	(0.025)
Tracking*class rank (Middle 1/3)	-0.008			
	(0.056)			
Tracking*class rank (Top 1/3)	0.016			
	(0.065)			
Tracking*Female		0.009		
		(0.044)		
Tracking*Boarding student			-0.078	
			(0.053)	
Tracking*Top half economic status				0.064
				(0.054)
Constant	-0.336	-0.768***	-0.786***	-0.756***
	(0.260)	(0.260)	(0.265)	(0.261)
Observations	9061	9061	9061	9061
R^2	0.275	0.264	0.264	0.264
Baseline score	YES	YES	YES	YES
Student characteristics	YES	YES	YES	YES
Household characteristics	YES	YES	YES	YES
Teacher characteristics	YES	YES	YES	YES
School characteristics	YES	YES	YES	YES
Teacher incentive	YES	YES	YES	YES

Note:
1Student characteristics: gender (1=female); boarding status (1=yes); and age (in years).
2Household characteristics: mother graduated junior-high school (1=yes); mother absent for both semesters (1=yes); father absent for both semesters (1=yes); household assets (out of 7); and students discuss homework with parents - more than once a week (1=yes).
3Teacher characteristics: gender (1=female); years of teaching experience; and the teacher has a junior high school math teaching certificate (1=yes).
4School characteristics: school asset (out of 5); and, student-teacher ratio.
5Teacher incentive (1 =yes)
6SEs clustered at school level
7*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

and principals disproportionately invest their time in high-ability classrooms and students (Persson, 1998). Similarly, teachers in China focus on high-ability students (Li et al., 2018). Having more engaged teachers in the classroom is associated with positive students learning experience (Klusmann et al., 2008). Access to such resources means that students in high-ability classes learn better and thus, suffer from lesser anxiety. Second, not only do the high-ability students benefit from better resources, the general social well-being of low-ability class students in China lags behind those of high-ability students in that they are more likely to suffer from mental health problems and drop out (Yi et al., 2012; Mo et al., 2013; Shi et al., 2015; Wang et al., 2015). Moreover, high-ability classroom students in Chinese rural schools have higher interpersonal trust, greater confidence in the educational institutions, and more faith in financial and government systems than their low-ability counterparts (Li et al., 2018). These factors could also positively influence high-ability classroom students' general anxiety levels. Last but not least, being in high-ability classes means that students incur the positive effects of having higher-quality peers. Studies have shown that access to better-quality peers in China increases academic learning (Ding and Lehrer, 2007; Lai, 2007), and thus may reduce anxiety.

We acknowledge five limitations of the study. First, the study identified ability tracking by comparing differences in baseline scores between classrooms within each school (at 10% confidence, or $p < 0.1$) as opposed to direct confirmation by school administrators. Is it possible that the differences in baseline scores between classes may be due to factors other than ability tracking? We argue this is highly unlikely for two reasons: a.) The schools and students were sampled from the same

Table 6
Heterogeneous effect of ability tracking across students in high and low-ability classrooms in schools that practice ability tracking.

	(1)	(2)	(3)	(4)
PANEL A:	***Value added: Math score***			
Tracking (High-ability)	-0.082	-0.110	-0.083	-0.122
	(0.083)	(0.081)	(0.084)	(0.092)
Tracking (Low-ability)	-0.081	-0.037	0.052	-0.077
	(0.058)	(0.070)	(0.076)	(0.059)
Baseline math score (in SD)	-0.304***	-0.407***	-0.406***	-0.407***
	(0.026)	(0.013)	(0.013)	(0.013)
Class rank (Middle 1/3)	-0.176***			
	(0.037)			
Class rank (Top1/3)	-0.292***			
	(0.059)			
Top half economic status				-0.024
				(0.031)
Female	-0.015	-0.009	-0.012	-0.013
	(0.018)	(0.021)	(0.018)	(0.018)
Boarding student	0.049*	0.063**	0.088***	0.064**
	(0.028)	(0.030)	(0.034)	(0.030)
Tracking (High-ability)*class rank (Middle 1/3)	-0.090			
	(0.070)			
Tracking (Low-ability)* class rank (Middle 1/3)	-0.017			
	(0.048)			
Tracking (High-ability)*class rank (Top1/3)	-0.121			
	(0.095)			
Tracking (Low-ability)*class rank (Top1/3)	0.076			
	(0.062)			
Tracking (High-ability)*Female		0.020		
		(0.070)		
Tracking (Low-ability)*Female		-0.040		
		(0.046)		
Tracking (High-ability) *Boarding student			-0.029	
			(0.082)	
Tracking (Low-ability)*Boarding student			-0.168**	
			(0.073)	
Tracking (High-ability)* Top half economic status				0.042
				(0.089)
Tracking (Low-ability)*Top half economic status				0.038
				(0.075)
Constant	1.768***	1.513***	1.485***	1.521***
	(0.279)	(0.287)	(0.288)	(0.288)
Observations	9061	9061	9061	9061
R^2	0.223	0.215	0.216	0.215
Panel B:	***Value added: Math self-concept score***			
Tracking (High-ability)	0.033	-0.007	0.040	-0.009
	(0.051)	(0.068)	(0.073)	(0.067)
Tracking (Low-ability)	0.061	0.019	0.084	0.027
	(0.056)	(0.065)	(0.069)	(0.054)
Baseline self-concept score (in SD)	-0.461***	-0.429***	-0.429***	-0.429***
	(0.010)	(0.010)	(0.010)	(0.010)
Class rank (Middle 1/3)	0.126***			
	(0.022)			
Class rank (Top1/3)	0.343***			
	(0.028)			
Top half economic status				-0.029
				(0.030)
Female	-0.128***	-0.157***	-0.143***	-0.143***
	(0.020)	(0.023)	(0.020)	(0.020)
Boarding student	0.037	0.033	0.043	0.034
	(0.025)	(0.025)	(0.028)	(0.025)
Tracking (High-ability)*class rank (Middle 1/3)	-0.025			
	(0.054)			
Tracking (Low-ability)* class rank (Middle 1/3)	-0.013			
	(0.071)			
Tracking (High-ability)*class rank (Top1/3)	0.008			
	(0.076)			
Tracking (Low-ability)*class rank (Top1/3)	-0.029			
	(0.065)			
Tracking (High-ability)*Female		0.066		
		(0.060)		
Tracking (Low-ability)*Female		0.058		
		(0.059)		
Tracking (High-ability) *Boarding student			-0.024	
			(0.090)	
Tracking (Low-ability)*Boarding student			-0.056	
			(0.063)	
Tracking (High-ability)* Top half economic status				0.066
				(0.069)
Tracking (Low-ability)*Top half economic status				0.041
				(0.072)
Constant	0.004	0.522**	0.504**	0.524**
	(0.238)	(0.232)	(0.235)	(0.230)
Observations	9061	9061	9061	9061
R^2	0.236	0.215	0.215	0.215
Panel C:	***Value added: Math anxiety score***			
Tracking (High-ability)	-0.130**	-0.135**	-0.070	-0.136**
	(0.059)	(0.064)	(0.065)	(0.058)
Tracking (Low-ability)	-0.015	-0.010	0.038	-0.058
	(0.048)	(0.053)	(0.059)	(0.043)
Baseline anxiety score (in SD)	-0.532***	-0.513***	-0.512***	-0.513***
	(0.011)	(0.012)	(0.011)	(0.012)
Class rank (Middle1/3)	-0.116***			
	(0.024)			
Class rank (Top1/3)	-0.280***			
	(0.027)			
Top half economic status				-0.028
				(0.029)
Female	0.116***	0.132***	0.135***	0.134***
	(0.018)	(0.021)	(0.018)	(0.018)
Boarding student	-0.056**	-0.054**	-0.036	-0.052**
	(0.025)	(0.025)	(0.029)	(0.025)
Tracking (High-ability)*class rank (Middle1/3)	0.027			
	(0.083)			
Tracking (Low-ability)* class rank (Middle1/3)	-0.028			
	(0.054)			
Tracking (High-ability)*class rank (Top1/3)	0.058			
	(0.098)			
Tracking (Low-ability)*class rank (Top1/3)	-0.011			
	(0.060)			
Tracking (High-ability)*Female		0.066		
		(0.063)		
Tracking (Low-ability)*Female		-0.033		
		(0.061)		
Tracking (High-ability) *Boarding student			-0.053	
			(0.062)	
Tracking (Low-ability)*Boarding student			-0.099	
			(0.072)	
Tracking (High-ability)* Top half economic status				0.066
				(0.063)
Tracking (Low-ability)*Top half economic status				0.061
				(0.078)
Constant	-0.339	-0.768***	-0.790***	-0.758***
	(0.260)	(0.261)	(0.265)	(0.261)
Observations	9061	9061	9061	9061
R^2	0.276	0.264	0.264	0.264
Baseline score	YES	YES	YES	YES

(*continued on next page*)

Table 6 (*continued*)

	(1)	(2)	(3)	(4)
Student characteristics	YES	YES	YES	YES
Household characteristics	YES	YES	YES	YES
Teacher characteristics	YES	YES	YES	YES
School characteristics	YES	YES	YES	YES
Teacher incentive	YES	YES	YES	YES

Note:
1Student characteristics: gender (1=female); boarding status (1=yes); and age (in years).
2Household characteristics: mother graduated junior-high school (1=yes); mother absent for both semesters (1=yes); father absent for both semesters (1=yes); household assets (out of 7); and students discuss homework with parents - more than once a week (1=yes).
3Teacher characteristics: gender (1=female); years of teaching experience; and the teacher has junior high school math teaching certificate (1=yes).
4School characteristics: school asset (out of 5); and student-teacher ratio.
5Teacher incentive (1=yes)
6SEs clustered at school level
7*p < 0.1; **p < 0.05; ***p < 0.01

region and shared common socio-economic backgrounds, and b.) Given that ability tracking is banned at the junior-high-school level and the class composition in our sample is generally based on an S-shape policy, it is less likely that these differences in scores could be attributed to anything but ability tracking. Second, the treatment and comparison groups were different in the baseline survey in terms of some student, teacher, and school characteristics, which might pose selection bias. The analysis tried to minimize the potential selection bias by controlling for these baseline characteristics in the regression analyses. However, we recognize that it does not eliminate the selection bias. So caution still needs to be taken in interpreting the findings as causal. Third, the sample was unequally distributed across the treatment and comparison groups (Table 1) which may negatively affect the power of the study to detect effects (higher likelihood of Type-2 error). Such a concern remains prevalent in quasi-experimental studies (Gopalan et al., 2020). However, given that our sample distribution is not extreme, alongside our large sample size, means that some of challenge of the unequal sample distribution may be mitigated. Fourth, while the study finds that ability tracking reduced math anxiety, it could not test for underlying mechanisms due to data constraints. Lastly, the study tracked academic and non-academic outcomes for only a year, which may not have been sufficient time to see the effects of ability tracking to manifest. This is even confounded by the fact there may be less variation in ability amongst students in rural schools in China, as compared to students in the US (Ding and Lehrer, 2007), which may have made it difficult to capture the differences in scores over eight months.

Despite its limitations, the paper contributes to the growing literature around the contentious debate on ability tracking. To address the paucity of empirical analysis on ability tracking in a developing context, the study employed a quasi-experimental design and showed that ability tracking was associated with a statistically significant reduction in math anxiety of high-ability classroom students. However, at the same time, the heterogeneous analysis revealed that low-ability boarding students experienced a negative effect on their test scores relative to their non-boarding counterparts. As far as we know, this is the first attempt to study the associations of ability tracking on both academic and non-academic outcomes using a longitudinal data set from rural China. Unlike other studies, it did so by comparing schools with ability tracking to non-ability tracking schools and thus can assess the effect of the practice. In this way, the paper attempts to contribute to the growing body of economics of education literature and is also relevant to the larger body of work on the study of peer effects in education.

Our results strengthen the evidence that high-ability classroom students are likely to draw benefits from ability grouping, and these benefits are in the form of non-academic outcomes (i.e., math anxiety), though not necessarily in the form of test scores. Simultaneously, we found that while there is no direct loss to low-ability students, certain subgroups (i.e., low-ability boarding students) suffer negative consequences. This implies that while ability tracking can be considered a viable option in low-resource settings to further student learning, it must be carefully complemented with other measures to ensure an expansion of its benefits and limits its shortcomings. This could be considered in multiple ways. First, given teacher's incentive to prioritize high-ability students in rural China, there is a need to re-align teacher's incentives to ensure they do not neglect low-ability students while employing ability tracking. Without such protective measures, ability tracking is likely to exacerbate the learning inequality between low and high-ability students, something previous studies have warned against in other contexts (Gamoran et al., 1995; Kerckhoff and Glennie, 1999). Incentive programs such as pay-for-percentile have shown to be effective in increasing teacher performance while ensuring that they pay attention to all students across all ability groups in China (Loyalka et al., 2019). Second, there is a need to pay special attention to vulnerable groups especially low-ability boarding students who are likely to be overlooked while practicing ability tracking. Boarding students already score lower, in both, health and learning outcomes than their non-boarding peers (Luo et al., 2009; Wang et al., 2016). Thus, even a minor oversight over their learning is likely to have an adverse implication, including (as evident from our results) the practice of ability tracking. Afterschool remedial programs have been shown to have a highly positive effect, especially on low-performing students, in other developing contexts (Banerjee et al., 2007). Similarly, policies such as encouraging peer interaction (Li et al., 2014) and introducing computer-assisted learning (Mo et al., 2015) have been shown to have a positive effect in China. Supplementing ability tracking with such initiatives can ensure that its negative implications are mitigated.

The study adds to the debate on ability tracking. However, there is still no consensus on whether it is effective in improving student learning. While the study did find that ability tracking reduced the anxiety of high-ability students, it is unclear whether the results could persist over a longer period. Previous studies have shown that the benefits of non-academic outcomes may disappear or switch from high-ability to low-ability students (Mulkey et al., 2005). Therefore, future study is needed to assess the effect of ability tracking over a longer period. Second, future work would need to investigate why ability tracking reduced the math anxiety of high-ability students. Further, it would need to investigate why ability tracking reduced math anxiety but did not have any influence on academic self-concept, and math scores. While the association between math academic self-concept and test scores is well documented (Marsh et al., 2008; Marsh and Martin, 2011), work on cross-interaction between anxiety, self-concept, and test scores needs to be done, especially with regard to peer effects. Last, ability tracking is legally allowed in high schools in China and is also widely practiced. Studies would need to investigate the effects of the policy at the high-school level on academic and non-academic outcomes. Moreover, attempts must be made to disentangle whether there are long-term associations between students being tracked in junior high schools, and its effect on performance in high schools.

Data disclosure declaration

The authors are not supposed to disclose the data.

Funding

The authors acknowledge the financial support of the National Natural Science Foundation of China (grant numbers 71861147003 and 71925009) and the Yenching Academy, Peking University. The funders had no role in the research design, analysis or production of the article.

CRediT authorship contribution statement

Conceived and designed the study: **Shaoping Li**, **Shriyam Gupta**, **Chengfang Liu**. Data collection: **Fang Chang**, **Chengfang Liu**, **Yaojiang Shi**. Data analysis: **Shriyam Gupta**, **Shaoping Li**. Wrote the paper: **Shriyam Gupta**, **Shaoping Li**, **Chengfang Liu**.

Declaration of Competing Interest

None.

Data availability

The authors are not supposed to share the dataset or the code. If anyone is interested, please feel free to contact the corresponding author.

Appendix

Table A1
Survey questions to measure math self-concept and anxiety as described in PISA (2012)[1].

Math Self-concept	1. I am just not good at math
	2. I get good grades in math
	3. I learn math quickly
	4. I have always believed that math is one of my best subjects
	5. In my math class, I understand even the most difficult work
Math anxiety	1. I often worry that it will be difficult for me in math classes
	2. I get very tense when I have to do math homework
	3. I get very nervous doing math problems
	4. I feel helpless when doing a math problem
	5. I worry that I will get poor grades in math

[1]Students answer on a 4-point Likert of strongly agree, agree, disagree, and strongly disagree.

Table A2
Distribution of treatment group across high-ability and low-ability students.

	Treatment (1)	High-ability (2)	Low-ability (3)
Panel A: Distribution of students across high-ability and low-ability groups (with students' that are missing outcome variables)			
Number of students	2285	939	1346
Panel B: Final sample of treatment group across high-ability and low-ability groups (without students' that are missing outcome variables)			
Number of students	2164	884	1280
Proportion	100%	40.85%	59.15%

Table A3
Baseline characteristics by attrition status[1].

	Non-attrited (N = 9170) (1)	Attrited (N = 462) (2)	p-value H0: (1)= (2) (3)
OUTCOME VARIABLES			
Standardized math score	0.0762 (1.00)	-0.306 (0.988)	< 0.01 * **
Standardized self-concept score	0.021 (0.997)	-0.182 (1.04)	< 0.01 * **
Standardized anxiety score	-0.0268 (1.01)	0.143 (1.05)	< 0.01 * **
STUDENT CHARACTERISTICS			
Female (1 =yes)	0.499 (0.500)	0.387 (0.488)	< 0.01 * **
Boarding at school (1 =yes)	0.597 (0.491)	0.604 (0.490)	0.76
Age (in years)	13.0 (0.940)	13.6 (1.15)	< 0.01 * **
HOUSEHOLD CHARACTERISTICS			
Mother graduated junior high school (1 =yes)	0.251 (0.434)	0.232 (0.423)	0.35
Household assets (score 0–7)	2.75 (1.70)	2.69 (1.78)	0.49
Student discussed homework with parents - more than once a week (1 =yes)	0.633 (0.482)	0.578 (0.494)	< 0.05 * *
MATH TEACHER CHARACTERISTICS			
Female (1 =yes)	0.358 (0.479)	0.377 (0.485)	0.39
Years of teaching experiences	9.61 (7.07)	10.7 (7.02)	< 0.01 * **
Teacher has junior high school math teaching certificate (1 =yes)	0.723 (0.447)	0.732 (0.443)	0.69
SCHOOL CHARACTERISTICS			
School assets (score 0–5)	4.69 (0.620)	4.72 (0.621)	0.29
Student/teacher ratio	8.75 (2.87)	8.93 (3.23)	0.26

Note:
1Outcome variable standardized relative to comparison group
2 N = number of observations
3Mother and father missing both semesters not reported here as variables were created from endline survey and are not available for attrited students.

References

Angrist, J.D., 2014. The perils of peer effects. Labour Econ. 30, 98–108. https://doi.org/10.1016/j.labeco.2014.05.008.

Antonovics, K., Black, S.E., Cullen, J.B., Meiselman, A.Y., 2022. Patterns, Determinants, and Consequences of Ability Tracking: Evidence from Texas Public Schools. National Bureau of Economic Research,.

Argys, L.M., Rees, D.I., Brewer, D.J., 1996. Detracking America's schools: Equity at zero cost? J. Pol. Anal. Manag. 15, 623–645. https://doi.org/10.1002/(SICI)1520-6688(199623)15:4<623::AID-PAM7>3.0.CO;2-J.

Banerjee, A.V., Cole, S., Duflo, E., Linden, L., 2007. Remedying education: evidence from two randomized experiments in India. Q. J. Econ. 30.

Belfi, B., Goos, M., De Fraine, B., Van Damme, J., 2012. The effect of class composition by gender and ability on secondary school students' school well-being and academic self-concept: a literature review. Educ. Res. Rev. 7, 62–74. https://doi.org/10.1016/j.edurev.2011.09.002.

Betts, J.R., 2011. The economics of tracking in education. Handb. Econ. Educ. 3 341–381. https://doi.org/10.1016/B978-0-444-53429-3.00007-7.

Betts, J.R., Shkolnik, J.L., 2000. The effects of ability grouping on student achievement and resource allocation in secondary schools. Econ. Educ. Rev. 19, 1–15. https://doi.org/10.1016/s0272-7757(98)00044-2.

Booij, A.S., Leuven, E., Oosterbeek, H., 2016. Ability peer effects in University: evidence from a randomized experiment. Rev. Econ. Stud. 84, 547–578. https://doi.org/10.1093/restud/rdw045.

Carman, K.G., Zhang, L., 2012. Classroom peer effects and academic achievement: evidence from a Chinese middle school. China Econ. Rev. 23, 223–237. https://doi.org/10.1016/j.chieco.2011.10.004.

Chetty, R., Friedman, J.N., Rockoff, J.E., 2014. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. Am. Econ. Rev. 104 (9), 2633–2679.

Cheung, C.K., Rudowicz, E., 2003. Academic outcomes of ability grouping among junior high school students in hong kong. J. Educ. Res. 96, 241–254. https://doi.org/10.1080/00220670309598813.

Chu, J.H., Loyalka, P., Chu, J., Qu, Q., Shi, Y., Li, G., 2015. The impact of teacher credentials on student achievement in China. China Econ. Rev. 36, 14–24. https://doi.org/10.1016/j.chieco.2015.08.006.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L., 1966. Equality of educational opportunity Washington. US Government Printing Office,, *DC*, pp. 1–32.

Collins, C.A., Gan, L., 2013. Does sorting students improve scores? An analysis of class composition. National Bureau of Economic Research,.

Compulsory education law of the People's Republic of China, 2006. Ministry of Education, Government of China.

Ding, W., Lehrer, S.F., 2007. Do peers affect student achievement in china's secondary schools? Rev. Econ. Stat. 89, 300–312. https://doi.org/10.1162/rest.89.2.300.

Duflo, E., Dupas, P., Kremer, M., 2011. Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. Am. Econ. Rev. 101, 1739–1774. https://doi.org/10.1257/aer.101.5.1739.

Feng, H., Li, J., 2016. Head teachers peer effects and student achievement. China Econ. Rev. 41, 268–283. https://doi.org/10.1016/j.chieco.2016.10.009.

Figlio, D.N., Page, M.E., 2002. School choice and the distributional effects of ability tracking: does separation increase inequality? J. Urban Econ. 51, 497–514. https://doi.org/10.1006/juec.2001.2255.

Fu, Chao, Mehta, Nirav, 2018. Ability tracking, school and parental effort, and student achievement: a structural model and estimation. J. Labor Econ. 36 (4), 923–979.

Fuligni, A.J., Eccles, J.S., Barber, B.L., 1995. The long-term effects of seventh-grade ability grouping in Mathematics. J. Early Adolesc. 15, 58–89. https://doi.org/10.1177/0272431695015001005.

Gamoran, A., 1992. The variable effects of high school tracking. Am. Sociol. Rev. 57, 812–828. https://doi.org/10.2307/2096125.

Gamoran, A., Nystrand, M., Berends, M., LePore, P.C., 1995. An organizational analysis of the effects of ability grouping. Am. Educ. Res. J. 32, 687–715. https://doi.org/10.3102/00028312032004687.

Gopalan, M., Rosinger, K., Ahn, J.B., 2020. Use of Quasi-experimental research designs in education research: growth, promise, and challenges. Rev. Res. Educ. 44 (1), 218–243. https://doi.org/10.3102/0091732×20903302.

Guo, Y., Li, S., Chen, S., Tang, Y., Liu, C., 2022. Health benefits of having more female classmates: quasi-experimental evidence from China. Econ. Educ. Rev. 91, 102330.

Hanushek, E.A., Rivkin, S.G., 2010. Generalizations about using value-added measures of teacher quality. Am. Econ. Rev. 100 (2), 267–271.

Hanushek, E.A., Wößmann, L., 2006. Does educational tracking affect performance and inequality? Differences- in-differences evidence across Countries. Econ. J. 116, C63–C76. https://doi.org/10.1111/j.1468-0297.2006.01076.x.

Hattie, J.A.C., 2002. Classroom composition and peer effects. Int. J. Educ. Res. 37, 449–481. https://doi.org/10.1016/S0883-0355(03)00015-6.

Hong, G., Corter, C., Hong, Y., Pelletier, J., 2012. Differential effects of literacy instruction time and homogeneous ability grouping in kindergarten classrooms: who will benefit? Who will suffer? Educ. Eval. Policy Anal. 34, 69–88. https://doi.org/10.3102/0162373711424206.

Ireson, J., Hallam, S., 2009. Academic self-concepts in adolescence: relations with achievement and ability grouping in schools. Learn. Instr. 19, 201–213. https://doi.org/10.1016/j.learninstruc.2008.04.001.

K. Thiemann Does Impact Abil. Group. vary Cult. Competitiveness ? - Evid. PISA 2017 1 46.

Kerckhoff, A.C., Glennie, E., 1999. The Matthew effect in American education. Res. Sociol. Educ. Social. 12, 35–66.

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., Baumert, J., 2008. Teachers' occupational well-being and quality of instruction: the important role of self-regulatory patterns. J. Educ. Psychol. 100, 702–715. https://doi.org/10.1037/0022-0663.100.3.702.

Koedel, C., Rockoff, J.E., 2015. Value-added modeling: a review. Econ. Educ. Rev. 47, 180–195.

Lai, F., 2007. How do classroom peers affect student outcomes? Evidence from a natural experiment in Beijing's middle schools. Unpublished manuscript.

Lai, G., Wong, O., Feng, X., 2015. Family, school, and access to social capital among high school students in urban Nanjing. Am. Behav. Sci. 59 (8), 946–960. https://doi.org/10.1177/0002764215580589.

Lei, W., Li, M., Zhang, S., Sun, Y., Sylvia, S., Yang, E., Ma, G., Zhang, L., Mo, D., Rozelle, S., 2018. Contract teachers and student achievement in rural China: evidence from class fixed effects. Aust. J. Agric. Resour. Econ. 299–322. https://doi.org/10.1111/1467-8489.12250.

Li, F., Loyalka, P., Yi, H., Shi, Y., Johnson, N., Rozelle, S., 2018. Ability tracking and social trust in China's rural secondary school system. Sch. Eff. Sch. Improv. 29, 545–572. https://doi.org/10.1080/09243453.2018.1480498.

Li, T., Han, L., Zhang, L., Rozelle, S., 2014. Encouraging classroom peer interactions: evidence from Chinese migrant schools. J. Public Econ. 111, 29–45. https://doi.org/10.1016/j.jpubeco.2013.12.014.

Liu, C., Zhang, L., Luo, R., Rozelle, S., Sharbono, B., Shi, Y., 2009. Development challenges, tuition barriers, and high school education in China. Asia Pac. J. Educ. 29, 503–520. https://doi.org/10.1080/02188790903312698.

Liu, C., Zhang, L., Luo, R., Rozelle, S., Loyalka, P., 2010. The effect of primary school mergers on academic performance of students in rural China. Int. J. Educ. Dev. 30, 570–585. https://doi.org/10.1016/j.ijedudev.2010.05.003.

Liu, W.C., Wang, C.K.J., Parkins, E.J., 2005. A longitudinal study of students' academic self-concept in a streamed setting: the Singapore context. Br. J. Educ. Psychol. 75, 567–586. https://doi.org/10.1348/000709905×42239.

Lou, Y., Abrami, P.C., Spence, J.C., 2000. Effects of within-class grouping on student achievement: an exploratory model. J. Educ. Res. 94, 101–112. https://doi.org/10.1080/00220670009598748.

Loyalka, P., Chu, J., Wei, J., Johnson, N., Reniker, J., 2017. Inequalities in the pathway to college in China: when do students from poor areas fall behind? China Q. 229, 172–194. https://doi.org/10.1017/S0305741016001594.

Loyalka, P., Zakharov, A., Kuzmina, Y., 2018. Catching the big fish in the little pond effect: evidence from 33 countries and regions. Comp. Educ. Rev. 62, 542–564. https://doi.org/10.1086/699672.

Loyalka, P., Sylvia, S., Liu, C., Chu, J., Shi, Y., 2019. Pay by design: teacher performance pay design and the distribution of student achievement. J. Labor Econ. 37, 621–662. https://doi.org/10.1086/702625.

Lu, M., Shi, Y., Chang, F., Kenny, K., Rozelle, S., 2017. The Gender gap in math performance, self-concept, and anxiety: rural and urban China in an international context (No. Working Paper 312).

Luo, R., Shi, Y., Zhang, L., Liu, C., Rozelle, S., Sharbono, B., 2009. Malnutrition in China's rural boarding schools: the case of primary schools in Shaanxi Province. Asia Pac. J. Educ. 29 (4), 481–501. https://doi.org/10.1080/02188790903312680.

Marsh, H.W., Martin, A.J., 2011. Academic self-concept and academic achievement: Relations and causal ordering: academic self-concept. Br. J. Educ. Psychol. 81 (1), 59–77. https://doi.org/10.1348/000709910×503501.

Marsh, H.W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.T., O'Mara, A.J., Craven, R. G., 2008. The big-fish-little-pond-effect stands up to critical scrutiny: implications for theory methodology and future research. Educ. Psychol. Rev. 20, 319–350. https://doi.org/10.1007/s10648-008-9075-6.

McEwan, P.J., 2003. Peer effects on student achievement: evidence from Chile. Econ. Educ. Rev. 22, 131–141. https://doi.org/10.1016/S0272-7757(02)00005-5.

Mo, D., Zhang, L., Yi, H., Luo, R., Rozelle, S., Brinton, C., 2013. School dropouts and conditional cash Transfers: evidence from a randomised controlled trial in Rural China's Junior High Schools. J. Dev. Stud. 49, 190–207. https://doi.org/10.1080/00220388.2012.724166.

Mo, D., Huang, W., Shi, Y., Zhang, L., Boswell, M., Rozelle, S., 2015. Computer technology in education: evidence from a pooled study of computer assisted learning programs among rural students in China. China Econ. Rev. 36, 131–145. https://doi.org/10.1016/j.chieco.2015.09.001.

Mulkey, L.M., Catsambis, S., Steelman, L.C., Crain, R.L., 2005. The long-term effects of ability grouping in mathematics: a national investigation. Soc. Psychol. Educ. 8, 137–177. https://doi.org/10.1007/s11218-005-4014-6.

Oakes, J., Guiton, G., 1995. Matchmaking: the dynamics of high school tracking decisions. Am. Educ. Res. J. 32, 3–33. https://doi.org/10.3102/00028312032001003.

OECD. 2012. Education at a Glance, 2012. Highlights from Education at a Glance. OECD. https://doi.org/10.1787/eag_highlights-2012-en.

Paloyo, A.R., 2020. Peer effects in education: Recent empirical evidence. In: Bradley, S., Green, C. (Eds.), The Economics of Education. Elsevier, pp. 291–305. https://doi.org/10.1016/B978-0-12-815391-8.00021-5.

Persson, R.S., 1998. Paragons of virtue: Teachers' conceptual understanding of high ability in an egalitarian school system. High. Abil. Stud. 9 (2), 181–196.

Sacerdote, B., 2011. Peer Effects in Education: How might they work, how big are they and how much do we know Thus Far? Handbook of the Economics of Education, 1st ed.,. Elsevier B.V,. https://doi.org/10.1016/B978-0-444-53429-3.00004-1.

Shi, Y., Zhang, L., Ma, Y., Yi, H., Liu, C., Johnson, N., Chu, J., Loyalka, P., Rozelle, S., 2015. Dropping out of Rural China's Secondary Schools: a mixed-methods analysis. China Q. 224, 1048–1069. https://doi.org/10.1017/S0305741015001277.

Slavin, R.E., 1990. Achievement effects of ability grouping in secondary schools: a best-evidence synthesis. Rev. Educ. Res. 60, 471–499. https://doi.org/10.3102/00346543060003471.

Slavin, R.E., 1993. Ability grouping in the middle grades: achievement effects and alternatives. Elem. Sch. J. 93, 535–552. https://doi.org/10.1086/461739.

Steenbergen-Hu, S., Makel, M.C., Olszewski-Kubilius, P., 2016. What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: findings of two second-order meta-analyses. Rev. Educ. Res. https://doi.org/10.3102/0034654316675417.

Tsang, M.C., 2000. Education and National Development in China since 1949: oscillating policies and enduring Dilemmas. China Rev. 579–618.

Van Houtte, M., 2005. Global self-esteem in technical/vocational versus general secondary school tracks: a matter of gender? Sex. Roles 53, 753–761. https://doi.org/10.1007/s11199-005-7739-y.

Wang, A., Medina, A., Luo, R., Shi, Y., Yue, A., 2016. To board or not to board: evidence from nutrition, health and education outcomes of students in rural China. China World Econ. 24, 52–66. https://doi.org/10.1111/cwe.12158.

Wang, H., Yang, C., He, F., Shi, Y., Qu, Q., Rozelle, S., Chu, J., 2015. Mental health and dropout behavior: a cross-sectional study of junior high students in northwest rural China. Int. J. Educ. Dev. 41, 1–12. https://doi.org/10.1016/j.ijedudev.2014.12.005.

Wang, L.C., 2015. All work and no play? The effects of ability sorting on students' non-school inputs, time use, and grade anxiety. Econ. Educ. Rev. 44, 29–41. https://doi.org/10.1016/j.econedurev.2014.10.008.

Wang, S., 2009. Key school policy hurts education equity. China Education Daily, March 8th, 2009. ⟨http://www.jyb.cn/xwzx/jcjy/sxkd/t20070308_68693htm⟩ (in Chinese).

Wang, X., Liu, C., Zhang, L., Luo, R., Glauben, T., Shi, Y., Rozelle, S., Sharbono, B., 2011. What is keeping the poor out of college? Enrollment rates, educational barriers and college matriculation in China. China Agric. Econ. Rev. 3, 131–149. https://doi.org/10.1108/17561371111131281.

Xinhua, 2010. Setting up the "key schools", exams for tracking placement [in Chinese: Qiao Li Ming Mu She "Zhong Dian Ban", Fen Ban Kao Shi Pai Man Kai Xue Ji].

Xinhua Daily, 2021. The key to "do not set up heavy shifts" is to put an end to "any name" ⟨https://www.thepaper.cn/newsDetail_forward_14280090⟩.

Yang, D., 2005. A research into the senior high school students' social-class-delamination and education acquisition. Res. Educ. Tsinghua Univ. 26 (3), 52–59.

Yi, H., Zhang, L., Luo, R., Shi, Y., Mo, D., Chen, X., Brinton, C., Rozelle, S., 2012. Dropping out: why are students leaving junior high in China's poor rural areas? Int. J. Educ. Dev. 32, 555–563. https://doi.org/10.1016/j.ijedudev.2011.09.002.

Zhang, Y., Chen, D., Wang, W., 2014. The heterogeneous effects of ability grouping on national college entrance exam performance - evidence from a large city in China. Int. J. Educ. Dev. 39, 80–91. https://doi.org/10.1016/j.ijedudev.2014.08.012.