ARTICLES
https://doi.org/10.1038/s41562-021-01062-3

nature human behaviour

Check for updates

Skill levels and gains in university STEM education in China, India, Russia and the United States

Prashant Loyalka[1,2]✉, Ou Lydia Liu[3], Guirong Li[4], Elena Kardanova[5], Igor Chirikov[5,6]✉, Shangfeng Hu[7], Ningning Yu[8], Liping Ma[9], Fei Guo[10], Tara Beteille[11], Namrata Tognatta[11], Lin Gu[3], Guangming Ling[3], Denis Federiakin[5], Huan Wang[2], Saurabh Khanna[2], Ashutosh Bhuradia[2], Zhaolei Shi[1] and Yanyan Li[4]

Universities contribute to economic growth and national competitiveness by equipping students with higher-order thinking and academic skills. Despite large investments in university science, technology, engineering and mathematics (STEM) education, little is known about how the skills of STEM undergraduates compare across countries and by institutional selectivity. Here, we provide direct evidence on these issues by collecting and analysing longitudinal data on tens of thousands of computer science and electrical engineering students in China, India, Russia and the United States. We find stark differences in skill levels and gains among countries and by institutional selectivity. Compared with the United States, students in China, India and Russia do not gain critical thinking skills over four years. Furthermore, while students in India and Russia gain academic skills during the first two years, students in China do not. These gaps in skill levels and gains provide insights into the global competitiveness of STEM university students across nations and institutional types.

A major goal of undergraduate STEM programs is to help students to develop academic knowledge, competencies and skills (hereafter, skills) as well as higher-order thinking skills such as critical thinking[1–7]. Equipping individuals with such skills contributes to human capital development and promotes innovation, helping nations grow and compete in the global knowledge economy[8–12].

Past studies show that there is a positive relationship between a nation's human capital, as measured by years of schooling, and its growth[13–16]. Recent studies show that skills measured by international assessments of primary and secondary school students are a closer proxy for country-level human capital and a more robust determinant of growth[17–21]. Research on the role of cognitive skills in economic growth acknowledges that cognitive skill measures may also capture non-cognitive or higher-order cognitive dimensions that also explain productivity and growth[18,22–24]. Such evidence supports a rich line of inquiry into educational reforms that can produce skills[24].

However, in emphasizing the importance of human capital for productivity and growth, researchers have largely focused on skills acquired in pretertiary education rather than in higher education. In particular, despite the tens of billions of dollars spent on undergraduate STEM programs each year, little is known about the extent to which students in these programs develop critical thinking and academic skills during university.

Attempts to measure skill acquisition—for example, by collecting data on the short-term employment outcomes of graduates—have been too indirect to provide actionable insights for education policymakers or university administrators[1,25]. Direct approaches using standardized assessments have rarely been applied[26,27]. In the few cases in which studies have collected standardized assessment data—such as the Assessment of Learning Outcomes in Higher Education of the Organisation for Economic Co-operation and Development—they have generally not assessed nationally representative (random) samples of students and institutions and have therefore been unable to contextualize skill levels and gains in an international perspective[28].

Loyalka et al.[29] compared skills among representative (random) samples of STEM undergraduates across countries. However, that study is limited in that it collects and analyses only cross-sectional data on computer science (CS) skills among CS majors at one point in time—that is, at the end of university. The results of this study therefore do not speak directly to: (1) skills learned during university (skill gains); (2) higher-order thinking skills, which are believed to be critical for workforce productivity; or (3) foundational academic skills, such as maths and science, which are largely covered in the first two years of university and which are the basis for success in later years. Loyalka et al.[29] also uses a relatively small sample of 1,593 students for China, India and Russia; furthermore, the sample from India is from only three states and is not strictly nationally representative.

The lack of evidence on skill acquisition in higher education is a major omission. Nations spend a substantial and growing proportion of their GDP on higher education (2.6% in the United States[30]). Higher education is also associated with greater returns compared with primary and secondary schooling[31]. Although private returns may be due in varying degrees to the contribution of higher education to skill development or its signalling value[32], the skills produced

[1]Graduate School of Education, Stanford University, Stanford, CA, USA. [2]Freeman Spogli Institute for International Studies, Stanford University, Stanford, CA, USA. [3]Educational Testing Service, Princeton, NJ, USA. [4]International Center for Action Research on Education, School of Education, Henan University, Henan, China. [5]Institute of Education, National Research University Higher School of Economics, Moscow, Russia. [6]Center for Studies in Higher Education, Goldman School of Public Policy, University of California Berkeley, Berkeley, CA, USA. [7]Faculty of Education, Sichuan Normal University, Sichuan, China. [8]Institute of Higher Education Research, University of Jinan, Jinan, China. [9]Graduate School of Education, Peking University, Beijing, China. [10]Institute of Education, Tsinghua University, Beijing, China. [11]World Bank, Washington, DC, USA. ✉e-mail: loyalka@stanford.edu; chirikov@berkeley.edu

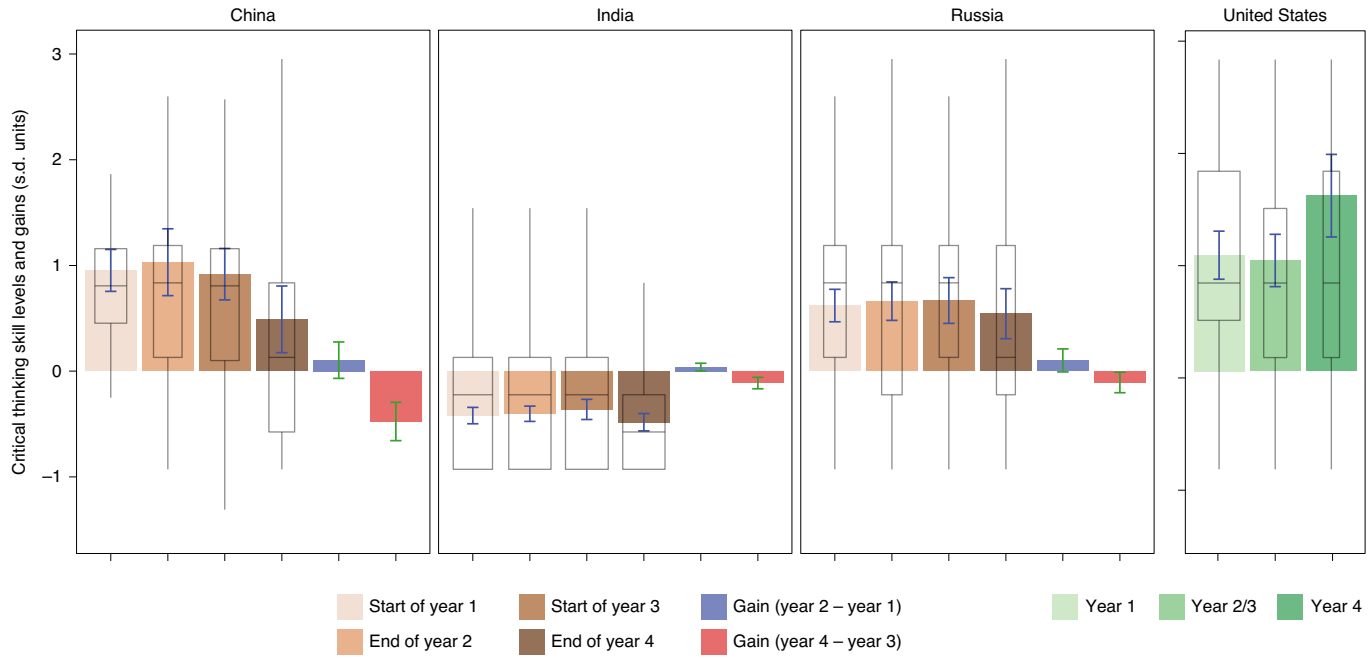

Fig. 1 | Critical thinking skill levels and gains (s.d. units) across China, India and Russia with benchmarks from the United States. Relevant statistical information and notes are provided in Table 1. The blue error bars indicate the 95% CI of the estimates of skill levels. CIs for the sample mean skill level estimates pertain to population mean skill levels for each country. The green error bars indicate the 95% CI of the estimates of skill gains. The box-and-whisker plots show the distribution of skill levels and gains for each country. The solid horizontal line shows the median, the box shows the interquartile range, and the whiskers show the upper and lower bounds (the most extreme value less than 1.5× the interquartile range beyond the first or third quartile).

through higher education may also have substantial externalities[33]. For example, higher education may lead to increased innovation[13,14,34] and knowledge transfer[35,36].

Especially relevant to today's economy, higher education is meant to help individuals to acquire advanced skills that are required to keep up with rapid technological change[11,37,38]. Advanced skills include higher-order cognitive skills, such as critical thinking and creativity, as well as academic skills, such as university-level maths and science[10,23]. However, direct, generalizable evidence on the degree to which higher education imparts these advanced university-level skills is lacking.

In this Article, we seek to address these gaps by providing direct, representative and longitudinal evidence of how the higher-order thinking and academic skills (skill levels and skill gains) of STEM undergraduates compare across national systems, as well as how they differ by selectivity of institution and by student gender. To do so, we collected internationally standardized assessment data on the critical thinking and academic skills of STEM undergraduates (students in four-year programs in CS and electrical engineering) in elite and non-elite institutions in China, India and Russia. In addition to being key political and economic actors, China, India and Russia produce approximately half of the world's STEM graduates[39]. Furthermore, we benchmarked the critical thinking skill levels and gains of STEM students in these three major countries against those of STEM students in the United States.

In regard to institutional selectivity, higher education systems are increasingly differentiated into elite and non-elite institutions[40–43]. Elite institutions, which are characterized by higher levels of public and private investment, limited quotas and selective admissions and, therefore, higher-scoring students and greater prestige, are generally thought to be of higher quality compared with the non-elite institutions that train the vast majority of university students in a country[40–42]. The growing bifurcation of higher education systems into elite and non-elite institutions has also been notable in emerging economies such as China, India and Russia, where policymakers have actively pushed elite institutions to become world-class, research universities that raise up highly qualified scientific and managerial cadres[43].

We used strict sampling procedures to randomly select institutions and students in China, India and Russia (Methods and Supplementary Information). By paying close attention to survey implementation, we also achieved high total response rates among institutions and enrolled students. Our exams were designed to be, and were validated as, culturally neutral. We trained hundreds of enumerators to proctor exams in the same way. All of the sampled students were provided with the same incentives to participate. We also tested the sensitivity of the results for potential differences in student motivation (Supplementary Information D).

Our estimates of skill gains are multidimensional and robust (Methods and Supplementary Information). Our strict sampling and survey procedures enabled us to examine cross-cohort skill gains in a relative sense—that is, across higher education systems and institutions. We also used vertically scaled test scores (using tests with sufficient anchor items) to examine cross-cohort skill gains in an absolute sense—whether students make positive, zero or negative changes in skills over time. We measured relative and absolute gains in both domain-general higher-order thinking skills (critical thinking) as well as in domain-specific academic skills (such as maths and physics—the primary science subject in our sampled majors). Controlling for the family background of students and their out-of-university activities, we provide evidence that differences in skill level gains are attributable to the in-university experiences of students and not to differences in their family background or out-of-university activities (Supplementary Information F). Thus, the skill gains that we measured probably reflect the value-added associated with participating in undergraduate STEM programs.

Previewing the main results, we found stark differences in skill levels among countries and between elite versus non-elite institutions.

Table 1 | Critical thinking skill levels and gains (s.d. units) across China, India and Russia with benchmarks from the United States

	China	India	Russia	US
Start of year 1				
s.d. units	0.95	−0.42	0.62	1.10
P	0.00	0.00	0.00	0.00
95% CI	0.75–1.15	−0.50 to −0.34	0.47–0.77	0.87–1.33
n	1,233	1,853	614	894
End of year 2				
s.d. units	1.03	−0.40	0.66	
P	0.00	0.00	0.00	
95% CI	0.71–1.35	−0.48 to −0.33	0.48–0.84	
n	966	2,154	512	
Start of year 3 (US year 2/3)				
s.d. units	0.92	−0.36	0.67	1.05
P	0.00	0.00	0.00	0.00
95% CI	0.67–1.16	−0.46 to −0.27	0.45–0.88	0.80–1.30
n	992	2,202	446	269
End of year 4				
s.d. units	0.49	−0.48	0.54	1.66
P	0.00	0.00	0.00	0.00
95% CI	0.17–0.81	−0.57 to −0.40	0.31–0.78	1.27–2.05
n	796	2,153	430	435
Gain (year 2 – year 1)				
s.d. units	0.10	0.04	0.10	
P	0.25	0.05	0.08	
95% CI	−0.08–0.28	0.00–0.08	−0.01–0.22	
Gain (year 4 – year 3)				
s.d. units	−0.48	−0.11	−0.11	
P	0.00	0.00	0.04	
95% CI	−0.66 to −0.29	−0.17 to −0.06	−0.21 to −0.01	

Data for China, India and Russia are from national random samples of four-year undergraduate CS-related and electrical-engineering-related majors. US data are from undergraduate (Bachelor's degree) STEM majors, from a representative range of Doctoral research, Masters and Baccalaureate institutions. Students in China, India and Russia took the critical thinking skills exam in the first semester of their freshman year and the second semester of their second year. Students in the United States took the critical thinking skills exam at different points during the academic year. For the effect sizes in s.d., year-equated exam scores were converted into *z* scores using the baseline mean and s.d. of the China, India and Russia cross-national sample of exam takers. Analytical estimates for China, India and Russia were calculated using sampling weights and are therefore representative of well-defined populations. To adjust for exam motivation, estimates were calculated using data for students who attempted at least 75% of the items on the test. The results were substantively the same with and without adjustment. The gain estimates for China, India and Russia that were unadjusted for attrition were substantively the same as the gain estimates that were adjusted for attrition using multiple imputation (Supplementary Information). The gain estimates for the United States are regression-adjusted estimates that control for gender, minority status (yes or no) and scaled SAT/ACT scores. s.e. values were adjusted for clustering at the institutional level. *P* values and 95% CIs are shown.

At the start of university, students in China and the United States score approximately 1.4 to 1.5 s.d. higher in critical thinking than students in India and approximately 0.3 to 0.5 s.d. higher than students in Russia. Furthermore, students in China score approximately 1 s.d. higher in academic skills than students in India and Russia. Students from elite institutions in China and India score much higher in academic and critical thinking skills compared with students from non-elite institutions. Female students start university with the same level of critical thinking scores and slightly lower maths and physics scores compared with male students. During the first two years of university, the gender gap closes in maths but not in physics.

These substantial gaps in skill levels provide insights into the university readiness of STEM undergraduates from different countries and types of institutions. We also present gaps in academic skill levels after two years of university and gaps in critical thinking skills after two and four years of university, which provide further insights into the global competitiveness of STEM graduates from each country. We later contextualize these gaps in skill levels by discussing differential selection into STEM majors.

Importantly, to focus on university quality, we show substantial differences in skill gains among countries. Students in India and Russia experience significant academic skill gains during the first two years (0.1 to 0.4 s.d.), whereas students in China experience no gains or significant, absolute academic skill losses (approximately −0.3 to 0 s.d.). This contributes to a closing of the academic skills gap between China and other countries. Whereas longitudinal gains reveal that students in China, India and Russia experience slight gains in critical thinking during the first two years and losses in critical thinking over the last two years of university, regression-adjusted results across multiple cohorts in the United States (although non-representative) align closely with previous literature that suggests that US students experience substantial gains within four years (approximately 0.5 s.d.)[2].

Results

Critical thinking skills levels and gains. We found that critical thinking skill levels differ considerably across countries (Fig. 1 and Table 1). First-year university students (hereafter, freshmen) in

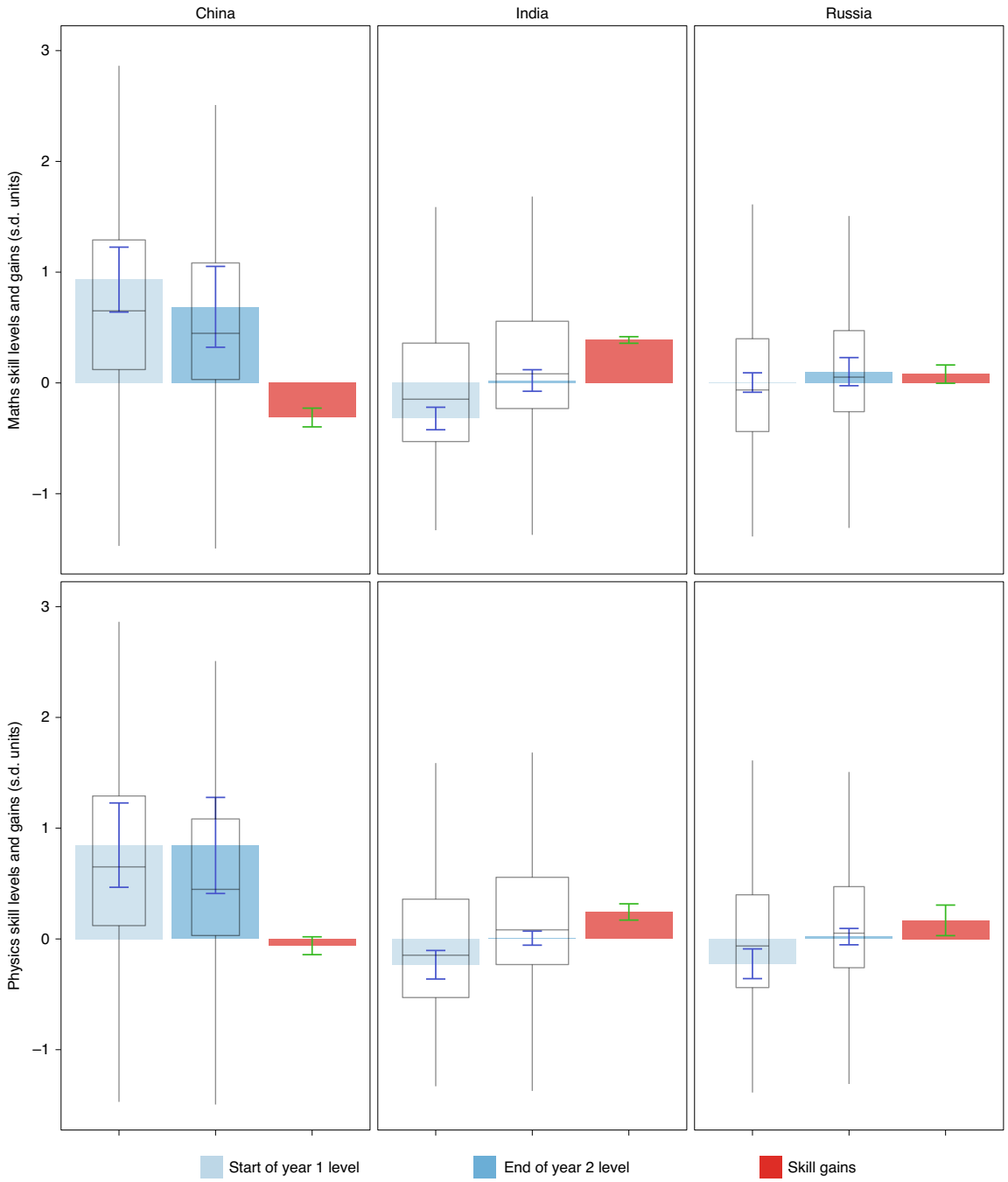

Fig. 2 | Maths and physics skill levels and gains from the start of the first year to the end of the second year (s.d. units). Relevant statistical information and notes are provided in Table 2. The blue error bars indicate the 95% CI of the estimates of skill levels. CIs for the sample mean skill level estimates pertain to population mean skill levels for each country. The green error bars indicate the 95% CI of the estimates of skill gains. The box-and-whisker plots show the distribution of skill levels or gains for each country. The solid horizontal line shows the median, the box shows the interquartile range, and the whiskers show the upper and lower bounds (the most extreme value less than 1.5× the interquartile range beyond the first or third quartile).

China exhibit similar levels of critical thinking skills as freshmen in the United States (difference = −0.146 s.d., $P = 0.319$, 95% confidence interval (CI) = −0.435–0.143) but much higher levels than freshmen in India (1.373 s.d., 95% CI = 1.168–1.579) and moderately higher levels than freshmen in Russia (0.332 s.d., $P = 0.007$, 95% CI = 0.090–0.573). Freshmen in Russia also exhibit much higher levels of critical thinking skills compared with freshmen in India (1.042 s.d., $P < 0.001$, 95% CI = 0.876–1.207). At the end of year 2, second year university students in China still score much higher in critical thinking compared with second-year students in India (1.433 s.d., $P < 0.001$, 95% CI = 1.123–1.744), moderately higher than second-year students in Russia (0.368 s.d., $P = 0.039$, 95% CI = 0.019–0.718) and comparably with their counterparts in the United States (−0.019 s.d., $P = 0.922$, 95% CI = −0.405–0.367; because students in the United States took the critical thinking exam at different times of the year, and because the relatively equal sample sizes for year 2 and year 3 are small, we combined the year 2 and year 3 observations). However, by the end of their fourth year, while students in China still scored much higher than students in India (0.973 s.d., $P < 0.001$, 95% CI = 0.661–1.286), their

Table 2 | Maths and physics skill levels and gains from the start of the first year to the end of the second year (s.d. units)

	China	India	Russia	Difference (China - India)	Difference (China - Russia)	Difference (India - Russia)
Panel A: maths						
Start of year 1						
s.d. units	0.933	−0.321	0.004	1.254	0.929	−0.325
P	0.000	0.000	0.935	0.000	0.000	0.000
95% CI	0.640–1.226	−0.423 to −0.220	−0.084–0.091	0.954–1.554	0.634–1.224	−0.455 to −0.194
n	2,435	3,742	1,132			
End of year 2						
s.d. units	0.687	0.022	0.102	0.665	0.585	−0.080
P	0.001	0.652	0.112	0.000	0.002	0.309
95% CI	0.322–1.052	−0.075–0.119	−0.025–0.229	0.302–1.028	0.214–0.956	−0.235–0.075
n	1,969	4,485	966			
Maths gains						
s.d. units	−0.312	0.388	0.079			
P	0.000	0.000	0.068			
95% CI	−0.400 to −0.224	0.357–0.418	−0.006–0.165			
n	1,844	3,472	760			
Panel B: physics						
Start of year 1						
s.d. units	0.847	−0.233	−0.224	1.080	1.071	−0.009
P	0.000	0.001	0.002	0.000	0.000	0.921
95% CI	0.466–1.228	−0.362 to −0.104	−0.358 to −0.090	0.691–1.469	0.681–1.460	−0.189–0.171
n	2,423	3,916	722			
End of year 2						
s.d. units	0.844	0.007	0.021	0.837	0.823	−0.014
P	0.000	0.821	0.568	0.000	0.000	0.774
95% CI	0.410–1.278	−0.056–0.071	−0.053–0.095	0.416–1.258	0.401–1.246	−0.108–0.081
n	1,983	4,510	627			
Physics gains						
s.d. units	−0.061	0.244	0.168			
P	0.145	0.000	0.024			
95% CI	−0.145–0.022	0.168–0.319	0.024–0.312			
n	1,845	3,568	515			

Students from China, India and Russia took exams during the first semester of their freshman year and then again at the end of the second semester of their second year. Gains were estimated for students who were present in both the baseline and follow-up phases. Alternative and closely aligned gain estimates using multiple imputation and either (1) all students in the baseline (regardless of whether they were in the follow up); or (2) all students in the follow up (regardless of whether they were in the baseline) are provided in Supplementary Information E and Supplementary Table 3a,b. Level and gain estimates are reported as effect sizes (in s.d. units). In the case of maths and physics, scaled (year-equated) exam scores were divided by the subject-specific baseline mean and s.d. of the entire cross-national sample of exam takers. China, India and Russia data are from national random samples of four-year undergraduate CS-related and electrical-engineering-related majors. To adjust for exam motivation, estimates were calculated using data for students who attempted at least 75% of the items on the test. Results were substantively the same with or without adjustment. The final number of observations used for each estimate are indicated (*n*). Analytical estimates were calculated using sampling weights and are therefore representative of well-defined national populations. s.e. were adjusted for clustering at the institution level. *P* values and 95% CIs are shown.

scores were statistically indistinguishable from students in Russia (−0.053 s.d., $P = 0.780$, 95% CI = −0.431–0.324), and much lower than year 4 students in the United States (−1.173 s.d., $P < 0.001$, 95% CI = −1.654 to −0.692).

Gaps in critical thinking skill levels at the end of university are in largely due to cross-national differences in critical thinking skill gains during the final two years of university. Students in China, India and Russia make minimal gains in critical thinking skills from the start of their first year to the end of their second year of university (0.04–0.10 s.d.; Fig. 1). Furthermore, cross-cohort regression-adjusted gains in the United States suggest that there are no significant gains in critical thinking skills during the first two years; the lack of gains in the first two years also aligns with the estimates from previous studies[2,6]. However, students experience significant declines in critical thinking skills during the final two years of university in China (−0.48 s.d., $P < 0.001$, 95% CI = −0.66 to −0.29), India (−0.11 s.d., $P < 0.001$, 95% CI = −0.17 to −0.06) and Russia (−0.11 s.d., $P = 0.037$, 95% CI = −0.21 to −0.01). By contrast,

Table 3 | Skill levels and gains: elite versus non-elite institutions (s.d. units)

	China			India			Russia		
	Elite	Non-Elite	Difference	Elite	Non-Elite	Difference	Elite	Non-Elite	Difference
Panel A: critical thinking									
Start of year 1									
s.d. units	1.612	0.741	0.871	0.509	−0.462	0.971	0.833	0.544	0.289
P			0.000			0.000			0.070
95% CI			0.523–1.218			0.780–1.162			−0.025–0.603
End of year 2									
s.d. units	1.905	0.735	1.170	0.542	−0.438	0.980	0.810	0.608	0.202
P			0.001			0.000			0.435
95% CI			0.519–1.822			0.845–1.115			−0.318–0.722
Year 1 to year 2 gains									
s.d. units	0.417	0.008	0.409	−0.014	0.041	−0.057	0.054	0.119	−0.065
P	0.170	0.899	0.124	0.855	0.040	0.453	0.799	0.002	0.734
95% CI	−0.237–1.071	−0.120–0.136	−0.119–0.937	−0.192–0.163	0.002–0.081	−0.204–0.092	−0.462–0.569	0.046–0.191	−0.450–0.320
Start of year 3									
s.d. units	1.602	0.665	0.937	0.596	−0.406	1.002	0.730	0.651	0.079
P			0.000			0.000			0.793
95% CI			0.565–1.309			0.795–1.209			−0.530–0.688
End of year 4									
s.d. units	1.339	0.234	1.104	0.232	−0.519	0.751	0.767	0.491	0.276
P			0.002			0.000			0.445
95% CI			0.446–1.762			0.537–0.964			−0.451–1.004
Year 3 to year 4 gains									
s.d. units	−0.397	−0.505	0.108	−0.343	−0.102	−0.241	−0.024	−0.126	0.102
P	0.171	0.000	0.663	0.019	0.001	0.035	0.802	0.033	0.317
95% CI	−1.037–0.242	−0.681 to −0.330	−0.394–0.610	−0.612 to −0.075	−0.159 to −0.046	−0.464 to −0.018	−0.256–0.208	−0.240 to −0.011	−0.102–0.306
Panel B: maths									
Start of year 1									
s.d. units	1.850	0.641	1.209	0.851	−0.366	1.217	0.111	−0.024	0.135
P			0.000			0.000			0.301
95% CI			0.678–1.739			1.063–1.371			−0.126–0.396
End of year 2									
s.d. units	1.767	0.342	1.425	1.166	−0.021	1.187	0.327	0.050	0.276
P			0.000			0.000			0.044
95% CI			0.743–2.108			1.014–1.360			0.008–0.545
Year 1 to year 2 gains									
s.d. units	−0.209	−0.333	0.125	0.306	0.387	−0.081	0.257	0.043	0.213
P	0.061	0.000	0.217	0.000	0.000	0.055	0.000	0.351	0.000
95% CI	−0.430–0.013	−0.437 to −0.229	−0.077–0.326	0.210–0.402	0.356–0.419	−0.164–0.002	0.178–0.335	−0.050–0.137	0.105–0.322
Panel C: physics									
Start of year 1									
s.d. units	1.936	0.504	1.433	1.419	−0.301	1.720	−0.029	−0.288	0.259
P			0.002			0.000			0.058
95% CI			0.557–2.308			1.478–1.962			−0.009–0.527

Continued

Table 3 | Skill levels and gains: elite versus non-elite institutions (s.d. units) (continued)

	China			India			Russia		
	Elite	Non-Elite	Difference	Elite	Non-Elite	Difference	Elite	Non-Elite	Difference
End of year 2									
s.d. units	2.089	0.455	1.634	1.008	−0.029	1.037	0.036	0.017	0.019
P			0.001			0.000			0.881
95% CI			0.690–2.577			0.901–1.173			−0.240–0.278
Year 1 to year 2 gains									
s.d. units	0.044	−0.076	0.120	−0.403	0.272	−0.675	−0.168	0.244	−0.413
P	0.718	0.081	0.313	0.006	0.000	0.000	0.230	0.000	0.003
95% CI	−0.241–0.329	−0.163–0.010	−0.119–0.359	−0.650 to −0.156	0.202–0.342	−0.885 to −0.465	−0.484–0.148	0.129–0.360	−0.672 to −0.153

For critical thinking, one cohort of students took exams during the first semester of their first year and then again at the end of the second semester of their second year, while another cohort of students took exams during the first semester of their third year and then again at the end of the second semester of their fourth year. For maths and physics, students took exams in first semester of their freshman year and then again at the end of the second semester of their second year. Gains were estimated for students who were present in both the baseline and follow-up phases. Alternative and closely aligned gain estimates using multiple imputation and either (1) all students in the baseline (regardless of whether they were in the follow up) or (2) all students in the follow up (regardless of whether they were in the baseline) are provided in Supplementary Information E and Supplementary Table 3b. Level and gain estimates are reported as effect sizes (in s.d. units). Scaled exam scores were divided by the subject-specific baseline mean and s.d. of the China, India and Russia cross-national sample of exam takers. China, India and Russia data are from national random samples of four-year undergraduate CS-related and electrical-engineering-related majors. Analytical estimates from China, India and Russia were calculated using sampling weights such that they are representative of well-defined national populations. To adjust for exam motivation, estimates were calculated using data for students who attempted at least 75% of the items on a test. The results were substantively the same with and without adjustment. Definitions of elite: for China, all 985 and 211 institutions; India, IITs, NITs and other top-100 MHRD ranked universities; Russia, all national research and federal universities. s.e. values were adjusted for clustering at the institution level. *P* values and 95% CIs are shown.

across-cohort regression-adjusted gains in the United States show significant increases in critical thinking skills from the middle to the end of university (0.46 s.d., $P<0.001$), which again aligns with estimates from several previous studies[2] (regression-adjusted gains from year 1 to year 4 are only slightly and not significantly higher (0.53 s.d., $P<0.001$)).

Academic skills levels and gains. As with critical thinking, freshmen in China have the highest levels of maths and physics skills, much higher than freshmen in India (maths difference = 1.254 s.d., $P<0.001$, 95% CI = 0.954–1.554; physics difference = 1.080 s.d., $P<0.001$, 95% CI = 0.691–1.469) and Russia (maths difference = 0.929 s.d., $P<0.001$, 95% CI = 0.634–1.224; physics difference = 1.071 s.d., $P<0.001$, 95% CI = 0.681–1.460; Fig. 2 and Table 2). The differences are all statistically significant at the 1% level. Freshmen in Russia further have significantly higher levels of maths skills, but not physics skills, compared with freshmen in India (maths difference = −0.325 s.d., $P<0.001$, 95% CI = −0.455 to −0.194; physics difference = −0.009 s.d., $P=0.921$, 95% CI = −0.189–0.171).

China's advantage in academic skills narrows considerably after two years due to cross-country differences in skill gains (Fig. 2 and Table 2). According to the unadjusted estimates, skill gains from the start of the first to the end of the second year in China are negative and significant in magnitude in maths (−0.312 s.d., $P<0.001$, 95% CI = −0.400 to −0.224) and negative but not statistically significant in physics (−0.061 s.d., $P=0.145$, 95% CI = −0.145–0.022). By contrast, the skill gain estimates are positive and significant in India for maths (0.388 s.d., $P<0.001$, 95% CI = 0.357–0.418) and physics (0.244 s.d., $P<0.001$, 95% CI = 0.168 to 0.319). Results are also positive and significant, albeit smaller, in Russia for maths (0.079 s.d., $P=0.068$, 95% CI = −0.006–0.165) and physics (0.168 s.d., $P=0.024$, 95% CI = 0.024–0.312). Taken together, the results show that students in India and Russia make significant gains in maths and physics during the first two years of university. By contrast, students in China experience a decrease in the maths skills that they had acquired before entering university.

The results hold whether or not we convert the item response theory (IRT)-scaled scores into *z* scores. Furthermore, when we limited the start-of-year 1 and end-of-year 2 maths tests to the anchor items (that were exactly the same across the two maths tests and comprised approximately 40% of each test), we found that students in China also score 0.27 s.d. lower at the end of year 2 compared with at the start of year 1 (which is quite similar to the decrease in maths skills of 0.31 s.d. reported in Table 2). This is in contrast to significant maths score gains on the common items in India and Russia. Despite the loss in maths skills in absolute terms from the start of year 1 to the end of year 2 in China, the maths skill levels of students at the end of year 2 remain high in China compared to India and Russia.

Skills in elite and non-elite institutions. There are also stark cross-country differences in critical thinking and academic skill levels by institutional type (Table 3 and Supplementary Table 3b). Regarding critical thinking skills, students of all four years of study in elite institutions in China score approximately 0.5–1.3 s.d. higher than students in elite institutions in India and Russia; students of all four years of study in non-elite institutions in China every year score 0.7–1.2 s.d. higher than students in non-elite institutions in India and 0.2 s.d. higher than students in non-elite institutions in Russia (except for year 2 and year 3, when they score at the same level). Regarding maths and physics skill levels at the start of the first year and end of the second year, students in elite institutions in China score approximately 0.5–2 s.d. higher than students in elite institutions in India and Russia; students in non-elite institutions in China score approximately 0.3–1.0 s.d. higher than students in non-elite institutions in India and Russia. Notably, freshmen in non-elite institutions in China exhibit substantially higher levels of critical thinking skills compared with freshmen in elite institutions in India (this gap closes by year 4), and higher levels of maths and physics skills compared with freshmen in elite institutions in Russia (the gap in maths but not physics skills closes by year 2).

There are large differences in critical thinking and academic skill gains among students in elite and non-elite institutions both within and across countries (Table 3). Students in elite institutions in China do not experience any skill gains in critical thinking and maths and physics skills from the start of the first year to the end of the second year. Students in non-elite institutions in China experience a significant decrease in their critical thinking skills from the start of

Table 4 | Skill levels and gains for female and male students (s.d. units)

	China			India			Russia		
	Female	**Male**	**Difference**	**Female**	**Male**	**Difference**	**Female**	**Male**	**Difference**
Panel A: critical thinking									
Start of year 1									
s.d. units	0.871	0.986	−0.116	−0.451	−0.401	−0.049	0.594	0.626	−0.032
P			0.123			0.232			0.794
95% CI			−0.264–0.033			−0.131–0.032			−0.279–0.215
End of year 2									
s.d. units	1.009	1.039	−0.030	−0.429	−0.386	−0.042	0.700	0.652	0.048
P			0.801			0.295			0.802
95% CI			−0.271–0.211			−0.124–0.038			−0.339–0.435
Year 1 to year 2 gains									
s.d. units	0.151	0.083	0.068	0.036	0.040	−0.005	0.125	0.096	0.029
P	0.360	0.220	0.577	0.178	0.060	0.866	0.340	0.111	0.832
95% CI	−0.181–0.483	−0.052–0.218	−0.178–0.315	−0.017–0.088	−0.002–0.082	−0.060–0.051	−0.139–0.388	−0.023–0.215	−0.248–0.307
Start of year 3									
s.d. units	0.700	1.006	−0.306	−0.401	−0.335	−0.066	0.881	0.604	0.277
P			0.029			0.090			0.137
95% CI			−0.578 to −0.033			−0.144–0.011			−0.093–0.647
End of year 4									
s.d. units	0.226	0.588	−0.362	−0.496	−0.474	−0.023	0.687	0.498	0.189
P			0.003			0.564			0.319
95% CI			−0.593 to −0.130			−0.102–0.056			−0.191–0.568
Year 3 to year 4 gains									
s.d. units	−0.438	−0.490	0.052	−0.080	−0.135	0.055	−0.109	−0.104	−0.004
P	0.000	0.000	0.711	0.007	0.000	0.056	0.171	0.065	0.959
95% CI	−0.572 to −0.304	−0.745 to −0.234	−0.230–0.333	−0.137 to −0.022	−0.200 to −0.071	−0.001–0.112	−0.268–0.050	−0.216–0.007	−0.176–0.168
Panel B: maths									
Start of year 1									
s.d. units	0.827	0.978	−0.151	−0.384	−0.280	−0.104	−0.004	0.006	−0.010
P			0.099			0.017			0.836
95% CI			−0.331–0.030			−0.189 to −0.019			−0.112–0.091
End of year 2									
s.d. units	0.704	0.681	0.023	0.023	0.021	0.002	0.196	0.075	0.122
P			0.830			0.962			0.019
95% CI			−0.194–0.240			−0.094–0.099			0.021–0.222
Year 1 to year 2 gains									
s.d. units	−0.193	−0.358	0.165	0.472	0.328	0.144	0.161	0.053	0.108
P	0.001	0.000	0.022	0.000	0.000	0.000	0.002	0.277	0.058
95% CI	−0.304 to −0.083	−0.469 to −0.247	0.026–0.304	0.432–0.512	0.290–0.367	0.086–0.201	0.066–0.256	−0.044–0.150	−0.004–0.220
Panel C: physics									
Start of year 1									
s.d. units	0.630	0.938	−0.307	−0.300	−0.189	−0.111	−0.295	−0.206	−0.089

Continued

Table 4 | Skill levels and gains for female and male students (s.d. units) (continued)

	China			India			Russia		
	Female	Male	Difference	Female	Male	Difference	Female	Male	Difference
P			0.001			0.059			0.482
95% CI			−0.480 to −0.135			−0.226–0.004			−0.343–0.166
End of year 2									
s.d. units	0.689	0.905	−0.216	−0.030	0.033	−0.063	0.039	0.016	0.023
P			0.040			0.041			0.695
95% CI			−0.421 to −0.011			−0.123 to −0.002			−0.094–0.140
Year 1 to year 2 gains									
s.d. units	0.057	−0.106	0.163	0.267	0.227	0.040	0.205	0.158	0.047
P	0.239	0.032	0.002	0.000	0.000	0.463	0.107	0.053	0.727
95% CI	−0.040–0.154	−0.202 to −0.010	0.064–0.262	0.217–0.318	0.114–0.340	−0.069–0.150	−0.048–0.458	−0.002–0.318	−0.227–0.322

For critical thinking, one cohort of students took exams in the first semester of their first year and then again at the end of the second semester of their second year, while another cohort took exams in the first semester of their third year and then again at the end of the second semester of their fourth year. For maths and physics, students took exams in the first semester of their freshman year and then again at the end of the second semester of their second year. The start of years 1 and 3 results were estimated using the sample of students present in baseline phase, while the end of years 2 and 4 results were estimated using the sample of students who were present in follow-up phase. By contrast, gains were calculated on the basis of students who were present in both the baseline and follow-up phases. As such, the difference between the years 2 and 4 and years 1 and 3 level estimates are not strictly the same as the gain estimates. Level and gain estimates are reported as effect sizes (in s.d. units). Scaled exam scores were divided by the subject-specific baseline mean and s.d. of the China, India and Russia cross-national sample of exam takers. China, India and Russia data are from national random samples of four-year undergraduate CS-related and electrical-engineering-related majors. Analytical estimates from China, India and Russia were calculated using sampling weights such that they are representative of well-defined national populations. To adjust for exam motivation, estimates were calculated using data for students who attempted at least 75% of the items on a test. Results were substantively the same with and without adjustment. s.e. values were adjusted for clustering at the institution level. *P* values and 95% CIs are shown.

the third year to the end of the fourth year (−0.505 s.d., $P<0.001$, 95% CI = −0.681 to −0.330) and in maths skills from the start of the first year to the end of the second year (−0.333 s.d., $P<0.001$, 95% CI = −0.437 to −0.229), and experience no gains in physics skills. Whereas the physics skills of students in elite institutions in India decrease during the first two years (−0.403 s.d., $P=0.006$, 95% CI = −0.650 to −0.156) and critical thinking skills decrease in the final two years (−0.343 s.d., $P=0.019$, 95% CI = −0.612 to −0.075), they make substantial gains in maths skills (0.306 s.d., $P<0.001$, 95% CI = 0.210–0.402). Furthermore, students in non-elite institutions in India make significant gains in both maths (0.387 s.d., $P<0.001$, 95% CI = 0.356–0.419) and physics (0.272 s.d., $P<0.001$, 95% CI = 0.202–0.342) skills but experience a decrease in critical thinking skills during the final two years (−0.102 s.d., $P=0.001$, 95% CI = −0.159 to −0.046). Students in elite institutions in Russia appear to make gains in maths (0.257 s.d., $P<0.001$, 95% CI = 0.178–0.335) but not in physics and critical thinking, whereas students in non-elite institutions make gains in physics (0.244 s.d., $P<0.001$, 95% CI = 0.129–0.360) but not in maths and critical thinking.

Skills by gender. There are small differences in skill levels and gains by gender (Table 4). At the start of university, female students exhibit similar levels of critical thinking skills as male students in China, India and Russia. Female freshmen in China and India have slightly lower maths and physics scores compared with male freshmen (0.1–0.3 s.d.). Female freshmen in Russia score at the same level as male freshmen in maths and physics.

During the first two years of university, female and male students make similar gains in critical thinking. By the end of year 4, female students in India and Russia have similar scores in critical thinking while female students in China score 0.3 s.d. lower compared with male students.

During the first two years of university, female students in China, India and Russia make higher gains in maths compared with male students, closing the gender gap in China and India and outperforming male students in Russia by 0.1 s.d. by the end of year 2. By contrast, the gender gap in physics persists in China and India during the first two years—female students score 0.1–0.2 s.d. lower than male students by the end of year 2. Female students in Russia score at the same level as male students in physics at the end of year 2.

Discussion

Regarding how well students are prepared at the start of university, freshmen in China and the United States have a large head start over freshmen in India and Russia in critical thinking. Freshmen in China also have a large head start over freshmen in India and Russia in maths and physics. Freshmen in India are also far behind freshmen in Russia in critical thinking but are competitive in maths and physics. The especially low levels of critical thinking skills in India may be due not only to lower economic levels, a higher prevalence of health impairments that impede early cognitive development and fewer inputs per student in pretertiary schooling, but also to an overemphasis on rote academic learning at the expense of higher-order cognitive skills[44–47].

China's high levels of critical thinking and academic skills at the start and middle of university are noteworthy given the large number of computer scientists and engineers that it produces (approximately eight times more than the United States)[25,43]. China's elite institutions, which score at an even higher level, produce almost 1.5 times as many computer scientists and engineers as all institutions in the United States combined[25]. The fact that China produces so many highly skilled individuals has implications for the global labour market for university STEM graduates. However, the fact that China has such high levels of skills does not necessarily imply that its pretertiary education system has a greater capacity compared with that of other countries to prepare students for university. The percentage of 18–22-year-old individuals who attend universities that offer STEM undergraduate (Bachelor's) programs is relatively small in China (8–10%; similar to India) compared with Russia (35–40%)[43].

In interpreting skill level differences across countries, it is also important to consider selection into majors. For China, Russia and

the United States, achievement differences between CS and engineering major students versus other major students at the start of university are modest within each country. In China, in the science track (roughly two-thirds of all four-year university students), CS and engineering major students scored approximately the same on the university entrance exam as non-CS and engineering major students[43]. In Russia, CS and engineering major students scored 0.26 s.d. higher on the maths module of the university entrance exam (and 0.22 s.d. lower on the language module) compared with students in other majors. In the United States, freshmen planning to enter CS and engineering majors score ~0.25 s.d. higher on a 12th grade maths exam compared with students planning to enter other majors (calculated using the nationally representative High School Longitudinal Study 2016 V1.0 dataset). National-level standardized data to examine between-major differences were unavailable for India.

For skill gains, whereas students in China, India and Russia make no gains or even losses in critical thinking during university, students in the United States make significant gains. The latter evidence is consistent with a handful of non-representative studies from the United States that explored critical thinking skill gains in a wide range of majors[2,48]. Although further research is needed to explain the lack of improvement in critical thinking in China, India and Russia both in absolute terms and relative to the United States, one possible reason may be that STEM undergraduates in these countries are required to take fewer courses in the humanities and social sciences compared with students in the United States[43]. Another potential reason is that university instruction tends to be less active in these countries, especially during the final two years of study[49,50].

The substantial losses in academic skills among students in elite and non-elite institutions in China—as opposed to the gains in India and Russia—are striking and perhaps unexpected. The results are robust even after accounting for (negligible) differences in test-taking motivation in each assessment wave (Supplementary Information D). A possible contributor to the skill losses is that students in China are rarely forced out of courses or programs for poor performance and may therefore be less motivated to study[25]. Another possible reason is that Chinese instructors, despite a similar maths and physics course load[43], tend to assign less homework and reading outside of class, which could also be associated with students' limited learning[51,52]. In contrast to students in China, students in India and Russia may exhibit gains because they are held accountable—through regular assessment and risk of failure—for learning skills[43].

Skill gains seem to be due to time spent attending classes or doing schoolwork directly related to classes rather than time spent on receiving tutoring or mentoring outside of class. Supplementary Table 2 shows the number of hours spent on in-university and out-of-university activities at elite and non-elite institutions in the three countries. Only 1–7% of students' total study time is spent on receiving tutoring or mentoring outside of class. These are most likely upper-bound estimates, as we cannot distinguish between outside tutoring and mentoring (which may also be from in-university peers or faculty). As the vast majority of time spent studying is on class-related studies rather than outside tutoring or mentoring, skill gains probably reflect value-added associated with receiving a university education.

Furthermore, the observed differences in skill levels and gains across countries and between elite and non-elite institutions can be attributed to higher education systems and institutions, and not to differences in family background and out-of-university activities (such as outside tutoring, internships, a paid job and volunteering) among countries and institutional types. Specifically, in estimating differences in skill levels and gains between countries, we controlled for family background and out-of-university activities (details are provided in Supplementary Information F). The magnitude of the large differences in skill levels between China and India are similar whether or not we control for family background and out-of-university activities. By contrast, Russia falls further behind China and India in skill levels after adjusting for family background and out-of-university activities. These changing score gaps between Russia and the other two countries are predictable, as students in Russia have higher levels of family wealth and are much more likely to have parents who are university educated (and such socioeconomic factors are almost always positively correlated with test scores). Controlling for family background and out-of-university activity does not substantively change cross-country differences in skill gains nor within country differences between elite and non-elite universities.

Finally, according to our results, universities seem to be closing gender gaps in maths (in China, India and Russia) and critical thinking (in India and Russia), which can have implications for increasing the equal representation of women in the STEM workforce[53]. Our study complements earlier research from China[48] that suggests that female STEM students exhibit higher learning gains compared with male students despite a lower or the same level of academic achievement at the start of university. That being said, a moderate gender gap in physics persists through the first two years of study in China and India. The persistence of this gap as well as initial gender gaps in maths and physics at the start of university indicate that countries need to invest more in improving student achievement in maths and science at the secondary level or that STEM programs in these countries have room to attract higher achieving female students[29].

A limitation of our data and analysis, due to resource constraints, is that we focus on two majors. Thus, although our findings on skill levels do highlight differences in the abilities of students in two important STEM fields across countries and institutions, they should not be misinterpreted as proxies for the quality of entire education systems. Furthermore, our findings on skill gains, which proxy for university quality, may not necessarily generalize to other fields of study. That being said, the findings represent cross-national, representative information on skill acquisition in university.

Specifically, our findings contribute to the literature on human capital development and its relationship with productivity and growth in several ways. First, the large variation in skill gains across countries and institutions underscores the need for more research concerning skill development in university. The fact that, on net, China, India and Russia experience no gains in critical thinking and China experiences absolute losses in academic skills indicates that higher education systems, including elite and non-elite institutions, often do not prepare students for skill-biased technological change. Although a large microeconomic literature is concerned with issues of university access and completion[54–56] and skill development in pretertiary education[57,58], it rarely considers skill development in university.

Second, by using only cognitive skill measures of primary and secondary school-age students, recent studies on human capital and economic growth implicitly assume that the skills gained by nations in pretertiary education are comparable to skills gained in tertiary education[59]. However, the evidence presented here reveals that substantial heterogeneity—absolute gains, no gains and even absolute losses in university skills—exists across countries. A closely related point is that understanding the production and availability of a country's human capital requires understanding how pretertiary and higher education systems interact to produce economically relevant skills. We also provide indirect evidence on the signalling value of a university degree[32]. Recent studies have suggested that there is little or no signalling value in a high school diploma[60]. However, overall negative learning gains combined with high economic returns to a university or elite university degree in China[43,61]

suggest that a university diploma may have a large signalling value in certain contexts. In such contexts, the social return to university is much lower than the private return, calling into question the efficiency of public investments in higher education.

Third, we provide insights into the stock of human capital regionally and globally. Although evidence is available on the state of pretertiary education in China, India and Russia[62–64], the lack of evidence about skill levels in tertiary education has led to an incomplete picture about human capital development in these countries. Our results shed light on the ability of these major world powers to produce skilled graduates in STEM fields, which may be critical for economic development and global competitiveness[8]. Given the propensity of students from China, India and Russia to migrate, our results also provide context for trends in the global migration of highly skilled STEM workers to developed countries such as the United States[65–67].

Methods

The Institutional Review Board approval for this research project was approved by Stanford University (IRB#31585). Informed consent was obtained from all of the participants. No compensation was provided to the participants. Data collection and analysis were not performed blind to the hypotheses.

Sampling and analysis in China, India and Russia. We sampled CS and electrical engineering major students who together comprise a large proportion of STEM undergraduates in China (34%), India (24%) and Russia (24%). We first identified all undergraduate (Bachelor's degree) CS and electrical engineering programs from China, India and Russia that had comparable course requirements and content with undergraduate CS and electrical engineering programs in the United States. Using the population frame of all higher education institutions with these undergraduate CS and electrical engineering programs, we then randomly sampled institutions from these countries. In brief, from China, we took a simple random sample of six institutions from each of six representative provinces. In India and Russia, we took stratified national random samples of 50 and 34 universities, respectively. Together, we sampled 7 elite and 29 non-elite institutions in China, 8 elite and 42 non-elite institutions in India, and 6 elite and 28 non-elite institutions in Russia. Further information about the sampling of institutions is provided in the Supplementary Information.

Next, we randomly sampled administrative units within the sample institutions. In each randomly selected administrative unit, we sampled all of the freshmen and third-year students. We randomly assigned half of the students in each year to take year-specific maths and physics exams, one quarter of the students to take a critical thinking exam and one quarter of the students to take a quantitative literacy exam. All electrical engineering programs and the vast majority of CS programs in China, India and Russia teach maths and physics courses, and almost entirely during the first two years. However, as a minority of CS programs do not teach physics classes during the first two years, a small proportion of the sampled third-year students in Russia (18.6%) and China (0.7%) took an informatics exam rather than the physics exam. Our estimates of physics skills are therefore based on the sample of students who were required to take physics courses in their programs. Response rates in the baseline were high with 95% of enrolled students taking the exams in China, 95% in India and 87% in Russia. Together, 5,102 freshmen and 4,145 third-year students from China, 8,232 freshmen and 9,223 third-year students from India, and 2,607 freshmen and 2,096 third-year students from Russia participated. Among the freshmen, 36% of the participants were female, 64% of participants were male; the average age was 18.4 years (further details are provide in Supplementary Table 1). Among the third-year students, 39% of participants were female, 61% of the participants were male; the average age was 20.5 years. No statistical methods were used to predetermine the sample sizes. To the best of our knowledge, our sample sizes are substantially larger than those of previous studies that assess skills in university using standardized assessments and nationally representative (random) samples[27].

We conducted follow-up testing after almost two years with the different subsets of freshmen and third-year students from the baseline (when they were at the end of their second and fourth years). Freshmen who had taken maths and physics tests in the baseline took an end-of-year-2-appropriate maths and physics test in the follow up, while freshmen and third-year students who took critical thinking in the baseline took the critical thinking test in the follow up (at the end of year 2 and the end of year 4, respectively). Response rates in the follow up were again relatively high, with 80% of enrolled students taking the exams in China, 95% in India and 90% in Russia.

We generated estimates of skill levels and gains in several steps. To estimate the skill level for a particular country or institutional type, for a particular year in university (start of year 1 or end of year 2 for cohort 1; start of year 3 or end of year 4 for cohort 2), and for a particular subject test (maths, physics or critical thinking), we calculated the mean score for students in that country, institutional type, year and test.

To estimate skill gains for a particular country (and, when applicable, institutional type), subject test and student cohort (start of year 1 to end of year 2 for cohort 1; or start of year 3 to end of year 4 for cohort 2), we ran the following regression:

$$Y_{ijt} = \beta_0 + \beta_1 F_{ijt} + \varepsilon_{ijt} \quad (1)$$

where Y_{ijt} is a subject-specific test score (for example, maths) for student i in university j at time t (baseline or follow up); F_{ijt} is a dummy variable indicating follow up (as opposed to baseline) and ε_{ijt} is an error term.

Our estimates of start of year 1 (as well as start of year 3) skill levels use the sample of students present in baseline phase. Our primary estimates of end of year 2 (as well as end of year 4) skill levels used the entire sample of students present in the follow-up phase. Gains were calculated on the basis of students present in both the baseline and follow up phases. As such, the difference between year 2 and year 1 level (and, similarly, year 4 and year 3 levels) estimates are not strictly the same as the gain estimates. We also calculated two alternative sets of gain estimates using multiple imputation and (1) all of the students in the baseline (regardless of whether they were in the follow up); or (2) all students in the follow up (regardless of whether they were in the baseline). The unadjusted gain estimates and adjusted gain estimates were all substantively the same (Supplementary Table 3a,b).

To compare skill levels across countries, we ran the following regression on the sample students of a particular cohort at a particular point in time (the start of year 1 or the end of year 2 for cohort 1; the start of year 3 or the end of year 4 for cohort 2) who took a particular test (maths, physics or critical thinking):

$$Y_{ij} = \alpha_0 + \mathbf{C}'_{ij}\alpha + \varepsilon_{ij} \quad (2)$$

where $\mathbf{C}'_{ij}$ is a vector of country dummies (binary indicators for India, Russia when the dependent variable is a student's maths or physics score and binary indicators for India, Russia and the United States when the dependent variable is a student's critical thinking score). Coefficient estimates on the country indicators indicate pairwise differences in skill levels between India, Russia and the United States on the one hand and the left-out country (China) on the other; we used the Stata 15.1 command --lincom- to compute point estimates and s.e. values for the remaining pair-wise comparisons.

To compare skill levels across elite and non-elite institutions (or across female and male students), we ran a regression similar to that of equation (2), but substituted the country dummies with a single binary indicator of elite versus non-elite institutional status (or female versus male).

To compare skill gains across countries, we ran the following regression on the entire sample students of a particular cohort (the start of year 1 to the end of year 2 cohort or the start of year 3 to the end of year 4 cohort) who took a particular test (maths, physics or critical thinking):

$$Y_{ijt} = \gamma_0 + \gamma_1 F_{ijt} + \mathbf{C}'_{ijt}\gamma + F_{ijt} \times \mathbf{C}'_{ijt}\delta + \varepsilon_{ijt} \quad (3)$$

Similarly, to compare skill levels across elite and non-elite institutions (or across female and male students), we ran a regression similar to that of equation (2), but substituted the country dummies with a single binary indicator of elite versus non-elite institutional status (or female versus male).

Finally, to examine the extent to which differences in skill levels and gains are explained by differences in country and institutional type versus other factors, we ran the various iterations of equations (1), (2) and (3) with different sets of baseline control measures (Supplementary Table 5). These sets of baseline control measures included socioeconomic status (mother went to university, father went to university and a wealth index based on household assets) and the degree of participation in out-of-university activities (tutoring and part-time work as well as participation in internships, entrepreneurial activities, community service or volunteer work, and religious organizations).

To ensure national representativeness, we adjusted all of our analytical estimates and s.e. values for survey design features including multistage sampling and probability sampling weights (Supplementary Information). We also estimated both unadjusted (using listwise deletion) and adjusted (using multiple imputation; Supplementary Information) estimates of skill gains. As skill gain estimates were substantively the same in either case, we reported only unadjusted estimates in the main text (adjusted estimates are provided in the Supplementary Information).

Exams and exam administration in China, India and Russia. The critical thinking exam is part of the HEIghten suite of assessments from Educational Testing Service (ETS). The construct that the exam measures was defined according to a systematic review of research on critical thinking in higher education; it reflects the ability to develop sound and valid arguments, evaluate evidence and its use, understand implications and consequences, and differentiate between causation and explanation[68] (Supplementary Information). The exam was designed to be culturally neutral, such that it could be given to students in different national contexts. The same critical thinking exam was given to first- and third-year students in the baseline. It was also given, almost two years later, to the same students in the follow up.

The maths and physics exams were specially designed to examine skills among first year and end of second year (equivalently start of third year) CS and electrical engineering students across countries and institutions[69] (Supplementary Information). Exams were year-specific, testing students on the maths and physics skills that they were supposed to have learned by the start of their first and end of their second years of university. The year-specific exams for each subject contained a substantial number of anchor items that enabled scores to be equated across years. The year-specific exams were also identical across countries, testing students on content areas that were validated to be common and important across countries and across years (and across elite and non-elite institutions).

For each type of test, scores were scaled to be comparable across countries and years (further details are provided in Supplementary Information B). Scaled scores were further converted into z scores (with a mean of 0 and a s.d. of 1) for the sake of interpretability. To create z scores for critical thinking, we used the survey-weighted mean and s.d. of critical thinking scores from the baseline survey across China, India and Russia. To create z scores for maths and physics, we used the survey-weighted mean and s.d. of IRT-scaled scores from both the baseline and follow-up phases across China, India and Russia.

We took steps to ensure that exam-taking conditions were as similar as possible across countries and institutions. First, exams were given approximately halfway through the first semester of the academic year in each country. Specifically, late November and early December 2015 for China and Russia and late October and early November 2017 for India. Whereas the academic year typically begins in August in India, it typically begins in September in China and Russia. Second, as previously mentioned, we had high and comparable student participation rates in each country—well above the PISA 2015 minimum participation rate requirement of 80%. Given its very low rate, non-response bias did not change the main conclusions of the paper. Third, we followed a rigorous multistage translation, adaptation and review process for the exams (Supplementary Information). Fourth, the exams were introduced and proctored in the same way by trained enumerators. Fifth, proctors provided students with the same incentives to participate—in particular, all of the students were given the option of receiving an individualized report of their exam performance after the completion of the study (we also consider exam motivation; Supplementary Information D).

After both exams were completed, students responded to a questionnaire. In the questionnaire, students were asked about their age, gender, father's education level, mother's education level and whether they took the university entrance exam in their own country. Summary statistics for these student background variables, adjusted for sample weights, are presented in Supplementary Table 1. We also asked a random subset of students (third-year students who took the critical thinking or quantitative literacy tests in the baseline) about the time that they spent studying through attending class, doing schoolwork directly related to classes and receiving tutoring or mentoring outside of class.

Sampling, exam administration and analysis for the United States. Data on the critical thinking skills of students in universities in the United States were collected from 2016 to 2018 by ETS. We used a subsample of STEM Bachelor's degree program students from a range of institutions in the United States to create comparative benchmarks of critical thinking skill levels. The sample of STEM major Bachelor's degree students was identified by asking students their prospective or actual major. In terms of Carnegie classifications, the sample includes 12 Doctoral research institutions (1035 students or 65% of the sample), 22 Masters institutions (473 students or 30% of the sample) and 9 Baccalaureate institutions (90 students or 6% of the sample). Approximately 45% of the sampled students were in fact from the highest ranking R1 institutions—Doctoral universities, institutions with the highest research activity. As the distribution of STEM Bachelor's degree program students in the United States is 67%, 24% and 9% across Doctoral research, Masters and Baccalaureate institutions (with 44% in R1 institutions), the across-institution distribution of students in the sample is similar to that of STEM students in Bachelor's degree programs in the United States.

We estimated regression-adjusted gains to account for potential inconsistencies in sampling students across years as well as much higher rates of dropout in STEM programs in the United States. We controlled for ACT/SAT equivalent scores (total and maths separately) as well as information on age, gender, minority status (yes or no), whether English is spoken at home (yes or no) and high-school GPA to obtain the adjusted skill gain estimates for students in the United States. GPA values were divided into high (3.5 to 4.0), medium (3.0 to 3.5), low (under 3.0) and 'not reported' (16% of the sample) categories. ACT/SAT equivalent scores were available for 51% of the sample. We dealt with missingness by including missing value dummies in the regression. The results were substantively the same when using listwise deletion and including only ACT/SAT total scores, only ACT/SAT maths scores, both or neither.

We further validated our across-cohort estimates of critical thinking skill gains in the United States during four years of university by comparing them with skill gain estimates from other major studies based on longitudinal data[2,6]. Our reported effect size of critical thinking skills over four years in university is similar to effect sizes reported in these two major studies.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data have been deposited at the Open Science Framework (https://osf.io/4t8cu/).

Code availability

Stata do-files used to perform the analyses have been deposited at the Open Science Framework (https://osf.io/4t8cu/).

Received: 6 June 2020; Accepted: 27 January 2021;
Published online: 1 March 2021

References

1. National Academies of Sciences, Engineering, and Medicine. *Indicators for Monitoring Undergraduate STEM Education* (National Academies Press, 2017).
2. Mayhew, M. J. et al. *How College Affects Students Volume 3: 21st Century Evidence that Higher Education Works* (Jossey-Bass, 2016).
3. *Criteria for Accrediting Engineering Technology Programs* (ABET, 2015); www.abet.org/wp-content/uploads/2015/10/T001-16-17-ETAC-Criteria-10-16-15.pdf
4. Passow, H. J. Which ABET competencies do engineering graduates find most important in their work? *J. Eng. Educ.* **101**, 95–118 (2012).
5. National Research Council. *Education for Life and Work: Developing Transferable Knowledge and Skills in the Twenty-First Century* (National Academies Press, 2012).
6. Arum, R. & Roksa, J. *Academically Adrift: Limited Learning on College Campuses* (Univ. Chicago Press, 2011).
7. AAC&U. *How Should Colleges Assess and Improve Student Learning? Employers' Views on the Accountability Challenge* (AAC&U, 2008).
8. National Academy of Science. *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future* (National Academy of Science, 2005).
9. Goldin, C. D. & Katz, L. F. *The Race Between Education and Technology* (Harvard Univ. Press, 2009).
10. Autor, D. H., Levy, F. & Murnane, R. J. The skill content of recent technological change: an empirical exploration. *Q. J. Econ.* **118**, 1279–1333 (2003).
11. Bresnahan, T. F., Brynjolfsson, E. & Hitt, L. M. Information technology, workplace organization, and the demand for skilled labor: firm-level evidence. *Q. J. Econ.* **117**, 339–376 (2002).
12. Katz, L. F. & Krueger, A. B. Computer science inequality: have computers changed the labor market? *Q. J. Econ.* **113**, 1169–1213 (1998).
13. Lucas, R. E. On the mechanics of economic development. *J. Monetary Econ.* **22**, 3–42 (1988).
14. Romer, P. M. Endogenous technological change. *J. Polit. Econ.* **98**, S71–S102 (1990).
15. Mankiw, N. G., Romer, D. & Weil, D. N. A contribution to the empirics of economic growth. *Q. J. Econ.* **107**, 407–437 (1992).
16. Murnane, R. J., Willett, J. B. & Levy, F. The growing importance of cognitive skills in wage determination. *Rev. Econ. Stat.* **77**, 251–266 (1995).
17. Hanushek, E. A. & Kimko, D. D. Schooling, labor-force quality, and the growth of nations. *Am. Econ. Rev.* **90**, 1184–1208 (2000).
18. Hanushek, E. A. & Woessmann, L. The role of cognitive skills in economic development. *J. Econ. Lit.* **46**, 607–668 (2008).
19. Hanushek, E. & Woessmann, L. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *J. Econ. Growth* **17**, 267–321 (2012).
20. Hanushek, E. & Woessmann, L. Schooling, educational achievement, and the Latin American growth puzzle. *J. Dev. Econ.* **99**, 497–512 (2012).
21. Kaarsen, N. Cross-country differences in the quality of schooling. *J. Dev. Econ.* **107**, 215–224 (2014).
22. Heckman, J., Stixrud, J. & Urzua, S. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Labor Econ.* **24**, 411–482 (2006).
23. Deming, D. & Kahn, L. B. Skill requirements across firms and labor markets: evidence from job postings for professionals. *J. Labor Econ.* **36**, S337–S369 (2018).
24. World Bank. *World Development Report 2018: Learning to Realize Education's Promise* (World Bank, 2018).
25. Loyalka, P. et al. Factors affecting the quality of engineering education in the four largest emerging economies. *High. Educ.* **68**, 977–1004 (2014).
26. Musekamp, F. & Pearce, J. Assessing engineering competencies: the conditions for educational improvement. *Stud. High. Educ.* **40**, 505–524 (2015).
27. Zlatkin-Troitschanskaia, O., Shavelson, R. J. & Kuhn, C. The international state of research on measurement of competency in higher education. *Stud. High. Educ.* **40**, 393–411 (2015).
28. OECD. *Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study Report—Volume 2 Data Analysis and National Experiences* (OECD, 2012).

29. Loyalka, P. et al. Computer science skills across China, India, Russia, and the United States. *Proc. Natl Acad. Sci. USA* **116**, 6732–6736 (2019).
30. OECD. *Education at a Glance 2018* (OECD, 2018).
31. Montenegro, C. E. & Patrinos, H. A. *Comparable Estimates of Returns to Schooling Around the World* (World Bank, 2014).
32. Spence, M. Job market signaling. *Q. J. Econ.* **87**, 355–374 (1973).
33. Moretti, E. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *J. Econ.* **121**, 175–212 (2004).
34. Aghion, P., Howitt, P., Howitt, P. W., Brant-Collett, M. & García-Peñalosa, C. *Endogenous Growth Theory* (MIT press, 1998).
35. Nelson, R. R. & Phelps, E. S. Investment in humans, technological diffusion, and economic growth. *Am. Econ. Rev.* **56**, 69–75 (1966).
36. Benhabib, J. & Spiegel, M. M. Human capital and technology diffusion. *Handb. Econ. Growth* **1**, 935–966 (2005).
37. Katz, L. F. & Murphy, K. M. Changes in relative wages, 1963–1987: supply and demand factors. *Q. J. Econ.* **107**, 35–78 (1992).
38. Autor, D. H., Katz, L. F. & Krueger, A. B. Computing inequality: have computers changed the labor market? *Q. J. Econ.* **113**, 1169–1213 (1998).
39. National Science Board. *Science and Engineering Indicators* 2016 (National Science Foundation, 2016).
40. Bourdieu, P. *Homo Academicus* (Stanford Univ. Press, 1988)
41. Shavit, Y., Arum, R. & Gamoran, A. (eds) *Stratification in Higher Education: A Comparative Study*. (Stanford Univ. Press, 2007).
42. Altbach, P. G. & Salmi, J. (eds) *The Road to Academic Excellence: The Making of World-Class Research Universities* (The World Bank, 2011).
43. Carnoy, M. et al. *University Expansion in a Changing Global Economy: Triumph of the BRICs?* (Stanford Univ. Press, 2013).
44. World Bank. *World Development Indicators* 2018 (World Bank, 2018); http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators
45. Muralidharan, K., Das, J., Holla, A. & Mohpal, A. The fiscal cost of weak governance: evidence from teacher absence in India. *J. Publ. Econ.* **145**, 116–135 (2017).
46. Burdett, N. *Review of High Stakes Examination Instruments in Primary and Secondary School in Developing Countries* Research on Improving Systems of Education (RISE) Working Paper 17/018 (RISE Programme, 2017).
47. Cheney, G. R., Ruzzi, B. B. & Muralidharan, K. *A Profile of the Indian Education System* (National Center on Education and Economy, 2005).
48. Guo, C. B., Tsang, M. & Ding, X. H. Gender difference in the education and employment of science & engineering students in HEIs. *J. High. Educ.* **11**, 89–101 (2007).
49. Dobryakova, M. & Froumin, I. Higher engineering education in Russia: incentives for real change. *Int. J. Eng. Educ.* **26**, 1032–1041 (2010).
50. Pal, Y. *Report of the Committee to Advise on Renovation and Rejuvenation of Higher Education* (Department of Human Resources, Government of India, 2009).
51. Zha, Q., Hu, L. & Wang, Y. The characteristics of university students' extracurricular time assignment and its impact on learning outcome: based on China college student survey in J University in 2016. *High. Educ. Explor.* **7**, 44–49 (2017).
52. Lv, L. & Zhang, H. The characteristics of undergraduates' learning engagement of Chinese research-oriented university: based on the comparison of 12 research-oriented universities in the world. *Educ. Res.* **36**, 51–63 (2015).
53. Mandel, H. & Semyonov, M. Gender pay gap and employment sector: sources of earnings disparities in the United States, 1970-2010. *Demography* **51**, 1597–1618 (2014).
54. Bound, J., Lovenheim, M. F. & Turner, S. Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. *Am. Economic J. Appl. Econ.* **2**, 129–157 (2010).
55. Bettinger, E. P., Long, B. T., Oreopoulos, P. & Sanbonmatsu, L. The role of application assistance and information in college decisions: results from the H&R Block FAFSA experiment. *Q. J. Econ.* **127**, 1205–1242 (2012).
56. Stinebrickner, T. & Stinebrickner, R. Learning about academic ability and the college dropout decision. *J. Labor Econ.* **30**, 707–748 (2012).
57. Glewwe, P. & Kremer, M. Schools, teachers, and education outcomes in developing countries. *Handb. Econ. Educ.* **2**, 945–1017 (2006).
58. McEwan, P. Improving learning in primary schools of developing countries: a meta-analysis of randomized experiments. *Rev. Educ. Res.* **85**, 353–394 (2013).
59. Hanushek, E. A. Will more higher education improve economic growth? *Oxford Rev. Econ. Pol.* **32**, 538–552 (2016).
60. Clark, D. & Martorell, P. The signaling value of a high school diploma. *J. Polit. Econ.* **122**, 282–318 (2014).
61. Li, H., Meng, L., Shi, X. & Wu, B. Does attending elite colleges pay in China? *J. Comp. Econ.* **40**, 78–88 (2012).
62. Muralidharan, K. *Priorities for Primary Education Policy in India's 12th Five-Year Plan* India Policy Forum Vol. 9, 1–61 (National Council of Applied Economic Research and Brookings Institution, 2013).
63. Li, H., Loyalka, P., Rozelle, S. & Wu, B. Human capital and China's future growth. *J. Econ. Perspect.* **31**, 25–48 (2017).
64. Kouzminov, Y., Ovcharova, L. & Yakobson, L. (eds) *How to Increase Human Capital and its Impact on Economic and Social Development* (HSE Publishing House, 2018).
65. Bound, J., Braga, B., Golden, J. & Khanna, G. Recruitment of foreigners in the market for computer scientists in the United States. *J. Labor Econ.* **33**, S187–S223 (2015).
66. Bound, J., Khanna, G. & Morales, N. Reservoir of foreign talent. *Science* **356**, 697 (2017).
67. Hanson, G. H. & Slaughter, M. J. in *Education, Skills, and Technical Change: Implications for Future US GDP Growth* (eds Hulten, C. R. & Ramey, V. A.) 465–494 (Univ. Chicago Press, 2017).
68. Liu, O. L., Frankel, L. & Roohr K. C. *Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment* ETS Research Report Series (ETS, 2014).
69. Kardanova, E. et al. Developing instruments to assess and compare the quality of engineering education: the case of China and Russia. *Assess. Eval. High. Educ.* **41**, 770–786 (2016).

Acknowledgements

We thank M. Carnoy, J. Cohen, T. Dee, B. Domingue, A. Eble, R. Fairlie, E. Hanushek, B. Kim, S. Loeb, K. Muralidharan, S. Reardon, S. Rozelle, D. Schwartz, S. Sylvia and C. Wieman, and participants at the demography workshop at the University of Chicago, the economics of education workshop at the Teachers College, the South Asia Region Knowledge Exchange Group at the World Bank, KDI School and technical reviewers at ETS for their feedback. We appreciate research funding from E. Li, the Basic Research Program of the National Research University Higher School of Economics and Russian Academic Excellence Project 5–100, and the All India Council for Technical Education. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

P.L., O.L.L., G.Li, I.C., E.K., N.Y., F.G., L.M., S.H., A.B., T.B. and N.T. designed research. P.L., O.L.L., G.Li, I.C., E.K., N.Y., F.G., L.M., S.H., H.W., Y.L., A.B. and S.K. performed research. P.L., O.L.L., E.K., D.F., L.G., G.Ling, S.K. and Z.S. analysed data. P.L., O.L.L. and I.C. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01062-3.

Correspondence and requests for materials should be addressed to P.L. or I.C.

Peer review information *Nature Human Behaviour* thanks Alex Eble, Thomas Luschei and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Charlotte Payne.

Reprints and permissions information is available at www.nature.com/reprints.

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

nature research

Corresponding author(s): Prashant Loyalka; Igor Chirikov

Last updated by author(s): December 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected in the field. Data used to perform the analyses have been deposited in Open Science Framework (https://osf.io/4t8cu/)
Data analysis	Stata do-files used to perform the analyses have been deposited in Open Science Framework (https://osf.io/4t8cu/)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and Stata do-files used to perform the analyses have been deposited in Open Science Framework (https://osf.io/4t8cu/)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Colleges contribute to economic growth and national competitiveness by equipping students with higher order thinking and academic skills. Despite large investments in college STEM education, little is known about how STEM undergraduates' skills compare across countries and by institutional selectivity. We provide direct evidence on these issues by collecting and analyzing longitudinal data on tens of thousands of computer science and electrical engineering students in China, India, Russia and the United States. We find stark differences in skill levels and gains among countries and by institutional selectivity. Compared to the United States, students in China, India, and Russia do not experience critical thinking skill gains over four years. While students in India and Russia experience academic skill gains in the first two years, students in China do not. These gaps in skill levels and gains provide insights into the global competitiveness of STEM college students across nations and institutional types.
Research sample	Nationally representative (random) samples of college STEM students (undergraduate students in four-year programs in computer science and electrical engineering) in elite and non-elite institutions in China, India, and Russia. A national but non-representative sample of college STEM students from a range of four-year undergraduate programs in the United States.
Sampling strategy	(1) Sampling in China, India, and Russia: We sampled computer science (CS) and electrical engineering (EE) major students that, taken together, comprise a large proportion of STEM undergraduates in China (34%), India (24%), and Russia (24%). We first identified all undergraduate (bachelor's degree) CS and EE programs from China, India, and Russia that had comparable course requirements and content with undergraduate CS and EE programs in the United States. Using the population frame of all higher education institutions with these undergraduate CS and EE programs, we then randomly sampled institutions from these countries. Briefly, from China, we took a simple random sample of six institutions from each of six representative provinces. In India and Russia, we took stratified national random samples of 50 and 34 universities, respectively. Altogether, we sampled 7 elite and 29 non-elite institutions in China, 8 elite and 42 non-elite institutions in India, and 6 elite and 28 non-elite institutions in Russia. For more information about the sampling of institutions, see the SOM. We next randomly sampled administrative units within the sample institutions. In each randomly selected administrative unit, we sampled all first year (freshmen) and third year (junior) students. We randomly assigned half of the students in each year to take grade-specific math and physics exams, one quarter of the students to take a critical thinking exam, and one quarter of the students to take a quantitative literacy exam. Response rates in the baseline were high with 95% of enrolled students taking the exams in China, 95% in India, and 87% in Russia. Altogether, 5,102 freshmen and 4,145 juniors from China, 8,232 freshmen and 9,223 juniors from India, and 2,607 freshmen and 2,096 juniors from Russia participated. We conducted follow-up testing after almost two years with the different subsets of freshmen and junior students from the baseline (when they were at the end of their sophomore and senior years). Freshmen that had taken math and physics tests in the baseline took end-of-year 2-appropriate math and physics test in the follow-up, while freshmen and juniors that took critical thinking in the baseline took the critical thinking test in the follow-up. Response rates in the follow-up were again relatively high with 80% of enrolled students taking the exams in China, 95% in India, and 90% in Russia. To ensure national representativeness, we adjusted our analytical estimates and standard errors for survey design features including multi-stage sampling and probability sampling weights (see the SOM). We also estimated both unadjusted (using listwise deletion) and adjusted (using multiple imputation—see the SOM) estimates of skill gains. Because skill gains estimates are substantively the same in either case, we only report unadjusted estimates in the main text (for adjusted estimates, see the SOM). (2) Sampling in the United States.—Data on the critical thinking skills of students in colleges in the United States were collected from 2016 to 2018 by Educational Testing Service (ETS). We use a subsample of STEM bachelor's degree program students from a range of institutions in the United States to create comparative benchmarks of critical thinking skill levels. In terms of Carnegie classifications, the sample includes 11 doctoral research institutions (672 students or 69% of the sample), 17 masters institutions (245 students or 25% of the sample), and 8 baccalaureate institutions (56 students or 6% of the sample). Approximately 53% of the sampled students were in fact from the highest ranking R1 institutions: Doctoral Universities – Highest Research Activity. Since the distribution of STEM bachelor's degree program students in the United States is 67%, 24%, and 9% across doctoral research, masters, and baccalaureate institutions (with 44% in R1 institutions), the across-institution distribution of students in the sample is similar to that of STEM students in bachelor's degree programs in the United States.
Data collection	Critical Thinking Exam: The critical thinking exam is part of the HEIghten® suite of assessments from Educational Testing Service (ETS). The construct the exam measures was defined according to a systematic review of research on critical thinking in higher education; it reflects the ability to develop sound and valid arguments, evaluate evidence and its use, understand implications and consequences, and differentiate between causation and explanation. The exam was designed to be culturally neutral, so that it could be given to students in different national contexts. The same critical thinking exam was given to first and third year students in the baseline. It was also given, almost two years later, to the same students in the follow-up. Scores were scaled to be comparable across countries and years and were further converted into z-scores for the sake of interpretability. Data on the critical thinking skills of students in

colleges in the United States were collected from 2016 to 2018 by Educational Testing Service (ETS).

Math and Physics Exams: The math and physics exams were specially designed to examine skills among first year and end of second year (equivalently start of third year) CS and EE students across countries and institutions. Exams were year-specific, testing students on the math and physics skills they were supposed to have learned by the start of their first and end of their second years of college. The year-specific exams for each subject contained a substantial number of anchor items which allowed scores to be equated across years. The year-specific exams were also identical across countries, testing students on content areas that were validated to be common and important across countries and across years (and across elite and non-elite institutions). We create scaled scores for comparing skill levels and gains across countries and over time. For the sake of interpretability, the scaled exam scores were again converted into z-scores.

Details of the math and physics test development and validation process is explained in more detail in Kardanova et al. (2016). The content of the start of first year math and physics exams for were aligned with common and core content that students cover in high school curricula and on high-stakes college entrance exams; the content of end of second year math and physics exams were aligned with the common and core content that students cover in the first two years of their undergraduate programs. The content validity, appropriateness, and translation of large pools of exam items were confirmed, item-by-item, with dozens of experts at elite and non-elite universities from the different countries. The larger pools of exam items were also piloted with approximately 4,000 start of first and start of third year CS and EE students in China, India, and Russia. Afterwards, the psychometric properties (item quality, reliability, unidimensionality, validity, scalability and cross-national comparability) of the exams were validated. The final math and physics exams, for freshmen and juniors separately, each contained 35 items and lasted for 40 minutes.

Exam conditions: We took steps to ensure that exam-taking conditions were as similar as possible across countries and institutions. First, exams were given approximately halfway through the first semester of the academic year in each country. Second, as previously mentioned, we had high and comparable student participation rates in each country—well above the PISA 2015 minimum participation rate requirement of 80%. Third, we followed a rigorous multi-stage translation, adaptation, and review process for the exams (see the SOM). Fourth, the exams were introduced and proctored in the same way by trained enumerators. Fifth, proctors provided students with the same incentives to participate—in particular, all students were given the option of receiving an individualized report of their exam performance after the completion of the study.

Survey questionnaire: After exams were completed, students responded to a questionnaire. In the questionnaire, students were asked about their age, gender, father's education level, mother's education level, and whether they took the college entrance exam in their own country. Summary statistics for these student background variables, adjusted for sample weights, are presented in Supplementary Table 1. We also asked a random subset of students (juniors that took the critical thinking or quantitative literacy tests in the baseline) about the time they spent studying through: attending class, doing schoolwork directly related to classes, and receiving tutoring or mentoring outside of class.

Timing

Baseline tests/surveys for grade 1 and 3 students were conducted in late November and early December of 2015 for China and Russia and late October and early November of 2017 for India. We conducted follow-up testing after almost two years with the same freshmen and junior students from the baseline (when they were at the end of their sophomore and senior years); some students that were not present in the baseline (for various reasons - they were in general either enrolled but absent during the baseline survey or they transferred into the college between the baseline and follow-up surveys) also participated in the follow-up survey.

Data exclusions

None for the estimation of critical thinking and math/physics skill levels and gains. We did exclude a random subsample of data on students that took the quantitative literacy exam (and we mention this in the paper). The reason the results of quantitative literacy exam were not included in the paper is the lack of longitudinal data in China and Russia. This exam was only administered in the baseline to year 1 students. It was not administered in the endline. All other exams (critical thinking, math and physics) were administered in both baseline and endline.

Non-participation

Participation/response rates in the baseline were high with 95% of enrolled students taking the exams in China, 95% in India, and 87% in Russia. Participation/response rates in the follow-up were again relatively high with 80% of enrolled students taking the exams in China, 95% in India, and 90% in Russia. As we note in the paper, given its very low rate, non-response bias does not change the main conclusions of the paper.

Randomization

The study examines and compares skill levels and gains across higher education systems and institutional types. It is therefore a descriptive and not causal study. We do not randomize students to groups therefore. We do, however, use strict multi-level survey sampling procedures (random/representative sampling at each level) and construct appropriate survey weights (in consultation with statistics experts) to ensure the representativeness of the results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	College STEM students from China, India, Russia, and the United States. Among freshmen, 36% of participants were female, 64% of participants were male; average age is 18.4 years (see Supplementary Table 1 in the SOM for more details). Among juniors, 39% of participants were female, 61% of participants were male; average age is 20.5 years.
Recruitment	Students were recruited from their college STEM programs.
Ethics oversight	The Institutional Review Board approval for this research project was approved by Stanford University (IRB#31585). We provide this information in the text.

Note that full information on the approval of the study protocol must also be provided in the manuscript.