Supplementary Information for

Who Trains the Best Computer Scientists? Cross-National Evidence from China, India, Russia, and the U.S.

Authors: Prashant Loyalka, Ou Lydia Liu, Guirong Li, Igor Chirikov, Elena Kardanova, Lin Gu, Guangming Ling, Ningning Yu, Fei Guo, Liping Ma, Shangfeng Hu, Angela Sun Johnson, Ashutosh Bhuradia, Saurabh Khanna, Isak Froumin, Jinghuan Shi, Pradeep Kumar Choudhury, Tara Beteille, Francisco Marmolejo, Namrata Tognatta

Prashant Loyalka
Correspondence to: loyalka@stanford.edu

This PDF file includes:

Supplementary text
Figs. S1-S2
Tables S1-S4
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1814646116

Materials and Methods

Sampling and Implementation in China, India, and Russia

We drew nationally representative samples of broadly, yet comparably, defined computer science (CS) students in China, India, and Russia. The sampling procedure was comprised of four steps.

In the first step, we identified a broad set of cross-nationally comparable CS majors or departments. In China, the CS departments included *Computer Science and Technology, Computer Software, Software Engineering, Network Engineering, Internet of Things, Information Security, Information Engineering,* and *Informational and Computational Science*. In India, the CS departments included *Computer Science* and *Computer Science Engineering, Information Science and Engineering,* and *Information Technology*. In Russia, the CS departments included *IT and Computer Facilities, Informational Systems and Technologies, Applied IT, Program Engineering, Mathematics and Computer Science, Fundamental IT and Information Technologies, Software and Administration of Information Systems,* and *Information Security*.

In the second step, we selected higher education institutions ("universities" in China and Russia and "colleges" in India) from each country that had these comparable CS programs. To do so, we used administrative data on (the population frame of) all higher education institutions with CS programs in each country From China, we randomly selected institutions from six provinces that represent the country's northern (Beijing and Shandong), central (Henan and Shaanxi), and southern (Guangdong and Sichuan) regions; the six provinces also represent economically more developed (Beijing and Guangdong ranked 2nd and 8th in GDP per capita out of 31 provincial-level administrative divisions), mid-developed (Shandong and Shaanxi ranked 10th and 14th) and less developed (Sichuan and Henan ranked 22nd and 24th) regions (*1*). From each of the six provinces, we took a simple random sample of 6 universities that offered bachelor's degree programs in CS. In India, we purposefully sampled 5 institutions each from the north (Delhi region), center (Madhya Pradesh), and south (Bangalore region) that offered CS bachelor's degree programs. The Indian institutions were chosen to reflect a representative range of institutions by rank and public versus private ownership.[1] In Russia, we took a simple random sample of 27 universities from all universities that offered bachelor's programs in CS and electrical engineering (EE) and a simple random sample of 7 universities from all universities that offered bachelor's programs in CS but none in EE.[2]

The national samples were representative of elite and non-elite institutions in each country. The definitions of elite and non-elite institutions are particular to each country and yet are standard and generally accepted in the literature *(2)*. In China, "elite" universities were defined on the basis of their designation as Project 985 or 211 universities. Project 985 universities are the top 39 universities in China whereas Project 211 universities are the top 112 universities in China, and both have received preferential government funding. In India, elite institutions were defined as the India Institutes of Technology (IITs), the Indian Institutes of Information

[1] Specifically, we sampled 1 Indian Institute of Technology (IIT), 1 National Institute of Technology (NIT), one "elite" (top 100 NIRF ranking) private college, and 12 "non-elite" private and government colleges.

[2] After the initial random sampling of institutions, one institution in China and one institution in Russia decided not to participate in the study. Using the same sampling procedures, we replaced these single institutions in each country with another randomly sampled institution. Subsequently, all selected institutions and their subsidiary administrative units participated in the study.

Technology (IIITs), the National Institutes of Technology (NITs), and other institutions that ranked in the top 100 of the National Institutional Ranking Framework (NIRF) rankings developed by the Ministry of Human Resource Development, Government of India. In Russia, "elite" universities were defined as National Research Universities, "5-100" universities, and Federal universities. These universities receive additional funding from the Russian government. Elite programs teach a small proportion of the total number of CS undergraduates in each country: 21% in China; 4% in India; and 26% in Russia (authors calculations using administrative data with complete national coverage in each country).[3]

In the third step, we randomly sampled students within CS programs in the selected universities. In China, we randomly sampled up to 2 CS departments from each sample university. If there was only one CS department in the university, we sampled only that department. We then randomly sampled 1 administrative group or "class" of fourth and final-year (senior) students within each department. In India, we randomly sampled 1 CS department from each college. Finally, in Russia, we randomly sampled up to 3 CS departments from each university. If there were less than three CS departments in the university, we sampled all departments. We then randomly sampled 1 administrative group or class of fourth year students within each department. All students within the sampled classes in China and Russia, and all students within the sampled departments in India, were selected for participation in the study.

Fourth, in each randomly sampled class (or randomly sampled department in the case of India), half of the students were randomly assigned to take a CS exam. Altogether, 678 students from China, 364 students from India, and 551 students from Russia took the CS exam. Participation rates (the percentage of enrolled students that took the CS exam) were reasonably high (62% in China, 80% in India, 84% in Russia). Nonetheless, because we had prior test score data from China and Russia for the vast majority of the remaining students (413 from China and 106 from Russia), we used multiple imputation to test the sensitivity of the China and Russia results to missing data. In particular, we used standardized junior year math and physics exam scores, gender, age, parental education levels, and X to impute the missing CS scores for the seniors that did not participate in China and Russia. We find that the results are substantively similar even when accounting for these missing scores. For example, country mean scores only decline slightly from -0.602 to -0.624 SDs for China and from -0.616 to -0.649 SDs for Russia.

In regards to the representativeness of the China sample, we first note that the population of our 6 out of 31 sample provinces comprises 32% of China's total population *(5)*. The proportion of entrants into CS majors in 2010 from those 6 provinces is 37% (authors calculation using administrative data from 30 provinces). Therefore our sample, although drawn from only 6 provinces, represents more than a third of CS students in China.

We also note that we chose our 6 sample provinces so that they would be nationally representative in at least two ways: (a) by region: the sample includes two Northern, two Central, and two Southern provinces; (b) by income level: the sample includes two provinces in each tercile of GDP per capita.

Finally, we investigate the national representativeness of the China sample using college entrance exam (CEE) data on the entire population of undergraduate CS program entrants from 30 provinces in 2010 (the latest date for which such administrative data exist). By way of

[3] The number of CS graduates in each country is approximately 185,000 from China (in 2014-2015), 215,000 from India (in 2015-2016), 17,000 from Russia (in 2014-2015), and 64,405 from the United States (in 2015-2016) *(3, 4)*. The China and Russia numbers were calculated by authors using administrative data with complete national coverage.

background, the administration and content of the CEE in China differ by province: all the students of a particular province take their province-specific CEE and, based on their scores and submitted choices, are matched to institutions (one college and one major) across the country. Furthermore, the vast majority of prospective CS students take the "science-track" version of the CEE exam in each province.

Because these CEE data are province-specific, we compare the average CEE scores of CS entrants from a given province who attend a university in one of our six sample provinces against the average CEE scores of CS entrants from that same province who attend a university in one of the other 24 non-sample provinces. We calculate this average difference for each province and then take the mean of the average differences across provinces. Although this method is decidedly not perfect (because, given the different CEE scales in each province, the meaning of a standard deviation gap in each province is different), in the absence of nationally standardized data, it is the only (albeit crude) way to test the national representativeness of our China sample. Using these data, we find that the CEE scores of students entering CS programs in our six sample provinces are, on average (and when weighted by the number of students entering CS programs in each province) 0.26 SDs higher than students enter CS programs in other provinces. Since our sample represents CS students of slightly higher CEE ability than the national average, the estimates of CS skills for our China sample may be slightly higher than that of the national average.

We investigate the national representativeness of the India sample using two additional datasets of CS students that took the same math and physics achievement tests. The first dataset includes data from 517 end-of-second year CS students in our sample 15 colleges. The second dataset includes data from a national random sample of 2,458 start-of-third year CS students in 50 colleges in India. The national random sample was obtained using a stratified, multi-stage sampling procedure. Specifically, within the first strata, simple random sampling was used to select 8 elite colleges; within the second strata, probability proportional to size (PPS) sampling was used to select 42 non-elite colleges. Only one (non-elite) college chose not to participate in the national study, and this college was replaced using PPS sampling. Probability sampling weights for the second dataset were created in consultation with expert statisticians to ensure the national representativeness of estimates from this dataset.

After standardizing the math and physics scores (using the SDs of subject-specific scores from either the first or second dataset, as they are very similar), we find that the math and physics scores of end-of-second year CS students in our sample 15 colleges are approximately 0.25 SDs higher in math and 0.17 SDs higher in physics than the national sample of start-of-third year students. While the difference could be due to CS students in our sample 15 colleges being of slightly higher ability than CS students nationwide, the discrepancy may also be due in part to students in the national sample having already forgotten some of their math and physics knowledge during the summer *(6)*. Taken together, since the math and science achievement levels of CS students in our 15 college sample are quite close to those of a national random sample of CS students, our sample CS senior are roughly representative of the national population of CS seniors.

Sampling and Implementation in the United States

Our sample is representative of the US population of higher education institutions in terms of student entrance exam scores and degrees awarded in CS majors. Figure S1 shows the distribution of average ACT/SAT equivalent scores of admitted students in 2015-2016 for colleges in the sample and the population. The red bars representing the sample closely approximates the blue bars representing the population. Both are normally distributed with means around 1050. The sample has a slightly narrower range than the population, which is to be expected since fewer schools are included.

The sample and the population of US higher education institutions are also similar in the number of degrees awarded in CS majors and the percentage of total CS bachelor's degrees awarded. Figure S2 shows the number of degrees awarded in 2015-2016. The sample and the population distributions are very similar, except the population has a longer right tail. A closer look at only institutions that awarded 200 or fewer computer-related degrees shows that the sample does indeed resemble the population. As shown in Table S2, the sample and population distribution of CS bachelor's degrees, as a percentage of all bachelor's degrees awarded, are also similar. The main difference is that the population distribution has a larger range of percentages, which is not surprising.

Sample Weights and Basic Descriptive Statistics

We accounted for the sampling design in all of our analyses. For all students in China and Russia, we created and used sample weights that reflected the inverse probability of being sampled at the university, department, (in the case of China and Russia) class, and student levels.[4] We also used the svy command suite in Stata version 13 to adjust our standard estimates for the clustering of students within universities.

Exam

The Major Field Test (MFT) in Computer Science was developed by ETS and was first administered in 1989. The test assesses mastery of concepts, principles, and knowledge by graduating Computer Science students. To ensure fairness and content relevance, the test is revised approximately every four or five years. Experienced faculty members representing all the relevant areas of the discipline determine test specifications, questions and types of scores reported. ETS assessment experts subject each question to rigorous tests of reliability and fairness. Every effort is made to include questions that assess the most common and important topics and skills. In addition to factual knowledge, the test evaluates students' abilities to analyze and solve problems, understand relationships, and interpret material. Questions that require interpretation of graphs, diagrams, and charts are included. The test consists of 66 multiple-choice questions, some of which are grouped in sets and based on materials such as diagrams, graphs and program fragments. It is designed to take two hours and may be split into two sessions.

Following a similar procedure that is employed by Programme of International Student Assessment (PISA) to link scores of students who took the same assessment in different languages and countries *(7)*, scores on the CS assessments for students from China, India, Russia, and the United States were linked using a statistical procedure called concurrent calibration (and using a 2-parameter item response (2-PL IRT) model. The procedure assumes that all items function the same across countries and uses a threshold (RMSEA=0.15) to evaluate whether this assumption holds for each individual item. The items that meet this criterion are referred to as anchor items, as their item parameters are specified to be the same across countries.

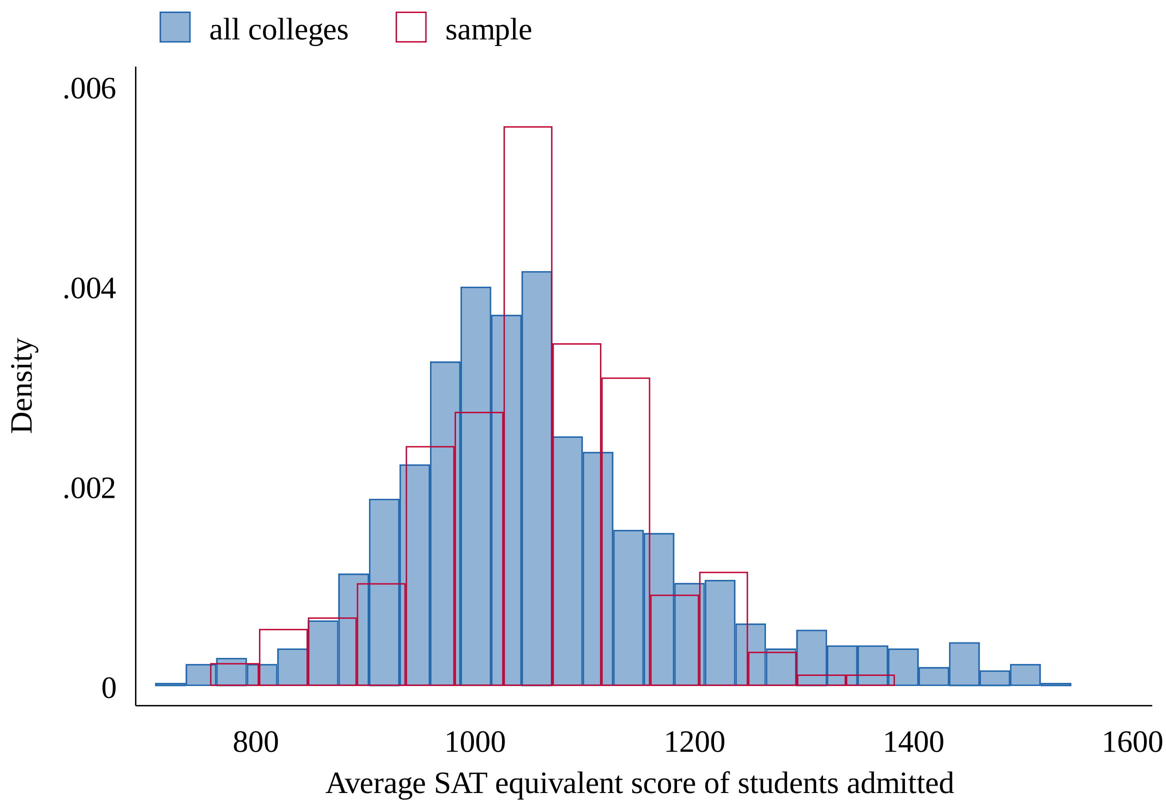
[4] The creation of the sample weights was done in consultation with trained statisticians. After the sample weights were created, we validated that the sample represented essential characteristics of the population of CS students (in particular, total size and percentage of elite students).

The items that do not meet this criterion are considered country-specific items and their parameters are specified as country-specific (varying among countries). As a result of this iterative process, all anchor items and country-specific items were identified. The ability of students estimated based on the anchor items and country-specific items was made comparable through concurrent calibration. These ability estimates were then put on a common scale (integers between 40 and 100 points) following a widely used true score to scale score linking procedure *(8)*. The software program MLTDM was used to conduct the concurrent calibration *(9)*.

Exam Translation and Translation Review

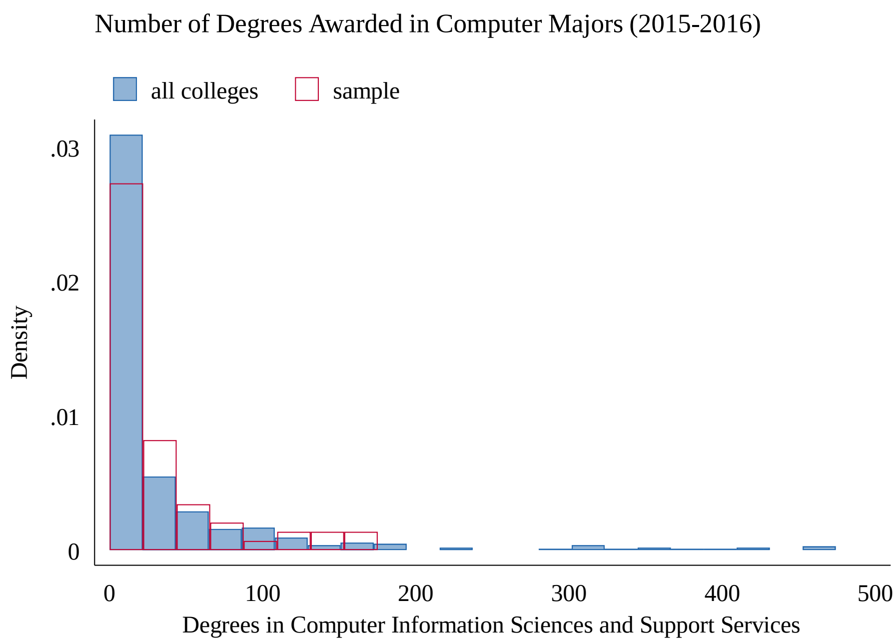
The translation and adaption of the computer science assessment into Chinese and Russian was conducted by Capstan, a linguistic quality control company based in Belgium that has undertaken translation work for large-scale programs such as the Programme for International Student Assessment (PISA) and Programme for International Assessment of Adult Competencies (PIAAC). We adopted the most rigorous, three-step translation and adaptation model that Capstan offers with double translation, reconciliation, and verification. Our internal team of experts also worked with Capstan to ensure that translations were of the highest quality. The computer science assessment was given in English (the language of instruction throughout the four years of college) to students in India and the United States.

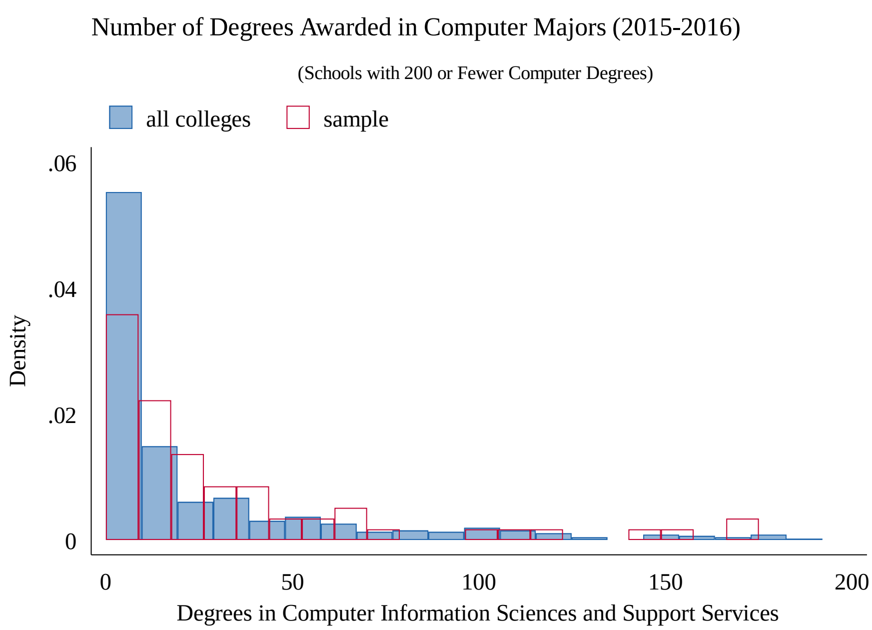
Figure S1. Average SAT Equivalent Scores of Admitted Students (2015-2016): Sample versus Population of Colleges



1) Data source: College Scorecard 2015-2016.
2) 195 colleges in the sample and 1279 colleges nationwide reported average SAT equivalent scores for admitted students.
3) The unit of observation is college/university.
4) Scores are not weighted by enrollment.

Figure S2. Number of Computer Science Degrees Awarded 2015-2016





1) Data source: College Scorecard 2015-2016.
2) The unit of observation is college/university.

Table S1. Detailed Content of the Major Field Test in Computer Science

I. Discrete Structures (15–21%)	A. Functions, relations and sets B. Basic logic C. Proof techniques D. Basics of counting and number theory E. Graphs and trees F. Discrete probability
II. Programming (21–27%)	A. Programming Fundamentals: fundamental programming constructs, basic algorithms and problem solving, fundamental data structures, recursion, event-driven programming, object-oriented programming B. Programming Languages: features, paradigms, implementation techniques
III. Algorithms and Complexity (16–22%)	Advanced data structures and algorithms (including graph algorithms), algorithmic strategies, distributed algorithms, basic computability and complexity, automata theory and formal languages
IV. Systems (16–24%)	A. Architecture: digital logic and digital systems, machine level representation of data, assembly level machine organization, interfacing and communication B. Operating Systems: operating system principles, concurrency, scheduling and dispatch, and memory management C. Networking
V. Software Engineering (3–9%)	Software requirements, specifications, design, validation and management
VI. Information Management (3–8%)	Database systems and data modeling
VII. Other (3–8%)	Human-computer interaction, graphics, intelligent systems, social and professional issues, web computing and security

Table S2. Comparison of CS bachelor's degrees between the United States sample and the United States Population in 2015-2016

	All Doctoral, Masters, and Baccalaureate Institutions	Institutions in Sample
Number of Institutions	1,717	229
Mean (% of CS BA out of total BA degrees)	2.87	2.57
Standard Deviation (% CS out of total BA degrees)	5.87	2.85
Percentile Values in Population and Sample Distribution of Institutions		
1%	0	0
5%	0	0.50
10%	0	0.74
25%	0	1.21
50%	1.67	2.02
75%	3.19	2.92
90%	5.91	5.15
95%	9.02	6.34
99%	25.0	14.01

1) Data source: College Scorecard 2015-2016

Table S3. Comparison of Content Covered by ETS's CS Major Field Test and ACM/IEEE-CS Curriculum Guidelines in Various Years

	CS Major Field Test	CS 2013 Guidelines	CS 2008 Guidelines	CS 2003 Guidelines
I. Discrete Structures	15-21%	13.31%	14.8%	15.4%
II. Programming	21-27%	23.1%	23.4%	21.1%
III. Algorithms and Complexity	16-22%	14.0%	10.7%	11.1%
IV. Systems	16-24%	22.1%	23.8%	24.6%
V. Software Engineering	3-9%	9.1%	10.7%	11.1%
VI. Information Management	3-8%	3.2%	3.8%	3.6%
VII. Other	3-8%	14.9%	12.8%	13.2%

1) Source: Computer Science Curricula 2013

Table S4. Comparison of Content Covered by ETS's CS Major Field Test and National Guidelines for CS Programs in China, India, and Russia

	CS Major Field Test	China Guidelines	India Guidelines	Russia Guidelines
I. Discrete Structures	15-21%	17%	8%	20%/21%
II. Programming	21-27%	17%	17%	27%/19%
III. Algorithms and Complexity	16-22%	9%	20%	16%/16%
IV. Systems	16-24%	28%	23%	22%/20%
V. Software Engineering	3-9%	7%	8%	2%/5%
VI. Information Management	3-8%	13%	12%	9%/14%
VII. Other	3-8%	9%	12%	4%/5%

1) Sources: Teaching Guidance Committee in Discipline of Computer Science, China Ministry of Education (2006); All-India Council for Technical Education (2018); Federal Curricular Association in Computer and Information Science, Russia (2009); Federal Curricular Association in Informatics and Computer Engineering, Russia (2009) *(10-13)*.
2) National Guidelines for China are for the Computer Science and Technology major (the most popular and only CS major for which national guidelines are available).
3) National Guidelines for India are for the Computer Science and Engineering major (the most popular and only major for which national guidelines are available).
4) In Russia all state-accredited CS programs must comply with the Federal State Educational Standard which includes guidelines developed by Federal Curricular Associations (FCA) for particular majors. Guidelines developed by FCA in Computer and Information Science (before the "/" in the "Russia Guidelines" column) and FCA in Informatics and Computer Engineering (after the "/" in the "Russia Guidelines" column) cover 87% of academic programs within CS. Detailed guidelines by which to allocate curricular coverage are unavailable for the remaining 13% of students in the Information Security major.
5) The percentages of CS curricular attributed to the seven knowledge areas (I-VII) are calculated using the ratio of hours of relevant professional (CS) coursework to the total number of hours of professional (CS) coursework.

References and Notes:

1. National Bureau of Statistics (2015) *China statistical yearbook*. Available at: http://www.stats.gov.cn/tjsj/ndsj/2015/indexeh.htm [Accessed Juy 24, 2018]
2. Carnoy M, Loyalka P, Dobryakova M, Dossani R, Froumin I, Kuhns K, Tilak J B, Wang R (2013) *University Expansion in a Changing Global Economy: Triumph of the BRICs?* (Stanford University Press, Palo Alto, CA).
3. National Science Board (2018) *Science and engineering indicators 2018* (National Science Foundation, Alexandria, VA).
4. Government of India (2016) *All India survey on higher education, 2015-2016* (Ministry of Human Resource Development, Department of Higher Education, New Delhi).
5. National Bureau of Statistics (2018) *China Statistical Abstract 2018* (China Statistics Press, Beijing).
6. Cooper H, Nye B, Charlton K, Lindsay J, Greathouse S (1996) The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research* 66(3):227–268. Available at: http://journals.sagepub.com/doi/10.3102/00346543066003227 [Accessed December 27, 2018]
7. Organization for Economic Cooperation and Development (2014) The PISA 2012 technical report. Available at: https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf [Accessed August 18, 2018]
8. Kolen M J, Brennan R L (2004) *Test Equating, Scaling, and Linking: Methods and Practice* (Springer, New York).
9. von Davier M (2006) *Multidimensional latent trait modelling (MDLTM) [Software program]* (NJ: Educational Testing Service, Princeton).
10. Teaching Guidance Committee in Discipline of Computer Science, Ministry of Education (2006) Professional Norms for Computer Science and Technology Undergraduate Program (Computer Science Area). Research Report and Professional Norms for the Development Strategy of Computer Science and Technology Discipline in Higher Education Institutions. *Higher Education Press*: 13-92.
11. All India Council for Technical Education (2018) *Model Curriculum for Undergraduate Degree Courses in Engineering and Technology, Volume I* (Ministry of Human Resource Development, Government of India) Available at: https://www.aicte-india.org/education/model-syllabus [Accessed December 30, 2018]
12. Federal Curricular Association in Computer and Information Science, Russia (2009). *Exemplary Curriculum for the Bachelor Degree in Software and Administration of Information Systems* (Ministry of Education and Science of the Russian Federation): 6-10.
13. Federal Curricular Association in Informatics and Computer Engineering, Russia (2009). *Exemplary Curriculum for the Bachelor Degree in Information Systems and Technologies* (Ministry of Education and Science of the Russian Federation): 5-11.