Routledge
Taylor & Francis Group

Journal of Development Effectiveness

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjde20

Stuck in the middle school rut: can anything improve academic achievement in rural Chinese middle schools?

Fei Qin, Huanmin Hu, Prashant Loyalka, Sarah-Eve Dill & Scott Rozelle

To cite this article: Fei Qin, Huanmin Hu, Prashant Loyalka, Sarah-Eve Dill & Scott Rozelle (2022): Stuck in the middle school rut: can anything improve academic achievement in rural Chinese middle schools?, Journal of Development Effectiveness, DOI: 10.1080/19439342.2022.2067890

To link to this article: https://doi.org/10.1080/19439342.2022.2067890

JOURNAL OF DEVELOPMENT EFFECTIVENESS
https://doi.org/10.1080/19439342.2022.2067890

Routledge
Taylor & Francis Group

Stuck in the middle school rut: can anything improve academic achievement in rural Chinese middle schools?

Fei Qin[a], Huanmin Hu[b], Prashant Loyalka[b,c], Sarah-Eve Dill[c] and Scott Rozelle[c]

[a]Department of Agricultural Economics, Purdue University West Lafayette Indiana US; [b]Graduate School of Education, Stanford University, Stanford, CA, USA; [c]Freeman Spogli Institute for International Studies, Rural Education Action Program (REAP), Freeman Spogli Institute for International Studies, Stanford University, Stanford, CA, USA

ABSTRACT

Academic achievement in middle schools in rural China remains poor for many students. This study examines whether programmes and interventions can improve academic achievement by reviewing rigorous experimental evaluations of nine programmes (11 interventions) on 47,480 rural middle school students in China. The results find none of the interventions improved academic achievement. Moreover, we find no evidence for heterogeneous treatment effects by student gender, age or previous academic achievement. These results may be due in part to the academically-demanding nature of the middle school curriculum, which is applied universally to students with varying levels of cognitive ability.

ARTICLE HISTORY
Received 29 January 2020
Accepted 14 April 2022

KEYWORDS
Junior high school student; academic achievement; rural China

1. Introduction

Education is essential for the well-being of modern-day China. As the country transitions from a manufacturing-based economy to a knowledge-based one, its labour force needs to become highly-skilled and highly-educated. There are concerns that, without high levels of human capital, the economy may fail to compete with higher-income countries and, as a result, growth may stagnate (Khor et al. 2016). Education is also important on an individual level as returns to schooling, especially at the tertiary level, have been on the rise with the country's economic transition (Carnoy et al. 2013; Heckman and Li 2004; Li 2003). Students who fail to attend higher levels of education miss out on these returns.

Unfortunately, the education system in poor rural areas has been unable to produce high school graduates at a rate considered healthy at this point in China's development path. Approximately 29 million students – or three-fourths of China's relevant middle school-age population – enrol in rural schools every year, but less than half of them advance to academic high school (grades 10 to 12), either because they do not pass competitive admission tests or because they choose not to pursue high school at all (C. Liu et al. 2009; Loyalka et al. 2017).[1] Low levels of achievement among rural middle school students (grades 7 to 9) may be an important reason why the education system in rural areas fails to produce higher rates of matriculation to academic high schools.

To solve this problem, efforts have been made to provide a better academic learning environment in China's rural middle schools. China's government, for example, implemented a large national teacher training programme with the explicit goal of increasing student achievement (H. Liu et al. 2016). The government also sought to incentivise teachers by implementing a national scheme

CONTACT Sarah-Eve Dill sedill@stanford.edu Rural Education Action Program, Stanford University, 616 Jane Stanford Way, Stanford, CA 94305

where teacher promotion, and consequently teacher pay, became highly dependent on how well students performed academically (Karachiwalla and Park 2017). Likewise, international NGOs and foundations have introduced various programmes to improve student achievement in rural middle schools (Clinton Foundation 2009). Unfortunately, in the absence of rigorous evaluation, it was unclear what types of interventions or programmes could successfully improve student outcomes.

Fortunately, the last few years have witnessed a growing body of research studying rural schools, in general, and rural middle schools, in particular. In carrying out their research, scholars and their implementing partners (such as local school districts and NGOs) have begun to use rigorous evaluation methods to test the impact of a variety of interventions on student outcomes. Interventions include programmes to reduce financial constraints, improve social-emotional learning, provide information about returns to schooling, improve teacher quality, and address health-related barriers such as poor eyesight (F. Li et al. 2017a; H. Wang et al. 2016; Loyalka et al. 2013, Forthcoming; ; Nie et al. 2016; Mo et al. 2013; Yi et al. 2015). Findings from these studies have shown significant impacts from some interventions, especially when it comes to reducing dropout behaviour and increasing attendance (e.g. dropout rates decreased by 22% for a single semester in programmes that taught students social-emotional learning skills (H. Wang et al. 2016); 44% in programmes that gave myopic students free glasses (Nie et al. 2016); and 60% in programmes that provided a student's household cash transfers conditional on the student's attendance (Mo et al. 2013).

Nevertheless, as it pertains to student achievement, the results have been less clear, and no overall picture emerges from the literature. One reason behind this lack of clarity may be that recent studies have appeared in the literature somewhat idiosyncratically. Some have been published in the education literature; others in development economics and public health. Not all of the studies have made studying the impact on academic achievement the focus of their work (and some have not included analysis of the impacts on achievement, despite having data from standardised academic assessments). As such, no clear pattern has emerged regarding what types of programmes or interventions are able to positively affect student achievement. Moreover, studies that have looked at the impacts of interventions on achievement have not always been designed with enough power to estimate heterogeneity in treatment effects.

This study attempts to overcome these shortcomings by providing a comprehensive synthesis of existing, large-scale and in-the-field experimental studies in rural Chinese middle schools over the past eight years. In this paper, our approach is to use the size and strength of this entire body of research to investigate the impact potential of a broad range of interventions on academic achievement. To do so, we created a large pooled dataset from nine well-identified, randomised controlled trials (RCTs). The nine programmes included eleven interventions, each falling under one of four broad categories: financial support, information-based interventions, teacher-based interventions, and glasses for students with poor vision. Because the pooled sample is large (containing information on 47,480 rural middle school students in 713 schools), the pooling strategy affords us a high degree of statistical power, a greater ability to generalise the findings, and an ability to detect treatment effect heterogeneity among different student types.

With this approach, the study addresses three primary questions: which interventions, if any, can improve academic achievement in Chinese middle schools? Do the interventions in question have heterogeneous effects on particular sub-populations of students? What are some of the barriers that may hinder the success of the interventions? To measure the success of the interventions, we utilise academic achievement, and more specifically scores from standardised maths scales, as the main outcome variable of interest. We examine whether the different interventions had heterogeneous effects on student achievement based on student gender, previous academic performance and age. Finally, we use subsets of the data that included additional variables (variables used in one or more, but not all, of the full set of nine programmes), as well as previous literature, to further explore structural factors that may help explain the success or failure of the interventions to raise student achievement.

With the possible exception of the glasses-based intervention (which may or may not have an impact, because, at most, the impact is relatively small and is only statistically significant at the 10% level), we find no evidence that any of the interventions improved academic achievement. We also find no evidence for heterogeneous effects that could be masking the zero average effects. Unlike the results of interventions that were run in rural elementary schools in the same regions (which often did have large & statistically significant impacts), the findings in this paper suggest that academic achievement in rural middle schools is unresponsive to interventions that are designed to improve student achievement. And, it is for this reason that the title of the paper is that academic achievement in rural Chinese middle schools in poor rural areas appears to be 'stuck in a rut'.

Based on evidence that the interventions were well-implemented (except potentially one of the interventions, which will be discussed below), we explore reasons for the lack of impact. To do so, we examine several potential hypotheses about how institutional factors may be responsible for the absence of impacts on academic achievement. In briefest terms, we reject three hypotheses that interventions fail because: (a) teachers in rural Chinese middle schools are not qualified; (b) teachers in rural Chinese middle schools lack incentives to teach well/exert effort; or (c) students are suffering from high levels of anxiety and this is undermining their ability to learn. In contrast, we do find some support for the hypothesis that the absence of improved achievement is rooted in the fast-paced and academically-demanding nature of the curriculum in rural Chinese middle schools. We also make the conjecture that curriculum-associated challenges are particularly difficult to overcome because the curriculum is applied to a student body that has a mix of both highly-motivated students (being taught by equally motivated teachers) and students whose levels of cognitive development may make it difficult to increase achievement in response to new interventions.

We briefly discuss other reasons that may have some explanatory power. Some of the interventions were not fully focused on improving achievement, and instead were also focused on raising attendance. It also may be that the high opportunity cost of labour in China makes it difficult to keep some students focused on improving achievement; students may have been anxious to leave school and work in the labour market.

Overall, however, our results show that educational interventions (at least the types we are studying) cannot easily improve student achievement. programmes that seek to address traditional barriers to improved achievement in other educational contexts, such as financial constraints, lack of information on returns to schooling, and teachers who are not qualified or incentivised enough to produce learning gains, are clearly ineffective. Rather, any major effort to improve achievement in rural middle schools in China may have to address fundamental structural issues. For current or soon-to-be matriculating middle school cohorts, this may entail addressing the structure and speed of the curriculum and making it more flexible to match the needs of the students and the pace of their learning. For future cohorts, policy reforms may need to better prepare students in elementary school, preschool, and earlier stages. For example, educational policymakers may wish to collaborate with paediatric health and early childhood education providers to ensure that children do not fall behind developmentally.

In addition to its direct relevance to rural middle schools in China, this study complements a large literature that seeks to identify how and when to invest in human capital in order to yield the highest returns. The work of James Heckman suggests that cognitive abilities become stable by age 10, and as such, the earlier the timing of intervention, the higher the returns (Cunha and Heckman 2007; Heckman 2008). Even though interventions targeted at adolescents can sometimes produce gains in schooling, earnings, and crime prevention, these gains are at best modest when compared to those from interventions targeted in utero (e.g. ones for a healthy pregnancy) and/or the first three years of life (e.g. programmes for healthy child development; Heckman 2008). This study does not review human capital interventions targeted at earlier stages of life; however, our findings are in keeping with Heckman's work in that investments targeted at middle school may be too late and, consequently, ineffective. In fact, in recent years researchers have conducted and evaluated early

childhood programmes, of the type that Heckman has been advocating, in rural China. These pilot programmes have been effective at raising cognition in early childhood but were done too recently to know if there are longer run impacts on middle school achievement (Sylvia et al. 2018).

The rest of the paper is organised as follows. Section 2 provides basic background information on middle school in China and the interventions in question. In section 3, we describe the data, the sample, and the empirical strategy. We then present the results in section 4, discuss the structural factors that may impede achievement gains in section 5, and conclude in section 6.

2. Background

2.1 Middle school in China

In China, academic achievement during middle school is critical. Serving as the last stage of compulsory education, middle school is three years long. During these three years, students are supposed to learn a number of abstract concepts and concepts that require integrated thinking––for example, in maths where students study algebra, geometry, and trigonometry; in Chinese language class where there are integrated essay writing assignments and critical thinking skills; and in English as a foreign language (Norton and Zhang 2013; Wu 2015). It is also during these years that students acquire important content and skills that often become the foundation for high school and college.

In addition to the importance of its instructional content, middle school represents a critical stage because of the gated nature of the Chinese educational system. At the end of middle school, students that want to continue on in school (as opposed to joining the labour force) must pass a high school admission exam in order to advance to academic high school. The exam itself, a prefecture-wide exam in most provinces, is designed to be a difficult assessment that provides policymakers a way to sort students by their academic achievement levels.

For students from rural areas, the exam is the bottleneck for advancing into higher levels of education (Khor et al. 2016). A study by Loyalka et al. (2017), for example, shows that only three out of five students (60%) attempt the exam; middle school students from rural areas are furthermore much less likely than their urban counterparts to take the exam. Although official rates are not published, the rate of passing is lower in rural areas than urban.[2] As such, if a student wants to advance further in school, strong academic achievement during middle school is necessary. Once students pass the high school admission exam, the likelihood of advancing into higher stages of education rises sharply and university becomes attainable for most academic high school students in most regions (Loyalka et al. 2017).

Because of the importance of the high school admission exam, the curriculum in middle school is highly structured, difficult, and fast-paced. It is also regulated at a higher level of administration (e.g. the county, the prefecture, or the province), so as to be fair for all students in the jurisdiction (D. Wang 2011). Because everyone in the jurisdiction takes the same exam, everyone needs to cover the same material for the exam in the same time frame and at the same depth. As a result, the pace of the class is often out of the control of teachers and principals – and does not depend on the differences in student ability. The materials taught and the pace at which they are taught are also typically independent of the rate at which different students can learn. In simplest terms, teachers cannot slow down their teaching to fit the needs of students (or subsets of students) because the ultimate purpose of middle school is to get students ready for the highly competitive high school admission exam.

If middle school performance holds this much importance, the question then arises: why is poor academic achievement common among rural students (relative to urban students)? Previous studies in rural China and elsewhere have identified four potential factors. The first is financial constraints; if students perceive they may not be able to pay for high school tuition fees or perceive these fees as too high, they may decide not to pursue further education, which precludes the need for high academic achievement in middle school (Brown and Park 2002). Second, students may not try hard

because they lack information about returns to high school and university, especially if they are not exposed to individuals in their immediate environment who are earning these returns (Jensen 2010; Nguyen 2008). Third, teachers who lack training or incentives may be the weak link (McEwan 2015; Muralidharan and Sundararaman 2011). Finally, almost one in four rural elementary students suffer from vision problems, but only a small fraction of them own glasses, which may impede learning (Congdon et al. 2008; He et al. 2007). The share of students that have uncorrected myopia is, in fact, higher in middle school than primary school or high school (Ma et al. 2018). As such, interventions that directly address these issues with the purpose of increasing student achievement are of substantial interest.

2.2 An overview of the interventions

The present study examines eleven randomised interventions (within nine programmes) targeted at middle school students (Table 1). All nine programmes were designed and executed by the Rural Education Action programme (REAP), an impact evaluation organisation which (among other research projects) uses RCTs to evaluate educational policy in rural China. Each intervention can be classified under one of four broad categories: providing financial aid to help poorer students pay for high school; providing information to students about the returns to schooling; raising the effectiveness of teachers; and providing glasses to aid with uncorrected myopia. This section describes the interventions and their theories of change. Readers interested in the details of each programme can refer to the listed studies for more information (Table 1).

Four programmes/interventions (P1 – P4 described in Table 1) provided financial aid. Due to common liquidity constraints in rural China, there are still households that cannot afford the tuition fees for academic high school, which are among the highest in the developing world (C. Liu et al. 2009). Further, the existence of off-farm labour markets – which still today often have low educational requirements for workers,[3] means that staying in school has a high opportunity cost (Heckman and Li 2004; H Li 2003) . Because of this, it is thought that a subset of students in junior high school (especially those from lower SES groups) may not be exerting effort to learn as much as they would if academic high school were free and mandatory.

In the four financial aid programmes, interventions were carried out according to a common procedure. First, we identified the poorest students in each class through our survey-based data on student household assets as well as information provided by homeroom teachers and principals. The second step entailed giving the identified poor students in the treatment arms financial aid in the form of a cash transfer. The actual nature of the condition varied among the four programmes. In P1, the transfer was given to the grade 7 middle school student if he was still in middle school at the end of the school year (and had maintained a low absentee rate). In P4, the transfer was given if the student matriculated into high school, either academic or vocational. We refer to the intervention in these two programmes as 'Conditional Cash Transfers' programmes or *CCTs*. In P2 and P3, the interventions took the form of an early commitment for financial aid. In these two programmes, the middle school student (grade 7 in P2 and grade 9 in P3) was guaranteed financial aid that fully covered tuition for all three years of high school, if he matriculated into any type of high school (and stayed in school). We refer to the intervention here as an 'Early Commitment to Financial Aid' programme or an *ECFA* programme.

We also included two information-based interventions. In these two interventions the research team aimed to addressed a recurrent problem that occurs in many developing countries where students often possess inaccurate information about returns to schooling or lack career planning skills (Jensen 2010; Nguyen 2008). As a result of this inaccurate or deficient knowledge, students may elect to pursue less education than what is optimal (for the individual) or put relatively less intensive effort into their academic activities.

Table 1. Summary of the Nine Middle School Programs (and 11 Interventions) Examined in this Study

Program		Number of schools	Number of students	Location	Grade	Citation
Pooled sample		713	**47,480**	5 provinces	7th-9th	
Financial aid interventions						
P1	**CCT 1**	10	**268**	Shanxi	7th	(Mo et al. 2013)
	P1-T: CCTs	9	140			
	P1-C: Control	10	128			
P2	**ECFA 1**	132	**1,254**	Shaanxi, Hebei	7th	(Yi et al. 2015)
	P2-T: ECFA	66	407			
	P2-C: Control	66	847			
P3	**ECFA 2**	30	**280**	Shaanxi	9th	(Yi et al. 2015)
	P2-T: ECFA	28	141			
	P3-C: Control	30	139			
P4	**CCT 2**	94	**443**	Shaanxi, Hebei	7th	(F. Li et al. 2017a)
	P4-T: CCTs	49	164			
	P4-C: Control	45	279			
Information-based interventions						
P5	**Educational returns and career**	131	**11,529**	Shaanxi, Hebei	7th	(Loyalka et al. 2013)
	P5-T1: returns	44	2,740			
	P5-T2: returns + career	43	2,724			
	P5_C: Control	46	6,065			
P6	**Social-emotional learning**	70	**6,173**	Shaanxi	7th, 8th	(H. Wang et al. 2016)
	P6-T: Social emotional learning	35	3,075			
	P6-C: Control	35	3,098			
Teacher-based interventions						
P7	**Teacher training**	298	**14,838**	Henan	7th,8th,9th	(Loyalka et al. Forthcoming)
	P7-T1: NTTP	99	5,020			
	P7-T2: NTTP + follow-up	99	4,914			
	P7-C: Control	100	4,904			
P8	**Teacher incentives**	200	**12,095**	Shaanxi, Gansu	7th	(Loyalka, Sylvia, etal. Forthcoming)
	P8-T: Teacher incentive	100	6,310			
	P8-C: Control	100	5,785			
Glasses						
P9	**Free glasses**	31	**600**	Shaanxi	7th,8th	(Nie et al. 2016)
	P9-T: Free glasses	15	307			
	P9-C: Control	16	293			

To address this information gap, programme 5 (P5) provided two types of interventions. In one subset of P5 schools, homeroom teachers were trained to give a single one hour lesson to their grade 7 students about: (a) the average wage levels that were associated with students that finished different levels of schooling; (b) the wage gap between middle school graduates and high school graduates; and (c) the levels of tuition needed for different levels of schooling and schools. This intervention, which we refer to as the 'returns arm', was similar to the intervention that Jensen (2010) implemented in his study in the Dominican Republic. In the other subset of schools, homeroom teachers were trained to give a set of four one-hour-lectures that included the returns arm plus three one-hour-classes focused on helping each student think about one's future career. These career-focus classes sought to get students thinking about the skills that they would need in the future for career planning, identifying career interests, and understanding how to navigate China's education system after middle school so they would be able to achieve their career goals. We refer to this intervention as the 'returns + career arm'.

programme 6 (P6––the second programme in the information category) provided a different type of information. In essence, P6 aimed at equipping students with social-emotional learning skills to address learning anxiety. Other work has shown that learning anxiety is high in many rural Chinese middle schools. In P6, non-core course teachers (e.g. art

or music teachers) received five days of training on how to execute a two-semester-long, one-day-per-week, fully-scripted set of sessions to students in grades 7 and 8 about topics such as emotional-management, self-awareness, and building positive relationships with both teachers and fellow students.

The third group of interventions addressed issues related to both the quality of the teaching skills of teachers and the incentives that teachers face. Teacher quality has been consistently causally linked to student achievement (Aaronson, Barrow, and Sander 2007; Rockoff 2004). The literature also has documented the importance of teacher contracts and how they are paid. In particular, research has demonstrated that teacher pay programmes which do not tie teacher bonus payments to achievement gains for all students can lead teachers to focus their efforts on some students more than others (Cochran-Smith 2005; Neal and Schanzenbach 2010).

Two programmes in our meta-study sought to directly influence teachers to raise the academic achievement of students. programme 7 (P7) evaluated the impact of China's National Teacher Training (NTTP) programme. There were two treatment arms: 'NTTP only' arm where teachers participated in the regular government-planned 15-day on-site NTTP and supplemental online training, and 'NTTP + follow-up' arm, which included continuous post-training *follow-up* with teachers, alerting them of supplementary materials, assignments, and progress reports through text messages and phone calls. programme 8 (P8) introduced an incentive-based payment system to treatment teachers. Under this scheme, implemented by the research team, teachers received bonus pay based not on the average achievement level of their students, but rather on the achievement gains of each student––regardless of where the student was on the achievement distribution––compared to students in other schools that had an equal level of achievement at baseline. In other words, in P8 the research team wanted this form of bonus design to make every student count in the determination of rewards and consequently make teachers attentive to all of their students rather than the top-achieving ones only.

programme 9 (P9) provided the fourth kind of intervention, free glasses. Previous research demonstrated that as high as one half of rural Chinese students who needed glasses did not own them (Congdon et al. 2008; He et al. 2007). Because visually-impaired students without the necessary corrections have been shown to fall behind academically (Ma et al. 2014), P9 aimed to enhance academic performance by identifying students with vision problems, providing them with quality prescriptions, and then giving each student a free pair of glasses (Nie et al. 2016).

3. Method

We use data from nine programmes targeted at middle school students. All of them took place between 2009 and 2016 and span rural areas in five Chinese provinces: Gansu, Shaanxi, Henan, Hebei, and Shanxi. Because two of these programmes include two different treatment arms, we thus examine a total of eleven interventions. To our knowledge, with the exception of this set of programmes, there have been no other RCT-based evaluations that have examined academic outcomes in rural Chinese middle schools.

In this section, we describe four aspects of the data and how we use them. We describe: (a) sampling and randomisation; (b) data collection; (c) the attrition, balance, and summary statistics of student/school characteristics and the outcome measures; and (d) the empirical strategy we employ in the present study.

3.1 Sampling and randomisation

The nine programmes discussed here all used similar six-step random sampling strategies. First, the research teams obtained a list of all counties in each sample province or prefecture. Second, the teams selected counties from those meeting each study's criteria, the most important of

which was GDP per capita. In all of the interventions – except the teacher training intervention (P7), researchers limited their selection of counties to ones in which GDP per capita was either below the national average or was low enough so that the county was qualified to be a 'national poverty county'. Third, in each sample county, officials in the bureau of education provided a list of all middle schools. Fourth, research teams used this list and called principals of the sampling-frame schools to identify whether these schools met specific criteria. For example, we excluded schools with low middle school enrolment (i.e. schools that were mostly primary schools); schools in prefecture or county seats which typically catered to urban students; etc. Fifth, the research teams randomly selected schools from the resulting sampling frame of eligible schools.[4] Sixth and finally, within each school, researchers randomly selected classes of students in the targeted age groups for inclusion in the programmes. Unlike the other programmes which randomised the treatment on a school level, the financial aid programmes (P1-P4) were targeted at less wealthy students and thus randomised the treatments among individual students who were identified as poor.

3.2 Data collection

All nine programmes used similar data collection procedures including both a baseline survey (before randomisation and before any treatment) and an endline survey. The baseline and endline surveys were nearly identical. Each was administered in three blocks. In the first block, sample students were asked to take a 30-minute standardised maths test, the score of which served as the primary outcome in this study.[5] The tests were comprised of maths questions from the curriculum so that they would be suitable for the academic levels of the students. Each maths test contained 30–35 maths questions and the difficulty level varied with the school grade.[6] To ensure credibility, the research teams administered the test themselves, and the test was strictly timed. To minimise cheating, the enumeration team also closely proctored the whole class. We converted test scores into z-scores by subtracting the mean and dividing by the standard deviation (SD) of the maths score distribution of students tested within a given grade.

In the second block, enumerators collected data on demographic and socioeconomic characteristics of students and their families. Because the demographic subsections of the surveys in all nine programmes were nearly identical, we could generate control variables for all observations on student gender and age as well as on education level and migration status of the parents.

In the third block of the survey, the research team collected information on the characteristics of the teachers in the sample schools. Specifically, information was collected on each teacher's teaching experience, professional rank, education level and major. Information was also collected on each teacher's monthly wage.

3.3 Attrition, balance, and summary statistics

Our pooled sample (obtained by simply combining all data from the nine programmes) contains 53,808 students at baseline. However, in the time between the baseline survey and the endline survey, 6,328 students (11.8%) had attrited (for a total final sample size that was used in the analysis of 47,480). To ensure that this attrition did not affect the integrity of the impact evaluation, we need to examine whether such attrition was related to the treatment assignment. We do so, first in each individual programme, second in the pooled dataset, and, finally, in the four subsets of interventions.

In the case of seven out of the nine programmes, attrition was statistically indistinguishable across the treatment arm(s) and the control arm (Table 2). For P1, which provided financial aid to students in the treatment group, attrition was less likely to take place in the treatment group. The difference, however, is not surprising and can be explained simply by considering the nature of this programme. Besides trying to raise student achievement, P1 also was also designed to reduce

dropout rates and was indeed successful in achieving that goal. The repercussion relevant to our analysis, however, is that students who dropped out did not take the maths test. Hence, attrition, defined by the absence of a maths score, is more pronounced in the control arms for this programme. For P2, attrition was significantly more likely in the treatment group––but only at the 10% level.

As seen in Table 2, in the case of the pooled sample, we find that the attrition was not significantly different between students who received any of the 11 interventions (from the nine programmes described above) and students who were assigned to a control arm.

Finally, when we further divide the pooled sample according to the intervention type, we find that for three out of the four categories, attrition was also random. In the teacher-based interventions, attrition was significantly higher at the treatment group––but only at the 10% level and with a small magnitude of one percentage point. Overall, we are confident that this analysis shows that treatment-induced (or treatment-reduced) attrition does not constitute a concern.

We find that the balance post-attrition also remained largely intact both within the pooled sample and the four categories of interventions (Table 3).[7] When we examine balance across treatment and control students in the pooled sample (Table 3, Column 1), three minor differences emerge: students

Table 2. Attrition rates.

programme(s)	(1)	(2)	(3)	(4)
	Overall	Treatment	Control	Difference
Individual interventions				
P1 – CCT	0.107	0.067	0.147	−0.080**
	[0.309]	[0.250]	[0.355]	(0.025)
P2 – ECFA	0.116	0.141	0.103	0.039*
	[0.320]	[0.349]	[0.304]	(0.020)
P3 – ECFA	0.263	0.258	0.268	−0.011
	[0.441]	[0.439]	[0.444]	(0.037)
P4 – CCT	0.656	0.635	0.667	−0.033
	[0.475]	[0.482]	[0.471]	(0.053)
P5 – T1 Returns	0.102	0.100	0.102	−0.002
	[0.302]	[0.300]	[0.303]	(0.014)
P5 – T2 Returns + Career	0.108	0.120	0.102	0.018
	[0.310]	[0.326]	[0.303]	(0.013)
P6 – Social Emotional Learning	0.176	0.168	0.185	−0.017
	[0.381]	[0.374]	[0.388]	(0.021)
P7 – T1 NTTP	0.105	0.105	0.105	0.000
	[0.307]	[0.307]	[0.307]	(0.013)
P7 – T2 NTTP + follow-up	0.111	0.118	0.105	0.013
	[0.315]	[0.322]	[0.307]	(0.014)
P8 – Teacher incentive	0.048	0.048	0.048	0.001
	[0.214]	[0.214]	[0.213]	(0.006)
P9 – Glasses	0.096	0.084	0.109	−0.026
	[0.295]	[0.277]	[0.313]	(0.032)
Pooled sample	0.118	0.113	0.123	0.010
	[0.322]	[0.316]	[0.329]	(0.007)
Grouped interventions				
Financial aid	0.337	0.325	0.344	−0.018
	[0.473]	[0.469]	[0.475]	(0.027)
Information	0.132	0.132	0.132	0.000
	[0.338]	[0.338]	[0.339]	(0.011)
Teacher	0.083	0.088	0.075	0.013*
	[0.276]	[0.283]	[0.263]	(0.007)
Glasses	0.096	0.084	0.109	−0.026
	[0.295]	[0.277]	[0.313]	(0.032)

Note. Columns 1–3 show attrition rates with standard deviations displayed between brackets. Column 4 show the estimate from a simple linear regression with attrition as the dependent variable and the corresponding treatment arm as an independent variable.

For Column 4, standard errors are displayed between parentheses.

Table 3 Balance Across Arms of Non-attrited Students in our Pooled Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Panel A Difference between treatment group and corresponding control group (SE)				*Panel B* Descriptive statistics [SD]		
	Pooled Treatment	Financial Aid	Information	Teacher quality	Glasses	Overall	Treatment	Control
Male student	0.010*	0.034	0.017*	0.001	0.013	0.502	0.507	0.497
	(0.005)	(0.022)	(0.010)	(0.007)	(0.044)	[0.499]	[0.499]	[0.498]
Student age, years	0.059	0.005	0.028	0.220**	-0.094	13.491	13.518	13.459
	(0.080)	(0.079)	(0.135)	(0.095)	(0.144)	[1.315]	[1.301]	[1.330]
Math test score, (SD)	-0.005	0.030	-0.049	0.007	0.211	0.051	0.048	0.054
	(0.034)	(0.076)	(0.056)	(0.046)	(0.152)	[0.989]	[0.979]	[1.000]
Mother completed middle school	0.027*	0.013	-0.002	0.032	-0.017	0.338	0.350	0.323
	(0.014)	(0.030)	(0.021)	(0.020)	(0.041)	[0.473]	[0.477]	[0.468]
Father completed middle school	0.023**	0.018	0.010	0.030*	0.023	0.484	0.495	0.472
	(0.012)	(0.030)	(0.018)	(0.017)	(0.056)	[0.500]	[0.500]	[0.499]
Mother migrated	-0.016	-0.019	0.017	0.002	0.003	0.277	0.270	0.286
	(0.018)	(0.031)	(0.041)	(0.011)	(0.037)	[0.383]	[0.364]	[0.405]
Father migrated	-0.023	-0.021	0.014	-0.003	-0.081	0.546	0.536	0.558
	(0.019)	(0.023)	(0.043)	(0.017)	(0.052)	[0.416]	[0.398]	[0.437]
Observations	47,480	2,245	17,702	26,933	600	47,480	25,942	21,538

Note. Column 1 compares students who were assigned to any intervention arm and students who were assigned to any control group. Columns 2-5 compare students who were assigned to an intervention arm in a given intervention category and students assigned to a control arm in the same programs.

* Significant at a 10% level. ** Significant at the 5% level. *** Significant at the 1% level.

in treatment arms are one percentage point more likely to be male; three percentage points more likely to have a mother who completed middle school and two percentage points more likely to have a father who completed middle school.

Importantly, one rather large difference, however, arises in the baseline maths scores in one of the programmes. Specifically, students who were in the eyeglasses treatment group scored 0.27 SD higher, a difference that is statistically significant at the 10% level. As it turns out, we will argue that it is precisely this large difference in baseline outcome scores (which is greater than the difference that we find due to the intervention) that will support a final conclusion that none of the interventions, including the free eyeglasses intervention, affect student achievement in China's middle schools.[8] To further minimise any bias resulting from these differences, in all of the analyses in this paper, we control for all characteristics displayed in Table 3.

The final size of the pooled sample is thus 47,480 students. As Table 3 (Panel B) shows, the average age is 13.5 years and half are male students (50.2%). Parental education was generally low; only one third of mothers (33.8%) and less than one half of fathers (48.4%) completed middle school or higher levels of education. We also find that 27.7% of students' mothers had migrated to another county or city. For fathers, this figure is around 54.6%.[9]

3.4 Empirical strategy

Each programme examined in this study relies on randomisation, and because the treatment arm(s) and control arm within each programme were, to a large extent, comparable at baseline (with the exception of the eyeglasses study), any differences in academic achievement at endline can be causally linked to the intervention. For this reason, our primary specification is a simple ordinary least squares (OLS) regression model:

$$Y_{ijc} = \beta_o + \beta_1 T_{ijc} + \gamma X'_{ijc} + \tau_c + \varepsilon_{ijc} \tag{1}$$

where Y_{ijc} represents the endline standardised maths test score for student i in school j in county c; T_{ijc} is a dummy variable which takes on a value of one if student i is in the treatment arm and zero otherwise[10]; τ_c is a set of county fixed effects; and X'_{ijc} represents a vector of baseline variables, including student gender, whether their father and/or mother completed middle school, whether their father and/or mother migrated, as well as their baseline maths scores. For the school-level interventions, standard errors are clustered at the school level.[11] Our parameter of interest is β_1 which measures the treatment impact on student maths achievement.

We use this model in two different ways. First, we provide a simple robustness check to the estimates produced by the authors in each study. To do this, we use Equation (1) uniformly but separately on each programme. We then compare the results to the estimates reported by the authors of each study, which, although substantively the same, differ slightly from one study to the next and from the present study (through the inclusion or exclusion of controls, fixed effects, and/or weights to address attrition). Second, we test whether there is an average effect of the eleven interventions. To do this, we use the specification in Equation (1) with the only difference being that T_{ijc} takes on a value of 1 if the student is in *any* treatment group and zero otherwise.

We use a slightly different model to understand the effects of each type of intervention. To do so, we pool data from all programmes together and test whether intervention type (out of the four large categories) could change maths achievement. Equation (2) shows the model we use:

$$\begin{aligned} Y_{ijc} = \beta_o + \beta_1 Aid_{ijc} + \beta_2 Information_{ijc} + \beta_3 Teacher_{ijc} \\ + \beta_4 Glasses_{ijc} + \alpha_l + \gamma X'_{ijc} + \tau_c + \varepsilon_{ijc} \end{aligned} \tag{2}$$

where the four treatment categories are represented by the four dummy variables: *Aid, Information, Teacher* and *Glasses*.

Finally, we test whether the interventions have different effects on different subgroups of students. The model we use for this purpose is as follows:

$$Y_{ijc} = \beta_o + \beta_1 T_{ijc} + \beta_2 D_{ijc} + \beta_3 T_{ijc} \times D_{ijc} + \gamma X'_{ijc} + \tau_c + \varepsilon_{ijc} \quad (3)$$

where D_{ijc} is a dummy indicator representing a specific baseline characteristic of student i; $T_{ijc} \times D_{ijc}$ is an interaction between the characteristic and the treatment assignment; and all other variables are the same as model (1). We use this model for the pooled sample as well as the four category-based samples obtained by dividing the sample according to intervention type. For heterogeneity analyses, we present the interaction term as well as the effect on each subgroup.

4. Results

Table 4 shows the treatment impacts of all nine programmes on maths achievement. Column 1 summarises the estimates produced from the same specifications used in the original evaluation of each programme.[12] In other words, these are almost precise replications of the results in published papers or working papers provided that the papers examined achievement as an outcome (reminder: see Table 1 for brief descriptions of the studies and the original citations). All estimates (like the estimates of the original studies) are small in magnitude––ten of the 11 estimates are below one tenth of an SD––and none are statistically significant. The one exception to this pattern is the glasses intervention (P9). Among students who did not have glasses at baseline, treated students scored 0.20 SD higher than their control

Table 4. Impact of each treatment on students' maths performance.

programme Number	Grade(s)	Intervention	(1)	(2)
			Replicated impact on Maths (SD)	Estimated: impact on Maths SD
P1	7th	CCT	0.010	−0.011
			(0.220)	(0.116)
P2	7th	ECFA	−0.016	−0.025
			(0.056)	(0.051)
P3	9th	ECFA	−0.019	−0.019
			(0.100)	(0.098)
P4	7th	CCT	−0.047	−0.013
			(0.090)	(0.101)
P5-1	7th	Returns	−0.005	0.030
			(0.046)	(0.041)
P5-2	7th	Returns + Career	−0.073	−0.008
			(0.046)	(0.044)
P6	7th	Social emotional learning	0.013	0.016
			(0.056)	(0.077)
P7-1	7th, 8th, 9th	NTTP	−0.006	0.022
			(0.034)	(0.033)
P7-2	7th, 8th, 9th	NTTP + online	0.005	0.026
			(0.035)	(0.034)
P8	7th	Teacher incentives	−0.004	0.007
			(0.032)	(0.031)
P9	7th, 8th	Free glasses	0.196**	0.196**
			(0.083)	(0.081)

Note. This table presents regression estimates where the dependent variable is the standardised maths score.
Column 1 shows estimates resulting from the specifications chosen by the authors of each study as well as their code which runs the regressions. Column 2 shows estimates from our model which is uniform for all interventions and which include county dummies and controls for gender, age, baseline maths score, parental migration status, and parental education. P9 includes additional strata fixed effects.
Robust standard errors are displayed in parentheses and are clustered at school level for all programmes except P2-P4.
* Significant at 10%. **Significant at 5%. *** Significant at 1%.

counterparts (significant at the 5% level). Although the estimated coefficient (0.20 SD) is significantly different than zero, it is important to remind the reader that this estimated effect is actually smaller than the observed difference between the treatment and the control group at baseline (0.27 SD).

In short, then, with the (possible) exception of the free eyeglasses intervention (and maybe not even this), the literature (in which these results are all published) is clear: the programmes that the research teams have been implementing in rural Chinese middle schools are not working. The results of Column 1 clearly show that research teams working to identify ways to raise maths scores in rural China's middle school are *stuck in a rut*.

To verify the robustness of these estimates, we used a single, standardised model, the model specified in Equation (1), and re-estimated the impact of the nine interventions on maths test scores (Table 4–Column 2).[13] To be clear, in this exercise (when estimating the impacts of the interventions) the inclusion of control variables and use of fixed effects are identical in all nine regressions (as opposed to the specifications of the models used in producing the estimates of the coefficients in Column 1 – which varied somewhat from paper to paper). As seen in Column 2, the point estimates and standard errors are small and statistically insignificant – just like those from the literature (Column 1). In other words, using a single standard specification verifies that interventions are not improving academic achievement in rural middle schools.[14]

In the next part of the analysis, we re-organise our data and produce four subsamples. To increase statistical power, we combine the datasets of interventions that are similar. Specifically, we combine interventions P1 to P4 to create a single financial aid intervention; we combine P5 and P6 to create a single information/training intervention; we combine P7 and P8 to create a teacher intervention; and P9 remains its own category. When doing so, and running the model from Equation (2), we show that on average (and controlling for demographic characteristics and baseline achievement), none of the meta-intervention types improved achievement (Table 5, Panel A). All of the estimates are

Table 5. Intervention effects by intervention type and average treatment effect.

	(1)
	Endline Maths Score (SD)
Panel A: Effect of each intervention type	
Treatment 1-Financial aid (P1-P4), 1 = *yes*	−0.046
	(0.044)
Treatment 2-Information (P5-P6), 1 = *yes*	0.017
	(0.041)
Treatment 3-Teacher (P7-P8), 1 = *yes*	0.011
	(0.025)
Treatment 4-Glasses (P9), 1 = *yes*	0.106
	(0.131)
Student characteristics	Yes
Family characteristics	Yes
County dummies	Yes
Observations	47,480
Panel B: Average treatment effect	
Pooled treatment, 1 = *yes*	0.011
	(0.021)
Student characteristics	Yes
Family characteristics	Yes
County dummies	Yes
Observations	47,480

Note. Panel A uses equation (2) displayed in Section 3.4 and Panel B uses equation (1). Student characteristics include their gender, age and baseline maths score. Family characteristics include the migration status and education of the parents. Robust standard errors are displayed in parentheses and are clustered at school level.

* Significant at 10%. **Significant at 5%. *** Significant at 1%.

small––none, with the exception of the coefficient on the glasses intervention, is larger than 0.05 SDs. None, including the coefficient on the glasses intervention treatment variable (now statistically insignificant), is statistically significant.

To further increase precision, we use the entire pooled sample (47,480) and create a single intervention to assess if there is an overall average effect for being in *any* treatment arm. The results in Table 5, Panel B show that students in the treatment arms did not, on average, improve their scores.

Finally, we consider whether this average zero effect is masking any heterogeneous treatment effects according to student's baseline achievement, gender, or age. Table 6 (all rows; Columns 1 to 4) shows that the four pooled interventions did not have different effects on students with different baseline achievement, genders, or ages. When we consider this heterogeneity for each intervention category, we similarly find no evidence of treatment effects. One exception emerges with student age in the teacher-based interventions, where the youngest one fifth of the students scored higher in maths. However, the effect size is small (0.06 SD). And, given that we estimated 48 coefficients in the four columns of the table, by chance, one would think that several of the coefficients would turn up significant, even if the true underlying distribution was centred in zero. Similarly, Column 5 shows that there were no heterogeneous impacts of being in any treatment arm when examining the entire pooled sample (Table 6).

4.1 Is the absence of success due to poor programme design/implementation?

One explanation of these results is that the interventions were not designed or implemented correctly. However, as seen in Appendix Table A1, the research teams involved in each study took measures to ensure the intervention was implemented with a high degree of fidelity. Moreover, two of the interventions examined here, P8 (Teacher Incentives) and P9 (Free Glasses), were also implemented and evaluated in rural elementary schools. The interventions and the research teams

Table 6. Heterogeneous effect on students with different characteristics.

	(1)	(2)	(3)	(4)	(5)
	Endline Maths Score (SD)				
	Financial aid	Information	Teacher	Glasses	Pooled sample
Panel A: Baseline Maths test score					
1/3 top * treatment	−0.016	−0.073	−0.007	0.147	0.006
	(0.088)	(0.058)	(0.027)	(0.159)	(0.029)
Treatment on middle and bottom	−0.019	0.037	0.014	0.040	0.009
	(0.051)	(0.044)	(0.024)	(0.132)	(0.024)
Treatment on 1/3 top	−0.035	−0.036	0.007	0.187	0.015
	(0.071)	(0.056)	(0.027)	(0.162)	(0.027)
Panel B: Gender					
Male * treatment	−0.032	−0.012	−0.005	0.101	−0.018
	(0.078)	(0.034)	(0.023)	(0.132)	(0.019)
Treatment on male	−0.039	0.007	0.008	0.159	0.002
	(0.057)	(0.045)	(0.025)	(0.148)	(0.024)
Treatment on female	−0.007	0.019	0.013	0.057	0.020
	(0.056)	(0.040)	(0.024)	(0.131)	(0.022)
Panel C: Age					
1/5 oldest * treatment	−0.036	−0.025	0.057**	0.047	0.012
	(0.095)	(0.045)	(0.028)	(0.162)	(0.026)
Treatment on 4/5 youngest	−0.018	0.020	0.003	0.094	0.009
	(0.048)	(0.036)	(0.022)	(0.134)	(0.021)
Treatment on 1/5 oldest	−0.054	−0.005	0.061*	0.141	0.021
	(0.080)	(0.061)	(0.032)	(0.145)	(0.034)
Observations	2,245	17,702	26,933	600	47,480

Note. Regression contains county fixed effect.
Robust standard errors are displayed in parentheses and are clustered at the school level.
* Significant at 10%. ** Significant at 5%. *** Significant at 1%.

implementing them were the same, and the programmes had large and significant impacts on academic outcomes at the elementary level (; Ma et al. 2014).[15] In addition, many other interventions in rural elementary schools (which were not duplicated at the middle school level) – such as computer-assisted learning (Mo et al. 2015); improving nutrition (Kleiman-Weiner et al. 2013; Luo et al. 2012); reading programmes and book corners (Gao et al. 2017) – also had significant impacts on achievement. In other words, although – according to the literature – many interventions succeed in raising academic achievement in rural elementary schools in China, the results in Tables 4 and 5 suggest that almost nothing is working in middle schools.

There is one programme in which the content and/or implementation of the interventions, in fact, may have been responsible (at least in part) for the absence of an impact. In the case of the teacher training intervention, trainees found the training to be overly theoretical in content and rote in delivery (Loyalka et al. Forthcoming). The absence of positive effects may have had less to do with the context of rural middle schools than with the programme itself.

5. Discussion

In this section, we will attempt to address the issue of why interventions in middle schools in rural China are not working to improve student achievement. The final subsection of the previous section tried to show that ineffectiveness was, by in large, not due to the poor design or implementation of the interventions. The results of this study are also consistent with other post-primary intervention studies internationally. A review of post-primary school intervention studies by J-PAL, for example, states that evaluations of programmes and interventions in junior high school have shown much less success than interventions in primary school, and that in the cases where positive effects are observed, these effects tend to fade quickly over time (J-PAL 2013). Despite this consistency with the international literature, our findings lead us to the question of why interventions in rural middle schools were unable to change the behaviour or increase effort of students or educators.

This section addresses this question by presenting a discussion of possible hypotheses behind this failure. To the extent that the data allow, we examine four hypotheses, each pointing to a different explanation.[16] Namely, we examine: teacher quality, teacher incentives, student anxiety, and the competitive fast-paced environment that characterises middle school in China.

5.1 Teacher quality

One reason why our interventions may not have worked is that teachers are not qualified enough to teach in rural middle schools. In other words, teachers may be expending effort in the classrooms (and they may try to implement the interventions) but the effort that they put out is ineffective because the average quality of teachers is too low. Because of this, their students do not make achievement gains.

Studies have consistently shown that differences in the quality of teachers do matter in being able to improve student achievement (Chetty, Friedman, and Rockoff 2014; Rockoff 2004). As described by Hanushek (2011), two identical students placed in two different classes can perform vastly differently in a year's time due solely to the teachers they are exposed to. Hence, the literature is clear that if teacher quality is too low, students may not be able to make achievement gains.

So how does the literature measure quality? The main approach is to identify attributes of teachers that have been shown to contribute to better teaching quality and measure the extent of (or the prevalence of) these attributes. In most developing countries, these attributes include teaching credentials (Rice 2003; Clotfelter, Ladd, and Vigdor 2010) experience (Ladd and Sorensen

2017), and teacher knowledge of subject content (Metzler and Woessmann 2012). These are all considered meaningful determinants of teacher quality because when teachers rank highly in terms of these attributes, their students have been shown to perform better.

In China, the literature has identified a similar set of attributes that are correlated with teaching quality. Adams (2012) found that, in rural Gansu, students who are taught by teachers who have accumulated three to five years of experience are usually the highest achieving. The educational background of teachers is also associated with teaching quality in China (Park and Hannum 2001). In addition to these two attributes, the literature has pointed to a more particular feature of Chinese education pertaining to the teacher's rank. Teachers in China are subject to a system of promotions where they advance through five ranks depending on their performance (with standards of performance enumerated within the ranking protocol). Teacher rank/credentials have been shown to be a rather important determinant in student academic performance (Park and Hannum 2001; Chu et al. 2015). Therefore, if teachers in our sample do not have teaching experience, are not credentialed, and/or have poor educational backgrounds, it may mean that their teaching ability is low, and it may be that poor-quality teaching is undermining the effectiveness of the interventions and preventing gains in student achievement.

We find that by all measures mentioned above (experience/educational background/rank or credentialing), teachers in rural China are of good quality (Table 7). On average, a middle school teacher has about 13.6 years of experience behind her and 69% have completed at least a junior college degree, which is roughly six times more than in the general population (China Statistical Yearbook 2015). About 45% of the teachers in our data attained either the first or highest rank in the credentialing system and nearly 40% majored in maths (when they were in college).[17] As such, to the extent that these three attributes are measuring the teacher quality, it is unlikely that poor teacher quality is impeding impacts of the rural middle school interventions.

5.2 Teacher incentives in rural China middle schools

Another reason why interventions did not work may be that teachers have misaligned incentives (Levačić 2009). In other words, if the incentives of teachers in rural China's middle schools are not aligned with the goal of raising academic achievement, it may be that teachers have little or no reason to perform well or help their students learn.

The byproducts of the absence of teacher motivation have been documented in the education literature developing countries. High teacher absenteeism is one (Bold et al. 2017). Low teaching effort is another. Such phenomena contribute to what some have called a 'learning crisis' where students are in school but are not learning and so their academic performance is stagnant (Bold et al. 2017).

To overcome this problem of misaligned incentives, the literature shows that teachers need to have a contract (or promotion process) that is at least in part based on student outcomes and/or the academic achievement of their class. For example, in India, a randomised evaluation that tied teacher pay to student achievement gains improved achievement by as much as 0.27 SDs (Muralidharan and

Table 7. Descriptive statistics of teacher quality.

	(1)	(2)
	Mean	SD
Teaching experience, years	13.585	8.748
Teacher have first or highest rank	0.449	0.498
Teacher completed junior college degree (*dazhuan*)	0.685	0.465
Major match	0.399	0.490

Note. Data obtained from P7 and P8 and contain 810 observations.
Major match is a dummy that equals 1 if the teacher teaches the same subject that was his major.

Sundararaman 2011). Another evaluation by Duflo, Hanna, and Ryan (2012) tied financial rewards to teacher attendance, causing teachers to reduce their absences and students to improve their achievement. Other evidence is summarised in Levačić (2009).

Although the existing teacher contracting system in China's education may not be optimal, it acknowledges the importance of incentives and, to a large extent, provides them. It does so through a clear system of teacher promotion that is tied to the teacher's rank mentioned earlier (Ministry of Human Resources and Social Security 2015). According to the protocol used in almost all schools in rural China, a teacher starts with no ranking and advances towards third rank, second rank, first rank, and finally the highest rank. Theoretically, and according to government policy, this ladder of seniority is supposed to be the largest determinant of pay.

To verify this is true in our sample, we use detailed information on teachers from a subset of our data (Table 8).[18] Using these data, we try to find the correlation between the professional ranks of teachers and their salaries while controlling for their experience and education. According to the analysis, whereas certain characteristics of the teachers are unrelated to the level of pay of the teachers, higher ranks are strongly correlated with higher salaries. Indeed, the findings show that the correlation estimates exhibit higher precision for higher ranks. Beyond the findings in our own dataset, other studies in the literature on China have examined this relationship and found consistent results (Ding and Lehrer 2001; Karachiwalla and Park 2017). The findings in these papers clearly show the extent to which teacher rank matters; a salary rise associated with a promotion from no rank to third rank is higher than the increase associated with gaining twenty years of experience (Ding and Lehrer 2001).

An educational system that merely ties pay level to rank promotions, however, does not guarantee that teachers will be responsive to or incentivised by it. One reason is that any bureaucratic system as large as the educational system in China, is subject to corruption. Even if the system does not display high levels of corruption, teachers may still be unresponsive to incentives if they believe it is corrupt. However, both of these propositions were shown to be false in a study by Karachiwalla and Park (2017). Evaluation scores used for promotions, although not solely based on student outcomes, were shown to be correlated with them. The amount of time that teachers spend teaching (one measure of effort) was also positively correlated with promotion. In other words, the Karachiwalla and Park (2017) study demonstrated that teachers who spend more time on work

Table 8. How are teachers' wages and their professional ranking related?.

	(1)	(2)	(3)
	Log of monthly wage		
Third rank	0.049	0.047	0.042
	(0.068)	(0.067)	(0.067)
Second rank	0.110***	0.095**	0.092**
	(0.039)	(0.043)	(0.043)
First rank	0.272***	0.212***	0.210***
	(0.042)	(0.049)	(0.049)
Highest rank	0.486***	0.367***	0.365***
	(0.041)	(0.051)	(0.051)
Experience		0.000	0.000
		(0.003)	(0.004)
Experience-squared		0.000**	0.000*
		(0.000)	(0.000)
Teacher's education characteristics	Yes	Yes	Yes
County fixed effects	Yes	Yes	Yes
R-squared	0.774	0.784	0.785
Observations	810	810	810

Note. Data obtained from P-7 and P-8: Teacher training and teacher incentives.
Teacher's education characteristics include their high school type and college type.
Cluster-robust standard errors adjusted for clustering at the school level in parentheses.
*Significant at 10%. **Significant at 5%. *** Significant at 1%.

and improve student achievement are paid more. This study also showed that teachers are largely responsive to the system. In the years leading up to promotion eligibility, teachers increase their effort (Karachiwalla and Park 2017). Taken together, these points suggest that the system in place is largely successful in incentivising teachers.[19]

Is it the case, then, that the absence of further incentives for teachers is one reason for the inability of the interventions to produce improved student achievement? In fact, it might be the case that the current system is already working so well that Chinese teachers are already fully incentivised and are already exerting maximum effort (and so when they are asked to do something in addition to their current workload, they are simply unable to respond). This explanation is considered by the authors of the original paper written about the teacher incentive programme (P8; Loyalka et al. 2018). In this programme, teachers were offered a large additional financial reward (up to 2 months of salary) if they improved the achievement of their students. However, as the results in Tables 4 to 6 show, the incentive intervention did not improve student achievement. Loyalka et al. (2018) find that one reason student achievement did not improve was that treatment teachers did not increase their effort or change their teaching behaviour. The authors thus propose that the absence of response from teachers may have been due to the fact that the current incentive system that teachers faced was more powerful (or was already eliciting maximum effort) than the one that the intervention offered them.

5.3 Student anxiety

Beyond the absence of responsiveness from teachers, an alternative explanation for the absence of intervention effects in rural middle schools may be that students are being prevented from learning by psychological barriers. For example, students may be experiencing a high degree of mental anxiety that is serious enough to dampen the impacts of the interventions in raising academic achievement.

In fact, the literature on rural China has consistently documented both high rates of anxiety among rural China's youth and the negative association between anxiety and schooling––especially at the middle school level. According to the dataset (which is used in Liu, Shi, and Rozelle 2017), 7% of the study students exhibit symptoms of overall anxiety. When examining specific types of anxiety, however, 54% of students exhibit serious symptoms of at least one type of anxiety. The work of M. Zhou et al. (2018), using an alternative dataset, showed similarly high rates of mental health issues among students. Moreover, these mental health issues can be limiting. Hesketh and Ding (2005), for example, found that the symptoms of anxiety among Chinese students are sufficient to interfere with enjoyment of life, relaxation, and sleep.

Of even more relevance to this discussion of why interventions to improve achievement are not doing so, the literature––on rural China and elsewhere––has confirmed the idea that mental health conditions can be negatively associated with schooling (Currie 2009). Currie and Stabile (2006), for example, showed that in developed countries, mental health conditions can be more detrimental to school attainment than physical conditions. Similarly, in rural China, H. Wang et al. (2016) showed that high levels of student anxiety may be pushing students to eventually drop out of school. Thus, given the high levels of anxiety among rural Chinese students and its ability to hamper academic success, it is possible that student anxiety might be behind the inability of the interventions to increase achievement in our sample middle schools.

Given this background, we examined whether the middle school interventions have heterogeneous effects on students with different levels of anxiety. Specifically, we examined the social-emotional learning programme (P6) in H. Wang et al. (2016) for which mental health (anxiety assessment) data were collected (Table 9). We used two variables to indicate the anxiety levels of the students: The Mental Health Test (MHT) score (a continuous variable) and the Learning Anxiety Index (LAI) score (a dummy variable).[20] The idea of the analysis is that if anxiety is one of the barriers to achievement gains, then, if one were to divide the students

Table 9. Heterogeneous effects with respect to student anxiety.

	(1)
	Endline Maths Score (SD)
Panel A: MHT score	
1/3 top * treatment	−0.021
	(0.066)
Treatment on middle and bottom	0.037
	(0.100)
Treatment on 1/3 top	0.015
	(0.118)
Panel B: Student learning anxiety	
Learning anxiety * treatment	−0.065
	(0.059)
Treatment on non-anxiety student	0.070
	(0.102)
Treatment on anxiety student	0.003
	(0.108)
Observations	5,958

Note. Data obtained from P-6. Regression contains county fixed effect.
Cluster-robust standard errors adjusted for clustering at the school level in parentheses.
*Significant at 10%. **Significant at 5%. *** Significant at 1%.

into two groups – those with high anxiety scores and those without, the interventions might work on the latter group (those without anxiety) but not in the former (those showing symptoms of anxiety).

According to the analysis (Table 9, Panel A), there is no evidence that anxiety is the reason for the absence of achievement gains from the interventions. Specifically, the coefficient for the interaction term (representing the anxiety penalty/premium) is statistically insignificant. In other words, this finding suggests that there are no heterogeneous treatment effects among students with different levels (terciles) of MHT scores. Similarly, the coefficient for the interaction term in Panel B indicates that the treatment did not have heterogeneous effects for students with different LAI scores. Hence, the results of this analysis (although only run on the sample of one of our interventions – because the data collection teams in the other interventions did not collect information on student anxiety) suggest that the absence of measurable impact is not being masked by heterogeneity along anxiety lines.

5.4 Inappropriate curriculum

In our search for reasons that hamper interventions in middle schools, we now consider the nature of China's mandated curriculum. International research in education has documented that overambitious curricula, which are prevalent in some developing countries, can limit student learning rather than increase it (Pritchett and Beatty 2012; Siddaiah-Subramanya, Smith, and Lonie 2017). This paradoxical effect can take place in multiple ways. For example, if student baseline achievement levels do not match what the curriculum requires, students find it difficult to catch up and eventually fall behind (Siddaiah-Subramanya, Smith, and Lonie 2017). Similarly, if the curricular pace exceeds the learning pace, students fall behind (Pritchett and Beatty 2012). This subsection examines whether these two features––mismatch of presumed and actual achievement levels and inappropriate pace––characterise the curriculum in rural Chinese middle schools and whether such a curriculum may be responsible for the failure of interventions.

Before answering this question, to understand the full impact of the middle school curriculum, we believe one also needs to consider the context of the school system within which the curriculum is being taught. In rural China – especially at the middle school level – the education system is

extremely competitive. One of the main sources of the competitiveness is that, although there is a high demand to go to academic high school (because this is the route to college, which today has a high return; Li et al. 2017b), places in academic high school are limited. All rural counties are different, but it has often been the case that the number of places in academic high school per year are less than half the number of graduating middle school students (Hansen and Woronov 2013; Woronov 2016).[21] How is it decided who goes and who does not? For almost all rural students, there is one and only one criterion: they must score sufficiently high on the admission exam. Because of this, for students who have the desire to go to college, they must work very hard to earn a competitive score on the high school admission exam to secure a position in academic high school.

It is therefore unsurprising that China's curriculum content might be overambitious in scope and speed – at least for a subset of students. In nearly all middle schools in China, including rural China, the curriculum is primarily focused on teaching the students the materials that they will need to pass the high school admission exam (F. Liu 2004). Moreover, because all students in a prefecture take the same exam, including students in highest performing urban areas who often enjoy better resources and are competing to be able to test into key high schools, the curriculum is designed to be taught at an extremely high and fast-paced level (F. Liu 2004).

Therefore, in some sense, we believe that to understand the real problem involved with the nature of rural China's middle school curriculum, one might imagine (in admittedly simplified terms) that there are two types of students in middle school. There are students that are capable of learning the curriculum at its high level and fact pace. And, there are students who struggle with the materials and keeping up with the pace. As mentioned above, in the case of the students capable of learning the curriculum, there is a huge incentive to work as hard as they can. The higher the score, the better the high school, and the higher the chance there will ultimately be to attend a good college/university.

There is thus some reason to believe that in the same way that teachers are already working so hard that there is little scope for improving teaching, this may also be true in the case of better-performing students. Earlier in this section, we have shown that teachers already face strong incentives to teach in ways that will produce higher academic outcomes for their students. If the incentives are strong enough, it may be that teachers are already doing everything they can and so when there is an intervention designed to improve academic achievement, there is little room for response. The high incentives already in place for high-achieving students in the form of admission to quality high schools may create the same effect in the case of these students; the better-performing students may already be working so hard that there is little scope for improving achievement.

But what is happening to the other half of students? To examine this question, let us return to the question of whether Chinese middle schools suffer from an overambitious curriculum – characterised by mismatch of presumed and actual achievement levels and inappropriate pace – and think about how this might affect the lower-achieving subset of students. Evidence from qualitative studies shows that the curriculum in rural China's middle schools often assumes prior knowledge that many rural students do not have. For example, the middle school English curriculum assumes that instruction in English started in grade three (Lou 2011; Yiu 2017). Although this is adhered to by urban schools––indeed, many urban parents have their children start learning English in preschool, the scarcity of English teachers in remote rural areas frequently forces schools to start the curriculum a year or two later, causing students to fall behind from the very beginning (Hu 2002; Lou 2011). But, the curriculum is uniform and is focused on the level of English of the top students in the school system. The same is true for maths and Chinese language. In many rural middle schools, the gaps between urban and the better rural students are so wide that in their efforts to narrow them that the school day is often prolonged to over 14 hours as part of efforts to shrink the gap (D. Wang 2011).

The burden of catching up is also exacerbated by the fast pace and rigidity of the curriculum. Each week teachers are required to cover a specific amount of material, following a strict, mandated timeline (D. Wang 2011). Subsequently, the curriculum enforces a trade-off between adhering to the

mandated timeline and ensuring that students are learning the new material at a pace appropriate to their learning style. Many factors may push teachers to forego the latter goal. School principals, for example, pressure teachers to meet time requirements (D. Wang 2011). In fact, some of the greatest pressure to keep on track actually comes from the parents of the top students in class (Law 2014). They want their children to be able to compete with the best students in the prefecture on the high school entrance exam. Moreover, school district-wide, county-wide and prefecture-wide exams are administered frequently and the exam content depends to a large extent on what the timeline dictates (Law 2014). This practice forces teachers to follow the timeline so that they are not giving students exams on topics that have not been covered. But in following the timeline, D. Wang (2011) documented that some teachers gear their instruction towards higher-achieving students so that at least *some* of their students receive good grades. Other symptoms of the time pressure associated with the curriculum appeared in the teacher training programme (P8) which we examined earlier. Qualitative feedback from teachers pointed to the time pressure and the need to follow the curriculum as barriers towards implementing the techniques they had learned.

Given all these factors, consider the scenario for a below average or even average-performing middle school student. She has to rise up to curricular expectations that are suited to urban students/top rural students, who have a much stronger background, better preparation, and more resources available to them. The below-average student has to keep up with a fast pace of learning while possibly being ignored by her teachers. On top of all of that, she has to prepare for the high school entrance exam which is highly competitive, allowing only a small number of students from poor rural areas to matriculate into the top academic high schools (Loyalka et al. 2017). As such, the student may foresee (or even be told by her teacher) that her odds of passing are low. In addition, because there are opportunities for low-skill workers to find jobs without going to academic high school and/or college, she may just decide that they need not keep up with the curriculum. As a rational decision-maker, the student may then decide not to expend effort to overcome the challenging curriculum.

Although an overambitious curriculum is detrimental in any population, it is possible that the poorer performing rural Chinese students are particularly vulnerable and it may be not only that they choose not to learn, it may be that they are unable to learn. According to recent findings, infants and toddlers growing up in rural China are often exposed to multiple risk factors that are associated with compromised cognitive development. Specifically, it has been shown that around half of rural infants suffer from iron-deficiency anaemia (Luo et al. 2017). Moreover, parents and caregivers in rural China frequently do not provide the necessary cognitive stimulation to, and interaction with, their infants/toddlers, potentially contributing to the cognitive delays prevalent among 48% of children under three (Yue et al. 2017). These disadvantages in cognition and health matter because the first few years of life, due to rapid brain development and brain malleability characteristic of these early years, comprise a critical developmental period that has implications for lifelong outcomes (Center on the Developing Child at Harvard University 2010). If half of rural Chinese children are developmentally delayed when they were young children, there is a high likelihood that a significant share of them will continue to suffer from low levels of cognition and language skills when they are middle school students. If this is the case, then it would be difficult for such children to learn certain parts of the curriculum under any circumstance. If it were taught at a blisteringly fast pace, it would be nearly impossible.

To have a better understanding of the capacity of China's rural middle school students to learn, we used two IQ tests that research teams gave a sample of students in two provinces. Specifically, we utilised data from a sample of 2,525 students in P8 that provided teachers with incentives and gave students an IQ test using the Raven's Standard Progressive Matrices. In addition, another subset of students (also from P8; 480 sample students in total), were administered an IQ test in the form of the Wechsler Intelligence Scale for Children (WISC).

To examine our hypothesis that many students in rural China are not able to keep up with the ambitious, fast-paced curriculum, we first examine the prevalence of students that have cognitive deficiencies. For the 2,525 students for which Raven IQ data are available, there appear to be a large number of students with relatively low IQs. In a healthy Chinese population, the average Raven IQ is 100 (Lynn and Cheng 2013). In our sample, the original average stands at 93.5. The difference is statistically significant at the 1% level ($p < 0.001$). An even larger deviation can be seen in the WISC IQ test results, where the healthy average is again 100 and the sample average is 88.2. The difference is also statistically significant at the 1% level ($p < 0.001$).

Worth noting, however, is that the Raven deviation may be conservative. According to the Flynn Effect documented by James Flynn (Flynn 1984, 1987, 1998), there has been a substantial and long-sustained increase in intelligence test scores measured in many parts of the world. In China, between 1986 and 2012, this increase has been equivalent to 6.19 points on the WISC test (J. Liu and Lynn 2013). Whereas the WISC test has been updated to reflect the new norm, the Raven has not.[22] Thus, if we were to use the Flynn effect-adjusted scores on our Raven data (given that both tests are standardised to have an average of 100 and a standard deviation of 15), we would add 6.19 points to the healthy average and our sample average would be approximately 12.7 points lower.[23] Because we do not know the exact magnitude for the Flynn effect for the Raven IQ test, this result can only be suggestive. Yet, whether adjusted (as in the case of WISC or our adjusted Raven) or not (as in the case of unadjusted Raven), the results show that the average IQ in our sample is relatively low (Chen et al. 2010).

Further comparisons can be made using the overall IQ distribution, rather than just the average. Consider Figure 1, Panel A below, which plots the distribution of Raven IQ scores in our sample and a normal distribution which describes IQ scores in a healthy Chinese population––with the Flynn adjustment.[24] As seen from the figure, the distribution for the rural middle school students is skewed to the left. Whereas 16% of a healthy Chinese population would be expected to score below 91 (one SD below the Flynn-adjusted 106 mean), 37% of our subsample scored below this cut-off. Results using WISC consistently show that the IQ of the sample students is substantially lower than that in a healthy population and that the percentage of students whose WISC IQ is one SD below the healthy WISC average is 40% (Figure 1, Panel B).

Is this overrepresentation of cognitive deficiencies associated with achievement? Table 10 shows the result of a correlation exercise in which we examine the association between maths scores and IQ. According to the results, the two variables are strongly and significantly correlated: controlling for student and family characteristics and class fixed effects, we find that on average, a one SD fall in IQ is associated with a 0.38 SD fall in academic scores. Even when we account for the non-cognitive traits of the rural students, the analysis demonstrates that the interventions are not working in rural middle schools to improve achievement (that is, even when we hold the level of student grit constant in the

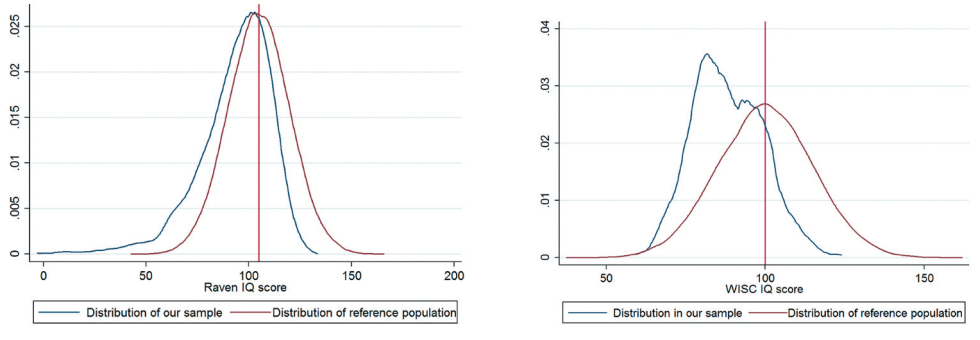

Figure 1. Panel A is for raven IQ distribution in a subsample of 2,525 students and an approximate distribution of the Chinese population with the Flynn-adjusted raven scores. panel B is for WISC IQ distribution in a subsample of 480 students and an approximate distribution of the Chinese population.

Table 10 Ordinary Least Squares Analysis of the Correlation between Raven IQ Score and Math Test Score

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: Baseline standardized math score					
	Whole sample		Normal IQ and above		Low IQ	
Z-score of Raven IQ	0.380***	0.379***	0.541***	0.539***	0.235***	0.235***
	(0.019)	(0.019)	(0.046)	(0.046)	(0.033)	(0.033)
Z-score of Grit		0.058***		0.064**		0.018
		(0.020)		(0.025)		(0.035)
Student characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Family characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Class fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.387	0.389	0.338	0.341	0.349	0.349
Observations	2,525	2,525	1,584	1,584	941	941

Notes. IQ scores correspond to the Raven's Standard Progressive Matrices test. Normal IQ and above is defined as IQ ≥ 85 which is equivalent to one SD below the mean or higher. Low IQ is defined as IQ < 85.
* Significant at 10%. ** Significant at 5%. *** Significant at 1%.

Table 11. Ordinary least squares analysis of the correlation between WISC IQ score and maths test score.

	(1)	(2)	(3)
	Dependent variable: Baseline standardised maths score		
	Whole sample	Normal IQ and above	Low IQ
Z-score of WISC IQ	0.506***	0.486***	1.030**
	(0.057)	(0.138)	(0.415)
Student characteristics	Yes	Yes	Yes
Family characteristics	Yes	Yes	Yes
Class fixed effect	Yes	Yes	Yes
R-squared	0.532	0.638	0.671
Observations	480	287	193

Note. IQ scores correspond to the Wechsler Intelligence Scale for Children (WISC) test.
Normal IQ and above is defined as IQ ≥ 85 which is equivalent to one SD below the mean or higher. Low IQ is defined as IQ < 85.
* Significant at 10%. ** Significant at 5%. *** Significant at 1%.

correlation analysis[25]), we find the correlation between IQ and maths test scores is equally positive and strong in both magnitude and precision (Column 2). In contrast, of course, students that have normal IQs are learning much higher levels of maths.[26] Table 11 shows that when using WISC IQ as opposed to Raven IQ, the same results hold: IQ and achievement are strongly and positively correlated.

5.5 Summary

In summary, we find that, among our hypotheses, the most compelling explanation for the inability of our interventions to have an impact is that performance among middle schoolers is confined by (a) the ambitious middle school curriculum; and (b) the competitive nature of the education's system with its focus on the high school admission exam as the only way to get one of the limited spaces in academic high school. On the one hand, in an educational system as competitive as China's, students who are able to make achievement gains are already highly incentivised to work as hard as they can – even without any of the interventions that attempted to improve achievement. Hence, even though some of interventions were designed to improve achievement, there was really not much room for improvement. On the other hand, for the other half of the students – those that were

cognitively delayed when they were young and, hence, were cognitively delayed in middle school, the curriculum may be hampering achievement gains. Students may already be at their maximum rate of learning or may not be making gains because the curriculum is taught much faster than they can digest it, rendering inefficient the set of interventions that were implemented in their schools. In short, for students who display cognitive deficiencies, it may be a matter of curricula and instruction being faster than appropriate.

6. Conclusion

In recent years, education experts have investigated a key question: which policies improve student achievement in resource-constrained settings? The answer to this question is particularly relevant to China, because returns to education have been rising and because prospects of economic growth rest on the ability to build a high-skilled labour force. At the same time, the academic achievement levels of many students in rural middle schools are low, leading them to forgo further schooling despite the high returns that individuals can get from advancing into academic high school and having a chance to pursue further education in college. In this study, we aim to answer this policy question by conducting a synthesis of recent, large-scale field experiments targeting rural middle schools in China. More specifically, we synthesised results from 11 randomised interventions that targeted different problems related to students' academic outcomes: liquidity constraints, lack of information on educational returns, teacher-related problems, and vision problems.

Out of the 11 interventions, we confirm that none of them were able to improve student achievement. One interpretation of this is that student achievement in rural Chinese middle school is not susceptible to simple policy changes. In the vernacular: It is stuck in a rut. Moreover, we find that this lack of susceptibility is generalisable across a variety of student demographics. In other words, students of different gender, ages, and levels of achievement all did not benefit from the intervention. After exploring a few hypotheses, we find suggestive evidence that the nature of China's high school matriculation policy and middle school curriculum are the best candidates to explain the lack of achievement gains. When tracing the source of the problem to its roots, the literature suggests that there is a large share of rural infants/toddlers that are suffering from cognitive development delays in young rural populations and that are likely set back in their ability to learn.

Our study is not without limitations. The first pertains to a subtle distinction between student learning and academic achievement. In our study, we aim to ultimately assess what makes students learn, assuming that what they score on a test is a good measure of what they learn. This assumption, however, may be unwarranted if achievement scores represent test-taking ability or rote memorisation. We attempt to make these two concepts of learning and achievement less distinct by choosing achievement on a maths test as a proxy for learning because maths evaluation is less likely to reflect memorisation.

The second limitation concerns our conjecture regarding how compromised cognitive ability may amplify the challenges associated with the curriculum. We had argued that students with development delays may learn at a slower pace and, given the competitive nature of middle school, choose not to expend as much effort on improving achievement. Although a student's level of cognition and other social emotional skills have critical roles in predicting success (in academic achievement and otherwise), they are not the only determinants of success. Studies show that other personality traits like self-motivation and self-perception predict academic achievement (Borghans et al. 2011). In fact, personality traits are sometimes found to be more important than intelligence in predicting academic success and as such, it is possible that students with high motivation may still be able to overcome curriculum-related difficulties, even if their cognitive ability had been compromised (Duckworth and Seligman 2005). We acknowledge the importance of discerning which of these two (IQ or motivation) helps students do better and the repercussions that the answer would have

on policy-making. Nevertheless, even if personality traits such as motivation are more important, the Chinese curriculum would still only benefit a specific niche of highly-motivated students, rather than the entire student population.

Though our findings are specific to the rural Chinese context, they complement a larger body of empirical work on improving education in developing countries and among students from disadvantaged backgrounds (McEwan 2015). That being said, existing work targeting middle school students has, for the most part, focused on attendance rather than achievement, with the consequence being rather limited understanding of what makes students learn better (J-PAL 2013; Krishnaratne, White, and Carpenter 2013). By providing a synthesis of studies, we show that student achievement––an arguably more meaningful outcome than attendance––is difficult to influence. Moreover, by shedding light on a relatively older cohort – middle school students, that is – the current study adds to existing scholarly work (e.g. Heckman 2008) which finds that later-age remediation strategies may be less effective.

Notes

1. A share of the middle school graduates in these poor rural areas do go on to vocational high schools (Yi et al. 2013). Yet, the literature shows that students learn few, if any, academic and vocational skills in these schools (Yi et al. 2018; Loyalka et al. 2015).
2. One study examines two rural schools in the early 2000's and finds that the rate of passing is between 40 to 46% (F. Liu 2004).
3. This is currently changing as wage rates are rising and the nature of tasks are beginning to change (Li et al. 2017b).
4. Because of research design considerations, the sampling strategies for the teacher incentive and teacher training interventions were slightly different, but similar in spirit.
5. For the purpose of this study, the maths test score serves as the key outcome variable (and a measure of the academic performance at large). There are at least two reasons for this choice. First, employers often rely on maths ability in their hiring decisions––this is documented outside of China as well (Koedel and Tyhurst 2012)––and thus instruction and learning in maths carry a lot of weight for students and educators alike. If the interventions are successful in improving academic outcomes, we expect that this success would be most pronounced in a maths assessment. Second, in comparison to other subjects like languages and/or humanities, maths serves as a better metric. If we were to use a Chinese test or a history test, for example, we could be simply considering the intervention effects on rote academic preparation (i.e. memorisation). Maths, on the other hand, often goes beyond such memorisation and uses advanced cognitive abilities. In a way, then, it allows us to assess whether the interventions in question have meaningful effects on the student's learning. Moreover, as argued by Ashcraft and Krause (2007), a student's performance in languages and humanities is to a large degree a function of her household socioeconomic status (in addition to schooling), whereas maths must be taught systematically through schooling. Although randomisation precludes this last point from being a concern about the validity of our estimates, it is still important when it comes to choosing the appropriate metric to assess what works in the educational institution of middle school.
6. All of the tests used in the nine studies in this paper were created and validated according to a multi-stage process. First, maths test items were selected from China's standardised mathematics curricula for each grade (7, 8 and 9). Maths curricula guidelines from the central government remained consistent during the years 2009 to 2016. Moreover, although provinces/localities can use different textbooks, they are all based on the detailed central government guidelines. Second, the content validity of these test items were repeatedly checked by expert teachers from each grade and from multiple localities across the country. Finally, the psychometric properties of the tests (reliability, unidimensionality, differential item functioning, lack of ceiling or floor effects) were repeatedly validated by trained psychometricians.
7. For purposes of brevity, we invite interested readers to examine the balance within each programme in its corresponding study. Because balance in programme P9 for students who did not own glasses prior to the intervention (the sample we focus on in this study) is not presented in the original study, we only verify it here. The results – unreported here for brevity – show that students who did not own glasses are comparable across the treatment and control arms.
8. This difference at baseline is actually fairly large, especially (as will be discussed later in the manuscript) compared to the effect size that is found in the evaluation (comparing differences between the treatment and control groups at the endline). It is for this reason that it we point out (especially given that the reported impact is only significant at the 10 percent level) that there may be no measurable impact of the free glasses programme.

9. We define maternal (paternal) migration in all programmes as whether the student's mother (or father) has been living away from home during the semester in which the survey is conducted.
10. The notation here implies that the treatment assignment was randomised on an individual level. It should be noted, however, that this is only the case for the financial aid programmes.
11. The only exceptions are regressions that involve one or more financial aid programmes and exclude the rest. If this is the case, we use simple robust standard errors. If a regression includes both financial aid programmes and other programmes, we use standard errors clustered at the school level.
12. To be clear, the coefficients in Column 1 that show that there is no impact in programmes P1-P6 are reporting the same coefficients as are published (Mo et al. 2013; Yi et al. 2015; F. Li et al. 2017a; Loyalka et al. 2013; H. Wang et al. 2016). The other programmes P7-P9 are from working papers (Loyalka et al. Forthcoming; ; Nie et al. 2016). In other words, then, our results in Column 1 are like a meta-analysis (that is, an analysis based on the results reported in the literature) which shows that ten of the 11 different interventions that were rolled out as RCTs in rural China's middle schools did not have any impact on academic achievement.
13. The difference is with programmes P1-P4 and P9, where the standard errors are robust and not clustered on school.
14. The one possible exception remains for the free glasses programme, P9, where our estimate (0.196 SDs) is statistically significant and both the coefficient and standard errors are close in both Columns 1 and 2. However, the gap between the treatment and control groups is smaller than the gap at the baseline. Hence, it is possible that the nature of the randomisation and not the intervention is (at least in part) producing the weakly positive results (weakly as the coefficient is significant only at the 10% level).
15. As the discussion above implies, it could be that the free eyeglasses intervention worked in junior high school (unlike the 10 other interventions). However, as discussed above also, it could be that the measured impacts were due to the imperfect randomisation and the intervention had either a little or no effect. At the very least compared to the measured impact of the free eyeglasses intervention in rural primary schools, the junior high results are clearly much smaller in magnitude and weaker statistically.
16. Of course, besides these four explanations, there may be other reasons that we do not directly address in this section, but which may have some explanatory power. As discussed above, it may be that some of interventions were not fully focused on improving achievement, but instead were really trying to raise attendance. In addition, it may be that the high opportunity cost of labour means that it was difficult to keep a subset of the students focused on schooling as they were anxious to get out of school and begin to work in the off-farm labour market. There may be other reasons that we have not thought of.
17. Rankings, from best to worst, are called 'highest rank', 'first rank', 'second rank', 'third rank', and finally 'no ranking' (teachers who have not yet earned any ranking).
18. We use data from P7 (teacher training programme) and P8 (teacher incentives) in this part. By combining these two programmes, we have 810 sample teachers in total.
19. Perhaps as a result of this success, teachers in rural China are rarely absent, unlike their counterparts in neighbouring India (Rao, Cheng, and Narain 2003; Karachiwalla and Park 2017).
20. The MHT score is the most widely used scale to measure the anxiety status of grade school students in China, with a reliability of 0.84 to 0.88 and a retest reliability of 0.78 to 0.86 (Gan, Bi, and Ruan 2007; B. Zhou 1991; Yao et al. 2011). The LAI is a set of 15 questions from the MHT, with a score above 7 indicating the student being at risk for learning anxiety. We constructed a dummy variable that equals 1 for students with LAI scores that are over 7.
21. In fact, the government seeks to achieve a balance such that half of middle school students are admitted to academic high schools and half are admitted to vocational training high schools, a goal which Woronov (2016) argues is shaping the number of positions available for academic high schools.
22. The Chinese edition of the Raven test we use was standardised in 1989 and thus the norms do not take into consideration the Flynn effect (Zhang and Wang 1989). The WISC test, on the other hand, corresponds to its fourth edition, published in 2003 but standardised in China in 2008 (Chen et al. 2010).
23. Alternatively, the healthy average would be 106.19. However, the difference would still hold.
24. The standard deviation is assumed to be 15 in the reference distribution.
25. Grit entails working strenuously to overcome challenges, maintaining effort and interest over years despite failure, adversity, and plateaus in progress (Duckworth et al. 2007). Individuals with a high grit score characteristically do not swerve from their goals, even in the absence of positive feedback (McClelland 1985).
26. In Columns 3–6 of Table 10, we run the same correlation dividing the sample along lines of IQ; those who score lower than 85 are considered low-IQ students and those who score 85 or above are considered in the 'normal' range. Similar to the results in the whole sample, IQ is strongly associated with achievement in both groups. Worth noting, however, is that the coefficient for grit becomes statistically insignificant among the low-IQ group, suggesting that hard work is not associated with academic success when cognitive ability is compromised. Yet, this interpretation should be considered with caution given the loss of sample size––and hence statistical power––associated with moving from the whole sample to the low-IQ group.

Acknowledgments

This work was supported by the 111 Project (grant number B16031); the National Natural Science Foundation of China (grant numbers 71110107028, 71033003, 71273237 and 70803047); the National Social Science Foundation of China (grant number 15CJL005); Ford Foundation (grant number 1110-0113); Plan International; the International Initiative for Impact Evaluation (3ie, grant numbers PW3.06.CH.IE.HP, PW3.05.CH.IE, and OW2.208); and the Poverty and Economic Policy Research Network (PEP). We also received the generous support of local government officials and educators throughout China, especially those from Ningshan County. Support from Eric Hemel and Eric Xu also is gratefully acknowledged.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Ford Foundation [1110-0113]; Higher Education Discipline Innovation Project [B16031]; National Natural Science Foundation of China [71033003,711110107028,71273237]; National Social Science Foundation of China [15CJL005]; International Initiative for Impact Evaluation [OW2.208,PW3.05.CH.IE,PW3.06.CH.IE.HP].

References

Aaronson, D., L. Barrow, and W. Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25 (1): 95–135. doi:10.1086/508733.

Adams, J. (2012). "Identifying the Attributes of Effective Rural Teachers: Teacher Attributes and Mathematics Achievement among Rural Primary School Students in Northwest China". *Gansu Survey of Children and Families Papers*, January. Retrieved from https://repository.upenn.edu/gansu_papers/32

Ashcraft, M. H., and J. A. Krause. 2007. "Working Memory, Math Performance, and Math Anxiety." *Psychonomic Bulletin & Review* 14 (2): 243–248. doi:10.3758/BF03194059.

Bold, T., D. Filmer, G. Martin, E. Molina, C. Rockmore, B. Stacy, and W. Wane. 2017. "What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa." *Policy Research Working Paper;* (7956): World Bank. Retrieved from. http://hdl.handle.net/10986/25964

Borghans, L., B. H. Golsteyn, J. Heckman, and J. E. Humphries. 2011. "Identification Problems in Personality Psychology." *Personality and Individual Differences* 51 (3): 315–320.Special Issue on Personality and Economics.10.1016/j.paid.2011.03.029.

Brown, P. H., and A. Park. 2002. "Education and Poverty in Rural China." *Economics of Education Review* 21 (6): 523–541. doi:10.1016/S0272-7757(01)00040-1.

Carnoy, M., P. Loyalka, M. Dobryakova, R. Dossani, I. Froumin, K. Kuhns, J. Tilak, and R. Wang. 2013. *University Expansion in a Changing Global Economy: Triumph of the BRICs?* Palo Alto, CA: Stanford University Press. doi:10.11126/stanford/9780804786010.001.0001.

Center on the Developing Child at Harvard University. (2010). "The Foundations of Lifelong Health are Built in Early Childhood". Boston, MA.: Harvard University. Retrieved from http://www.developingchild.harvard.edu

Chen, H., T. Z. Keith, L. Weiss, J. Zhu, and Y. Li. 2010. "Testing for Multigroup Invariance of Second-Order WISC-IV Structure across China, Hong Kong, Macau, and Taiwan." *Personality and Individual Differences* 49 (7): 677–682. doi:10.1016/j.paid.2010.06.004.

Chetty, R., J. N. Friedman, and J. E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 1352–1370. doi:10.1257/aer.104.9.2633.

China Statistical Yearbook. (2015). "Population by Sex, Educational Attainment and Region". Stats.Gov.Cn. 2015. Retrieved from http://www.stats.gov.cn/tjsj/ndsj/2015/indexeh.htm

Chu, J. H., P. Loyalka, J. Chu, Q. Qu, Y. Shi, and G. Li. 2015. "The Impact of Teacher Credentials on Student Achievement in China." *China Economic Review* 36: 14–24. doi:10.1016/j.chieco.2015.08.006.

Clinton Foundation. (2009). "Innovative Curriculum for Rural Chinese Primary Schools". *Clinton Foundation. 2009*. Retrieved from https://www.clintonfoundation.org/clinton-global-initiative/commitments/innovative-curriculum-rural-chinese-primary-schools

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor. 2010. "Teacher Credentials and Student Achievement in High School a Cross-subject Analysis with Student Fixed Effects." *Journal of Human Resources* 45 (3): 655–681. doi:10.1353/jhr.2010.0023.

Cochran-Smith, M. 2005. "The New Teacher Education: For Better or for Worse?" *Educational Researcher* 34 (7): 3–17. doi:10.3102/0013189X034007003.

Congdon, N., Y. Wang, Y. Song, K. Choi, M. Zhang, Z. Zhou, . . . B. Wu. 2008. "Visual Disability, Visual Function, and Myopia among Rural Chinese Secondary School Children: The Xichang Pediatric Refractive Error Study (X-pres)—report 1." *Investigative Opthalmology & Visual Science* 49 (7): 2888. doi:10.1167/iovs.07-1160.

Cunha, F., and J. Heckman (2007)."The Technology of Skill Formation". *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w12840.ack

Currie, J. 2009. "Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development". *Journal of Economic Literature* 47 (1): 87–122. doi:10.1257/jel.47.1.87.

Currie, J., and M. Stabile. 2006. "Child Mental Health and Human Capital Accumulation: The Case of ADHD." *Journal of Health Economics* 25 (6): 1094–1118. doi:10.1016/j.jhealeco.2006.03.001.

Ding, W., and S. F. Lehrer (2001, November). "The Optimal Policy to Reward the Value Added by Educators: Theory and Evidence from China". In *International Conference on Education Reform in China*, Conference conducted at the meeting of Harvard University, Cambridge, Mass. Processed.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly. 2007. "Grit: Perseverance and Passion for Long-term Goals." *Journal of Personality and Social Psychology* 92 (6): 1087. doi:10.1037/0022-3514.92.6.1087.

Duckworth, A. L., and M. E. Seligman. 2005. "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents." *Psychological Science* 16 (12): 939–944. doi:10.1111/j.1467-9280.2005.01641.x.

Duflo, E., R. Hanna, and S. P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–1278. doi:10.1257/aer.102.4.1241.

Flynn, J. R. 1984. "The Mean IQ of Americans: Massive Gains 1932 to 1978." *Psychological Bulletin* 95 (1): 29. doi:10.1037/0033-2909.95.1.29.

Flynn, J. R. 1987. "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure." *Psychological Bulletin* 101 (2): 171. doi:10.1037/0033-2909.101.2.171.

Flynn, J. R. 1998. "WAIS–III and WISC–III IQ Gains in the United States from 1972 to 1995: How to Compensate for Obsolete Norms." *Perceptual and Motor Skills* 86 (3_suppl): 1231–1239. doi:10.2466/pms.1998.86.3c.1231.

Gan, T., Z. Bi, and K. Ruan. 2007. "A Review and Outlook of Mental Health Measure Tools Used for Middle School Students." *Chinese Journal of School Health* 28: 191–192.

Gao, Q., D. Mo, Y. Shi, H. Wang, K. Kenny, and S. Rozelle (2017). "Can Reading Programs Improve Reading Skills and Academic Performance in Rural China?" Retrieved from https://www.semanticscholar.org/paper/Can-Reading-Programs-Improve-Reading-Skills-and-in-Gao-Mo/af005fab028979f006e1b4d1462d4e629ba988de

Hansen, M. H., and T. E. Woronov. 2013. "Demanding and Resisting Vocational Education: A Comparative Study of Schools in Rural and Urban China." *Comparative Education* 49 (2): 242–259. doi:10.1080/03050068.2012.733848.

Hanushek, E. A. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30 (3): 466–479. doi:10.1016/j.econedurev.2010.12.006.

He, M., W. Huang, Y. Zheng, L. Huang, and L. B. Ellwein. 2007. "Refractive Error and Visual Impairment in School Children in Rural Southern China." *Ophthalmology* 114 (2): 374–382.e1. doi:10.1016/j.ophtha.2006.08.020.

Heckman, J. J. 2008. "Schools, Skills, and Synapses." *Economic Inquiry* 46 (3): 289–324. doi:10.1111/j.1465-7295.2008.00163.x.

Heckman, J. J., and X. Li. 2004. "Selection Bias, Comparative Advantage and Heterogeneous Returns to Education: Evidence from China in 2000." *Pacific Economic Review* 9 (3): 155–171. doi:10.1111/j.1468-0106.2004.00242.x.

Hesketh, T., and Q. J. Ding. 2005. "Anxiety and Depression in Adolescents in Urban and Rural China." *Psychological Reports* 96 (2): 435–444. doi:10.2466/pr0.96.2.435-444.

Hu, G. 2002. "English Language Teaching in the People's Republic of China." *English Language Education in China, Japan, and Singapore* 1–77.

Jensen, R. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *The Quarterly Journal of Economics* 125 (2): 515–548. doi:10.1162/qjec.2010.125.2.515.

Karachiwalla, N., and A. Park. 2017. "Promotion Incentives in the Public Sector: Evidence from Chinese Schools." *Journal of Public Economics* 146: 109–128. doi:10.1016/j.jpubeco.2016.12.004.

Khor, N., L. Pang, C. Liu, F. Chang, D. Mo, P. Loyalka, and S. Rozelle. 2016. "China's Looming Human Capital Crisis: Upper Secondary Educational Attainment Rates and the Middle-income Trap." *The China Quarterly* 228: 905–926. doi:10.1017/S0305741016001119.

Kleiman-Weiner, M., R. Luo, L. Zhang, Y. Shi, A. Medina, and S. Rozelle. 2013. "Eggs versus Chewable Vitamins: Which Intervention Can Increase Nutrition and Test Scores in Rural China?" *China Economic Review* 24: 165–176. doi:10.1016/j.chieco.2012.12.005.

Koedel, C., and E. Tyhurst. 2012. "Math Skills and Labor-market Outcomes: Evidence from a Resume-based Field Experiment." *Economics of Education Review* 31 (1): 131–140. doi:10.1016/j.econedurev.2011.09.006.

Krishnaratne, S., H. White, and E. Carpenter, (2013). "Quality Education for All Children? What Works in Education in Developing Countries, Working Paper 20". New Delhi: International Initiative for Impact Evaluation (3ie)

Ladd, H. F., and L. C. Sorensen. 2017. "Returns to Teacher Experience: Student Achievement and Motivation in Middle School." *Education Finance and Policy* 12 (2): 241–279. doi:10.1162/EDFP_a_00194.

Law, W. W. 2014. "Understanding China's Curriculum Reform for the 21st Century." *Journal of Curriculum Studies* 46 (3): 332–360. doi:10.1080/00220272.2014.883431.

Levačić, R. 2009. "Teacher Incentives and Performance: An Application of Principal–agent Theory." *Oxford Development Studies* 37 (1): 33–46. doi:10.1080/13600810802660844.
Li, H. 2003. "Economic Transition and Returns to Education in China." *Economics of Education Review* 22 (3): 317–328. doi:10.1016/S0272-7757(02)00056-0.
Li, H., P. Loyalka, S. Rozelle, and B. Wu. 2017b. "Human Capital and China's Future Growth." *Journal of Economic Perspectives* 31 (1): 25–48. doi:10.1257/jep.31.1.25.
Li, F., Y. Song, H. Yi, J. Wei, L. Zhang, Y. Shi, S. Rozelle, N. Johnson, P. Loyalka, and S. Rozelle. 2017a. "The Impact of Conditional Cash Transfers on the Matriculation of Junior High School Students into Rural China's High Schools." *Journal of Development Effectiveness* 9 (1): 41–60. doi:10.1080/19439342.2016.1231701.
Liu, F. 2004. "Basic Education in China's Rural Areas: A Legal Obligation or an Individual Choice?" *International Journal of Educational Development* 24 (1): 5–21. doi:10.1016/j.ijedudev.2003.09.001.
Liu, H., C. Liu, F. Chang, and P. Loyalka. 2016. "Implementation of Teacher Training in China and Its Policy Implications." *China & World Economy* 24 (3): 86–104. doi:10.1111/cwe.12160.
Liu, J., and R. Lynn. 2013. "An Increase of Intelligence in China 1986–2012." *Intelligence* 41 (5): 479–481. doi:10.1016/j.intell.2013.06.017.
Liu, H., Y. Shi, and S. Rozelle (2017). "Mental Health in Rural China: Comparisons across Provinces and among Subgroups of Children and Adolescents". Working Paper. REAP Working Papers. Stanford, CA.: Rural Education Action Progam.
Liu, C., L. Zhang, R. Luo, S. Rozelle, B. Sharbono, and Y. Shi. 2009. "Development Challenges, Tuition Barriers, and High School Education in China." *Asia Pacific Journal of Education* 29 (4): 503–520. doi:10.1080/02188790903312698.
Lou, J. 2011. "Suzhi, Relevance, and the New Curriculum: A Case Study of One Rural Middle School in Northwest China." *Chinese Education & Society* 44 (6): 73–86. doi:10.2753/CED1061-1932440605.
Loyalka, P., J. Chu, J. Wei, N. Johnson, and J. Reniker. 2017. "Inequalities in the Pathway to College in China: When Do Students from Poor Areas Fall Behind?" *The China Quarterly* 229: 172–194. doi:10.1017/S0305741016001594.
Loyalka, P., X. Huang, L. Zhang, J. Wei, H. Yi, Y. Song, and J. Chu (2015). "The Impact of Vocational Schooling on Human Capital Development in Developing Countries: Evidence from China". doi:10.1093/wber/lhv050.
Loyalka, P., C. Liu, Y. Song, H. Yi, X. Huang, J. Wei, S. Rozelle, Y. Shi, J. Chu, and S. Rozelle. 2013. "Can Information and Counseling Help Students from Poor Rural Areas Go to High School? Evidence from China." *Journal of Comparative Economics* 41 (4): 1012–1025. doi:10.1016/j.jce.2013.06.004.
Loyalka, P., A. Popova, G. Li, C. Liu, and H. Shi. Forthcoming. "Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program." *American Economic Journal. Applied Economics*.
Loyalka, P. K., S. Sylvia, C. Liu, J. Chu, and Y. Shi. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." *Journal of Labor Economics* 37 (3): 621–662. doi:10.1086/702625
Loyalka, P., S. Sylvia, C. Liu, Y. Shi, Y. Qian, et al. (2018). "The Distributional Impacts of Pay for Percentile: Evidence from a Randomized Trial in Rural China". Working paper. Stanford, CA.: Rural Education Action Program.
Luo, R., Y. Shi, L. Zhang, C. Liu, S. Rozelle, B. Sharbono, R. Martorell, Q. Zhao, and R. Martorell. 2012. "Nutrition and Educational Performance in Rural China's Elementary Schools: Results of a Randomized Control Trial in Shaanxi Province." *Economic Development and Cultural Change* 60 (4): 735–772. doi:10.1086/665606.
Luo, R., A. Yue, H. Zhou, Y. Shi, L. Zhang, R. Martorell, S. Sylvia, S. Rozelle, and S. Sylvia. 2017. "The Effect of a Micronutrient Powder Home Fortification Program on Anemia and Cognitive Outcomes among Young Children in Rural China: A Cluster Randomized Trial." *BMC Public Health* 17 (1): 738. doi:10.1186/s12889-017-4755-0.
Lynn, R., and H. Cheng. 2013. "Differences in Intelligence across Thirty-one Regions of China and Their Economic and Demographic Correlates." *Intelligence* 41 (5): 553–559. doi:10.1016/j.intell.2013.07.009.
Ma, Y., X. Ma, Y. Shi, N. Congdon, H. Yi, S. Knotb, A. Medina, and M. Iyer. (2018). "Visual Impairment in Rural China: Prevalence, Severity, and Association with Income across Student Cohorts". Working Paper. REAP Working Papers. Stanford, CA.: Rural Education Action Program.
Ma, X., Z. Zhou, H. Yi, X. Pang, Y. Shi, Q. Chen, and Y. Liu. 2014. "Effect of Providing Free Glasses on Children's Educational Outcomes in China: Cluster Randomized Controlled Trial." *Bmj* 349: g5740. doi:10.1136/bmj.g5740.
McClelland, D. C. 1985. "How Motives, Skills, and Values Determine What People Do." *American Psychologist* 40 (7): 812. doi:10.1037/0003-066X.40.7.812.
McEwan, P. J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-analysis of Randomized Experiments." *Review of Educational Research* 85 (3): 353–394. doi:10.3102/0034654314553127.
Metzler, J., and L. Woessmann. 2012. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-teacher Within-student Variation." *Journal of Development Economics* 99 (2): 486–496. doi:10.1016/j.jdeveco.2012.06.002.
Ministry of Human Resources and Social Security. (2015). "Primary and Secondary School Teacher Credentials Reforms". Mohrss.Gov.Cn. August 2015. Retrieved from http://www.mohrss.gov.cn/SYrlzyhshbzb/ldbk/rencaiduiwujianshe/zhuanyejishurenyuan/201509/t20150902_219575.htm
Mo, D., W. Huang, Y. Shi, L. Zhang, M. Boswell, and S. Rozelle. 2015. "Computer Technology in Education: Evidence from a Pooled Study of Computer Assisted Learning Programs among Rural Students in China." *China Economic Review* 36: 131–145. doi:10.1016/j.chieco.2015.09.001.

Mo, D., L. Zhang, H. Yi, R. Luo, S. Rozelle, and C. Brinton. 2013. "School Dropouts and Conditional Cash Transfers: Evidence from a Randomised Controlled Trial in Rural China's Junior High Schools." *The Journal of Development Studies* 49 (2): 190–207. doi:10.1080/00220388.2012.724166.

Muralidharan, K., and V. Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1): 39–77. doi:10.1086/659655.

Neal, D., and D. W. Schanzenbach. 2010. "Left behind by Design: Proficiency Counts and Test-based Accountability." *The Review of Economics and Statistics* 92 (2): 263–283. doi:10.1162/rest.2010.12318.

Nguyen, T. (2008). "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar". *Unpublished manuscript*, 6. Retrieved from http://xxpt.ynjgy.com/resource/data/20091115/U/MIT20091115040/OcwWeb/Economics/14-771Fall-2008/Readings/PaperJM%20TRANG%20NGUYEN%2022jan08.pdf

Nie, J., X. Pang, S. Sylvia, L. Wang, and S. Rozelle (2016). "Seeing Is Believing: Experimental Evidence of the Impact of Eyeglasses on Academic Performance, Aspirations and Dropout among Junior High School Students in Rural China". Working Paper. REAP Working Papers. Stanford, CA.: Rural Education Action Program.

Norton, S., and Q. Zhang (2013). "Chinese Students' Engagement with Mathematics Learning". *International Journal for Mathematics Teaching & Learning*. Retrieved from https://researchrepository.griffith.edu.au/bitstream/handle/10072/58928/90837_1.pdf?sequence=1

PAL, J. (2013). "J-PAL Youth Initiative Review Paper". Cambridge, MA: Abdul Latif Jameel Poverty Action Lab.

Park, A., and E. Hannum (2001 June). "Do Teachers Affect Learning in Developing Countries? Evidence from Matched Student-teacher Data from China". In *Conference Rethinking Social Science Research on the Developing World in the 21st Century* June 7-11, 2001 Park City, Utah. Retrieved from http://www.academia.edu/download/37306958/teachers.pdf

Pritchett, L., and A. Beatty (2012). "The Negative Consequences of Overambitious Curricula in Developing Countries". Working Paper 293. Washington, D.C.: Center for Global Development. Retrieved from https://www.cgdev.org/publication/negative-consequences-overambitious-curricula-developing-countries-working-paper-293

Rao, N., K. M. Cheng, and K. Narain. 2003. "Primary Schooling in China and India: Understanding How Socio-contextual Factors Moderate the Role of the State." In Bray, M. *Comparative Education*, 153–176. Dordrecht: Springer. doi:10.1007/978-94-007-1094-8_9.

Rice, J. K. (2003). "*Teacher Quality: Understanding the Effectiveness of Teacher Attributes*". Economic Policy Institute, 1660 L Street, NW, Suite 1200, Washington, DC 20035. Retrieved from https://eric.ed.gov/?id=ED480858

Rockoff, J. E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2): 247–252. doi:10.1257/0002828041302244.

Siddaiah-Subramanya, M., S. Smith, and J. Lonie. 2017. "Mastery Learning: How Is It Helpful? an Analytical Review." *Advances in Medical Education and Practice* 8: 269. doi:10.2147/AMEP.S131638.

Sylvia, S., N. Warrinnier, R. Luo, A. Yue, O. Attanasio, A. Medina, and S. Rozelle (2018). "From Quantity to Quality: Delivering a Home-based Parenting Intervention through China's Family Planning Cadres". Working Paper. Rural Education Action Progam.

Wang, D. 2011. "The Dilemma of Time: Student-centered Teaching in the Rural Classroom in China." *Teaching and Teacher Education* 27 (1): 157–164. doi:10.1016/j.tate.2010.07.012.

Wang, H., J. Chu, P. Loyalka, T. Xin, Y. Shi, Q. Qu, and C. Yang. 2016. "Can Social–emotional Learning Reduce School Dropout in Developing Countries?" *Journal of Policy Analysis and Management* 35 (4): 818–847. doi:10.1002/pam.21915.

Woronov, T. 2016. "The High School Entrance Exam And/as Class Sorter: Working Class Youth and the HSEE in Contemporary China." *Handbook of Class and Stratification in the People's Republic of China* 178–196. doi:10.4337/9781783470648.00020.

Wu, Y. 2015. "The Examination System in China: The Case of Zhongkao Mathematics." In Selected Regular Lectures from the 12th International Congress on Mathematical Education, edited by Cho, S. 897–914. Cham: Springer. doi:10.1007/978-3-319-17187-6_50.

Yao, Y., Y. Kang, W. Gong, Y. Chen, and L. Zhang. 2011. "MHT Scale of Adolescent with Different Gender: A Meta-analysis." *Chinese Journal of Evidence-based Medicine* 11: 211–219.

Yi, H., G. Li, L. Li, P. Loyalka, L. Zhang, J. Xu, J. Chu, H. Shi, and J. Chu. 2018. "Assessing the Quality of Upper-Secondary Vocational Education and Training: Evidence from China." *Comparative Education Review* 62 (2): 199–230. doi:10.1086/696920.

Yi, H., Y. Song, C. Liu, X. Huang, L. Zhang, Y. Bai, and S. Rozelle. 2015. "Giving Kids a Head Start: The Impact and Mechanisms of Early Commitment of Financial Aid on Poor Students in Rural China." *Journal of Development Economics* 113: 1–15. doi:10.1016/j.jdeveco.2014.11.002.

Yi, H., L. Zhang, C. Liu, J. Chu, P. Loyalka, M. Maani, and J. Wei. 2013. "How are Secondary Vocational Schools in China Measuring up to Government Benchmarks?" *China & World Economy* 21 (3): 98–120. doi:10.1111/j.1749-124X.2013.12024.x.

Yiu, L. 2017. "Policy as a Context for Quality Teaching in China: Diversity, Teacher Adaptation, and Chinese Migrant Children." In *International Handbook of Teacher Quality and Policy*, edited by Akiba, Motoko, LeTendre, Gerald K. 632. New York: Routledge. doi:10.4324/9781315710068.ch17.

Yue, A., Y. Shi, R. Luo, J. Chen, J. Garth, J. Zhang, S. Rozelle, S. Kotb, and S. Rozelle. 2017. "China's Invisible Crisis: Cognitive Delays among Rural Toddlers and the Absence of Modern Parenting." *The China Journal* 78 (1): 50–80. doi:10.1086/692290.

Zhang, H., and X. Wang. 1989. "[Standardization Research on Raven's Standard Progressive Matrices in China.]."*ACTA PSYCHOLOGICA SINICA* 21 (2):113–121. 瑞文标准推理测验在我国的修订

Zhou, B. 1991. *Mental Health Test (MHT). Shanghai*. China: Department of Psychology, East Normal University.

Zhou, M., G. Zhang, S. Rozelle, K. Kenny, and H. Xue. 2018. "Depressive Symptoms of Chinese Children: Prevalence and Correlated Factors among Subgroups." *International Journal of Environmental Research and Public Health* 15 (2): 283. doi:10.3390/ijerph15020283.

Appendix

Table A1. Intervention fidelity in the 11 middle school interventions examined in this study.

programme	Intervention Fidelity	Similar intervention in primary school?	Significant impacts in primary school?
P1 – CCT 1	150 students in treatment group (and their guardian) were individually offered to participate in CCT. 100% of students/guardians accepted CCT offer. CCT were to be provided on condition of minimum 80% attendance rate over one school year 8 students withdrew from CCT intervention before endline due to dropout Of the remaining 142 students, all students who met the attendance requirement received cash transfers.	No	
P2 – ECFA 1 (7th Grade)	474 students in treatment group (and their guardian) were individually offered ECFA contract in a private meeting with school principal and research team representative. 100% of students/guardians accepted the offer and signed the contract. ECFA contract stipulated that research team would provide 1500 yuan/year in high school financial aid for three years if student was enrolled in academic or vocational high school. Funds would be wired to the post office nearest to student's high school on confirmation of enrolment in 2013. Students and parents understood that post offices in China can serve as banks. All treatment students/guardians were contacted by research team in April 2011 to remind them that the contract was still valid 67 treatment students withdrew from ECFA programme before endline (52 dropped out; 9 transferred; 6 withdrew for other reasons) Of the remaining 407 students, all students attending high school in September 2013 received ECFA funds	No	

(*Continued*)

Table A1. (Continued).

P3 – ECFA 2 (9th Grade)	190 students in treatment group (and their guardian) were individually offered ECFA contract in a private meeting with school principal and research team representative. 100% of students/guardians accepted the offer and signed contracts. ECFA contract stipulated that research team would provide 1500 RMB/year in high school financial aid for three years if student was actively enrolled in academic or vocational high school by September 2011. Funds would be wired to the post office nearest to student's high school on confirmation of enrolment. Students and parents understood that post offices in China can serve as banks. Research team tracked all treatment students in August 2011 to confirm high school enrolment via 3 methods: 1) student was given a pre-paid envelope to mail a signed and stamped high school matriculation letter to the research team by 20 August 2011; 2) student's 9th grade homeroom teacher was asked about their whereabouts; 3) enumerators visited student's reported high school to confirm their attendance in person After confirmation, all students attending high school received ECFA funds.	No
P4 – CCT 2	474 students in treatment group and their guardians were individually offered a CCT contract (printed on the letterhead of the Chinese Academy of Sciences) in a private meeting with school principal and research team representative 474 students/guardians (100%) accepted CCT offer and signed contracts in December, 2010. Both students and guardians signed the contract. The research team took a photograph of the contract signing ceremony and mailed the photograph to each student's family as a reminder of the agreement one week after the ceremony. ECFA contract stipulated that research team would provide 1500 RMB/year for three years if student was actively enrolled in academic or vocational high school by September 2013. Funds would be wired to the post office nearest to student's high school on confirmation of enrolment. Students and parents understood that post offices in China can serve as banks. During follow-up surveys in May 2011, May 2013 and October, 2013, the research team identified students who had dropped out of school. After the field survey was over, the enumerators called the relatives or neighbours of the students to confirm whether the students had actually dropped out of school (or were instead temporarily absent or had transferred schools). For treatment students, we also confirmed that the family still had the contract and confirmed the contract was still valid. In October 2013, Students who reported attending high school were visited by the research team to confirm attendance. 100% of CCT treatment students were successfully followed, and 443 students were attending high school. All treatment students attending high school received cash transfers	No
P5 – T1 Educational Returns	In all treatment schools, all grade 7 homeroom teachers and school principals attended a scripted, half-day training led by a professional counsellor in a central location within the province. All teachers and principals were trained to give a 45-minute scripted information intervention lesson to their students. At the end of the training, each teacher received a teacher's manual, a DVD of the lesson and workbooks for their students. All teachers agreed to conduct the information intervention lesson during the week of December 20–24, 2010.	No

(*Continued*)

Table A1. (Continued).

P5 – T2 Educational Returns + Career	In all treatment schools, all grade 7 homeroom teachers and school principals attended a scripted, 1.5 day training led by a professional counsellor in a central location within the province. All teachers and principals were trained to give four 45-minute scripted information intervention lesson to their students. At the end of the training, each teacher received a teacher's manual, a DVD of the lesson and workbooks for their students. all teachers agreed to conduct one lesson per week for four consecutive weeks in December 2010	No	
P6 – Social-Emotional Learning	Officials in the prefectural department of education implemented the intervention. Officials sent an official document in December 2012 to the principal of each treatment school explaining the SEL programme and instructing them to designate a music, art, or physical exercise teacher with previous experience as a homeroom teacher to serve as a part-time SEL teacher. All SEL teachers and school principals attended a scripted training in the prefectural seat led by a professional trainer from Beijing Normal University. Teachers received a five-day training and principals received a half-day training for the principals. All SEL teachers were trained to deliver 32 scripted, 45-minute lessons SEL teachers were instructed to teach the SEL class once per week, during one hour set aside for weekly homeroom class meetings	No	
P7 – Teacher training	Teachers were trained through China's National Teacher Training programme (NTTP). Policymakers gave us a list of the 300 rural junior high schools from 94 counties across a large province that were already slated to participate in the NTTP. Within each of the 300 schools, one grade 7–9 maths teacher was selected according to a standard process: each school nominated one teacher and this nomination was approved by the local education bureau. Selected teachers that were randomised to the treatment arms participated in the NTTP at the start of 2016. Selected teachers that were randomised to the control arm were told they would participate in the NTTP at the start of 2017. The research team randomised treatment teachers into one of 3 NTTP treatment arms: 1) in-person training, 2) post-training follow-up, and 3) post-training evaluation. Teachers in the in-person training treatment arm received a 15-day onsite training at a centralised location. Teachers in the post-training follow-up treatment arm received 3 messages per month about online materials/assignments and progress reports, and were asked to confirm the receipt of the text messages and reply with comments and questions if desired. Trainees in the post-training evaluation treatment arm were informed after the training that they would have to participate in an in-person evaluation (with a lesson plan and interview component) 2 months after the training. 91% of teachers in the in-person training treatment arm participated in the training; 98% of teachers in the follow-up treatment responded to the follow-ups; and 87% of teachers in the post-training evaluation treatment completed the evaluation	Yes	No

(*Continued*)

Table A1. (Continued).

P8 – Teacher incentives	The research team selected 200 rural schools for inclusion in the study. Half of the schools were randomly assigned to the treatment arm. All grade 7 maths teachers in each treatment school were included in the treatment, totalling 145 teachers in 100 schools. In November 2015, treatment teachers were invited by prefecture-level government officials to a central location within the prefecture, where they were given pay-for-percentile performance pay contracts and training on the pay-for-percentile performance pay scheme. During the training, each teacher also received a manual which detailed the pay-for-percentile programme in full. Teachers took a quiz at the end of the training to ensure that they had understood the contents of the programme. Approximately 98% teachers gained full marks on the quiz. Approximately 10% of treatment teachers could not attend the in-person training. Each of these teachers was approached separately at their school and individually received an in-person training and pay-for-percentile performance pay manual, and took (and passed) the associated quiz. As detailed in the training and the contract, treatment teachers were offered the opportunity to receive cash bonuses for raising student test scores based on a pay-for-percentile incentive scheme. Under this scheme, each teacher could earn from 0 to approximately 200 Chinese yuan per student. Since the average number of grade 7 students taught by each teacher was roughly 60, each teacher could approximately earn a maximum of 12,000 Chinese yuan (approximately three months of salary) and an average of 6,000 Chinese yuan (approximately 1.5 months of salary). From January to April 2016, treatment teachers received monthly text messages to remind them about the programme and pay-for-percentile performance pay scheme. Following an endline survey in June 2016, incentive bonuses were calculated for all treatment teachers. Of 145 treatment teachers, 129 received their incentive bonuses on time. There were 16 treatment teachers that transferred schools before endline. According to the pay-for-percentile contract, treated teachers that left the programme were not eligible to receive incentive bonuses.	Yes	Yes
P9 – Free glasses	All students were screened for myopia by a team of optometrists. Students who failed the screening were taken by bus to a clinic located in the central locations to undergo further vision testing. Only 2% of students who failed screening did not go. At the clinic, students underwent automated refraction by highly trained refractionists to determine the nature of their vision problem and whether vision could be improved with eyeglasses. Prescriptions were then determined for the 96% of cases where vision could be improved with eyeglasses. After refraction, eyeglasses were manufactured for students using high-quality equipment that was brought from the United States. Free eyeglasses were distributed in treatment schools to all students found to require eyeglasses, regardless of whether they already had eyeglasses. Refractionists visited the schools and dispensed the eyeglasses, adjusting them to make sure they fit well, and answered any questions students had about wearing and caring for their eyeglasses.	Yes	Yes