Effects of the Education and Training Programme for Excellent Physicians in China on medical students’ academic performance: a cross-sectional study

Zehua Shi, Chunqing Li, Hongbin Wu

ABSTRACT

Objectives To evaluate the effect of the Education and Training Programme for Excellent Physicians in China on medical students’ academic performance.

Design This study is a secondary analysis of data from a cross-sectional survey of China Medical Student Survey (CMSS, 2021) and administrative data of the first-stage Medical Licensing Examination—the Standardised Competence Test for Clinical Medicine Undergraduates (the Test, 2021). CMSS used a stratified sampling process, with all undergraduate clinical medicine students participating voluntarily.

Setting This programme is implemented at the class level within reformed medical schools, targeting undergraduate students in 5-year clinical medicine in China. The reformed medical schools run pilot classes and traditional classes simultaneously. The analytic sample was a total of 12243 observations from pilot and traditional classes in 34 medical schools which implemented the reform across 19 provinces in China.

Methods This study applied the propensity score matching method to estimate the effect of the reform by comparing the scores of the Test between pilot and traditional classes within the same medical schools. We further explored the potential mechanisms driving the effect from two facets of the Test: medical knowledge modules and cognitive levels.

Results Pilot classes outperformed traditional classes by 0.104 SD on the Test (95%CI 0.037 to 0.171). Improvements were seen in basic medicine and clinical medicine modules (0.109 and 0.101 SD, respectively) and cognitive levels of memorisation and application (0.116 and 0.111 SD, respectively).

Conclusion These results suggest that the reform had a significant positive impact on medical students’ academic performance. Based on the components of this reform and the potential mechanism analysis of the two facets, this study indicates that curriculum reform in integrated learning and teaching methods reform in the adoption of problem-based learning may have been the possible drivers of this positive impact.

INTRODUCTION

Medical education is crucial as it directly determines the academic performance of medical students, and more importantly, it affects healthcare systems in the near future. With the rapid development of China’s economy, healthcare system in China has undergone tremendous changes. People’s health awareness and requirements for medical and health services are continuously increasing, while the factors affecting health are becoming more complex. Moreover, medical knowledge and medical technology are updating and advancing rapidly. All of these bring significant challenges to medical education in China. To produce qualified and competent doctors with solid knowledge and skills in medicine, as well as adequate abilities in communication, collaboration, self-learning and critical thinking who can adapt and be competent to meet the needs of the rapidly changing healthcare system, the medical education reform named the ‘Education and Training Programme for Excellent Physicians’ was introduced in 2012 in China. This reform takes place in the classes at medical schools identified by the Ministry of Education and the National Health Commission of the People’s Republic of China, which are called pilot classes and is mainly oriented toward integrated curriculum reform and teaching methods reform for 5-year undergraduate clinical medical students.

Despite the vital importance of this reform, very few studies have empirically explored its effectiveness. On the one hand, most studies regarding the impact of the reform are normative studies. Existing studies have reviewed the
The performance of 5-year undergraduate clinical medical students. This study quantitatively evaluates the effectiveness of the reform by comparing the graduation examination scores for basic medicine and clinical skills between pilot classes and traditional classes at Qiqihar Medical School, a higher medical college in Heilongjiang Province, and found that the students in pilot classes performed better than those in traditional classes. However, this kind of evidence only applies to specific schools and cannot be extrapolated nationwide due to the unavoidable bias caused by the unique settings of individual schools. To quantitatively evaluate the effectiveness of the reform, empirical analysis based on high-quality representative microdata is necessary.

By using data from the 2021 China Medical Student Survey (CMSS), a nationally representative dataset, this study quantitatively explores the effects of the reform on the academic performance of 5-year clinical medical students. This study contributes to the literature in two ways. First, this study fills the gap in the literature on the impact of medical education reform in China by using standard, objective and nationwide indicators. We measured students’ academic performance by test scores on the Standardised Competence Test for Clinical Medicine Undergraduates in 2021 (the Test). The Test is the first stage of the National Medical Licensing Examination, which is a standardised nationwide examination that can effectively measure the quality of medical personnel. Second, our findings may provide insights into those countries that are facing similar issues in medical education and are in urgent need of reform. Prior studies have found that medical education in many countries, such as India, Indonesia, Brazil, has not been able to keep pace with these transitions, not only lagging in curriculum development and teaching and learning methods but also struggling to train doctors to provide appropriate healthcare to meet the needs of their population, much as is the case in China. Thus, this study on China’s medical education reform could provide policy implications for the development of medical education in these countries.

INSTITUTIONAL BACKGROUND

Education and Training Programme for Excellent Physicians in China

Based on the experience of the third-generation medical education reform developed by the Global Commission on Health Professional Education, the Education and Training Programme for Excellent Physicians was introduced in China in 2012. This reform in China successfully integrated the core principles and approaches of international medical education reform, namely Competency-Based Medical Education (PBL), summarised its achievements, such as strengthening clinical reasoning and clinical skills, and pointed out its problems, such as insufficient investment in medical education, a shortage of teaching resources and inequalities in teaching faculty. These normative studies provide insights into understanding the reform from a holistic perspective. On the other hand, prior empirical studies have been limited and are mainly case studies focused on the effect of the reform in individual medical schools. For example, Han et al. compared the graduation examination scores for basic medical courses, core clinical courses and clinical practice. Moreover, the traditional curriculum is mostly discipline-based curriculum, lacking the connection and communication between disciplines, and unilaterally focusing on the instillation of theoretical knowledge, ignoring the importance of practical skills training.

There are two components in this reform of the training model for 5-year undergraduate clinical medical students: the reform of curriculum and the reform of teaching methods. The traditional curriculum is a three-stage ‘discipline-based’ curriculum in which students are trained in a sequential manner: basic medical courses, core clinical courses and clinical practice. Moreover, the traditional curriculum is mostly discipline-based curriculum, lacking the connection and communication between disciplines, and unilaterally focusing on the instillation of theoretical knowledge, ignoring the importance of practical skills training. The innovative reformed curriculum, by contrast, is an organ-system-based integrated course curriculum in which students are taught in organ-system blocks that link body structure and function to signs and symptoms in clinical cases. Through the integration of teaching content, the redundancy and division of knowledge between disciplines can be reduced.

Medical schools in China generally use horizontal integration and vertical integration approaches in the reformed curriculum. In horizontally integrated courses, the disciplines are combined together around concepts or ideas in each year or level of the course. Commonly this is done using a body system approach, such as cardiovascular, respiratory, renal. In vertically integrated courses, the disciplines are organised into themes or domains which run throughout all years of the course, such as basic and clinical sciences. Meanwhile, the traditional teaching method is the ‘lecture-based learning’ in which students solely receive information from the lecturer. While the new teaching method focuses on student-centred learning in a small class (generally less than 30 students), such as the adoption of PBL, which is adapted to the mode of the integrated curriculum. PBL is a pedagogical approach that shifts the role of the teacher to the student (student-centred) and is based on self-directed learning. With the guidance of tutors, PBL uses clinical cases to stimulate inquiry, critical thinking and knowledge application and integration related to basic and clinical sciences. Although the details of the PBL implementation are not exactly the same in various medical schools, the overall implementation of PBL includes the following aspects: (1) present the case (classically on paper, but sometimes video, audio, patient data or a letter/phone call from a patient); (2) define the problem, key questions and an approach to inquiry; (3) divide into learning issues/tasks and collect

information; (4) self-directed learning (‘private study’); (5) report back to group to share learning; (6) generate hypotheses; (7) integrate knowledge; (8) evaluate group process and learning.  

According to the ‘Report on the Development of Undergraduate Medical Education in China’ in 2019, from the 2013–2014 school year to the 2018–2019 school year, the proportion of integrated courses increased from 4.55% to 7.00%, and the proportion of courses that applied PBL increased from 5.83% to 10.16%.  

Selection of pilot classes
Students enter medical schools through a centralised college admission system. The admission process contains two stages: in the first stage, students take a centralised examination called the national college entrance examination (NCEE) and receive their own scores; in the second stage, students submit their school and major preferences to the placement offices, and are matched to the schools according to their NCEE scores and preferences lists they just submitted. This process is organised separately at the province level.  

The reformed medical schools run pilot classes and traditional classes simultaneously. There are two reasons for running simultaneously. First, from the perspective of policy-makers, the Chinese government should, if possible, choose some local areas (such as units, departments, regions) to implement policy first, and then formulate an overall policy or reform on the basis of summarising experience to avoid major mistakes. Second, from the perspective of policy implementer, the deficient educational preparation in the field of medical education in China must be taken into consideration. The deficiency includes the necessity of strengthening of teaching faculty, and the construction and enrichment of teaching materials. Therefore, the limited education resources constrain the large-scale expansion of pilot classes in China. The pilot classes apply the abovementioned reforms, including the introduction of integrated courses and the PBL teaching method, while the traditional classes continue to use the old curriculum and teaching methods. Medical schools separate students into pilot classes in one of two ways. The first is based on the students’ NCEE score during the enrolment process. The second is based on students’ performance after enrolment at the end of the first school year. In both cases, eligible students can opt out of placement into the pilot class.

METHODS
Data
In this study, we used two datasets. The first was the 2021 CMSS, which is a large-scale nationally representative survey conducted by the National Centre for Health Professions Education Development, which was commissioned by the Ministry of Education. This survey aims to collect detailed and reliable information on undergraduate students in 5-year clinical medicine programmes under a sampling process stratified by geographical location, type of institution and reputation of the medical schools. Each participating medical school sent a formal survey invitation to all undergraduate clinical medicine students. Participation was completely voluntary. An online disclosure statement and a consent agreement were provided to them before the survey was started. In 2021, 121 medical schools across 30 provinces in China completed the CMSS with a response rate of 64.6%. The questionnaire focused on six aspects of students: their basic information, pre-university experience, academic activities at university, attitude toward teaching and university service, evaluation of clinical practice, and academic performance and employment intention after graduation.

The second dataset was from the first-stage National Medical Licensing Examination—the Standardised Competence Test for Clinical Medicine Undergraduates in 2021 (the Test). The Test is held at the end of the fourth year of undergraduate medical education, just before the students entering the clinical rotation phase. The Test is a unified examination in China jointly organised and carried out by the National Medical Examination Centre and National Centre for Health Professions Education Development. The purpose of the Test is to further promote a staged examination of the National Medical Licensing Examination, scientifically evaluate the academic performance of clinical medical students in medical colleges, and provide feedback on the medical education. The Test is conducted in the form of a computer-based test with a total of 300 multiple-choice questions (MCQs). The total possible score is 300 points, with 1 point per question. One hundred and eighty points are required to pass the examination. The questions are categorised into two facets according to the experts from the National Medical Examination Centre. One is divided into four medical knowledge modules with 40%–50% for basic medicine, 40%–45% for clinical medicine, 5%–10% for preventive medicine and 5%–10% for medical humanities. The other is classified into three cognitive levels with 20%–25% for memorisation, 25%–30% for comprehension and 45%–50% for application based on Bloom’s taxonomy of cognitive levels.

We matched the two datasets by the students’ ID numbers and further restricted our sample to reformed medical schools that had implemented reforms for 5-year clinical medical students.

Variables
The primary outcome was the test score on the Test among all students who participated in the Test, normalised to have a mean of 0 and an SD of 1. The secondary outcome measures were the normalised test scores on the two facets of the Test: the medical knowledge modules and the cognitive levels. Again, the test scores were normalised by the facet among all students who participated in the Test to have a mean of 0 and an SD of 1.
The interested independent variable was a binary variable indicating whether a student was in the pilot class (=1) or traditional class (=0) in a given reformed medical school. We controlled for three types of covariates related to whether students would be admitted into a pilot class for propensity score matching (PSM). First, we controlled for the students’ personal backgrounds: female (1 if the student is a female and 0 otherwise), one child (1 if the student was the only child in the family and 0 otherwise) and family background information: parents’ education (parents’ years of schooling) and parents’ occupation (International Socio-Economic Index (ISEI) of occupational status of parents), family member (1 if the student had a family member who was a health professional and 0 otherwise) and rural area (1 if the student’s hukou belonged to the rural area and 0 otherwise).

Second, we included the motivation to study medicine: career (1 if the student received career planning in high school and 0 otherwise), know the major (1 if the student had a general understanding of clinical medicine major before majoring in clinical medicine and 0 otherwise), environment (1 if the student thought that the environment of being a doctor in China is a positive environment to work in, 0 otherwise), intrinsic and extrinsic motivation (students’ intrinsic and extrinsic motivation to study medicine, see online supplemental appendix A).

Third, we controlled for the test score on the NCEE. The NCEE papers used by each province are not exactly the same, but in most provinces, they are scored out of a total of 750. We normalised the test scores of students from provinces without a total score of 750 to a total score of 750.

**Empirical strategy**

To estimate the effect of the medical education reform on students’ academic performance, we applied PSM to eliminate the imbalance between the two types of classes caused by the selection system mentioned earlier. We took the students who entered the pilot classes as the treatment group and the students who entered the traditional classes as the control group.

We estimated the propensity score after setting the treatment and control groups. A person was selected for a pilot class based on their baseline academic achievement (their NCEE or performance at the end of the first school year) and personal willingness. Therefore, we employed three types of covariates related to baseline achievement and willingness to enter a pilot class (their NCEE score, motivation to study medicine and background) to estimate the propensity score.

To estimate the propensity score, we applied a logistic regression model to regress the treatment binary variable on these three types of covariates. Next, we matched students based on their propensity scores and checked the balance of the matching. Specifically, we calculated the standardised mean differences (SMDs) between the treatment and control groups for all the covariates. This was based on the differences in means between the two groups, divided by the (pooled) SD:

\[ SMD = \frac{\bar{X}_{treatment} - \bar{X}_{control}}{\sqrt{\frac{s_{treatment}^2 + s_{control}^2}{2}}} \]

Alongside calculating SMDs, we assessed the unconfoundedness assumption and the overlap assumption for PSM (online supplemental appendix B).

Finally, we estimated the effect of treatment through a regression in which the school fixed effects were controlled. There are two reasons to control school fixed effects. First, students in China are sorted based on their NCEE scores. Second, each medical school may have slightly different strategies for implementing the reform depending on the school’s characteristics and the context in practice. Thus, we included school fixed effects to estimate the impact of the reform using the variation within schools. In other words, we restricted our sample to the reformed medical schools that implemented reforms in the training model for 5-year clinical medical students. Our identification, therefore, comes from the comparison of students in pilot classes with those in traditional classes at the same reformed medical school. Standard errors are clustered at the school level to account for any correlation in outcomes for students at the same school. For this, we used the following equation:

\[ y_i = \alpha \text{Pilot}_i + X_i \beta + \text{FE}_{sch} + \epsilon_i \]  

(1)

where \( y_i \) is the student i’s test score on the Test. \( \text{Pilot}_i \), our dummy variable of interest, is set to 1 when student \( i \) is enrolled in the pilot classes. Otherwise, it is set to 0. \( X_i \) is the set of student controls mentioned earlier. School fixed effects, \( \text{FE}_{sch} \), are also controlled in the regression. The \( \alpha \) coefficient is the main coefficient of interest, capturing the effect of the reform given any medical school.

In addition, we applied inverse probability of treatment weighting to estimate the effect of treatment as a robustness check (online supplemental appendix B).

To explore the potential mechanism that drives the effect of the reform, we further investigated the Test from two facets: medical knowledge modules and cognitive levels. For the first facet, the total score is divided into basic medicine, clinical medicine, medical humanities and preventive medicine modules. We examined the difference between pilot classes and traditional classes in terms of the different medical knowledge modules from this facet. For the second facet, the total score is divided into three cognitive levels: memorisation, comprehension and application. This facet was used to determine the disparity in cognitive levels between the students in pilot classes and those in traditional classes. In practice, we replaced the total score \( y_i \) in equation 1 with the corresponding subscores.

**Patient and public involvement**

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.
RESULTS

Descriptive statistics

The CMSS 2021 covered 31,933 fourth-year undergraduate students (i.e., students who were about to enter clinical rotation) in 5-year clinical medicine programmes across 121 medical schools in 30 provinces throughout China. The Test dataset contained the test score information of 44,567 medical students from 99 medical schools across 29 provinces. Ultimately, we obtained 12,243 observations from 34 medical schools across 19 provinces. A summary of the statistics is presented in Table 1.

Effects of pilot class enrolment on students’ academic performance

We first compared the background of students who entered a pilot class with those of students who did not enter a pilot class using an unmatched sample. Table 2 panel A and figure 1 panel A present the balancing test of the unmatched sample. We found that almost all SMDs of students’ background variables showed an SD of more than 0.1. Specifically, we found that students in pilot classes had parents with more years of schooling (10.28 and 9.50 for mothers of students in pilot classes or not, respectively; 11.23 and 10.58 for fathers of students in pilot classes or not, respectively) and higher job indexes based on the ISEI (32.32 and 29.57 for mothers of students in pilot classes or not, respectively; 34.87 and 32.57 for fathers of students in pilot classes or not, respectively). In addition, students who entered a pilot class were more likely to come from urban areas than those who did not and were more likely to be an only child. Finally, students in pilot classes also had a larger probability that there would be medical staff in their family.

We report the results of the treatment effects using this unmatched sample in panel A of Table 3. The test score of students in pilot classes was higher than that of students in traditional classes, without and with student controls, by 0.165 and 0.124 SD on average (95% CI 0.112 to 0.218; 0.073 to 0.176), respectively.

We then applied PSM using the predetermined characteristics. Table 2 panel B and figure 1 panel B document the balancing test for the matched sample. We found that SMDs were less than 0.1 for all control variables using the matched sample. Additionally, we present the results of the assessment of the unconfoundedness assumption and the overlap assumption in online supplemental appendix B. The results of the investigations of both assumptions made us more confident that the matched sample was well balanced.

After obtaining well-balanced treatment and control groups, we estimated the treatment effect using the balanced matched sample. As shown in panel B of Table 3, we found that the effect of both regressions (with and without student controls) using the matched sample decreased dramatically compared with the unmatched sample. The test scores of students in pilot classes were still higher than those of students in traditional classes by 0.107 and 0.104 SDs (95% CI 0.038 to 0.176; 0.037 to 0.171) without and with student controls, respectively.

Potential mechanism

In Table 4 panel A, with regard to the first facet of the Test, medical knowledge modules, we found that students in pilot classes had better performance on basic medicine...
and clinical medicine than students in traditional classes by 0.109 and 0.101 SD, respectively (95% CI 0.041 to 0.177 for basic medicine; 0.034 to 0.169 for clinical medicine). In addition, there was no evidence showing a difference in performance on preventive medicine and medical humanities.

In table 4 panel B, in terms of the second facet of the Test, cognitive levels, we found that students in pilot classes outperformed students in the traditional classes on memorisation and application by 0.116 and 0.111 SD, respectively (95% CI 0.048 to 0.184 for memorisation; 0.045 to 0.178 for application). However, no such significant improvement was found for comprehension.

**DISCUSSION**

The results from PSM validate the balance on the covariates in the matched sample. On the one hand, we found that the effects of pilot classes showed a considerable decrease from the unmatched sample to the matched sample, suggesting that any potential selection bias might have been successfully removed (0.124–0.165 to 0.104–0.107). On the other hand, it was also reassuring to find that the effects of the pilot classes were almost unchanged with or without controls in the matched sample (0.104 vs 0.107), implying that the pilot class dummy might be independent of student controls, thus achieving a well-balanced sample. (In addition, we provided the results of assessments of the overlap and unconfoundedness assumptions in online supplemental appendix B).

The effect sizes of the medical education reform are 0.104–0.107 SD, indicating that the introduction of integrated curriculum and PBL may have a positive effect on students’ academic performance. An effect larger than 0.1 SD should not be ignored from an educational perspective.28 Specifically, the students in the pilot classes outperform the students in the traditional classes in a fourth-year computer-based medical knowledge test, just before their clinical rotation phase. The effect has the following educational implications. First, previous studies have found that medical knowledge test scores with MCQs have significant relationships with the future clinical skill performance, although caution should be applied if using this assessment as predictive measure for clinical performance due to a small body of evidence and large variations in the predictive strength of the relationships identified.29 In our study, the increase in students’ academic performance may induce a rising clinical skill performance that can be expected to be revealed in (1) the clinical rotation phase at their fifth year and (2) postgraduate medical education, such as standardised residency

**Table 2** Balancing test of unmatched and matched samples

<table>
<thead>
<tr>
<th>Panel A: unmatched sample</th>
<th>No (11017)</th>
<th>Yes (1226)</th>
<th>SMD</th>
<th>Panel B: matched sample</th>
<th>No (1226)</th>
<th>Yes (1226)</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students’ background information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.54</td>
<td>0.58</td>
<td>0.08</td>
<td>Female</td>
<td>0.57</td>
<td>0.58</td>
<td>0.01</td>
</tr>
<tr>
<td>One child</td>
<td>0.40</td>
<td>0.45</td>
<td>0.11</td>
<td>One child</td>
<td>0.46</td>
<td>0.45</td>
<td>0.01</td>
</tr>
<tr>
<td>Rural area</td>
<td>0.36</td>
<td>0.30</td>
<td>0.12</td>
<td>Rural area</td>
<td>0.30</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>Father’s education</td>
<td>10.58</td>
<td>11.23</td>
<td>0.17</td>
<td>Father’s education</td>
<td>11.25</td>
<td>11.23</td>
<td>0.01</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>9.50</td>
<td>10.28</td>
<td>0.19</td>
<td>Mother’s education</td>
<td>10.20</td>
<td>10.28</td>
<td>0.02</td>
</tr>
<tr>
<td>Father’s ISEI</td>
<td>32.57</td>
<td>34.87</td>
<td>0.12</td>
<td>Father’s ISEI</td>
<td>34.81</td>
<td>34.87</td>
<td>0.00</td>
</tr>
<tr>
<td>Mother’s ISEI</td>
<td>29.57</td>
<td>32.32</td>
<td>0.16</td>
<td>Mother’s ISEI</td>
<td>32.13</td>
<td>32.32</td>
<td>0.01</td>
</tr>
<tr>
<td>Family members</td>
<td>0.27</td>
<td>0.32</td>
<td>0.11</td>
<td>Family members</td>
<td>0.32</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>Motivation to study medicine</td>
<td></td>
<td></td>
<td></td>
<td>Motivation to study medicine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Career</td>
<td>0.21</td>
<td>0.22</td>
<td>0.03</td>
<td>Career</td>
<td>0.21</td>
<td>0.22</td>
<td>0.04</td>
</tr>
<tr>
<td>Know the major</td>
<td>0.67</td>
<td>0.72</td>
<td>0.09</td>
<td>Know the major</td>
<td>0.71</td>
<td>0.72</td>
<td>0.01</td>
</tr>
<tr>
<td>Environment</td>
<td>0.59</td>
<td>0.61</td>
<td>0.03</td>
<td>Environment</td>
<td>0.62</td>
<td>0.61</td>
<td>0.02</td>
</tr>
<tr>
<td>Intrinsic motivation</td>
<td>3.56</td>
<td>3.60</td>
<td>0.08</td>
<td>Intrinsic motivation</td>
<td>3.59</td>
<td>3.60</td>
<td>0.01</td>
</tr>
<tr>
<td>Extrinsic motivation</td>
<td>3.40</td>
<td>3.41</td>
<td>0.02</td>
<td>Extrinsic motivation</td>
<td>3.42</td>
<td>3.41</td>
<td>0.01</td>
</tr>
<tr>
<td>Ability before enrolment</td>
<td></td>
<td></td>
<td></td>
<td>Ability before enrolment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCEE</td>
<td>542.35</td>
<td>542.59</td>
<td>0.01</td>
<td>NCEE</td>
<td>542.13</td>
<td>542.59</td>
<td>0.01</td>
</tr>
<tr>
<td>Observations</td>
<td>11017</td>
<td>1226</td>
<td>12243</td>
<td>Observations</td>
<td>1226</td>
<td>1226</td>
<td>2452</td>
</tr>
</tbody>
</table>

We calculated the mean of each variable for the treatment and control groups. We also present the standard mean difference, which is the difference in means between the groups, divided by the (pooled) SD, $\text{SMD} = \frac{\mu_{\text{treatment}} - \mu_{\text{control}}}{\sqrt{\frac{s^2_{\text{treatment}}}{n_{\text{treatment}}} + \frac{s^2_{\text{control}}}{n_{\text{control}}}}}$.

ISEI, International Socio-Economic Index of occupational status; NCEE, National College Entrance Examination; SMD, standardised mean difference.
training. Second, the improvement in the scores of the Test is also directly reflected in the difference between students in the pilot classes and the traditional classes in terms of the pass rate of the Test. The ability to pass the Test directly determines whether a medical student can pursue a career as a doctor in the future. We found that, in terms of pass rate, students in the pilot classes were significantly higher than those in the traditional classes by 0.062 (95% CI 0.026 to 0.098), meaning that learning in the pilot classes is positively correlated with a probability increase of passing the Test of 6.2%.

We further investigated the possible mechanisms from two facets. Our results present the significant effects on academic performance of the basic medicine and clinical medicine modules, and cognitive levels of the memorisation and application. These effects might be attributed to the integration of the curriculum and changes in teaching methods. The structural adjustment of the integrated curriculum can reduce the content of memorisation in courses, especially in the basic medicine and clinical medicine modules which involve a huge amount of memorisation of knowledge. When both the structure and function of a given organ are studied together, rather than at separate times, the need does not exist for each to be repeated when the other is studied. This is particularly important as medical knowledge grows exponentially even though there is limited curricular time to accommodate this expanding body of knowledge. Learning to

**Table 3** The effects of pilot class

<table>
<thead>
<tr>
<th></th>
<th>Unmatched sample</th>
<th>Matched sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test score</td>
<td>Test score</td>
</tr>
<tr>
<td>Pilot class</td>
<td>0.165*** (0.027)</td>
<td>0.124*** (0.026)</td>
</tr>
<tr>
<td></td>
<td>0.107** (0.035)</td>
<td>0.104** (0.034)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>School FE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>R²</td>
<td>0.215</td>
<td>0.255</td>
</tr>
<tr>
<td>Observations</td>
<td>12243</td>
<td>2131</td>
</tr>
</tbody>
</table>

Test scores on the Test were normalised among all students who participated in the Test (ie, 44567 students) to a mean of 0 and an SD of 1. In the matched sample, we further trimmed the sample according to the propensity score, namely, we kept the observations whose propensity scores were located in the range from the minimum value in the treatment group to the maximum value in the control group. Each cell presents the coefficient and SE (in parentheses) for the pilot class dummy. SEs are clustered at the school level. ***Significant at the 1% level, **% level, *10% level.
integrate knowledge provides the self-directed learner with a method of learning that helps them cope with this knowledge growth even after they graduate from medical school. By contrast, the traditional ‘discipline-based’ model emphasises the integrity and independence of the subject knowledge, which is not helpful for students to understand human bodies and diseases, nor for the formation of clinical reasoning. There are too many overlaps and repetitions between disciplines in the basic medicine and clinical medicine modules, resulting in many wasted learning hours. Knowledge of the same organ system is often imparted across different courses that are arranged one or two semesters apart in China.31 Students need to spend substantial time reviewing before learning a new course, which increases the burden of study.

In addition to the structural adjustment of the integrated curriculum, an educational environment that promotes integrated learning is also helpful for recalling knowledge.32 An integrated curriculum contributes to the improvement of the educational environment. Specifically, students were found to be satisfied with the new educational environment at medical schools in Iran33 and India,34 where reforms in the integration of basic medicine and clinical medicine were initiated. A conducive educational environment is essential for student learning. The physical context in which learning occurs, such as clinical or community settings, determines the extent to which students will be able to retrieve what they learned when required to do so in practice. The concept of ‘context specificity’ means that what is learned in one context is more readily retrieved in another similar context. This similarity influences the learner’s ability to retrieve the knowledge when called on to apply it.35

It is worth noting that the construction of knowledge in an integrated curriculum should be underpinned by a process of rigorous inquiry. According to Newmann and colleagues, the central elements of the process of inquiry are building on a prior knowledge base, providing for in-depth learning and providing for elaborated learning. These match the central elements of PBL. Thus, problem-based approaches can provide a strong foundation for authentic integrated learning.36

PBL, as the foundation for integrated learning, appears to be more effective for the application of knowledge than for the acquisition of knowledge.37 In Dochy et al.,37 a meta-analysis of the reforms in the adoption of PBL and its effects, the authors selected 43 empirical studies on PBL in tertiary education conducted in real-life classrooms, 33 of which measured knowledge effects and 25 of which measured application of knowledge effects (some studies measured both knowledge effects and application of knowledge effects). The findings demonstrated that assessment methods that focused more on the application of knowledge showed larger effects for PBL versus traditional teaching methods. In our study, there is a corresponding subscore that measures the degree of students’ mastery of the application of knowledge, which is identified by the National Medical Examination Centre according to the cognitive level. Therefore, we reasonably speculate that the positive significance effect on the application of knowledge may be attributed to the adoption of PBL.

However, we must realise that, with regard to preventive medicine and medical humanities, there was no evidence for a difference between the pilot classes and the traditional classes. First, compared with basic medicine, clinical medicine and other disciplines, investment in preventive medicine teaching has been shown to be insufficient in China.38 This focus on clinical practice and neglect of prevention is still a primary obstacle to the reform of
preventive medicine education in China.\textsuperscript{39} Second, with regard to medical humanities, studies have shown that the proportion of humanities courses in 5-year clinical medicine programmes is still inadequate in China.\textsuperscript{40} In 2013, medical humanities courses accounted for only 5\% of the total credits in China, compared with 20\%–30\% in foreign countries such as the USA and Germany.\textsuperscript{41} Therefore, given the grim situation of COVID-19, it is urgent to strengthen education in preventive medicine and medical humanities in China.

Moreover, we also found that there was no significant difference in the cognitive level of comprehension between pilot classes and traditional classes. The possible reason might be that the comprehension part of the Test had low discrimination. We evaluated the discrimination of the three parts according to the method of the American educational psychometrician Ebel. It is generally believed that a test with a discrimination index greater than 0.3 should be regarded as a high discrimination test.\textsuperscript{42} It was found that the discrimination index of the comprehension questions was 0.27, revealing room for improvement. As a comparison, the discrimination indexes of the memorisation and application questions were both 0.32, indicating good discrimination for both parts.

The effects of this reform should be interpreted with caution. While we mentioned the advantages of curriculum integration and the introduction of PBL in the previous sections, negative effects should be considered as well. For instance, in the case of curriculum integration, students are required to be fully prepared to draw on their entire knowledge base for integrated assessments. Reference points for solving problems or answering questions are blurred, and are not constrained by the boundaries of disciplines as students draw on information learned from across disciplines to answer questions. This may increase the difficulty of assessment tasks, resulting in lower academic scores, and higher failure rates.\textsuperscript{43} Furthermore, regarding PBL, it showed that teaching basic science within a clinical context may have the disadvantage that once basic science knowledge is contextualised, it is difficult to separate it from the particular clinical problems into which it has been integrated. They showed that students trained in a PBL curriculum failed to separate basic science knowledge from the specific clinical knowledge associated with particular patients.\textsuperscript{44}

Finally, we also conducted heterogeneity analysis based on the geographical regions of the medical schools (online supplemental appendix C). We divided the medical schools into three regions: East, Central and West, based on their geographical locations. The East region in China is located along the coast and is relatively economically developed. The Central and West regions in China are inland areas and are economically less developed. We found that the reform effects were the strongest in the East region, with a coefficient of 0.296 SDs (p<0.001). The weakest reform effects are observed in the Central region, with a magnitude of 0.031 SD, which is not statistically significant (p=0.639). The West region’s reform effects fall in the middle, with a magnitude of 0.104 SD, and it is also not statistically significant (p=0.227). This indicates that, in terms of reform effects on academic performance, medical schools in economically developed provinces outperformed those in less developed provinces. This difference may be attributed to the better implementation of reform measures in developed provinces, such as the ability to attract more excellent faculty and prepare enough teaching materials. This finding also serves as a warning for economically less developed countries facing similar challenges in medical education with China. It suggests that reforms such as curriculum integration and the introduction of PBL require a certain level of faculty expertise and teaching resources as a prerequisite. Without these resources, the policies may become ineffective or even counterproductive.

Our study has some limitations. First, despite using PSM, there is still a possibility of omitting some unobservable variables that may favour students in the pilot classes over those in the traditional classes. However, we addressed this concern by matching a large number of control variables based on available data. We also assessed the unconfoundedness assumption and examined the overlap assumption to ensure balance in the matched sample, aiming to mitigate selection bias as much as possible. Furthermore, the focus of this study was to examine the effect of the reform on the academic performance of medical students. However, medical education reform is not solely focused on academic performance. It also aims to cultivate physicians who possess competency in various dimensions, such as critical thinking, clinical reasoning and effective interaction with patients and the public. Unfortunately, due to data limitations, we were unable to investigate the multidimensional effects of the reform on a national scale in this study. Further research can explore the effects of the reform on various dimensions in more depth.

Conclusion

Our findings highlight the effects of a crucial medical education reform in China, which was found to have a significant positive impact on medical students’ academic outcomes as measured by the first stage of the National Medical Licensing Examination—the Standardised Competence Test for Clinical Medicine Undergraduates which is a fourth-year computer-based medical knowledge test with MCQs. The effects have influential educational implications for improving clinical skills and the pass rate of the Test given that both are positively associated with medical knowledge test scores. Our results regarding the potential mechanism further revealed that these significant effects come from the academic performance of the basic medicine and clinical medicine modules, and the cognitive levels of memorisation and application. Based on these findings, we inferred that the positive effects might be attributable to the integration of the curriculum and changes in teaching methods. The results could be
useful for those countries that are facing common issues with China in medical education that are in urgent need of reform.

Acknowledgements We extend our thanks to the Institute of Medical Education/ National Centre for Health Professions Education Development and to the colleges and students who participated in the survey.

Contributors HW and ZS conceived and designed the study. ZS and HW performed the empirical analyses, drafted the initial paper and revised the paper. HW acquired the data. HW and CL provided the comments and suggestions for the study. All authors read the final manuscript and approved its submission. HW is responsible for the overall content as guarantor.

Funding This work was supported by National Social Science Foundation of China (18ZDA10011).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The China Medical Student Survey (CMSS) obtained the ethical approval (IRB00001052-20069) from the Institutional Review Board of Peking University. The front-page of the questionnaire provided an information letter and informed consent with the purpose and main content of the survey, together with a commitment to keep the data anonymous and confidential. Participants voluntarily participated, which did not affect their study. Informed consent was obtained from all study participants. The first-stage National Medical Licensing Examination—the Standardised Competence Test for Clinical Medicine Undergraduates (the Test) was licensed by the National Medical Examination Centre and National Centre for Health Professions Education Development. All methods related to the study were performed in accordance with the relevant guidelines and regulations in the Declaration of Helsinki.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, any changes made indicated, and the use and license their derivative works on different terms, provided the original work is

REFERENCES


38 Lv Y, Pang WTS. Thoughts on preventive medicine teaching for students majoring in clinical medicine. *Acta Med Sin*.
Appendix A

Self-assessment

We note that the score on the Test before enrollment is created based on students’ self-assessment of four areas: their scientific skills and scholarship, their practical clinical ability, their awareness of health and society, and their professionalism. The self-assessment for each area is obtained by taking the average of its relevant sub-assessments, which are all measured using an 11-point integer scale from 0 (lowest level of assessment, representing no ability at all) to 10 (highest level of assessment, representing that the respondent believes that they meet the requirements for being a qualified doctor). We calculate Cronbach’s alpha to check the consistency of each dimension. Cronbach’s alphas are 0.863, 0.961, 0.929, and 0.919 for science and scholarship ability, practical clinical ability, awareness of health and society, and professionalism, respectively. All of them are larger than 0.7, which means these outcome variables have a nice consistency in measurement. Then, we present the sub-assessments for each of these four areas.


For practical clinical ability: 1. The ability to collect a medical history 2. The ability to conduct a physical examination 3. The ability to evaluate a patient’s mental state 4. The ability to write medical records 5. The ability to diagnose diseases based on medical history, physical examination, auxiliary examinations, and more 6. The ability to select appropriate clinical examination methods based on the actual situation 7. The ability to determine a patient’s treatment plan and explain the rationality 8. The ability to manage the doctor-patient relationship 9. The ability to apply knowledge of disease prevention and healthcare to clinical practice 10. The ability to evaluate a patient’s condition and change 11. The ability to provide emergency treatment 12. The ability to retrieve and interpret information in clinical data systems.

For awareness of health and society: 1. Willingness to protect and promote the health of individuals and the population 2. An understanding of the factors that affect population health, disease, and effective treatment 3. Familiarity with the medical quality assurance system and medical safety management system in China 4. Attentiveness to patient safety and the ability to identify risk factors affecting patients 5. Understanding of the structure and function of the medical and health systems in China 6. Understanding of global health issues.

For professionalism: 1. Familiarity with the "Chinese Physicians’ Ethics Code" and the relevant laws and regulations of the medical industry 2. Degree of

Control Variables
Assessments for students’ intrinsic and extrinsic motivation were obtained by taking the average of their related sub-statements, for which students gave responses on a scale ranging from 1 (strongly disagree) to 5 (strongly agree). The four sub-statements for intrinsic motivation were: 1. I performed well in related subjects at high school 2. I have a strong interest in medicine 3. I feel that being a doctor is respected 4. I am confident that I will succeed in this field. The four sub-statements for extrinsic motivation were: 1. Working as a doctor is a stable career 2. My parents, relatives, or teachers encouraged or asked me to choose this major 3. In the future, accessing medical resources will be convenient for individuals, my family, or friends 4. My salary as a doctor will be high in the future.

Appendix B

Additional Assessment of Assumptions

Unconfoundedness Assumption
Although the unconfoundedness assumption is not testable, researchers can calculate the plausibility of this critical assumption (Imbens, 2015). We estimated the effect of the treatment on pseudo-outcome, specifically “the score on the Test before enrollment.” If the treatment effect on the pseudo-outcome was close to zero, we estimated that it was more likely that the unconfoundedness assumption held.

The estimated effect of pilot class on the pseudo outcome is 0.006 (95% CI -0.019-0.031), which showing no statistical significance different from zero. This suggests that unconfoundedness may be a reasonable assumption in the setting of the matched sample.

We estimate the score on the Test before enrollment as follows. There is no test to measure the students’ medical ability before the enrollment in our CMSS data. Nevertheless, CMSS contains the self-assessments of four modules: science and scholarship; practical clinical ability; awareness of health and society; and professionalism. The four modules are the four domains in Accreditation Standards for Basic Medical Education in China (The 2016 Revision) as well. Respondents are asked to assess themselves twice for each module: the ability before the enrollment and current ability. We first regress the test score of the Test on current abilities of the four modules and the school fixed effects. Then
we use the estimates from the previous regression to multiply the correspond-
ing self-assessment abilities before the enrollment and school dummies, we can
obtain a fitted baseline test score for each respondent as our pseudo outcome.

Overlap Assumption

Then, we examined the overlap assumption by presenting the distribution of
propensity scores of the treatment and control groups before and after matching.
We show the distribution of the propensity score before and after matching
in Figure A1. We find that there are a lot of overlap both before and after
matching. This suggests that the overlap assumption is reasonable.

Figure A1: The Distribution of Propensity Score
Robustness checks

In the results section, we show the results based on the propensity score matching. In this section, we apply inverse probability of treatment weighting (IPTW) to estimate the treatment effect. As with the propensity score matching, a balanced treatment and control sample can be created using IPTW. IPTW assigns each individual a weight to create a pseudo-population in which the exposure is independent of the measured confounders. However, IPTW estimates the average treatment effect (ATE), which is different from the propensity score matching (PSM) that estimates the average treatment effect for the treated (ATT).

Table A1 and Figure A2 show that the weighted sample is well-balanced across the treatment and the control group. The SMDs for all the control variables showed a standard deviation less than 0.011.

We also examine the weight to assess the validity of the positivity assumption. We find that the weights range from 1.044 to 22.886, and the median and mean of the weights is 1.110 and 2.000, respectively.

Finally, we estimate the treatment effect, which shows that the estimates of IPTW are larger than the estimates of PSM. As shown in Table A2, students in pilot classes have a higher test score than the students in ordinary classes by 0.142 and 0.155 standard deviation (95% CI 0.084-0.200; 0.076-0.234). The reason may be the fact that we mentioned before, IPTW estimates the average treatment effect (ATE), while PSM estimates the average treatment effect for the treated (ATT). The increase from ATT to ATE is consistent with the concavity of the production function of human capital. We know that in our context, the treatment group (i.e., the students in the pilot classes) has a higher investment than the control group, as a result, the ATT should have a more gently slope in the human capital production function, which means that the extra treatment made by the pilot reform will provide the treatment group (ATT) with lower rewards than the whole groups (ATE).
### Table A1: Balancing Test of Sample by IPTW method

<table>
<thead>
<tr>
<th></th>
<th>No (12243)</th>
<th>Yes (12249)</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Students’ background information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.54</td>
<td>0.55</td>
<td>0.007</td>
</tr>
<tr>
<td>One Child</td>
<td>0.40</td>
<td>0.40</td>
<td>0.004</td>
</tr>
<tr>
<td>Rural Area</td>
<td>0.35</td>
<td>0.35</td>
<td>0.007</td>
</tr>
<tr>
<td>Father’s Education</td>
<td>10.64</td>
<td>10.63</td>
<td>0.002</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>9.57</td>
<td>9.54</td>
<td>0.008</td>
</tr>
<tr>
<td>Father’s ISEI</td>
<td>32.80</td>
<td>32.78</td>
<td>0.001</td>
</tr>
<tr>
<td>Mother’s ISEI</td>
<td>29.84</td>
<td>29.83</td>
<td>0.001</td>
</tr>
<tr>
<td>Family members</td>
<td>0.28</td>
<td>0.28</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Motivation to study medicine</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Career</td>
<td>0.21</td>
<td>0.21</td>
<td>0.002</td>
</tr>
<tr>
<td>Know the Major</td>
<td>0.68</td>
<td>0.67</td>
<td>0.011</td>
</tr>
<tr>
<td>Environment</td>
<td>0.59</td>
<td>0.60</td>
<td>0.005</td>
</tr>
<tr>
<td>Intrinsic Motivation</td>
<td>3.56</td>
<td>3.56</td>
<td>0.004</td>
</tr>
<tr>
<td>Extrinsic Motivation</td>
<td>3.40</td>
<td>3.40</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Abilities before enrollment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>College Entrance Examination</td>
<td>542.38</td>
<td>542.95</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Notes: We calculate the mean of each variable for the treatment and control groups. We also present the standard mean difference which is the difference in means between groups, divided by the (pooled) standard deviation, $SMD = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$.
Figure A2: SMD of IPTW Sample

Notes: We show the SMDs of Table A1 in this figure.
Table A2: The Effect of Pilot Class (IPTW)

<table>
<thead>
<tr>
<th></th>
<th>Test Score</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full sample</td>
<td>weight&lt;=10</td>
</tr>
<tr>
<td>Pilot Dummy</td>
<td>0.142***</td>
<td>0.155***</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>School FE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Pseudo Observations</td>
<td>24492</td>
<td>17709</td>
</tr>
</tbody>
</table>

Notes: Test score of the Test is standardized among all students participated the test (i.e., 44,567 students) to obtain a mean of zero and standard deviation of one. Each cell presents the coefficient and standard error (in parentheses) for the pilot class dummy. All specifications include all variables listed in Table 2 and school fixed effects. Standard errors are clustered at school level. ***significant at the 1% level, **5% level, *10% level.

Reference


Appendix C

In the Discussion section of the main text, we conducted an analysis of heterogeneity based on the geographical regions of the medical schools. We categorized the medical schools into three regions: East, Central, and West, according to their geographical locations. We used subsamples of these three regions to examine the effects of the reforms separately. The results are presented in Table A3.
Table A3. The Effects of Pilot Class by Geographical Regions of the Medical Schools

<table>
<thead>
<tr>
<th></th>
<th>East</th>
<th>Central</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot Class</td>
<td>0.296***</td>
<td>0.031</td>
<td>0.104</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>School FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.224</td>
<td>0.170</td>
<td>0.260</td>
</tr>
<tr>
<td>Observations</td>
<td>1052</td>
<td>702</td>
<td>377</td>
</tr>
</tbody>
</table>

Notes: Test scores on the Test were normalized among all students who participated in the Test (i.e., 44,567 students) to a mean of zero and a standard deviation of one. In this table, we presented the heterogeneous effects of pilot effects by geographical regions of the medical schools using subsample of each region. ***significant at the 1% level, **5% level, *10% level.