

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](https://www.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.