China Economic Review 87 (2024) 102225

Contents lists available at ScienceDirect

China Economic Review

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

Do teachers colleges prepare more effective teachers? Evidence from a top school district in China

Qi Zheng [a], Xin Xie [a,*], Xiaoyang Ye [b], Yi Wei [c]

[a] Department of Educational Leadership and Policy Analysis, University of Wisconsin-Madison, 1000 Bascom Mall, Madison, WI 53706, USA
[b] Annenberg Institute for School Reform, Brown University, 164 Angell St., Providence, RI 02906, USA
[c] China Institute for Educational Finance Research, Peking University, 5 Yiheyuan Road, Haidian District, Beijing 100871, PR China

ARTICLE INFO

JEL codes:
I21
I28

Keywords:
Teachers colleges
Traditional teacher education
Teacher effectiveness
Value-added model

ABSTRACT

This paper provides the first empirical evidence on the value of teacher education by comparing the effectiveness of traditionally trained teachers from teachers' colleges and non-traditionally trained teachers from comprehensive universities in China. We utilize the unique data of 424 traditionally and 40 non-traditionally trained teachers from a top school district in China and a cross-subject value-added model to address teacher-student sorting bias. Our results indicate that, on average, traditionally trained teachers contribute approximately 0.1 standard deviations higher value to students' high-stakes test scores over a three-year period than non-traditionally trained teachers. Behavioral analyses on the district's teacher survey suggest that while teachers from both routes are similar in time allocation and professional psychological characteristics, traditionally trained teacher report being better in applying multitudinous pedagogies and teaching strategies. We find similar results using two administrative data sources from another province. The findings of this study contribute to the ongoing debate about the effectiveness of alternative routes to teaching and the importance of teacher education.

1. Introduction

Teacher education in China has primarily been provided by normal universities since the 1950s, similar to teachers' colleges in Western contexts. These institutions train educators for K-12 schools (Hayhoe & Li, 2010; Zhu & Han, 2006). However, in recent years, there has been a marked increase in the number of teachers who graduated from comprehensive universities and entered the profession through alternative pathways, particularly in metropolitan areas such as Beijing and Shanghai (as shown in Fig. 1). These teachers typically have not received comprehensive teacher training prior to their entry into the field. Despite the rapid growth of this population of educators, there is limited knowledge regarding their teaching effectiveness on student academic achievement. It is of great interest whether these non-traditionally trained teachers have higher teaching effectiveness relative to their traditionally trained peers.

The comparison of the effectiveness of traditionally and non-traditionally trained teachers on student outcomes has been a topic of increasing interest in other countries. However, this has proven to be challenging. The results vary by training programs due to the lack of a standard definition of alternative routes to teaching (Goldhaber et al., 2013; Harris & Sass, 2011; Henry, Bastian, et al., 2014; von

* Corresponding author at: Department of Educational Leadership and Policy Analysis, University of Wisconsin-Madison, 253 Education Building, 1000 Bascom Mall, Madison, WI 53706, USA.
E-mail addresses: qzheng54@wisc.edu (Q. Zheng), xin.xie@wisc.edu (X. Xie), xiaoyang.ye@brown.edu (X. Ye), weiyipku@pku.edu.cn (Y. Wei).

https://doi.org/10.1016/j.chieco.2024.102225
Received 6 March 2023; Received in revised form 2 April 2024; Accepted 17 June 2024
Available online 18 June 2024

Hippel & Bellows, 2018). Furthermore, student and teacher sorting has largely hindered the evaluation of teaching effectiveness. Alternative pathways to teaching, such as *Teach For America* (TFA), are generally designed to address localized teacher shortages in hard-to-staff schools or subjects in lower educational levels (Clark & Isenberg, 2020; Xu et al., 2011). As a result, students in rural areas, poorer school districts, and students of color are more likely to have teachers with less professional training (Mason-Williams, 2015).

In China, the context is different from what is studied in the existing literature. Despite the lack of consensus on what constitutes alternative pathways to teaching in existing studies, non-traditionally trained teachers in China typically lack specialized teacher education, but can still enter the teacher labor market by passing certificate tests consisting of a knowledge test and an interview (Shi et al., 2022). Second, in China, non-traditionally trained teachers are increasingly recruited to highly-competitive prestigious senior high schools in developed metropolitan areas. These teachers are believed to possess better academic abilities, as indicated by their higher college entrance examination scores, and deeper understanding of their academic disciplines (Zhang, 2021; Zhao, 2018).

Based on this unique setting, this paper provides the first empirical evidence on two questions based on the data collected from 424 traditionally and 40 non-traditionally trained teachers in one of the most competitive school districts in China, Haidian. Firstly, which type of teacher - traditionally or non-traditionally trained - is more effective in value added to student high-stake test scores in China's senior high schools? Secondly, are there any differences in work behaviors between the two types of teachers?

Our study makes several important contributions to existing research. Firstly, it provides a unique comparison between traditionally trained and non-traditionally trained teachers in a highly competitive school district where teacher shortage is not the motivation for recruiting non-traditionally trained teachers. Using two administrative data sources from another province, we find similar results. This gives us a better understanding of the impact of teacher training on teacher effectiveness. Secondly, our findings support the value of pedagogical skills and strategies (Hill et al., 2005; Monk, 1994; Shulman, 1987). We find evidence that traditionally trained teachers may have an advantage over non-traditionally trained teachers due to their superior pedagogical skills and strategies.

The rest of the paper is structured as follows. In the second section, we review the relevant literature and teacher education system in China. The data and methodologies are described in Section 3. We present our findings and a set of robustness checks in Section 4, and the final section provides a discussion of the results, outlining the policy implications, limitations, and avenues for future research.

2. Existing literature and institutional backgrounds

2.1. Comparing traditional and alternative routes to teaching

In the context of most existing studies, alternative routes to teaching were created to address teacher shortages and increase diversity in the profession (Sass, 2015; West & Frey-Clark, 2019; Whitford et al., 2018). Although the definition of alternative routes may vary across contexts, they are typically viewed as a compromise option with a potentially negative impact on student achievement

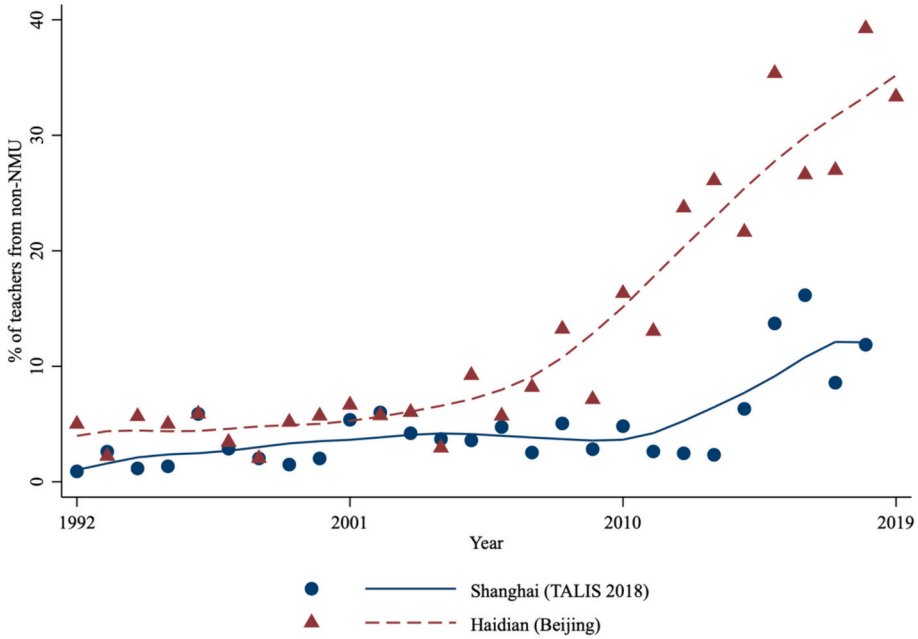

Fig. 1. Proportion of novice secondary school teachers without traditional training in Shanghai and Haidian (Beijing) Over Time.
Note: Non-parametric curves fitted using Lowess function with 0.5 bandwidth; Shanghai data from TALIS 2018; Haidian (Beijing) data collected from CIEFR teacher survey in 2019; both databases include teachers from middle and high schools.

(Goldhaber & Brewer, 2000). Previous studies have found that traditional teacher education programs have significant impacts on teachers' beliefs, attitudes, teaching skills, and education knowledge. These programs can foster teachers' critical thinking and decision-making abilities, as well as their understanding of effective teaching and learning (Graber, 1996; Leavy & Hourigan, 2018; Petek & Bedir, 2018; Rots et al., 2007). Through peer coaching and hands-on teaching experiences, traditional teacher education programs can enhance instructional skills and confidence, while also providing opportunities for personal and professional growth (Caires & Almeida, 2005; Goker, 2006; Ronfeldt, 2012). Furthermore, education courses in teacher education programs can increase teachers' knowledge about education and teaching, and provide opportunities to learn and adopt educational concepts in their teaching practices (Carroll et al., 2003; Lorente-Catalán & Kirk, 2016; Stürmer et al., 2013). The combination of these components in teacher education programs has been associated with stronger student outcomes (Lynch et al., 2019).

Contrary to the popular belief, empirical research in the last two decades has shown that teachers from alternative pathways tend to perform similarly or even better than traditionally trained teachers. For example, a widely-cited value-added analysis of around 40,000 public school teachers in New York City found little evidence of difference in student test scores between traditionally trained and non-traditionally trained teachers (Kane et al., 2008). A meta-analysis by Whitford et al. (2018) even found that non-traditionally trained teachers have a positive impact on student performance that was 0.03 standard deviations higher than traditionally trained teachers, and it is significant in statistics.

However, the reason behind this unexpected finding may stem from unequal comparisons. Recent discussions have focused primarily on comparing the effectiveness of alternative pathways to traditional teacher education in addressing teacher shortages and filling hard-to-staff subjects in disadvantaged schools. Programs such as TFA, *Teach For China* (TFC), and *Troops to Teachers* recruit and train individuals from specific backgrounds, such as top universities or diverse military experience, to become teachers (Backes et al., 2018; Coffman et al., 2019; Lam, 2017). Although previous studies have found positive effects on students' reading or math achievement from teachers of these programs (Clark & Isenberg, 2020; Henry, Purtell, et al., 2014; Xu et al., 2011), it is important to note that these teachers are often selected from highly selective universities and compared with traditionally trained teachers who may not share the same background and experience in low-income communities (Gershenson, 2021). This makes it challenging to draw generalizable conclusions about the effectiveness of teacher education, as the comparison groups may differ in important ways (Heilig & Jez, 2010).

To better inform the value of the traditional training regarding its effectiveness, it would be a good practice to compare teachers who go through different paths to enter the profession but otherwise possess similar qualities and backgrounds, which can be achieved in our sample as is discussed in the next section.

2.2. Teacher education in China

The teacher preparation system in China is standardized and homogeneous. The one and only traditional track roots in the distinct normal universities (NMUs) that specifically train teachers with regularized requirements by the central Department of Education administration (Zhu & Han, 2006). Compared with general comprehensive universities, NMUs exhibit the following unique characteristics in educating and training teachers. First, NMUs provide both subject-specific content and pedagogy courses (Zhou & Reed, 2005). About 47.1% to 52.0% of the total curriculum focuses on subject-related knowledge, and 10% of credits are designated for education, psychology, pedagogical content, and educational technology courses that are generally not provided in comprehensive universities (non-NMUs) (Zhou, 2014). Another 20% of the credits are dedicated to courses focusing on subject-specific teaching (Norton & Zhang, 2018). Second, in NMUs, all teacher candidates must complete teaching fieldwork and practice for at least eight weeks, typically in the form of full-time teacher internships (Norton & Zhang, 2018). This has been emphasized multiple times in official documents issued by the Ministry of Education of China, which suggest that the fieldwork practices should be incorporated throughout the entire teacher education and training process (2007, 2016).[1]

In the past two decades, the Chinese government has implemented policies to encourage high-ranking comprehensive universities to train teachers and to promote their graduates to pursue careers in teaching (*The State Council's Decisions about Deepening Education Reform and Fully Promoting Well-Rounded Education*, 1999; The Minstry of Education and Other Four Ministries' Notice on Printing and Distributing "Teacher Education Revitalization Action Plan (2018)", 2018–2022). The first endeavor of the government has not been well-received by the high-ranking comprehensive institutions, the vast majority of which have only established research-oriented education schools and programs, and for the very few that have established teacher education programs, their programs are extremely small and are not well-aligned with subjects and teaching practices in schools (Chen & Li, 2021; Li et al., 2021). Despite these challenges, the system still allows graduates from regular higher education institutions without teacher training to become teachers by passing the teacher certification tests (Guo, 2005; Shi & Englert, 2008).

On the other hand, school administrators and the public view comprehensive university graduates as having higher academic abilities and better understanding in their major disciplines, as these universities have higher admission bars, more comprehensive disciplines and resources, and a focus on research (Xun & Cao, 2021; Zhang, 2021). As a result, there has been an increase in the number of new teachers who have graduated from leading non-NMUs, particularly in high-performing school districts located in major cities.

[1] Please refer to http://www.moe.gov.cn/srcsite/A10/s7011/200707/t20070705_145953.html (in Chinese) and http://www.moe.gov.cn/srcsite/A10/s7011/201604/t20160407_237042.html (in Chinese) for more details.

2.3. Institutional context of Haidian

As one of the top-ranking school districts in China, Haidian has a highly competitive teacher labor market due to its strong economic position, abundant educational resources, and renowned public schools. In 2018, the district's GDP per capita was nearly 35% higher than the average in Beijing and three times that of the national average. Haidian is home to over one-fifth of China's most prestigious universities and is well-known for its skyrocketing housing prices that are linked to student enrollment in elite public schools. This is associated with a significantly higher chance of students getting into elite universities, better career paths, and a better quality of life (Han et al., 2021; Pan, 2016). These factors attract graduates from both top normal and non-normal universities, contributing to the district's highly competitive teacher labor market.

Fig. 1 illustrates the trend of teachers from non-NMUs in Haidian, with Shanghai included as a comparison. The proportion of secondary school teachers without traditional teacher education was below 5% in the 1990s and early 2000s, but this figure rapidly increased after 2010. In recent years, more than one-third of new teachers in Haidian were hired from non-NMUs.

3. Empirical design

3.1. Data and sample

The study utilized data from two distinct sources at both the student and teacher levels within the Haidian District. The primary source is a longitudinal administrative dataset encompassing four high school cohorts spanning from 2016 to 2019, comprising 42,728 unique students and 1102 subject teachers in mathematics and literacy. Provided by the Haidian Education Department, this dataset encompasses standardized scores from high school entrance examinations alongside information from four additional tests administered during the students' senior years. Furthermore, the dataset contains essential identifiers for the schools and subject teachers of the students, facilitating linkage with the second dataset.

The second dataset used in this study is a teacher survey conducted by the China Institute for Educational Finance Research at Peking University in the fall of 2019. This survey collected information about the teachers' professional and demographic characteristics and was distributed to all math and literacy teachers in the Haidian District. Out of the 1102 teachers, 59.62% (657 teachers) responded to the survey. However, 82 of them were novice teachers who had not completed their three-year cycle and were not yet recorded in the student dataset, and 102 chose to remain anonymous or provided fake names, making it impossible to link these 184 teachers to the student administrative dataset. In the end, 473 surveyed teachers are successfully linked to their students using their names, schools, and subjects, which represent 42.92% of the Haidian teacher population. This allows us to connect 23,615 students[2] (55.27% of the total student population) to at least one teacher in the CIEFR survey, and 10,102 of these students have records for both literacy and math.

To assess whether sample attrition was related to the main predictor, the teachers' training background, and the outcome measures of their effectiveness, we conducted logit regressions. After controlling for cohort and school clusters, the results show that teacher response to the CIEFR survey is not correlated with their students' high school entrance examination scores or senior year mock test scores (see Table A1 in the appendix). Since the CIEFR survey posed the same questions to all Haidian teachers regardless of their training backgrounds, this did not seem to lead to differential rates of attrition. Consequently, it is our assessment that the likelihood of teachers choosing to participate in the survey using their real names was not affected by their educational background or by their teaching effectiveness.

3.2. Measures

Table 1 offers a detailed summary of the variables used in our analysis, categorized by teacher type. Beyond the full sample, encompassing all students that could be matched with teachers, we applied a doubly-robust method (Bang & Robins, 2005; Funk et al., 2011). This involved the use of the Coarsened Exact Matching (CEM) technique to create a pre-treatment sample with balanced characteristics,[3] herein referred to as the "adjusted sample." This step was crucial for confirming the solidity of our results. The table details variables relevant to students, like test scores, and to teachers, including professional attributes.

Test score measures: the outcome of interest in this study is the students' academic growth. Our datasets contain several high-stakes tests, all tests were designed and administered by the Haidian Education Department and graded using the same rubrics, making the scores comparable across schools and classes. For main analyses, we utilized the two most important exams, the high school entrance examination and the second mock test of the National College Entrance Exam (NCEE, also named *gaokao*), as the pre- and post-test measures for student learning outcomes. The other three tests administered in the senior year of high school also mimic the format of the NCEE, which is only held once a year and determines which colleges a student can attend.

Both tests are high-stakes and mandatory for most students, with the high school entrance examination determining which high school(s) a student can be admitted to and the second mock test serving as the closest estimate of students' final performance in the

[2] Of the students matched and enrolled in this district, only 27 did not participate in the mock tests, representing a negligible fraction of 0.1%. Thus, concerns regarding student attrition over the three-year period should be minimal.

[3] For additional information regarding the sample adjustment process and the outcomes obtained from the adjusted sample, please see Appendix 1. This strategy is used for robustness checks.

Table 1
Descriptive statistics of analytic sample.

	Non-NMU Teachers		NMU Teachers	
	Mean	S.D.	Mean	S.D.
Test scores				
High school entrance exam	0.302	(0.80)	−0.024	(1.01)
First test in grade 12	0.313	(0.84)	−0.025	(1.01)
Second test in grade 12	0.248	(0.90)	−0.020	(1.00)
Mock test 1	0.280	(0.85)	−0.022	(1.01)
Mock test 2	0.302	(0.88)	−0.025	(1.00)
Elite class	11.76%	(0.32)	11.79%	(0.32)
Teacher characteristics				
Male	35.90%	(0.49)	23.88%	(0.43)
Advanced degree	82.05%	(0.39)	35.22%	(0.48)
Teaching experience	10.425	(8.31)	19.568	(8.42)
Subject				
Chinese	56.76%	(0.50)	50.59%	(0.50)
Math	43.24%	(0.50)	49.41%	(0.50)
Teacher survey outcomes				
Course load (log)	2.560	(0.37)	2.402	(0.32)
Course preparation[a]	87.18%	(0.34)	82.03%	(0.38)
School-based learning activity[a]	30.77%	(0.47)	23.88%	(0.43)
After-school direction[a]	25.64%	(0.44)	21.51%	(0.41)
Communication with parents[a]	17.95%	(0.39)	17.26%	(0.38)
Pedagogy application[b]	−0.340	(0.96)	0.031	(1.00)
Teaching strategy[b]	−0.265	(0.94)	0.024	(1.00)
Teaching design awards	66.67%	(0.48)	84.40%	(0.36)
Teaching research awards	64.10%	(0.49)	87.00%	(0.34)
Teaching quality awards	28.21%	(0.46)	60.05%	(0.49)
Career future time perspective[b]	0.103	(0.93)	−0.009	(1.01)
Professional identity[b]	0.066	(0.69)	−0.006	(1.02)
Sense of belonging[b]	0.026	(0.95)	−0.002	(1.01)
Job burnout[b]	0.018	(0.78)	−0.002	(1.02)
Work pressure[b]	0.074	(0.73)	−0.007	(1.02)
Number of unique teachers	40		424	

[a] Working time indicator, with 1 indicating spending more than 5 hours per week on it and 0 indicating less.
[b] Factor scores extracted from survey items.

formal standardized college entrance tests.[4] This mock test is a critical reference for students to evaluate their potential rankings in Beijing and inform their choices of institutions to which they apply. Moreover, performance on these mock exams is instrumental in qualifying students for the independent recruitments of selected elite universities, allowing high achievers to gain acceptance to one or a few elite universities with lower NCEE scores. The high-stakes nature of the mock test is further evidenced by its exceptionally high participation rate, with only 37 students not taking the exam, representing a mere 0.09%. Given their significance in academic and future educational trajectories, the results of these tests are robust indicators of students' learning capabilities and academic prowess.

It should be noted that the Chinese public school system implements the “looping” mechanism, where students are assigned to the same teachers throughout their high school education levels. In high schools, teachers complete a three-year loop and most students have only one teacher per subject throughout their education. Therefore, we defined the difference between the subject-specific scores of the high school entrance examination and the second mock test of the college entrance examination as the three-year learning attainment for that subject. To ensure comparability, we standardized the test scores to have a mean of zero and a variance of one.

Teacher characteristics: we also gathered information about the teachers' professional and demographic characteristics from the CIEFR survey, which includes whether they graduated from a normal university or not. Despite the growing presence of non-NMU teachers due to policy shifts and market trends, approximately 90% of teachers in our study obtained their degrees from NMUs, underscoring the continuing prominence of NMUs in teacher education. While the survey lacks direct indicators of the quality of pre-service training or individual competence, it reveals that non-NMU teachers generally possess higher academic qualifications than their NMU counterparts.

Teacher survey outcomes: in addition to basic professional and demographic characteristics, we also generated a series of measures regarding time allocation, teaching behaviors, and professional mental properties from the survey. These measures are used in supplementary analyses to explore potential mechanisms for differentiated teacher effectiveness. Specifically, the first five measures indicate the courses that teachers instruct each week and whether they spend more than five hours per week on these tasks. The next five measures depict their teaching behaviors. “Multifarious pedagogy” is extracted from a series of pedagogical frequency items that teachers use in classrooms, and “teaching strategy” is derived from items that describe how teachers emphasize and inspire student

[4] The data of the college entrance exam is not available from the district education department. It is confidential at the Beijing Ministry of Education.

learning strategically. The following three dummies indicate if teachers have won any awards in three aspects of teaching. The last five measures outline several typical professional psychological traits.

3.3. Analytical methods

We use value-added models (VAMs) to estimate teacher effectiveness and then examine how much of the difference in value-added among teachers can be attributed to teacher's educational backgrounds. VAMs have been broadly applied in an array of teaching effectiveness evaluations of teacher training programs (Goldhaber et al., 2013; Plecki et al., 2012), professional development programs (Biancarosa et al., 2010), and coaching programs (Harris & Sass, 2011). There are two types of VAMs: gain-score and lagged-score. While both have been used widely, lagged-score VAM is considered a better specification under a variety of estimation conditions (Andrabi et al., 2011), and is the norm in regions using VAMs for teacher evaluations (Chetty et al., 2014; Koedel et al., 2015).

As discussed earlier, we recognize that there might be bias from unmeasured confounders at the student-level influencing student-teacher assignment and test scores and adopt a cross-subject model based on previous research on similar topics. The cross-subject model was first introduced by Clotfelter et al. (2010) who included student dummy variables to address the sorting issues and compute teacher credential effects on academic achievement, and thence was widely used to compare teacher effectiveness. For instance, Xu et al. (2011) leveraged the model to gauge TFA teachers' effectiveness with student fixed effects, Schwerdt and Wuppermann (2011) employed the between-subject variation to control for unobserved student traits when comparing different teaching styles, and Hanushek et al. (2019) recently measured the effects of teacher cognitive skills by exploiting between-subject variation in teacher skills.

Previous studies reveal that students perform well in one subject are also likely to perform well in other subjects (Clotfelter et al., 2010; Xu et al., 2011), and schools typically assign students by a single dimension (Hanushek et al., 2019; Schwerdt & Wuppermann, 2011). Therefore, our cross-subject value-added model assumes that student's performances in different subjects are functions of the same underlying student ability, and that the student-teacher sorting would not be more in one subject than another.[5] To address similar challenges related to sorting, we introduce our empirical model as follows:

$$Y_{ijt} = \lambda_0 + \lambda_1 Y_{ij(t-3)} + \lambda_2 NMU_{ijt} + \lambda_3 T_{ijt} + \lambda_4 C_{ijt} + \mu_i + \varphi_j + e_{ijt} \quad (1)$$

where Y_{ijt} denotes the scores on the second mock test in subject j for student i in cohort t. The variable of interest, NMU_{ijt}, indicates whether the teachers are traditionally trained in a normal university. $Y_{ij(t-3)}$ refers to the high school entrance examination scores. T_{ijt} is a vector of teacher characteristics, including gender, education level, and teaching experience, with a quadratic term for teaching experience to account for potential non-linear effects as identified in previous studies (Papay & Kraft, 2015). C_{ijt} comprises classroom-level predictors specific to each subject, incorporating baseline mean scores and an indicator for classrooms that might be considered elite due to their significantly higher pre-treatment mean scores.[6] This approach assumes that these classes may receive systematically different treatment. Lastly, μ_i captures all unobserved student characteristics through dummy variables, and φ_j represents subject-specific fixed effects, and e_{ijt} is the idiosyncratic error term.

With these statistical adjustments in place, biases arising from students' preferences for specific types of teachers should be substantially mitigated. Additionally, using high school entrance examination scores as the pre-test scores can reduce bias, as they are a good measure of students' previous academic abilities which can influence their academic performance towards the end of high school. Studies have found that student-teacher sorting based on parental or family characteristics that are not justified in VAMs is limited (Chetty et al., 2014). A bare model with only one lagged test score and a full model with sufficient student demographic features and multiple lagged test scores both produce highly correlated estimates with a correlation index of over 0.9 (Ehlert et al., 2014).[7] Therefore, even if not all biases are excluded, the remaining should be minimal.

[5] Conditioning on student fixed effects aims to minimize bias from students' general preference for a particular type of (non-)NMU teachers, yet it may not fully address subject-specific sorting concerns. In addressing potential concerns regarding student preferences for specific (non-)NMU teachers in certain subjects, it is important to note that students may sort to particular classrooms but typically do not have the option to select their teachers, particularly not based on the (non-)NMU status across all subjects. To empirically test for such sorting, we analyzed whether students were assigned to one non-NMU teacher and one NMU teacher, using their pre-test scores as predictors within school-cohort-stream clusters. The results reveal no evidence of subject-based sorting tied to students' high school entrance exam scores, with a coefficient of "-0.0005" and a standard error of "0.003" ($p = 0.87$). This finding underscores the absence of specific subject sorting influenced by students' academic performance at entry.

[6] Please refer to Appendix 1 for details of how we identified potential elite classrooms.

[7] Potential concern may be raised regarding the assumption that prior test scores adequately address student-teacher sorting—a premise accepted in U.S. contexts—may not directly apply to China. However, it is important to underscore that our statistical models, which incorporate student fixed effects, are meticulously designed to mitigate biases arising from students' preferences for certain types of teachers. Furthermore, the looping mechanism inherent in the Chinese educational system significantly limits the scope for students to express preferences for specific subjects, further reinforcing the validity of our approach.

4. Findings

4.1. VAMs results

Table 2 presents the outcomes of the stepwise analysis, with the first four columns utilizing the full sample and the final two columns showcasing results for robustness and external validity using adjusted sample and supplementary data.

The analysis initiates with a basic model, devoid of additional teacher characteristics, school, and student fixed effects. This preliminary result indicates that teachers from normal universities contribute significantly more to student test scores, by an average of around 0.1 standard deviations. The robustness of this finding persists even with the inclusion of more covariates and fixed effects, maintaining the 0.1 standard deviation difference. Addressing the potential for student-teacher sorting within schools, the choice between school-stream-cohort FE or student FE influences the magnitude of the estimates slightly. Incorporating student FE slightly reduces the NMU teachers' coefficient, yet it remains statistically significant, hovering around 0.1 standard deviations. Moreover, the results hold steady when applying a doubly-robust method with a CEM adjusted sample, as shown in column (5), which features more balanced pre-treatment characteristics. Notably, this estimate stands at 0.095, marginally a little bit higher but still closely aligned with the full sample's outcomes presented in column (4). The results of column (6) will be discussed in Section 4.5.

4.2. Robustness tests

To evaluate the stability of our findings, we engaged in a series of sensitivity analyses, employing various model specifications. The first segment of this analysis focused on the robustness across all potential models derived from the main analysis variables. The subsequent segment investigated the sensitivity of our findings using different outcome measures and interactions to address specific econometric concerns.

Fig. 2 affirms the robustness of our results, displaying the range of model specifications applied to the full sample. Across models incorporating school-cohort-stream FE or student FE, all estimates remained positive, consistently averaging around a 0.1 standard deviation difference. Importantly, each combination of models revealed a statistically significant positive impact, distinct from zero, in presence of the school-cohort-stream FE. While some models with student FE show confidence intervals intersecting with zero, the overall pattern remains stable with estimates spanning from 0.08 to 0.1 standard deviation. It should be noted that models yielding insignificant results are likely omitting crucial confounders, such as teaching experience or classroom-subject cluster variables. As indicated in Table 2, with a comprehensive condition of all covariates and fixed effects, the results are significantly positive. Few insignificant estimates, particularly among the smaller subset of non-NMU teachers, may result from limited statistical power, suggesting a need for a larger sample in future research. Analyses using the CEM adjusted sample yield very similar results, reinforcing the consistency of the NMU teachers' value-added effect on student test scores, with details provided in the appendix.

Table 3 conducts additional robustness tests by employing alternative metrics or by formulating models inclusive of interactions. The initial rows alter the dependent variable to Mock Test 1 and the first test conducted in the senior year, both administered within the final year of school, while maintaining consistency with other covariates and fixed effects. Although both estimates show a positive impact, the magnitude of the effect is reduced. This reduction in effect size is sensible considering the timing of these tests, particularly the first Grade 12 test, which occurred nearly a year prior to the second mock test. In the third row, incorporating multiple tests from the senior year as control variables reflects the recommendation from previous research to utilize pre-test scores from multiple years (Goldhaber & Hansen, 2013; Koedel & Betts, 2010). Consequently, the coefficient becomes statistically insignificant, indicating that our primary variable of interest specifically assesses the value-added throughout the senior year only.

The analysis in the fourth row examines potential heterogeneous effects. As per Lin (2013), introducing interactions between the treatment of interest, NMU teachers, and other covariates, excluding student dummies, does not detract from model precision. Rather, it enables an evaluation of the average value-added by NMU teachers amidst varying levels of effectiveness. And the results are consistent with the main analysis.

In summary, our investigation consistently highlights that teachers from normal universities enhance students' academic performance, with a quantifiable added value averaging around 0.1 standard deviation over three years. This positive impact on effectiveness is robust, significant, and maintains across various models. Despite occasional variances and some non-significant outcomes across replication attempts, such discrepancies are understandable considering potential confounder omissions and the intrinsic differences in time length. Collectively, our research substantiates the substantial advantage conferred by NMU-trained educators in elevating student test scores, underscoring their significance to learning growth.

4.3. Heterogeneity analyses

Considering the favorable reputation surrounding of non-NMU teachers and existing literature indicating that teachers' impacts may vary among students of different ability levels (Penner, 2016), this study investigates whether non-NMU educators are especially effective for students with higher pre-test scores, who typically exhibit greater independence in learning. These students may benefit more from teachers who possess extensive subject knowledge rather than pedagogical skills. Furthermore, considering the increasing prominence of non-NMU teachers over the past decade, it is crucial to examine whether newly hired non-NMU teachers demonstrate enhanced effectiveness, suggesting a potential interaction between the experience of teachers and their effectiveness. To address these queries, we incorporate interaction terms related to NMU teacher status in our regression models, seeking to elucidate the complex relationships between teacher training backgrounds, student academic readiness, and teacher experience on student academic

Table 2
VAM results of NUM teacher effectiveness.

	(1)	(2)	(3)	(4)	(5)	(6)
NMU teacher	0.096**	0.098**	0.141***	0.084*	0.095*	0.116*
	(0.045)	(0.049)	(0.045)	(0.051)	(0.055)	(0.067)
Pre-Test	0.206***	0.206***	0.205***	0.252***	0.254***	0.444***
	(0.014)	(0.014)	(0.014)	(0.012)	(0.018)	(0.007)
Other teacher characteristics						
Male		−0.058	−0.043	−0.039	−0.070*	−0.051***
		(0.039)	(0.033)	(0.032)	(0.039)	(0.012)
Advanced Degree		−0.118***	−0.059	0.031	0.028	0.008
		(0.045)	(0.039)	(0.038)	(0.047)	(0.018)
Teaching experience		−0.010**	−0.001	0.000	−0.004	−0.001
		(0.004)	(0.004)	(0.004)	(0.005)	(0.002)
Teaching experience^2		0.000	0.000	0.000	0.000	−0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Classroom characteristics						
Class Average Pre-Test	0.649***	0.662***	0.406***	−0.033	−0.206*	0.064**
	(0.033)	−0.033	−0.053	−0.07	(0.123)	(0.026)
Elite Class	0.124**	0.125**	0.217***	−0.018	−0.003	0.008
	(0.050)	−0.048	−0.043	−0.052	(0.056)	(0.012)
Subject FE	√	√	√	√	√	√
School-Stream-Cohort FE			√			
Student FE				√	√	√
A-R^2	0.365	0.368	0.42	0.589	0.463	0.776
N	16,410	16,410	16,410	16,410	9160	83,338

Note: The observations pertain to subject-specific individuals, and standard errors, which are clustered at the classroom-subject level, are presented in parentheses. The first four columns utilize the full sample, comprising all students who could be paired with teachers from the survey. The fifth column presents the doubly-robust results derived from an adjusted sample, which has been refined using a coarsened exact matching approach. The results in the last column originate from supplementary data sourced from a capital city in Northeast China.
* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

outcomes.

Fig. 3 displays interaction plots comparing NMU teacher effects across student pre-test scores and teachers' years of experience. Notably, there appears to be no significant interaction between NMU teachers and either the foundational academic abilities of their students or their duration of teaching experience. In terms of pre-test scores, the trajectories remain parallel, indicating that NMU teachers consistently contribute greater value to their students' learning outcomes, irrespective of the students' initial academic levels, when compared to their non-NMU counterparts. Regarding teaching experience, although the trajectories converge slightly as experience accrues, they do not intersect within a meaningful range, suggesting NMU teachers maintain superior effectiveness regardless of their years of being a teacher.

Nonetheless, considering that students within the same schools are expected to have relatively similar pre-high school academic performances due to selection processes like the high school entrance exam, using the absolute pre-test scores may not adequately reflect the nuanced variances among students within smaller groupings such as classrooms or schools. In every school, there exist both higher and lower-performing students. To address this, we embarked on interaction analyses using students' pre-test rankings (percentiles) within specific clusters, the results of which are depicted in Fig. 4. Consistent with previous findings, no interaction was observed concerning students' relative academic standings within their classrooms or school-cohort clusters. NMU teachers consistently deliver higher value-added to student test scores, regardless of the students' comparative academic positions within their more immediate educational environments.

Table 4 presents the coefficients from these interaction regressions, corroborating the observations from Figs. 3 and 4. The interaction terms in all the columns are not only statistically insignificant but also negligible in effect size. Summarily, NMU teachers demonstrate a consistently higher impact on student test scores, unaffected by variations in students' pre-test or teaching experience.[8]

4.4. Disparity in teacher behavior

This section aims to understand the reason behind the higher effectiveness of NMU teachers by exploring potential mechanisms. To do this, we use teacher behavioral measures from the CIEFR teacher survey, as presented in Table 1. We run separate regression models

[8] Observing the notable increase in non-NMU teachers since 2010, as illustrated in Figure 1, concerns may arise regarding whether the profile and effectiveness of non-NMU teachers have undergone systematic changes during this period. To explore this possibility, we undertook supplementary analyses, segmenting our data into subsamples based on whether teachers commenced their careers before or after 2010. Additionally, we employed heterogeneity analyses incorporating interaction terms. The outcomes for the full sample are presented in Table A5 within Appendix 2. Our analysis reveals a consistent level of higher effectiveness among NMU teachers, irrespective of their career start date relative to 2010. We also conducted similar analyses for the adjusted sample. The results are similar and are documented in Table A4.

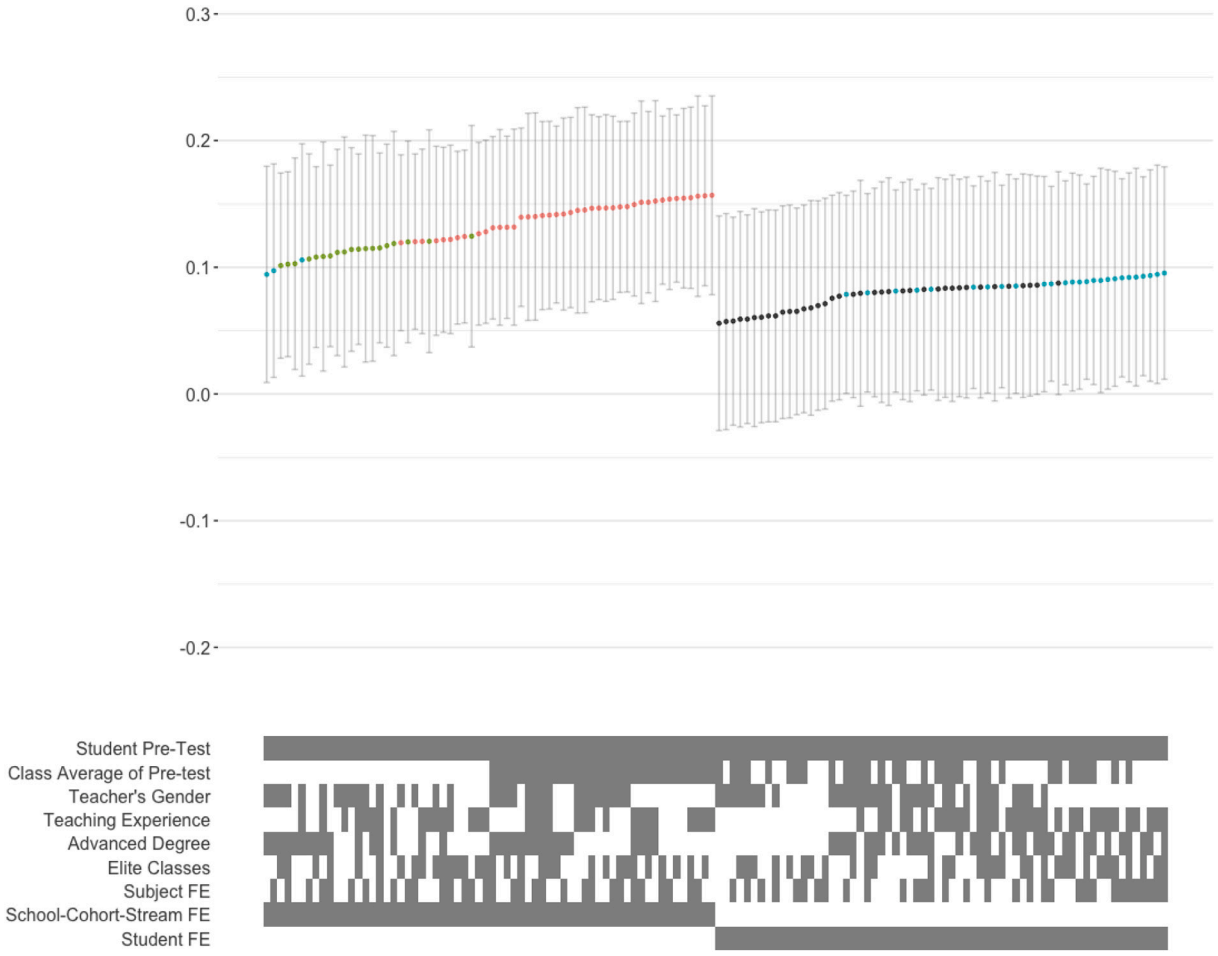

Fig. 2. Coefficient plots of NMU teacher effectiveness (sensitivity check).
Note: This chart shows the full sample analyses including all the students that could be matched with the teacher survey for all possible combinations of covariates and fixed effects, with estimates for NMU teachers indicated by solid nodes and 90% confidence intervals shown by gray line segments. The top panels display the estimates of NMU teachers for each model, while the bottom panels show the corresponding control variables. Standard errors are clustered at the classroom-subject level. The orange, green, and blue nodes indicate statistical significance levels of $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Robustness checks using additional specifications (full sample).

	β	S.E.
1. Mock test 1 as the outcome	0.077**	(0.032)
2. First test in grade 12 as the outcome	0.046	(0.062)
3. Condition on multiple pre-tests	0.065	(0.040)
4. Add interactions of covariates and NMU teacher	0.104***	(0.066)

Note: Standard errors are clustered at the classroom-subject level. The sample compromises all students who could be paired with teachers from the survey. For rows 1 and 2, covariates align with those used in the main effect value-added models presented in Table 2, but the dependent variables differ. Mock test 1, which precedes mock test 2 by two months, and the first test in grade 12, conducted at the onset of the final high school year, are used as outcomes. For rows 3 and 4, the outcome values correspond to those in the primary analysis, yet the estimations are expanded to incorporate additional predictors, specifically multiple pre-treatment values (other three tests collected in their senior years), as well as all interactions between covariates and the central predictor, the NMU teacher, to account for potential heterogeneity.
* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

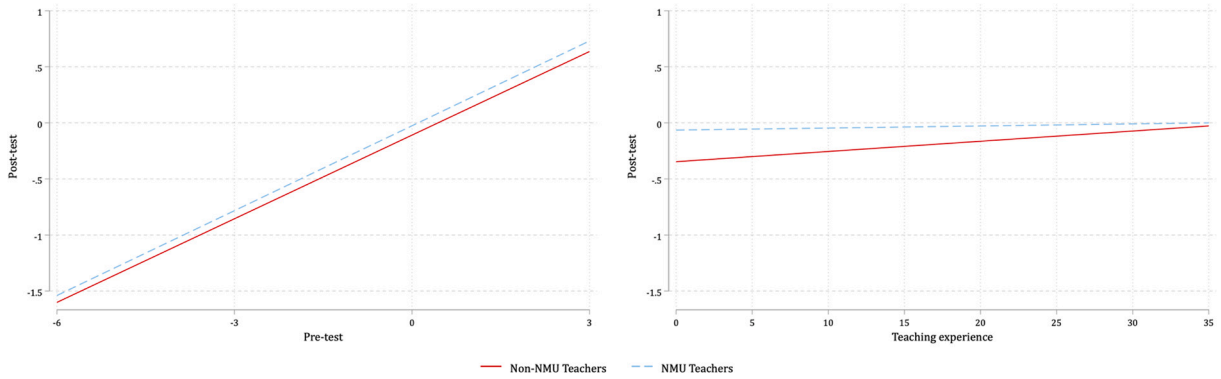

Fig. 3. Interaction plots of NMU teacher effectiveness by student's pre-test scores and teacher experience.

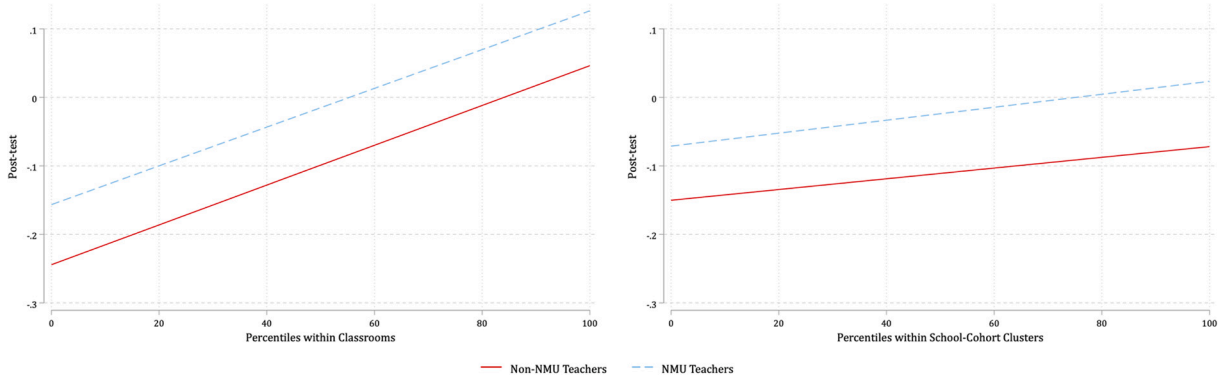

Fig. 4. Interaction plots of NMU teacher effectiveness by students' ranking within classroom and school-cohort clusters.

to predict each behavior using unique teacher observations. This allows us to compare the differences in behavior between teachers with different training.

Fig. 5 delineates the regression analyses concerning teacher behaviors aiming to illustrate the potential mechanism of higher value-added of NMU teachers. The findings indicate no significant differences in time allocation and professional psychological traits between teachers of different training backgrounds. Yet, it is observed that NMU teachers excel in diverse pedagogical practices, measured by the adoption of varied teaching methods, a focus on learning objectives, and the ability to influence students' attitudes towards learning—factors all positively correlated with academic outcomes (Gollwitzer & Sheeran, 2006; Lau & Lam, 2017). Furthermore, NMU teachers are more frequently recognized with awards for their outstanding instructional design and teaching quality. These observations are in line with existing literature, underscoring the benefits of conventional teacher education in enhancing pedagogical skills and readiness (Caires & Almeida, 2005; Carroll et al., 2003; Goker, 2006). While non-NMU teachers may hone their instructional skills over time, our analysis does not show them engaging more in professional development than their NMU counterparts. This suggests that experiential learning alone may not adequately bridge the gap created by a lack of formal educational training.

4.5. *External validity of the findings*

The limited sample size of non-NMU teachers in our primary analysis might impede the generalizability of our conclusions to larger population. To mitigate this concern and enhance the representativeness of our findings, we expanded our explorations to include supplementary data from two additional sources. These sources provide a larger pool of non-NMU teacher observations from other contexts, allowing for a more comprehensive validation of our results within the broader context of China's educational landscape.

The first dataset follows similar pattern of the Haidian data and consists of administrative records of student test scores from Grades 1 to 9 compulsory education schools across spring and fall semesters in 2023, paired with a teacher survey conducted at the end of the 2023 fall semester. This survey and test data were collected in a capital city in Northeast China, with the tests being the biannual final

Table 4
VAM results of heterogeneity analysis (full sample).

	(1)	(2)	(3)	(4)
NMU teacher	0.083	0.137**	0.084*	0.087*
	(0.056)	(0.668)	(0.051)	(0.051)
Pre-Test	0.249***	0.252***	0.137***	0.216***
	(0.044)	(0.012)	(0.023)	(0.032)
Pre-test * NMU teacher	0.004			
	(0.045)			
Teaching experience	−0.000	0.009	−0.000	−0.000
	(0.004)	(0.021)	(0.004)	(0.004)
		−0.007		
Teaching experience * NMU teacher		(0.022)		
Teaching experience^2	−0.000	−0.001	−0.000	−0.000
	(0.000)	(0.001)	(0.000)	(0.000)
		0.000		
Teaching experience^2 * NMU teacher		(0.001)		
Percentile in classrooms			0.003***	
			(0.001)	
Percentile in classrooms * NMU teacher			−0.000	
			(0.001)	
Percentile in school-cohort				0.001
				(0.001)
Percentile in school-cohort * NMU teacher				0.000
				(0.001)
Other covariates	√	√	√	√
Subject FE	√	√	√	√
Student FE	√	√	√	√
Adjusted R^2	0.589	0.590	0.591	0.590
N	16,410	16,410	16,410	16,410

Note: Standard errors, clustered at the classroom-subject level, are presented in parentheses. The sample compromises all students who could be paired with teachers from the survey. All models are conditioned on the covariates utilized in the primary analysis. For the last two columns, pre-test rankings have also been included as part of the analysis.
* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

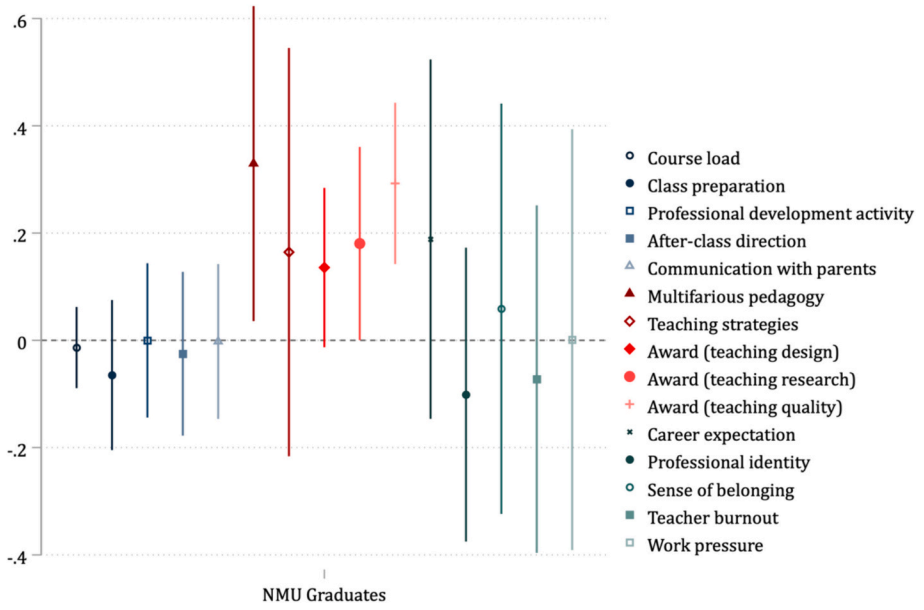

Fig. 5. Coefficient plots of teacher behavioral differences by NMU-teacher. Note: The nods suggest the coefficients of NMU teacher, and the bar denotes the 90% confidence intervals. In addition to the predictor of NMU teacher, teacher's age, gender, educational level, the subject they teach, and school fixed effects are included. We took the negative value of teacher burnout, so higher scores denote less burnout. Standard errors are clustered at the school level.

examinations administered by the city's education department. For the purposes of VAM analysis, we included students from grades 5, 6, 8, and 9, who had test scores from both periods under review.[9] Ultimately, our analysis focused on 20,100 unique students who could be matched with at least one subject teacher, accounting for 83.3% of the total student population. Of the 1119 teachers who completed the survey, 176 (15.7%) held degrees from non-normal universities. This resulted in 1982 classroom-subject clusters being analyzed out of a total of 3876.

The second dataset encompasses administrative records for all grade 9 students and their teachers from public schools in five counties of a province in Northeast China, collected by the local education department in fall 2021. It aggregates student test scores in the three core subjects (Chinese, English, and Mathematics) at the school-level, offering both mean and median values for each subject from 74 junior middle schools, accounting for 10,693 unique students in total. The dataset also details 593 unique educators, with 529 graduated from normal universities, and 10.8% as non-NMU teachers.[10] The independent structure of the student and teacher data sets, preventing a direct match, necessitated the aggregation of teachers' background information at the school-subject level, presented as percentages. It is important to note that there are no pre-test scores available in this dataset.

The outcomes derived from the first dataset are illustrated in the final column of Table 2, while the findings from the second dataset are showcased in Table 5. For the first dataset, NMU teachers demonstrate a significantly higher value-added to student test scores—approximately 0.12 standard deviations over three years,[11] closely paralleling the results of the main study. These conclusions are reached under the same covariates and fixed effects as employed in the main analysis.

Likewise, findings from the second supplementary dataset corroborate the advantages of employing more NMU teachers. Specifically, an increase of 10% in NMU teacher presence leads to an elevation in school-subject mean scores by approximately 0.4 standard deviation. When evaluating through ranking indicators, NMU teachers show a positive correlation with school standings within these five counties. Controlling for other variables, schools staffed entirely by NMU teachers are ranked higher by 4 positions (or 6 percentiles) compared to those without NMU educators. Although estimates for median score-related metrics (median scores within school-subject clusters, rankings by median, and percentiles by median) are not statistically significant, they trend in the same direction, favoring the effectiveness of NMU teachers. These supplementary findings bolster our primary analysis, suggesting that the observed benefits of NMU teachers have strong external validity and can be extrapolated to a broader scope within China's K-12 education system.

5. Discussion and conclusion

While non-traditionally trained teachers are gaining popularity in China's developed regions, our findings diverge from this trend, showing that NMU-trained educators specializing in teacher education are more effective in boosting high school students' high-stake test scores over three years, by an average of 0.1 standard deviation. This aligns with recent research endorsing NMU teacher effectiveness in diverse settings, including national middle school samples and rural areas (Shi et al., 2022), and is further validated by supplementary data from both a capital city and five counties in a Northeast province.[12] Our results contrast with studies positing that alternative-pathway teachers can match or surpass the performance of their traditionally trained counterparts in challenging educational environments (Clark & Isenberg, 2020; Whitford et al., 2018; Xu et al., 2011).

Our study is set in a competitive district-level teacher labor market in China, distinct from the scenarios in much of the existing literature, which often focuses on areas facing teacher shortages and consequently hiring less professionally trained educators. In contrast, developed regions in China, including Haidian, have attracted highly educated graduates from elite research universities into teaching, prioritizing cognitive abilities as a key metric of teacher quality. This approach is supported by research linking teachers' cognitive skills to student performance (Clotfelter et al., 2010; Hanushek et al., 2019; Jackson, 2012), suggesting that the cognitive skills of teachers, alongside their training, are crucial for educational outcomes. Hence, the success of programs like TFA in previous studies (Backes & Hansen, 2017; Xu et al., 2011) might stem from the cognitive strengths of non-traditionally trained teachers over the systematic pre-service training of traditionally trained teachers who are employed in such settings where the teacher selection is less rigorous and schools serve underprivileged communities.

Our study stands out from previous research by evaluating teacher effectiveness among teachers with similar backgrounds in a highly competitive teacher labor market. Although lacking direct data on cognitive abilities, it is noted that NMU entry scores are generally lower those for non-NMUs (Han & Xie, 2020; Zhao, 2018). Additionally, hiring data indicates that non-NMU teachers often graduate from China's top universities, surpassing NMU teacher selectivity (Zhang, 2021). Previous studies have also found that teachers from non-NMUs have at least similar levels of educational backgrounds to their NMU colleagues in China (Zhou, 2010). In our sample, a significantly larger fraction of non-NMU teachers (82.05%) have postgraduate qualifications compared to NMU educators (35.22%), highlighting their superior academic abilities. Thus, while our findings might not directly reflect on traditional teacher

[9] Grades 1 to 4 were excluded due to the absence of written tests, and grade 7, being newly enrolled in the fall semester of 2023, lacked baseline test scores for comparison.

[10] In addition to the grade 9 educators, the teacher dataset extends to include all teachers from other subjects and grades 7 and 8, totaling 3633. Among these, 12.1% (438 teachers) are identified as non-NMU teachers.

[11] The supplementary data calculates value-added for one semester. For consistent comparison with the primary analysis, the binary indicator of NMU status is adjusted by dividing by 6, rendering the coefficient for NMU teachers as reflective of a 6-semester, or three-year, value-added.

[12] We verified the representativeness of our Haidian sample against external databases and the study by Shi et al. (2022), which notes that non-NMU teachers comprise 10% to 15% of the teacher population nationally.

Table 5
Supplementary analysis of middle school performance from five counties.

	Mean	Median	Ranking by mean	Ranking by median	Percentile by mean	Percentile by median
NMU teacher %	0.039*	0.034	−0.042**	−0.021	−0.060**	−0.030
	(0.022)	(0.036)	(0.019)	(0.022)	(0.027)	(0.030)
Age	−0.144	−0.355*	0.039	0.262	0.068	0.377*
	(0.109)	(0.189)	(0.107)	(0.162)	(0.152)	(0.227)
Female %	−0.009	0.005	−0.015	−0.044*	−0.023	−0.063*
	(0.024)	(0.034)	(0.016)	(0.026)	(0.022)	(0.036)
Bachelor or higher degree %	0.022	0.026	−0.033	−0.040	−0.044	−0.054
	(0.028)	(0.040)	(0.026)	(0.028)	(0.036)	(0.038)
Teaching experience	0.187**	0.330**	−0.060	−0.190*	−0.092	−0.272*
	(0.093)	(0.158)	(0.081)	(0.114)	(0.113)	(0.159)
Full-time teacher %	0.051	0.070	−0.023	−0.017	−0.028	−0.019
	(0.061)	(0.088)	(0.039)	(0.046)	(0.053)	(0.063)
Senior teacher %	−0.003	0.021	0.009	0.008	0.011	0.009
	(0.023)	(0.036)	(0.022)	(0.026)	(0.030)	(0.036)
Student teacher ratio	0.065	0.114	0.027	−0.015	0.041	−0.017
	(0.056)	(0.082)	(0.039)	(0.043)	(0.053)	(0.059)
Number of subject teachers	1.621**	2.800***	0.498	−0.841	0.718	−1.107
	(0.740)	(1.046)	(0.599)	(0.663)	(0.825)	(0.911)
Non-teaching work hours per week	−0.011	−0.068	−0.009	0.088	−0.020	0.115
	(0.065)	(0.089)	(0.053)	(0.088)	(0.072)	(0.121)
Teaching load per week	−0.225	−0.226	0.153	0.031	0.209	0.041
	(0.210)	(0.298)	(0.128)	(0.165)	(0.177)	(0.228)
School FE	√	√	√	√	√	√
Subject FE	√	√	√	√	√	√
A-R^2	0.911	0.865	0.946	0.920	0.946	0.920
N	217	217	217	217	217	217

Note: Robust standard errors are presented in parentheses, with regressions weighted by the number of students who took the tests. The sample comprises data collected at the school-subject level, encompassing three main subjects—Chinese, Mathematics, and English—in 74 distinct middle schools. This dataset represents 10,694 unique students and 593 unique teachers, of which 64 are non-NMU teachers. Percentage-based predictors are multiplied by 100 to illustrate the effect of a one-percentage-point change. Additional covariates represent average values within school-subject clusters.
* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

training's impact, they likely present a conservative estimate of the teacher preparation effects, considering the overall higher (or at least similar) academic abilities of non-NMU teachers within the same districts.

Another strength of our study is our analysis of the corresponding work behaviors of teachers based on their educational backgrounds. To our knowledge, none of the previous studies suggesting that non-traditionally trained teachers have higher value-added on student test scores has comprehensively evaluated teacher behaviors accordingly. In our study, we find that while there is no difference in working time allocation or professional mental traits between the two types of teachers, NMU teachers are more proficient in applying diverse pedagogy and teaching strategies in classrooms. They are also more likely to win teaching awards regarding different aspects of their teaching performance. These findings are important in understanding the differences in teacher effectiveness in Haidian, as they are consistent with previous studies that focus on teacher training and teaching skills (Carroll et al., 2003; Darling-Hammond, 2006; Ronfeldt, 2012). Had previous studies conducted similar analyses, they could have found, as we did, that teachers with higher value-added have advantages in certain teaching behaviors and skills. Without this evidence, their findings may lead to misunderstandings regarding the value of intense teacher training.

Our results highlight the higher value-added of traditionally trained teachers in highly-competitive schools. However, given the trend of teacher source diversification worldwide and the prevalent issue of teacher shortage particularly in low-income regions, we do not suggest that schools should stop hiring teachers from alternative routes. Instead, we emphasize the irreplaceable value provided by traditional teacher education. We suggest that comprehensive universities should provide more corresponding curriculum and teaching practice opportunities for their students who want to be teachers to learn and grow effectively, similar to traditional teacher training institutes and programs. This is especially important in situations where top-performing graduates are willing to enter the teaching profession, as is the current case in China. Unlike the alternative pathways in other countries, such as TFA, which demand a certain period of teacher training and orientation (Clark & Isenberg, 2020; Xu et al., 2011), little training is provided in current Chinese non-normal research universities (Li et al., 2021). Combined with the fact that the teacher certification system in China only requires a knowledge test and an interview, this may be the reason behind the lower effectiveness of non-NMU teachers. Therefore, we suggest that stricter qualification requirements be implemented to ensure that certified teachers possess the necessary educator skills.

Our study acknowledges certain limitations that future research should aim to address. One potential concern is about the data limitations that prevented us from tracking changes in teacher and classroom assignments throughout the three-year study period, potentially introducing bias. However, the student-teacher sorting observed in the Haidian District provides a solid foundation that mitigates these concerns. In Haidian, students scoring higher in entrance exams are more likely to be taught by non-NMU teachers. This pattern is significant in light of the “Matthew effect” in learning, where students with lower initial abilities tend to fall further behind over time—a well-documented global trend (Chatterji, 2006; Heckman & Landersø, 2022; Kim et al., 2010; Morgan et al., 2008). This

phenomenon implies that assigning non-NMU teachers to academically stronger students might lead to an underestimation of NMU teachers' effectiveness, which would likely support rather than undermine our conclusion regarding NMU teacher effectiveness. Therefore, despite the outlined limitations, the fundamental conclusions of our study—that NMU teachers are highly effective—remain intact and are not compromised by these methodological and data constraints. Future studies, if possible, could gain more nuanced insights from higher quality datasets.

Second, while our dataset encompasses roughly half of the literacy and math educators in the Haidian District, it features only 40 teachers graduating from non-normal universities. Although the fraction of non-NMU teachers and findings align with previous research (Shi et al., 2022) and supplementary data within similar and varied Chinese contexts, the relatively modest sample size limits our capacity for deeper explorations into heterogeneity, such as subject-specific effects and the impact of specific educational institutions teachers graduated from. Future research would benefit from a broader sample of teachers and more background information to facilitate a more granular analysis.

In summary, this research adds valuable insights to the current debate on the impact of non-traditionally trained teachers on student outcomes in a school district that has been the subject of significant public discourse. Despite the constraints related to sample coverage, our findings offer significant implications for policymakers to consider the potential advantages of comprehensive teacher training amidst the policy shift towards diversifying the teaching workforce. As an initial foray into this area of research, our study underscores the need for further evidence across various contexts, encouraging a broader examination of the efficacy of teacher education programs.

Data availability

The authors do not have permission to share data.

Acknowledgement

This research is supported by the National Office for Education Sciences Planning, China [Grant No. BFA201173], received by Dr. Yi Wei. We have no conflicts of interests to disclose. We are grateful for the constructive feedback from two anonymous reviewers. We extend our thanks to Mark Johnson and Clifton Conrad at UW-Madison, and members of CEFP for their insightful advice, especially Xuehui An, Jian Zou, Jing Liu, Yang Song, Jinqing Liu, Rui Wang, and Chuanyi Guo.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chieco.2024.102225.

References

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics, 3*(3), 29–54. https://doi.org/10.1257/app.3.3.29

Backes, B., Goldhaber, D., Cade, W., Sullivan, K., & Dodson, M. (2018). Can UTeach? Assessing the relative effectiveness of STEM teachers. *Economics of Education Review, 64*, 184–198. https://doi.org/10.1016/j.econedurev.2018.05.002

Backes, B., & Hansen, M. (2017). The impact of teach for America on non-test academic outcomes. *Education Finance and Policy, 13*(2), 168–193. https://doi.org/10.1162/edfp_a_00231

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal, 111*(1), 7–34. https://doi.org/10.1086/653468

Caires, S., & Almeida, L. (2005). Teaching practice in initial teacher education: Its impact on student teachers' professional skills and development [article]. *Journal of Education for Teaching, 31*(2), 111–120. https://doi.org/10.1080/02607470500127236

Carroll, A., Forlin, C., & Jobling, A. (2003). The impact of teacher training in special education on the attitudes of Australian preservice general educators towards people with disabilities. *Teacher Education Quarterly, 30*(3), 65–79. http://www.jstor.org/stable/23478441.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the early childhood longitudinal study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology, 98*(3), 489–507. https://doi.org/10.1037/0022-0663.98.3.489

Chen, P., & Li, W. (2021). Distinguishing and comprehensive: The bottleneck in the development of education discipline in comprehensive universities. *Research in Educational Development, 19*, 41–44.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Clark, M. A., & Isenberg, E. (2020). Do teach for America corps members still improve student achievement? Evidence from a randomized controlled trial of teach for America's scale-up effort. *Education Finance and Policy, 15*(4), 736–760. https://doi.org/10.1162/edfp_a_00311

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (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.

Coffman, L. C., Conlon, J. J., Featherstone, C. R., & Kessler, J. B. (2019). Liquidity affects job choice: Evidence from teach for America. *The Quarterly Journal of Economics, 134*(4), 2203–2236.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes [article]. *Journal of Teacher Education, 57*(2), 120–138. https://doi.org/10.1177/0022487105283796

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy, 1*(1), 19–27. https://doi.org/10.1080/2330443X.2013.856152

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology, 173*(7), 761–767. https://doi.org/10.1093/aje/kwq439
Gershenson, S. (2021). *Identifying and producing effective teachers*. EdWorkingPapers.com. https://doi.org/10.26300/rzsy-7158.
Goker, S. D. (2006). Impact of peer coaching on self-efficacy and instructional skills in TEFL teacher education [article]. *System, 34*(2), 239–254. https://doi.org/10.1016/j.system.2005.12.002
Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319), 589–612.
Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review, 34*, 29–44. https://doi.org/10.1016/j.econedurev.2013.01.011
Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129–145. https://doi.org/10.3102/01623737022002129
Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology, 38*, 69–119.
Graber, K. C. (1996). Influencing student beliefs: The design of a "high impact" teacher education program. *Teaching and Teacher Education, 12*(5), 451–466. https://doi.org/10.1016/0742-051X(95)00059-S
Guo, S. (2005). Exploring current issues in teacher education in China. *Alberta Journal of Educational Research, 51*(1).
Han, J., Cui, L., & Yu, H. (2021). Pricing the value of the chance to gain admission to an elite senior high school in Beijing: The effect of the LDHSE policy on resale housing prices. *Cities, 115*, Article 103238.
Han, L., & Xie, J. (2020). Can conditional grants attract better students? Evidence from Chinese teachers' colleges. *Economics of Education Review, 78*, Article 102034. https://doi.org/10.1016/j.econedurev.2020.102034
Hanushek, E. A., Piopiunik, M., & Wiederhold, S. (2019). The value of smarter teachers international evidence on teacher cognitive skills and student performance. *Journal of Human Resources, 54*(4), 857–899.
Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009
Hayhoe, R., & Li, J. (2010). The idea of a normal university in the 21st century [article]. *Frontiers of Education in China, 5*(1), 74–103. https://doi.org/10.1007/s11516-010-0007-0
Heckman, J., & Landersø, R. (2022). Lessons for Americans from Denmark about inequality and social mobility. *Labour Economics, 77*, Article 101999. https://doi.org/10.1016/j.labeco.2021.101999
Heilig, J. V., & Jez, S. J. (2010). Teach for america: A review of the evidence. http://epicpolicy.org/publication/teach-for-america.
Henry, G. T., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., Thompson, C. L., & Zulli, R. A. (2014). Teacher preparation policies and their effects on student achievement. *Education Finance and Policy, 9*(3), 264–303.
Henry, G. T., Purtell, K. M., Bastian, K. C., Fortner, C. K., Thompson, C. L., Campbell, S. L., & Patterson, K. M. (2014). The effects of teacher entry portals on student achievement. *Journal of Teacher Education, 65*(1), 7–23.
Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406.
von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review, 64*, 298–312.
Jackson, C. K. (2012). School competition and teacher labor markets: Evidence from charter school entry in North Carolina. *Journal of Public Economics, 96*(5), 431–448. https://doi.org/10.1016/j.jpubeco.2011.12.006
Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615–631. https://doi.org/10.1016/j.econedurev.2007.05.005
Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*(3), 652–667. https://doi.org/10.1037/a0019643
Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy, 5*(1), 54–81.
Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195.
Lam, S. G. (2017). *Teach for America goes to China: Teach for China, educational equity, and public sphere participation in education*. Publication Number 10281110 [Ph. D.]. Madison: The University of Wisconsin.
Lau, K.-C., & Lam, T. Y.-P. (2017). Instructional practices and science performance of 10 top-performing regions in PISA 2015. *International Journal of Science Education, 39*(15), 2128–2149. https://doi.org/10.1080/09500693.2017.1387947
Leavy, A., & Hourigan, M. (2018). The beliefs of 'Tomorrow's teachers' about mathematics: Precipitating change in beliefs as a result of participation in an initial teacher education programme [article]. *International Journal of Mathematical Education in Science & Technology, 49*(5), 759–777. https://doi.org/10.1080/0020739X.2017.1418916
Li, L., Liu, Y., & Chang, B. (2021). Analysis of the present cultivation of masters of education at high-level comprehensive universities and improvement measures. *Journal of Graduate Education, 5*, 39-44+57.
Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics, 7*(1), 295–318. https://doi.org/10.1214/12-AOAS583
Lorente-Catalán, E., & Kirk, D. (2016). Student teachers' understanding and application of assessment for learning during a physical education teacher education course. *European Physical Education Review, 22*(1), 65–81.
Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260–293. https://doi.org/10.3102/0162373719849044
Mason-Williams, L. (2015). Unequal opportunities: A profile of the distribution of special education teachers. *Exceptional Children, 81*(2), 247–262. https://doi.org/10.1177/0014402914551737
Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125–145.
Morgan, P. L., Farkas, G., & Hibel, J. (2008). Matthew effects for whom? *Learning Disability Quarterly, 31*(4), 187–198. https://doi.org/10.2307/25474651
Norton, S., & Zhang, Q. (2018). Primary mathematics teacher education in Australia and China: What might we learn from each other? *Journal of Mathematics Teacher Education, 21*(3), 263–285. https://doi.org/10.1007/s10857-016-9359-6
Pan, Y. (2016). *Education in China: A Snapshot*. O. f. E. C.-O. A. Development. https://www.oecd.org/education/Education-in-China-a-snapshot.pdf.
Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics, 130*, 105–119.
Penner, E. K. (2016). Teaching for all? Teach for America's effects across the distribution of student achievement. *Journal of Research on Educational Effectiveness, 9*(3), 259–282. https://doi.org/10.1080/19345747.2016.1164779
Petek, E., & Bedir, H. (2018). An adaptable teacher education framework for critical thinking in language teaching [article]. *Thinking Skills & Creativity, 28*, 56–72. https://doi.org/10.1016/j.tsc.2018.02.008
Plecki, M. L., Elfers, A. M., & Nakamura, Y. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. *Journal of Teacher Education, 63*(5), 318–334. https://doi.org/10.1177/0022487112447110
Ronfeldt, M. (2012). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis, 34*(1), 3–26.

Rots, I., Aelterman, A., Vlerick, P., & Vermeulen, K. (2007). Teacher education, graduates' teaching commitment and entrance into the teaching profession. *Teaching and Teacher Education, 23*(5), 543–556.
Sass, T. R. (2015). Licensure and worker quality: A comparison of alternative routes to teaching. *The Journal of Law and Economics, 58*(1), 1–35.
Schwerdt, G., & Wuppermann, A. C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review, 30*(2), 365–379.
Shi, X., & Englert, P. A. J. (2008). Reform of teacher education in China. *Journal of Education for Teaching, 34*(4), 347–359. https://doi.org/10.1080/02607470802401537
Shi, Y., Zhang, X., Zheng, Q., & Zhang, X. (2022). Evaluate teaching effectiveness of normal universtiy graduates: An empirical study based on value-added model. *Educational Development Research, 42*(18), 27–37. https://doi.org/10.14121/j.cnki.1008-3855.2022.18.011
Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411
Stürmer, K., Könings, K. D., & Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. *British Journal of Educational Psychology, 83*(3), 467–483. https://doi.org/10.1111/j.2044-8279.2012.02075.x
The Minstry of Education and Other Four Ministries' Notice on Printing and Distributing "Teacher Education Revitalization Action Plan (2018–2022)". http://www.moe.gov.cn/srcsite/A10/s7034/201803/t20180323_331063.html, (2018).
The State Council's Decisions about Deepening Education Reform and Fully Promoting Well-Rounded Education.(1999).
West, J. J., & Frey-Clark, M. L. (2019). Traditional versus alternative pathways to certification: Assessing differences in music teacher self-efficacy [article]. *Journal of Music Teacher Education, 28*(2), 98–111. https://doi.org/10.1177/1057083718788035
Whitford, D. K., Zhang, D., & Katsiyannis, A. (2018). Traditional vs. alternative teacher preparation programs: A Meta-analysis [article]. *Journal of Child and Family Studies, 27*(3), 671–685. https://doi.org/10.1007/s10826-017-0932-0
Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of teach for America in high school. *Journal of Policy Analysis and Management, 30*(3), 447–469. https://doi.org/10.1002/pam.20585
Xun, Y., & Cao, H. (2021). The policy development and path of comprehensive universities' involvement in teacher education. *Research in Educational Development, 19*, 38–41.
Zhang, Y. (2021). Teacher education in high-level comprehensive universities: Possible advantages and practical transformation. *Journal of National Academy of Education Administration, 11*, 18-27+59.
Zhao, D. (2018). *Multiple Ways to Improve the Quality of Student Source in Normal University*. China Education Daily, 10/15/2018.
Zhou, J. (2014). Teacher education changes in China: 1974–2014. *Journal of Education for Teaching, 40*(5), 507–523. https://doi.org/10.1080/02607476.2014.956543
Zhou, J., & Reed, L. (2005). Chinese government documents on teacher education since the 1980s. *Journal of Education for Teaching, 31*(3), 201–213. https://doi.org/10.1080/02607470500169030
Zhou, W. (2010). The survey and policy analysis on teachers source structure in Shandong high schools. *Teacher Education Research, 22*(3), 61–65.
Zhu, X., & Han, X. (2006). Reconstruction of the teacher education system in China. *International Education Journal, 7*(1), 66–73.

Qi Zheng is a doctoral student in the Department of Educational Leadership and Policy Analysis and a research affiliate in the SSTAR Lab at the University of Wisconsin—Madison. His research focuses on the economics of education and educational policy evaluations, with particular interests in the teacher labor market, teacher effectiveness, college admissions, and financial incentives in higher education.

Xin Xie is a doctoral student in the Department of Educational Leadership and Policy Analysis, the Institute for Diversity Science Graduate Fellow, and an assistant researcher with the Wisconsin Center for Education Research at the University of Wisconsin—Madison. Her research focuses on mixed-method education program evaluation, educational equity, teacher labor markets, and organizational disparities. She holds a dual master's degree in language and literacy education and statistics, measurement, assessment, and research technology from the University of Pennsylvania.

Xiaoyang Ye, Ph.D. was a postdoc researcher at Brown University. His work focuses on applying behavioral economics and data science to optimize human capital decisions from school to the workforce. He received his Ph.D. from the University of Michigan and was a postdoc researcher at Princeton University.

Yi Wei, Ph.D. is an Associate Professor at the China Institute for Educational Finance Research at Peking University. Her research focuses on education economics and education finance. She received her Ph.D. from the Michigan State University.