



Think Twice before Seeking Private Supplementary Tutoring in Mathematics: A Data Set from China Questioned Its Effectiveness

Ping Ren¹ · Zhongchen Dou¹ · Xici Wang² · Simeng Li¹ · Lidong Wang¹

Accepted: 5 May 2022 / Published online: 13 June 2022
© De La Salle University 2022

Abstract Private supplementary tutoring (PT) has expanded significantly and attracted widespread public interest in China in recent years. It deserves systematic empirical research and reflection, given its crucial theoretical and practical implications. Based on the data of a large-scale survey and testing project, this study examined its effects on the students' academic achievements in mathematics. The results revealed that PT in the current semester did not have any significant positive effect on the students' academic achievement in mathematics. This suggested that households' huge financial investment in PT failed to achieve the desired result. Accordingly, these findings imply that parents should carefully select private tutoring for their children, and the government should provide more comprehensive professional guidelines to regulate the private tutoring industry.

Keywords Assessment · China · Educational governance · Private supplementary tutoring in mathematics · Student evaluations of teaching

Introduction

Private supplementary tutoring (PT) is widely referred to as shadow education in the educational research community because it mimics the regular school curriculum (Baker et al., 2001; Bray & Kwok, 2003). Recently, the PT industry has experienced rapid development across the world with the

expansion of various forms of its service. The PISA 2012 data indicated that in Shanghai, over 70 percent of the students received after-school mathematics tutoring (Hu et al., 2015). According to the Quality Monitoring Report on China's Compulsory Education released by the National Center of Education Quality (Zhang et al., 2021), 23.4 percent of eighth-graders participated in after-school mathematics classes. In the last two decades, PT has emerged in regions other than East Asia and has been increasingly discussed by researchers worldwide, including in Europe and the USA (Bray, 2014; Davies, 2004; Mischo & Haag, 2002; Zhang et al., 2021). This widespread social phenomenon impacts the daily lives of people involved in education (Liu & Bray, 2018) and opens new research areas (Bray, 2014; Davies, 2004) outside the focus on regular school or family education (He et al., 2021). PT is the basic strategy adopted by parents, students, teachers, principals, and school systems when students fail to meet the desired academic standards (Bray, 2014; Wang & Guo, 2017; Zeng & Zhou, 2012). Some Canadian coaching centers (also referred to as "cram" centers) claim, without supporting empirical evidence, that students can gain academic advantages by receiving long-term training on sufficiently complex and solid cognitive skills (Davies, 2004). The claim is still attractive in China today, even though PT institutions lack systematic evidence on the effectiveness of their services to support their aggressive advertising campaigns. The effectiveness of PT in mathematics in improving a student's learning career is a core question examined in relevant studies (He et al., 2021; Zhang et al., 2021). After investing considerable financial resources and time to PT in mathematics, families are entitled to know whether their investments bring any benefits to their children's academic achievement in mathematics (Bray, 2014).

✉ Lidong Wang
wanglidong@bnu.edu.cn

¹ Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing 100875, China

² Beijing Sanfan Middle School, Beijing 100120, China

Previous international academic achievement assessments in mathematics have shown a clear advantage for Chinese and other Eastern Asian students over their Western peers (Leung, 2001; Mullis et al., 2004; Wang & Lin, 2005). This advantage is potentially attributable to Chinese students' efforts in both in-school learning and PT outside of school (Wang & Guo, 2017; Bray et al., 2020). Therefore, the effect of PT on students' mathematics achievement should be an important aspect in the studies of mathematics education in China (Wang & Guo, 2017; Zhang et al., 2021). Additionally, an evaluation of PT effects on students' learning performance across China offers the basis for national educational governance and support related to PT. This could assist national educational authorities in developing and improving standards and systems for after-school education, training, and taking relevant actions in monitoring the industry.

This paper presents an empirical study of PT in mathematics due to its importance among other subjects. Leveraging data from a large-scale survey and testing project in China, it profiled the PT for middle school students and examined whether the tutoring facilitated the improvement of students' mathematics learning performance.

Literature Review

The Roles and Functions of PT in Mathematics Learning

Studies on various education-related factors affecting students' (mathematics) academic performance focus on quantitative educational indicators of formal schooling, including those related to teachers (Konstantopoulos & Sun, 2012; Wang & Cao, 2014) and schools (for example, Coleman et al., 1966). Both formal and outside-school mathematics instruction should be included when discussing the potential factors in promoting students' mathematics learning or in analyzing the variance in students' mathematics achievement (Wang & Guo, 2017; Wang et al., 2019; Zhang et al., 2021).

The purposes of PT include enhancing the results of formal schooling and expanding its curriculum (Zeng & Zhou, 2012). It requires students to learn ahead of schedule or review what they have learned in school. Students with high and low mathematical abilities are usually the targets of PT. For the former, the tutoring helps them build on their advantages or at least maintain their performance levels, while, for the latter, it offers the opportunity to catch up with their peers (Bray & Kobakhidze, 2014; Zhang, 2011). The second purpose often relates to subject competitions, for example, improving the competition results in mathematical Olympiads in China (Wang & Guo, 2017), as well as independent enrollment courses for domestic universities and preparatory courses for overseas ones (such as advanced placement

courses) that have become increasingly important in recent years.

Effects of PT on Students' Mathematics Learning Performance

Regarding the global PT research agenda, effectiveness studies have been the fundamental research issue. These studies provide fundamental empirical results that explain Asian students' achievements in international assessment projects and provide insights to parents for informed educational decisions. Considering PT's comprehensive theoretical effectiveness, previous studies have typically discussed its effectiveness under the Carroll's (1963) model of school learning (Guill et al., 2019; He et al., 2021; Zhang et al., 2021). This model posits that learning is a function of the ratio of the time spent on learning to the time required to learn, with the latter depending on cognitive factors and instruction quality. Specifically, PT potentially adds external time to students' learning and helps them access further learning opportunities (Bäumer, et al., 2011), but cognitive factors and instruction quality decide the effectiveness of the external time as well as the time required. Zhang et al (2021) applied Dunkin and Biddle's (1974) paradigm to determine the aspects of the instructional quality of PT that impact students' mathematics achievement, as illustrated in Table 1. Several presage or process variables discussed in previous studies as well as newly developed variables were included in the framework to explain the effect of PT on the product variable (mathematics achievement in school). The effectiveness was also explained by the time spent on PT, according to Carroll's (1963) model.

Therefore, according to the comprehensive theoretical framework, the effectiveness of PT must be evaluated based on carefully designed data analysis processes and comprehensive representative samples in the future.

The overall effectiveness of PT should be tested and the instructional quality analyzed to explain the corresponding results. Rigorous empirical research is necessary to dispel exaggerated or even misleading advertisements and inform the policy (supervision) decisions of education authorities and education decisions of households.

Few studies have examined the effects of PT in mathematics on students' academic achievement in mathematics using a sample from mainland China, which is insufficient to comprehensively identify complex PT effectiveness. Simple correlation analysis may not assist in understanding the effects, particularly given the complex and diverse teaching practices in PT. Bray (2014) contended that robust data analyses pointing to causation (rather than simple correlation analysis) are more suitable for tracing data, especially when experimental intervention studies are lacking. Further, some prior studies (e.g., Fang et al., 2018) are not specifically

Table 1 Theoretical framework

Dimension	Sub-dimension
Instructional quality	
Presage variables	Teacher qualifications; Time point; Forms of instructional organization
Process variables	Instructional contents
Product variables	Mathematics achievement
Cognitive factors	
Context variables	Pre-mathematics achievement; Self-expectation of education; Family socio-economic status; Demographic characteristics
Time spent	
Intensity	Hours spent on PT; Average hours spent on PT per week

designed for the analysis of PT in mathematics and mathematics performance. Some statistical methods are known to be quite efficient in the literature on the effectiveness of PT. Kuan (2011) used two sets of panel data and propensity score matching (PSM) to analyze the data collected by the Taiwan Education Panel Study in 2001 and 2003. The statistical analysis results indicated that, on an average, the mathematics tutoring programs had a small effect on participants. Zhang (2011) used instrumental variable to analyze the effects of PT on the mathematics scores of college entrance examinations in the megacities of China and found no significant results after controlling for the scores of high school entrance examinations. Sun et al. (2020) used a PSM model and found no significant impact of PT on normalized mathematics scores in a Chinese background. Furthermore, the heterogeneity of effectiveness is frequently addressed in the literature (Campbell et al., 2003; Choi & Park, 2016; Ömeroğulları, et al., 2020; Zhang et al., 2021). It should be noted that, several sorts of variables in the theoretical framework (see Table 1) could be applied to analyze the heterogeneity of effectiveness, especially for the context variables, such as students with different propensity scores (Choi & Park, 2016), or PT with different pre-achievements (Zhang et al., 2021), which can provide an empirical basis for educational theories and practices (Wang & Cao, 2014; Zhang et al., 2021). Zhang et al. (2021) suggested that students with low achievement in mathematics might benefit from PT, while others might not.

There is no straightforward evidence suggesting that PT impacts learning performance in all cases. Results from different countries do not support the argument that PT improves students' educational benefits (Choi & Park, 2016). Bray (2014) summarized several international studies and declared that the results on the effectiveness of PT were inconclusive or even contradictory. He attributed the results partly to different operational definitions and study foci, considering a wide spectrum of forms, organizations, and intensity of PT. Moreover, differences in teaching characteristics (quality), such as curriculum and teaching objectives, teaching methods, and teaching

organizational forms, have not been considered by extant studies. Of course, it is not easy to measure some variables quantitatively due to the lack of theoretical research. Zimmer et al., (2010) suggested the importance of accurately evaluating the effectiveness of various programs and estimating the effects of PT on mathematics performance. PT was informal, less regulated, and voluntary, and PT institutions are less formal than mainstream schools (for example, many PT activities are conducted in students' or teachers' homes; Zhang et al., 2021). These characteristics made the research complex, especially in data collection. Additionally, though there is not a chasm between PT instruction and formal school instruction, PT instruction has specific characteristics (Wang & Wang, 2021) and might not be identified through measurement instruction developed for formal school instruction.

Previous studies have focused on limited samples such as a single city and the results might not be generalizable across China. Although Sun et al (2020) and Zheng et al (2020) applied the same nationally representative survey of China, the achievement data they used included midterm exam scores for students provided by schools. The scores were not comparable across schools, and might not represent the students' achievement appropriately, which might have biased the evaluation of effectiveness (Lockwood et al., 2007; Shavelson et al., 1986).

Research Questions

Overall, PT can provide students with additional learning time. The mathematics teacher in shadow education (Wang & Wang, 2021) can potentially assist students in optimizing such time. Specifically, the researchers of the study asked, "What are the effects of PT in mathematics on students' mathematics learning performance in China?"

Materials and Methods

Participants and Procedures

The database used in the study was from a large-scale nationally representative survey and testing project. The comprised counties from 31 provinces (including autonomous regions and municipalities directly under the central government) and the Xinjiang Production and Construction Corps. The final data sample covered 120 districts and counties. 8 middle schools were sampled from each selected county. In each county, the schools were stratified based on their locations (i.e., city, town, and rural area). 30 eighth-grade students were randomly selected from each school. Each participant completed a mathematics test and a self-report questionnaire. Among them, 41 are in Eastern China, 35 in Central China, and 44 in the west. A total of 29,337 eighth-grade students were sampled, and 27,716 students were finally included in the data analysis. We used the multiple imputation for interpolation (Little & Rubin, 2002; Rubin, 1987) and iterated 50 times to get the data set with SPSS 24.0.

Measure

In this study, the dependent variable is the students' academic achievement in mathematics, which was measured by normalized z scores. These were converted from scaling scores generated based on the Item Response Theory (IRT). The test was developed by a team including mathematics expert teachers and researchers in mathematics education and educational measurement was based on the national curriculum standard of China according to a standardized process. The mathematics-testing instrument reflected the requirements of national mathematics curriculum standards and has good measurement indexes for evaluating students' mathematics academic achievement. Item analysis indicated that the difficulty index of the mathematics testing was moderate with better discrimination and high overall reliability and validity. Unidimensional IRT models were used to calibrate items and create the score scales for mathematics (see Jiang et al., 2018). In addition, this study also adopted the Modified Angoff method to analyze the students' testing data, supplemented by the Bookmark procedures, to allocate students' academic performance from low to high into several grades, and performance standards for each subject were specified according to the national curriculum. There are four performance levels that classify students into the levels of I, II, III, and IV, with Level I being the lowest and Level IV being the highest level (see Jiang et al., 2018). It should be noted that, though the database contained abundant variables and the sample was large enough, the current study is still a cross-sectional study, which may not well identify the causal impact of PT on students' achievement (Zhang et al.,

2021). To partly address the issue, the categories of levels were used as a proxy for students' pre-existing performance levels, since the results of previous studies indicated that PT might not enhance the students' mathematics achievement (Zhang et al., 2021), or at most enhance the students' mathematics achievement slightly within a semester. Therefore, we could reasonably assume that the students' grade level did not change within one semester, and the grade level could partly reflect the students' academic level at the beginning of the semester. The independent variable of this study is whether students receive PT in the current semester.

The covariates variables in this study include those at the individual level of students (e.g., gender, self-academic expectation, and single-child family), at the household level (e.g., parents' maximum education years, parents' academic expectation, and household economic status), and at the school level (e.g., student-teacher ratio, class size, school nature, and location). These variables were discussed in previous studies (He et al., 2021; Zhang et al., 2020, 2021) to solve the problem of sample selection bias. The specific classification and assignment of variables are presented in Table 2 below.

Data Analysis

As discussed in the introduction, the HLM regression model and the PSM were applied to assess the effectiveness of PT in student achievement. Two powerful statistical models were applied in this study to obtain triangulation. The Stata15.0 and SPSS24.0 were applied for the data analysis.

HLM

HLM can avoid the limitations of classical statistical techniques in analyzing hierarchical data and the potential misinterpretations of results. It is a suitable tool for in-depth analysis and interpretation of commonly occurring data with hierarchical levels (Raudenbush & Bryk, 2002).

Given the structural characteristics of testing data obtained by the random multistage stratified cluster sampling, and considering the significant differences in the educational quality between urban and rural areas and across regions, we assigned the student level characteristics and household background factors, school-level factors and province-level factors to the first, second, and third levels, respectively, for constructing a multi-level random intercept model. Then, the model was used to analyze the effects of private tutoring on students' academic performance. Specifically:

$$\text{Level 1: } Y_{ijk} = \beta_{0jk} + \beta_{1jk}X_{1ijk} + \beta_{2jk}X_{2ijk} + \dots + \beta_{njk}X_{nijk} + r_{ijk}, r_{ijk} \sim N(0, \sigma^2)$$

Table 2 Assignment and definitions of the variables

Variables	Description
Mathematics achievement scores	Standardized scores derived from the original scaling scores
Grade level	Grades, 1 = To be improved, 2 = Medium, 3 = Good, 4 = Excellent
Receiving mathematics PT or not	0 = No tutoring, 1 = Yes, (in the current semester)
Gender	0 = Female, 1 = Male
Single-child	0 = Non-single-child, 1 = Single-child
Self-expectation	1 = No specific expectation, 2 = Medium in class, 3 = Top 10 in class, 4 = Top 3 in class
Interest in learning	1 = Low interest in learning, 2 = Medium interest in learning, 3 = High interest in learning
Parents' expectation	1 = No specific expectation, 2 = Medium in class, 3 = Top 10 in class, 4 = Top 3 in class
Parents' educational background	Parents' maximum education years
Raising types	1 = Jointly raised by both parents and grandparents, 2 = Raised by grandparents, 3 = Raised by parents
Homework time	1 = No homework, 2 = Within 15 min, 3 = 15 to 30 min, 4 = 30 to 60 min, 5 = 1 to 2 h, 6 = More than 2 h
Economic status	1 = Low economic status, 2 = Medium economic status, 3 = High economic status
Boarding students	0 = No, 1 = Yes
Class size for Grade 8	Total number of eighth graders divided by the number of classes
Student-teacher ratio	The number of students divided by the number of teachers in a school
Weekly mathematics lessons	1 = 4 lessons or less, 2 = 5 lessons, 3 = 6 lessons, 4 = 7 lessons, 5 = 8 lessons, 6 = 9 lessons, 7 = 10 or more lessons
School nature	0 = non-public school, 1 = public school
School location	1 = city, 2 = county town, 3 = town, 4 = rural area
Region	1 = Eastern, 2 = Central, 3 = Western

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}s_{1jk} + \dots + \gamma_{0mk}s_{mjk} + \mu_{0jk}, \mu_{0jk} \sim N(0, \tau^2)$$

Level 2: $\beta_{1jk} = \gamma_{10k}$
 ...

$$\beta_{njc} = \gamma_{n0k}$$

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}s_{1jk} + \dots + \gamma_{0mk}s_{mjk} + \mu_{0jk}, \mu_{0jk} \sim N(0, \tau^2)$$

$$\beta_{1jk} = \gamma_{10k}$$

...

$$\beta_{njc} = \gamma_{n0k}$$

$$\gamma_{00k} = \pi_{000} + \varepsilon_{00k}$$

Level 3: $\gamma_{01k} = \pi_{010}$
 ...

$$\gamma_{n0k} = \pi_{n00}$$

In the above model, π_{000} was the average score of all students' academic performance; γ_{00k} was the average score of academic performance for students in province k ; β_{0jk} was the average score of academic performance for students in province k , where school j was located; $\chi_{1ijk} \sim \chi_{nijk}$ denoted variables for individual- and household-level factors affecting students' academic performance; $\beta_{1jk} \sim \beta_{njc}$ denoted the regression parameters of variables for the factors at the individual level; $s_{1jk} \sim s_{mjk}$ denoted variables for school-level factors affecting students' academic performance; $\gamma_{01k} \sim \gamma_{0mk}$ denoted the regression parameters of variables for the factors at the school level; r_{ijk} , μ_{0jk} , and ε_{00k} , respectively, represented random errors at the individual, school, and province

levels. Grand mean was adopted for the variables, and the ICC is 0.312, which indicated the hierarchical structure of the sample.

PSM

The factors that affect students' academic performance are complex and diverse, and the typical regression analysis model often cannot capture all the potential variables, especially some latent variables that cannot be observed, such as students' individual ability. In addition, because the samples were not completely homogeneous, self-selection bias may have occurred in whether to choose to receive tutoring. For example, poor performers and students with favorable household backgrounds might have been more likely to receive PT (Choi & Park, 2016), introducing some bias in the analysis results of the PT effects. To evaluate the causal effect of PT on students' academic development accurately, the ideal way is to compare the results of each individual student in the states of receiving and not receiving PT simultaneously, which are counterfactual and untenable. Therefore, based on a quasi-experimental research design, this study adopted the PSM method (Rosenbaum & Rubin, 1983) under the assumption of conditional independence. The purpose of this method was to match a student receiving PT with a student not receiving the service albeit having

similar or even identical endowment characteristics. Statistically, this method could control the interference caused by sample non-random selection bias by ensuring the independence of the two groups of students, making a more robust and reliable estimation of the causal effect of private tutoring on middle school students' development possible. We used Logit regression to fit the propensity score value of each sample, which reflected the probability of a sample participating in PT. Then, according to the propensity score, the treatment and control groups were matched with the put back, one-to-many nearest neighbor matching, and finally achieved matching the control group with the treatment group. We used software Stata15.0 for PSM analysis, and then checked the matching effect through balance and common support tests. The result shows that the matching effect is good (details described in the result section).

Results

HLM Analysis on the Effects of PT

For comparison, we assigned the student-level characteristics and household background factors, school-level factors, and province-level factors to the first, second, and third levels, respectively, for constructing a multi-level random intercept model. Then, the model was used to analyze the effects of PT on students' academic performance.

Table 3 presents the results of the analysis. Several covariates were found to have significant effect on students' mathematics achievement with a considerable effect size, such as self-expectation, parents' expectation, raised by parents, and school nature. Moreover, other covariates might be understood with caution because the effect size was small, and the sample size of the current study was large and easy to obtain a significant result with small effect size.

The results indicate that receiving PT in the current study did not improve students' learning performance significantly. Instead, it had significant negative effects on learning performance.

PSM Analysis on the Effects of PT

Given some problems of endogeneity and self-selection bias even for HLM regression, the result that PT negatively affected students' academic performance might be deemed as conclusive. Next, to address the problems, we performed a further analysis with PSM as Triangulation.

Based on typical Logit regression, the propensity matching scores of students receiving PT in mathematics were derived. Figure 1 and Table 4 present the test results of the matching quality. It demonstrates that the matching quality was satisfactory as the distribution of the experimental

Table 3 Analysis results of the three-level model for the effect of tutoring on academic performance

Variables	Coefficients	Standard errors
Mathematics Private tutoring	-0.181***	0.009
Gender	0.035***	0.009
Single-child families	0.022**	0.010
Self-expectation	0.350***	0.007
Parents' expectation	0.179***	0.007
Parents' educational background	0.029***	0.002
Jointly raised by both parents and grandparents	0.124***	0.010
Raised by grandparents	0.068***	0.018
Economic status	-0.020***	0.007
Boarding students	0.017	0.013
Class size	0.095***	0.012
Student-teacher ratio	0.003	0.013
Weekly class hours	0.027***	0.003
No homework	-0.272***	0.029
Within 15 min	-0.159***	0.015
15 to 30 min*		
30 to 60 min	0.085***	0.011
1 to 2 h	0.048***	0.013
More than 2 h	-0.038***	0.021
School nature	-0.189***	0.040
Intercept	-1.664***	0.077
LR chi2	-30,206.849	
Prob > chi2	0.000	
Sample size	27,716	
ICC	0.312	

*Reference group; significance is denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$

group (with PT) and the control group (without PT) after matching was consistent, and there was no statistically significant difference between the covariates.

Table 5 indicates that before matching, the average scores of students with PT were significantly lower than those of peers without PT. In addition, after matching, the difference in academic performance between the student groups was almost zero, although the result was significant. Since the sample size was indeed large, the significant result in such a small effect size could be ignored. That is, receiving PT had almost no significant effect on students' learning performance, which is paralleled to the results of the HLM analysis. The combined results of the two models suggest that there is no evidence to support the argument that PT in mathematics could improve students' mathematics academic performance.

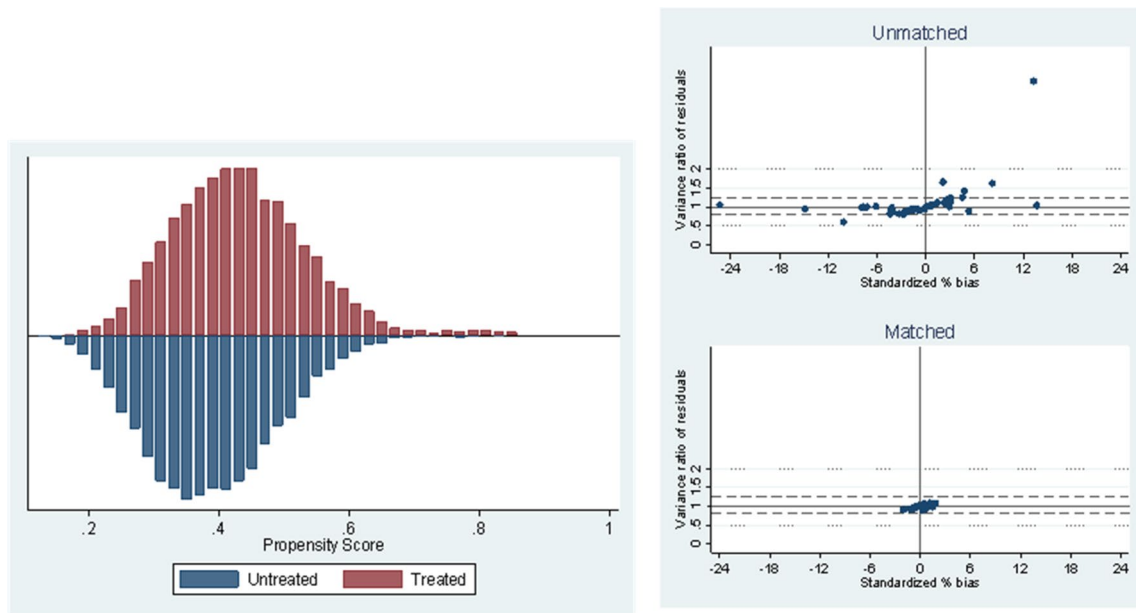


Fig. 1 Tests on matching quality

Table 4 The comparison of the treated and control groups

Variable	Mean		<i>t</i>
	Treated	Control	
Gender	0.515	0.545	-5.010***
Single-child families	0.307	0.377	-12.120***
Self-expectation	2.734	2.795	-5.910***
Grade level	2.459	2.732	-20.650***
Interest in learning	2.787	2.789	-0.150
Boarding students	2.720	2.786	-6.440***
Parents' expectation	9.927	9.849	2.320***
Parents' educational background	0.231	0.249	-3.490***
Raised by grandparents	0.058	0.068	-3.290***
Raised by parents	1.828	1.729	11.180***
Economic status	0.346	0.383	-6.300***
Class size for Grade 8	48.740	48.350	2.360***
Student-teacher ratio	13.900	12.957	1.750**
Mathematics lessons	3.996	4.071	-3.420***
School nature	0.902	0.886	4.300***

Discussion and Implications

This study aimed to investigate the contribution of PT to the academic performance of middle school students all

over China. In conclusion, the study did not identify any significant positive effect of receiving PT (in the current semester) in mathematics on academic achievement in the subject, contrary to parents' expectations. This is consistent with the results of many previous empirical studies with limited samples in China (e.g. Zhang, 2011; Zhang et al., 2021) as well as large scale survey in other countries (e.g. Guill et al., 2019; Ömeroğulları et al., 2020). The result questions the effectiveness of PT in China. Based on the framework of Carroll model of schooling discussed above, although receiving PT implies extended learning time and the acquisition of more learning resources, the activity alone does not necessarily improve students' academic achievement (He et al, 2021).

The learning time of PT might interfere with the time for formal school learning (Bray, 2009), such as time spent on homework. Moreover, the quality of PT instruction is a factor that impacts PT effectiveness (Guill et al., 2019; Zhang, et al., 2021), as PT instruction might not be coherent with the regular school instruction in content or objectives of learning (He et al, 2021). Previous studies tended to pay little attention on addressing variables related to specific instructional characteristics of PT (Wang & Wang, 2021), such as the learning processes and teaching behaviors because these variables cannot be measured readily. The arrangement of

Table 5 Effects of PT before and after matching

	Experimental group	Control group	ATT	Standard errors
Before matching	-0.153	0.105	0.258***	0.012
After matching	-0.153	-0.131	-0.023**	0.009

PT and formal schooling might interfere with each other and make it harder for the PT teacher to maintain coherence within PT instruction. This is especially the case when the class size is larger than one student from different schools or classes with different learning progressions. Simultaneously, students might face difficulties in dealing with two or even more incoherent trajectories of mathematics learning at the same time. Thus, PT might not benefit students' mathematical learning (He et al., 2021). Uniform professional standards and professional development program for PT teachers are lacking (Wang & Wang, 2021). The instructional quality could explain the results of evaluation of PT effectiveness, because in practice, PT activities in mathematics are not identical and inevitably lead to distinctive learning outcomes. This issue needs to be further adequately addressed in future teaching activities.

It should be noted that, though several variables were controlled to address the problem of the selection bias of PT and endogeneity (Zhang, 2013), this study was still a cross-section study, and longitudinal research design would be developed in the future study. In addition, this study only explored the PT within one semester, and if students would benefit from longer time PT should be tested in the future studies.

This study indicates that PT in mathematics may fail to improve students' mathematics academic achievement. As the PT could potentially overburden students, special attention must be paid to the results of the current study, which confirms an important educational paradigm—partnership with schools should be the priority for parents who care about their children's learning performance. To improve students' academic achievement, schooling may be a better option, especially considering schools' potential educational and teaching resources and their roles as providers of mainstream education. The partnership promotes students' participation in school activities to improve their learning performance. Seeking PT without providing further parental support would be ineffective or even harmful. Parents should motivate their children to take ownership of their learning careers. Educational authorities need to take measures to mitigate parents' anxiety about students' learning, and inform parents and the wider public about the empirical evidence regarding the effectiveness of PT to guide them to seek private tutoring carefully.

References

- Baker, D. P., Motoko Akiba, M., Letendre, G. K., & Wiseman, A. W. (2001). Worldwide shadow education: Outside-school learning, institutional quality of schooling, and cross-national mathematics achievement. *Educational Evaluation and Policy Analysis*, 23(1), 1–17. <https://doi.org/10.3102/01623737023001001>
- Bäumer, T., Preis, N., Roßbach, H. G., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. *Zeitschrift Für Erziehungswissenschaft*, 14(S2), 87–101.
- Bray, M. (2009). *Confronting the shadow education system: What government policies for what private tutoring?* UNESCO International Institute for Educational Planning.
- Bray, M. (2014). The impact of shadow education on student academic achievement: Why the research is inconclusive and what can be done about it. *Asia Pacific Education Review*, 15(3), 381–389.
- Bray, M., & Kobakhidze, M. N. (2014). Measurement issues in research on shadow education: Challenges and pitfalls encountered in TIMSS and PISA. *Comparative Education Review*, 58(4), 590–620.
- Bray, M., & Kwok, P. (2003). Demand for PT: Conceptual considerations, and socioeconomic patterns in Hong Kong. *Economics of Education Review*, 22(6), 611–620.
- Bray, M., Kobakhidze, M. N., & Suter, L. E. (2020). The challenges of measuring outside-school-time educational activities: Experiences and lessons from the programme for international student assessment (PISA). *Comparative Education Review*. <https://doi.org/10.1086/706776>
- Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education*, 29(3), 347–362.
- Carroll, J. B. (1963). A model of school learning. *Teacher College Record*, 64, 723–733.
- Choi, Y., & Park, H. (2016). Shadow education and educational inequality in South Korea: Examining effect heterogeneity of shadow education on middle school seniors' achievement test scores. *Research in Social Stratification and Mobility*, 44, 22–32.
- Davies, S. (2004). School choice by default? Understanding the demand for private tutoring in Canada. *American Journal of Education*, 110(3), 233–255.
- Dunkin, M., & Biddle, B. (1974). *The study of teaching*. Holt, Rhinehart & Winston.
- Fang, C., Hu, Y., & Zhang, P. (2018). 小学生能从课后学习时间内受益吗——基于教育基尼系数的测算[Can pupils of primary schools benefit from the study time after class]. *Journal of Educational Science of Hunan Normal University*, 17(1), 69–77.
- Guill, K., Ludtke, O., & Koller, O. (2019). Assessing the instructional quality of private tutoring and its effects on student outcomes: Analyses from the German National Educational Panel Study. *British Journal of Educational Psychology*. <https://doi.org/10.1111/bjep.12281>
- He, Y., Zhang, Y., Ma, X., & Wang, L. (2021). Does PT matter? The effect of PT on mathematics achievement. *International Journal of Educational Development*. <https://doi.org/10.1016/j.ijedudev.2021.102402>
- Hu, Y., Fan, W., & Ding, W. (2015). 影子教育是否扩大教育结果的不均等——基于PISA2012上海数据的经验研究 [Does shadow education aggravate inequality of educational outcomes? An empirical study on PISA2012 Shanghai data]. *Peking University Education Review*, 13(3), 29–46.
- Jiang, Y., Zhang, J., & Xin, T. (2018). Toward education quality improvement in china: A brief overview of the national assessment of education quality. *Journal of Educational and Behavioral Statistics*, 44(6), 733–751.
- Konstantopoulos, S., & Sun, M. (2012). Is the persistence of teacher effects in early grades larger for lower-performing students? *American Journal of Education*, 118(3), 309–339.
- Kuan, P. Y. (2011). Effects of cram schooling on mathematics performance: Evidence from junior high students in Taiwan. *Comparative Education Review*, 55(3), 342–368.

- Leung, F. K. S. (2001). In search of an East Asian identity in mathematics education. *Educational Studies in Mathematics*, 47(1), 35–51. <https://doi.org/10.1023/a:1017936429620>
- Little, R. A., & Rubin, D. B. (2002). *Rubin statistical analysis with missing data* (pp. 218–265). Wiley.
- Liu, J., & Bray, M. (2018). Evolving micro-level processes of demand for PT: Patterns and implications at primary and lower secondary levels in China. *Educational Studies*. <https://doi.org/10.1080/03055698.2018.1555452>
- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47–67. <https://doi.org/10.1111/j.1745-3984.2007.00026.x>
- Mischo, C., & Haag, M. L. (2002). Expansion and effectiveness of private tutoring. *European Journal of Psychology of Education*, 17(3), 263–273.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. TIMSS and PIRLS International Study Centre, Boston College.
- Ömeroğulları, M., Guill, K., & Köller, O. (2020). Effectiveness of private tutoring during secondary schooling in Germany: Do the duration of private tutoring and tutor qualification affect school achievement? *Learning and Instruction*, 66, 101306. <https://doi.org/10.1016/j.learninstruc.2020.101306>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.
- Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse In surveys* (pp. 75–112). Wiley.
- Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50–91). MacMillan.
- Sun, L., Shafiq, M. N., McClure, M., & Guo, S. (2020). Are there educational and psychological benefits from private supplementary tutoring in Mainland China? Evidence from the China education panel survey, 2013–15. *International Journal of Educational Development*, 71, 102144.
- Wang, C., & Wang, L. (2021). The mathematics teacher in shadow education: A new area of focus in teacher education. *Journal of Mathematics Teacher Education*. <https://doi.org/10.1007/s10857-021-09522-3>
- Wang, J., & Lin, E. (2005). Comparative studies on U.S. and Chinese mathematics learning and the implications for standards-based mathematics teaching reform. *Educational Researcher*, 34(5), 3–13.
- Wang, L., & Cao, Y. (2014). 教师对学生数学学业成就的影响研究述评 [Literature review of mathematics teacher effect on student achievement]. *Journal of Mathematics Education*, 23(3), 48–52.
- Wang, L., & Guo, K. (2017). Shadow education of mathematics in China. In Y. Cao & F. K. S. Leung (Eds.), *The 21st Century Mathematics Education in China* (pp. 93–103). Springer.
- Wang, L., Gong, X., & Pei, C. (2019). 数学教育视角下的影子教育研究的意义与未来方向 [Shadow education research with mathematics education: Its value and future research area]. *Journal of Mathematics Education*, 28(1), 79–82.
- Zeng, X., & Zhou, H. (2012). 北京市四、八年级学生课后补习的代价与收益 [A descriptive analysis on after-school tutoring in Beijing: Its costs and benefits]. *Journal of Educational Studies*, 8(6), 103–109.
- Zhang, Y. (2011). *The determinants of national college entrance exam performance in [People's Republic of] China—with an analysis of private tutoring*. (Unpublished doctoral dissertation). Columbia University
- Zhang, Y. (2013). Does private tutoring improve students' national college entrance exam performance? A case study from Jinan, China. *Economics of Education Review*, 32, 1–28.
- Zhang, Y., Dang, Y., He, Y., & Wang, L. (2021). Is PT effective? A longitudinally detailed analysis of private tutoring quality in China. *Asia-Pacific Education Review*, 22(4), 239–259.
- Zhang, Y., Ma, X., & Wang, L. (2020). The determinants of private tutoring participation for mathematics in China: Focusing on the role of student metacognition. *Frontiers in Psychology*, 11, 603. <https://doi.org/10.3389/fpsyg.2020.00603>
- Zheng, X., Wang, C., Shen, Z., & Fang, X. (2020). Associations of private tutoring with Chinese students' academic achievement, emotional well-being, and parent–child relationship. *Children and Youth Services Review*. <https://doi.org/10.1016/j.childyouth.2020.104934>
- Zimmer, R., Hamilton, L., & Christina, R. (2010). After-school tutoring in the context of no child left behind: Effectiveness of two programs in the Pittsburgh public schools. *Economics of Education Review*, 29(1), 18–28.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.