Developing instruments to assess and compare the quality of engineering education: the case of China and Russia

Elena Kardanova, Prashant Loyalka, Igor Chirikov, Lydia Liu, Guirong Li, Huan Wang, Ekaterina Enchikova, Henry Shi & Natalie Johnson

To cite this article: Elena Kardanova, Prashant Loyalka, Igor Chirikov, Lydia Liu, Guirong Li, Huan Wang, Ekaterina Enchikova, Henry Shi & Natalie Johnson (2016) Developing instruments to assess and compare the quality of engineering education: the case of China and Russia, Assessment & Evaluation in Higher Education, 41:5, 770-786, DOI: 10.1080/02602938.2016.1162278

To link to this article: http://dx.doi.org/10.1080/02602938.2016.1162278
Developing instruments to assess and compare the quality of engineering education: the case of China and Russia

Elena Kardanova*, Prashant Loyalka, Igor Chirikov, Lydia Liu, Guirong Li, Huan Wang, Ekaterina Enchikova, Henry Shi and Natalie Johnson

*Institute of Education, National Research University Higher School of Economics, Moscow, Russian Federation; bGraduate School of Education and Freeman Spogli Institute for International Studies, Stanford University, Stanford, CA, USA; cEducational Testing Service, Princeton, NJ, USA; dSchool of Education, Henan University, Kaifeng, China; eCenter for Experimental Economics of Education, Shaanxi Normal University, Xi’an, China

Relatively little is known about differences in the quality of engineering education within and across countries because of the lack of valid instruments that allow for the assessment and comparison of engineering students’ skill gains. The purpose of our study is to develop and validate instruments that can be used to compare student skill gains in mathematics and physics courses in undergraduate engineering programmes across countries. The approach includes procedures to establish construct validity and other necessary psychometric properties. Drawing on data collected from over 24 engineering experts and 3600 engineering students across Russia and China, we establish that it is possible to develop valid, equitable and cross-nationally comparable instruments that can assess and compare skill gains.

Keywords: engineering education; assessment; quality; China; Russia

Introduction

Raising the quality of engineering education is a high priority for researchers and policy-makers in developed and developing countries (NAS 2007; Lucena et al. 2008; Carnoy et al. 2013). Researchers and policy-makers believe that skilled engineering graduates can contribute towards economic productivity and innovation (ABET 1997; NAS 2007). They likewise fear that failing to develop skilled engineering graduates may undermine the capacity of countries to compete in the global knowledge economy (NAS 2007). This is especially true for the BRIC countries (Brazil, Russia, India, and China), which have greatly expanded engineering education in the last several decades, and now produce a large proportion of the world’s engineers (Carnoy et al. 2013).

Although engineering education has expanded rapidly worldwide over the last few decades (Freeman 2010; National Science Board 2010), and especially in the BRIC countries (Carnoy et al. 2013), relatively little is known about its quality. A few studies have used indirect methods to examine and compare the quality of engineering education within and across countries. These studies have, for example, gathered and analysed subjective employer feedback on the skill levels of engineering graduates from various types of programmes (Bondarenko, Krasilnikova, and

*Corresponding author. Email: ekardanova@hse.ru
Kharlamov 2006; Mooney and Neelakantan 2006; Gereffi et al. 2008; Klintsov, Shvakman, and Solzhenitsyn 2009; Borsch and Abramova 2010; Levin Institute 2010; Blom and Saeki 2011). They have also compared the inputs, processes and outputs of engineering programmes across countries using an educational production function approach (Loyalka et al. 2014). A couple of studies have further explored the feasibility of directly assessing the level at which students have mastered specific engineering competencies (OECD 2013; Musekamp and Pearce 2015). Few, if any, studies, however, have directly assessed the skill gains (how much students learn over time) of engineering students and compared these skill gains within and across countries.

While there are many ways to evaluate ‘quality’ in engineer training, we focus on evaluating gains in students’ cognitive skills. Understanding cognitive skill gains among engineering students within and across countries is important for several reasons. As the example of such international studies of school education as the Programme for International Student Assessment and the Trends in International Mathematics and Science Study have shown, comparative assessments of student competencies across countries have the potential to draw international consideration and prompt research-driven changes in policy and practices around the world (Zlatkin-Troitschanskaia, Shavelson, and Kuhn 2015). To date, very little comparable research has been conducted in higher education, partly due to the lack of valid, equitable and cross-nationally comparable assessment instruments. The lack of large-scale assessments is unfortunate, since assessing gains in student skills across a wide and representative range of engineering programmes is an important first step in helping policy-makers improve the state of engineering education.

A major means by which policy-makers in both developed and developing countries have striven to improve the quality of engineering education is through accreditation. The Accreditation Board for Engineering and Technology (ABET) in the United States and the Engineering Council in the UK, for example, have had a substantial impact on the reform of engineering education in developed countries (Volkwein et al. 2004; Augusti 2007). The Association for Engineering Education in Russia has signed the Washington Accords (on engineering education standards), has joined the European Network for Accreditation of Engineering Education (EUR-ACE) and has been establishing a national system for accrediting engineering programmes. The China Association for Science and Technology has also signed the Washington Accords and has started accrediting (mostly elite) engineering programmes (Loyalka et al. 2014). It is notable that these accreditation agencies have shifted their focus from inputs-based to outcomes-based assessment in recent years, perhaps reflecting a growing movement in the research and policy-making community to hold schools accountable for students’ learning and career outcomes (Volkwein et al. 2004; Augusti 2007).

Although accreditation programmes have become more widespread and important, they have generally failed to assess student learning outcomes directly or on a large-scale. This is surprising given the increasing attention being given to the direct assessment of learning outcomes in higher education (Coates 2014).

Indeed, very few studies even assess the skill levels of engineering students. One exception is the KOM-ING study, which focuses on assessing the competency of engineering mechanics through open-ended (construct response) items (Musekamp et al. 2013). Because the KOM-ING study was in a pilot stage, it was relatively small scale, assessing the skill levels of only 278 students from seven universities in
Germany. The only other effort to directly assess engineering student skill levels, and this time on a large, international scale, is the OECD’s Assessment of Higher Education Learning Outcomes (AHELO) project (OECD 2010, 2012, 2013). One of the main goals of the AHELO project has been to measure the major-specific skill levels of civil engineering students who are in the last year of their undergraduate programmes. The project has focused on measuring skill levels at the end of university, however, and does not measure skill gains. By measuring skill levels at only one point in time, it is difficult to separate out whether students learned skills during their time at university or before university. Developing instruments that can assess skill gains is thus critical for shaping policies and practices that can effectively target improvements in student learning.

Given the lack of research in this area, the purpose of this study is to provide empirical support for the development and validation of instruments that assess and compare engineering students’ skill gains within and across countries. Specifically, we seek to design and validate instruments that have the potential to assess and compare gains in the mathematics and physics skills of engineering students in different types of institutions and countries.

For the purposes of this study, we focus on students enrolled in majors in one of two categories within the field of engineering: electrical engineering and computer science. We limit our target sample since engineering is a broad and highly varied field that attracts students of differing initial skill levels, and prepares them for different types of work. These particular major-categories were selected because they produce highly skilled and sought-after graduates who often compete for jobs on an international stage. Understanding relative student learning across countries for this subset of the engineering field is, therefore, of particular importance.

Solid foundations in mathematics and physics are believed to be essential for developing strong engineering competencies (Mills and Treagust 2003). Indeed, most engineering majors – especially in the subfields of electrical engineering and computer science – require students to take multiple courses in mathematics and physics in the first 2 years of their undergraduate programmes (Carnoy et al. 2013). As such, our goal is to develop assessment instruments that can be used to assess and compare skill gains in mathematics and physics for engineering students in electrical engineering and computer science majors. We seek to do so for two BRIC countries that produce a large number of skilled engineering graduates for the global economy: China and Russia (Loyalka et al. 2014).

We develop and provide evidence in support of the assessment instruments in two stages. In the first stage, we conduct analyses of content and construct validity using cross-national expert evaluations of content areas, subcontent areas and test items for each subject. To conduct these analyses, we collected feedback from 24 experts from a range of elite and non-elite engineering programmes in China and Russia. In the second stage, we conduct a series of psychometric analyses, not only to ensure that the assessment instruments meet basic standards for educational measurement (AERA, APA, and NCME 2014), but also to make sure that they can be equated both between two grades and across two countries. The data for the psychometric analyses were collected from approximately 3600 first- and third-year students from 21 undergraduate engineering programmes in China and Russia. Using these data, we demonstrate that it is possible to develop instruments that are valid, reliable and cross-nationally comparable.
Methodology

In attempting to develop and validate assessment instruments for comparing skill gains among institutions within and across countries, researchers face multiple challenges. These include construct differences, instrument differences, administration differences, sample differences and response procedure differences (Ercikan and Lyons-Thomas 2013).

Our test development methodology included several stages, each of which sought to account for these challenges and thereby contribute to the comparability of results. Stage 1 aimed to develop assessment instruments in mathematics and physics. Stage 2 established the content and construct validity of these instruments. Stage 3 included a pilot survey of students in China and Russia, and a subsequent analysis of the psychometric quality of the instruments (dimensionality, reliability, cross-national comparability), as well as the construction of a common scale between the two grades and two countries.

The development of the assessment instruments

The assessment instruments were developed based on: (1) selecting comparable majors within the categories of electrical engineering and computer science across China and Russia, (2) rigorous expert evaluations of content maps to define content areas for each test (mathematics and physics, for grades 1 and 3) and (3) test item collection and verification by experts at elite and non-elite universities in China and Russia.

Selecting comparable electrical engineering and computer science majors across China and Russia

The project team selected majors in China and Russia that had consistent courses and curricula across universities within each country, as well as substantial overlap in courses and curricula across countries. We also focused on the majors that had courses and curricula that significantly overlapped with electrical engineering and computer science majors in the United States.

While the primary goal of this study was to develop assessment instruments that can be used to assess and compare skill gains between China and Russia, in the future we hope to extend this research to allow for comparative assessments of a more extensive sample of countries. Because the US is a global leader in electrical engineering and computer science related industries, curricular standards from the US serve as a convenient reference point to make sure our assessments are relevant around the world.

Specifically, we selected the electrical engineering majors of electronic and information engineering, electronic and information science and technology, communication engineering, electrical engineering and automation, and automation in China; and electrical energy and electric engineering, radio engineering, information and communication technology and communication systems, and design and technology of electronic instrumentation majors in Russia. These majors shared with each other (and with electrical engineering majors in the US) core course requirements of computer programming, circuits, analogue electronics, signal processing and digital systems. For computer science, we selected computer science and technology, and
software engineering majors in China; and information technology and computation, and informational systems and technologies majors in Russia. These majors share with each other (and computer science majors in the US) core course requirements of computer programming, computer architecture and organisation, data structure and algorithms, and operating systems.

Using expert feedback to select content areas for the tests

After selecting the core majors in electrical engineering and computer science for each country, we produced content maps in mathematics and physics. The content maps were based on the national curriculum standards for the sampled majors in Russia and China. These content maps covered areas that students should know at the start of their engineering studies, or at the end of high school for students planning to go into science and engineering majors (for the grade 1 test), and in the first two years of engineering studies (for the grade 3 test). We further weighted the importance of different content areas by examining the proportion of text devoted to them in popular textbooks and syllabi. Table 1 shows the number of content and subcontent areas included in the content maps.

We then asked 12 experts in each country to evaluate the content maps for electrical engineering and computer science, for grade 1 and grade 3 separately. Specifically, the experts were presented with the list of content and subcontent areas for each major-grade, and asked to rate the importance of each area for the academic progress of students. The experts were also invited to supplement our content map with any additional content areas they deemed important. The experts were professors of mathematics and physics who taught students in electrical engineering and computer science majors. In each country, there were six experts from elite universities and six experts from non-elite universities. In each country, five experts had experience teaching to undergraduate computer science students, five to undergraduate electrical engineering students and two taught both.

Collection and verification of test items

Our item collection procedure consisted of two steps. First, we collected tests items that reflected the top five content areas selected by experts for each subject and grade level. Second, we re-enlisted the help of our team of experts to evaluate the items to make sure they were valid, relevant, clear and of suitable difficulty.

Table 1. The total number of content and sub-content areas for the mathematics and physics tests (grades 1 and 3).

<table>
<thead>
<tr>
<th>Test</th>
<th>Number of content areas</th>
<th>Number of sub-content areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics, grade 1</td>
<td>8</td>
<td>36</td>
</tr>
<tr>
<td>Mathematics, grade 3</td>
<td>10</td>
<td>116</td>
</tr>
<tr>
<td>Physics, grade 1</td>
<td>10</td>
<td>49</td>
</tr>
<tr>
<td>Physics, grade 3</td>
<td>10</td>
<td>104</td>
</tr>
</tbody>
</table>
Sources of test items included items from each country’s past university entrance examinations (China’s Gaokao and Russia’s Unified State Examination), other standardised examinations and widely used exercise books in both countries.

To create comparable versions of the tests in both languages, we selected items which:

- fitted in the content areas of the content maps;
- used short and simple sentence structures and grammatical forms;
- used a multiple-choice format (as it is the most familiar format for our target populations).

A small number of items that reflected content areas not included in the content maps, but that had high ratings from the experts of one country, were also included. Altogether, we collected 85–90 items for each test. Approximately one-third of the items came from Russia, one-third from China and one-third from the United States.

After collecting the items, we translated them into Russian and Chinese and evaluated the quality of the translations. All items other than those from the United States were translated into English. Then the English versions were translated into the languages of our two target countries: Russian and Chinese. Next, all items were back-translated into English. The same team of experts (English speaking professors of mathematics and physics from each target country) were asked to compare the original and back-translated English versions of each item, and pick out those that had discrepancies between the two versions. They were further asked to offer suggestions for correcting the discrepancies if any were found. Most discrepancies in original translations were due to inaccurate translations of major-specific terminologies. Using suggestions from experts, we were able to correct for the few discrepancies.

The experts were asked to rate the items according to four criteria: (1) comprehensibility of wording, (2) appropriateness in measuring the content area of interest, (3) difficulty and (4) expected time required to answer. Out of a pool of 179 physics items and 174 mathematics items we initially collected, we selected 80 physics items and 80 mathematics items to be used in the clinical pilot, respectively. In selecting these items, we also took care to keep balance across weighted content areas.

**Methods to ensure the validity of the instruments**

One of the major challenges we faced was to ensure the cross-national equivalence of measurements. Without cross-national equivalence, it is unclear whether observed differences across groups are due to true differences in mathematics and physics ability, or to differences in the way items are perceived (Millsap 2010). Validity depends on the degree to which the test, given in different languages, actually measures the intended constructs and provides comparable measurements (Hambleton 2005; Ercikan and Lyons-Thomas 2013). To ensure the validity of measurements and cross-national equivalence, we followed the standards for Educational and Psychological Testing (AERA, APA, and NCME 2014), which provide guidelines for the process of test development for international comparative study. Our conceptualisation of validity is standard to the field, as reflected in Messick (1995) and the standards for Educational and Psychological Testing (AERA, APA, and NCME 2014). According to these sources, validity should be considered as a unitary
concept. In other words, most types of validity are subsumed under construct validity, and only content validity can be considered as a separate type.

We employed two major methods to ensure the content and construct validity of the assessment instruments. First, we analysed the evaluations given by our experts of each test item. We did this to reveal possible biases in specific items. To accomplish this we analysed the consistency of expert ratings using Cronbach’s Alpha, and looked at the correlations between the ratings of each expert against the ratings of the other experts. Additionally, to optimise and simplify the analysis of our expert evaluations, we used multi-faceted item response theory (IRT) analysis (Myford and Wolfe 2003). This analysis allowed us to determine whether the experts differed in how severely they evaluated each item and whether they demonstrated bias that might have affected the final ratings. Second, we analysed the difficulty of each item by conducting a clinical pilot with a small number of students. The clinical pilot tests were administered to 40 grade 1 students and 40 grade 3 students in each country. Detailed explanations of the consistency evaluation and the clinical pilot are presented in the results section.

Ensuring psychometric quality and a common scale between grades and countries

Based on the results of the expert evaluations and the clinical pilot, the instruments were modified in preparation for a large-scale pilot. For the large-scale pilot, the mathematics and physics tests included 45 items each. All items were in multiple-choice format with one correct answer from 4 to 5 options. The items were scored dichotomously: a student received 1 point for a correct response and 0 for an incorrect or missing response (with a maximum total of 45 points). The grade 1 and grade 3 tests for each subject had approximately 20 common items, thus making it possible to equate the test scores from different grades and place the results on a common scale.

Sample and data collection

We collected and analysed data from a large-scale pilot to evaluate whether the tests are reliable, cross-nationally valid, equitable and bias free. The target population for the pilot study was defined as students in the first and third years of electrical engineering and computer science undergraduate engineering programmes in Russia and China. In designing the sampling procedure for the pilot study, two factors were taken into account: university status (both elite and non-elite universities were to be selected) and location (both large and small cities across the country should be represented). In China, we sampled five elite and six non-elite universities, whereas in Russia we sampled five elite and five non-elite universities. The universities were spread fairly evenly between large and small cities in the two countries. Based on these criteria, we selected 11 universities in China and 10 universities in Russia for the pilot study. In each university, we sampled classes of grade 1 and grade 3 students until we reached at least 60 students from each major-grade (or until we sampled all the available students in that major-grade, even if it was less than 60). In each sampled class, two-thirds of the students were randomly selected to take the mathematics and physics tests, with mathematics and physics given in random order. Our final sample consisted of 1797 students in China and 1802 students in Russia.
The pilot was conducted at the end of October 2014. The tests were given in paper and pencil format. The testing was conducted during two 55-min sessions (one for mathematics and one for physics, in random order).

Analytic approach

We used IRT (Embretson and Reise 2000) modelling to conduct item analyses as well as tests of dimensionality and reliability. We also paid particular attention to differential item functioning (DIF) to provide evidence concerning the cross-national comparability of the test results and to ascertain the possibility of creating a common scale between the two grades and across the two countries.

The one-parameter dichotomous Rasch model (Wright and Stone 1979) was used for the IRT analysis. Under this model, each test item is characterised by one parameter (difficulty), and each student is also characterised by one parameter (ability level). The Rasch analysis places students and items on the same measurement scale in logit units.

The reasons for choosing the Rasch model are psychometric and practical. First, the Rasch model has optimal metric properties. Second, it is useful for data analysis: i.e. determining the quality of test items, constructing scales and carrying out test equating (Bond and Fox 2001). Winsteps software (Linacre 2011) was used for this process. Due to space constraints, the technical details of the data analysis methodology are presented in a supplementary document (https://reap.fsi.stanford.edu/publication/appendices-%E2%80%9Cdeveloping-instruments-assess-and-compare-quality-engineering-education-case).

The data analysis was performed for each subject, mathematics and physics, separately and in two stages. During the first stage, we treated the data from each grade separately. The purpose of this stage was to discover whether it would be possible to construct a common scale across countries within each grade. During the second stage, we equated the data sets for different grades using common items included in both grades for the link. The purpose of this stage was to determine whether it would be possible to place all parameters on a common scale between the two grades and across the two countries.

Results

Ensuring the validity of the assessment instruments for cross-national comparisons

Results from the expert evaluations of the content maps for the grade 1 and grade 3 tests showed that there was a large amount of overlapping content and subcontent areas in high school and university in both countries. Based on expert ratings of the importance of each content and subcontent area for the academic progress of students in electrical engineering and computer science majors, we selected the content and subcontent areas with the highest average ratings in each subject. Table 2 shows the list of content areas selected for grade 1 and grade 3 mathematics and physics tests.

Starting from a pool of 85–90 items, we selected items for the grade 1 and grade 3 tests taking into account expert ratings of item difficulty, total testing time and the need to keep balance between items from different countries. Additionally, a small
number of items that did not reflect the top five content areas, but had high ratings from experts of one country, were also included. In this stage, the number of items on each test was reduced to 55.

Consistency between experts was high. The Cronbach’s Alpha coefficient was over 0.8, and correlations between the ratings were over 0.6 and significant at the 0.01 level. There were, however, statistically significant differences in the level of ‘severity’ of the experts’ evaluations. In other words, the experts evaluated item difficulties differently, and our analysis showed that the ‘severity’ of an expert was not related to nationality. Despite this difference in ‘severity’, no experts demonstrated any other effects that might have affected the final ratings and brought bias into the evaluation procedure. We conducted extra analyses to minimise the risk of bias in the expert rating process. Using multifaceted Rasch measurement, we checked for such effects as central tendency, limitation of range or random assigning of ratings. None of our experts demonstrated these effects, so we are confident that the results of the rating process are bias free.

The results of the clinical pilot of the tests with 40 grade 1 students and 40 grade 3 students in each country showed the items were understandable and had an acceptable difficulty level for students in both countries. We did, however, find that the total time allotted was not enough for students to solve all 55 items. Based on the results of the clinical pilot, the number of the items for the subsequent pilot study was reduced to 45 per test.

Psychometrics and constructing common scale between countries

The results of our analysis show that the tests meet basic standards of psychometric quality.
The grade 1 mathematics test

Analysis of model fit showed that most items had good psychometric characteristics and fitted the model. Starting with the pool of 45 items, we deleted eight items because of poor psychometric quality (low discrimination and/or mis-fitting the model). For the rest of the analysis, we consider the reduced set of 37 items for the grade 1 mathematics test.

Based on the results of country-related DIF analysis, 13 items (out of 37) demonstrated DIF across countries: 7 items favoured students in China, while 6 items favoured students in Russia. Figure 1 shows the item difficulties separately for students from different countries. Most items are DIF-free as they demonstrate stable estimates of difficulty (the difference in item difficulty between countries is not significant). While the reasons for the observed DIF are not of immediate concern for this paper, to create comparable assessment instruments, we had to decide how to handle the items that exhibited DIF. One option would have been to eliminate these items, but this could have negatively impacted the precision of our measurement. For our analysis, we decided to keep these 13 items and considered them as unique items for each country. We used the other 24 DIF-free items for establishing the link between the two countries. The quality of the link was investigated by calculating the item-within-link statistic. Its value of 0.99 indicates a very good fit within the link (Wright and Bell 1984). Therefore, for further analysis the total data-set has 50 items (24 items are common for both countries, and 26 are country-specific items: 13 for China and 13 for Russia).

Our investigation of the dimensionality of the test supports the claim that the test can be recognised as essentially unidimensional. The eigenvalues of the residual correlation matrix for the five primary components of our PCA analysis ranged from

![Figure 1. The mathematics grade 1 item difficulties for each country.](image-url)
1.67 to 1.3. In addition, the variance accounted for in the distribution was roughly evenly split across components from 3.6 to 2.6%.

We find the test is also reliable. The person reliability is 0.85, which means that the proportion of observed student variance considered true is 85%. In addition, our analysis produced a person separation index of 2.39, indicating at least three statistically distinct groups of students along the continuum.

To show the relative distribution of items and students along a common metric, we constructed the variable map presented in Figure 2. The horizontal axis of Figure 2 is in logit units. The origin of the logit scale is set to average item difficulty. On the map, students are represented in the upper part and test items in the lower part. More difficult items and higher performing students are located on the right side of the map (positive logits), while easier items and lower performing students are located on the left side of the map (negative logits).

Figure 2. The mathematics grade 1 test variable map.
Note: All items (from various sources) are included on a common metric.
students are located on the left side of the map (negative logits). The distribution of students is wide and, for measurement purposes, reflects good differentiation between higher and lower scoring students.

The grade 3 mathematics test

The results for the grade 3 mathematics test are substantively similar. Six items were deleted because of poor psychometric quality. For further analysis, we considered the remaining set of 39 items. Of these, 11 demonstrated country-related DIF: 5 items favouring students in China and 6 favouring students in Russia. Figure 3 shows the item difficulties for students from each country. Following the same approach as for the grade 1 test, we used 28 DIF-free items for linking between the two countries. Altogether, the total data-set has 50 items, of these 28 items are common for both countries, and 22 items are country-specific: 11 for China and 11 for Russia.

Further analysis showed that all items had good psychometric characteristics, fit the model and could be considered unidimensional. The person reliability is 0.80 and the person separation index is 2.03, indicating three statistically distinct groups of students along the continuum. Figure 4 presents the variable map for the grade 3 test. We see that the distribution of students is wide and the student sample is well located relative to the items.

The results of the psychometric analysis for grades 1 and 3 physics tests are substantively similar.

To conclude, we find that the grade 1 and grade 3 mathematics and physics tests have good psychometric quality, and are essentially unidimensional, reliable and valid instruments for measuring and comparing the mathematics and physics skills.
of grade 1 and grade 3 engineering students within and across the two countries. Although some test items demonstrated DIF across countries, a sufficient number of DIF-free items allowed us to construct a common scale for both countries.

**Constructing the common scale between different grades**

Up until this point, item analyses were carried out separately for the two grades. The next step was to ascertain whether it would be possible to create a common scale between the two grades. This step required performing data analysis for both grades together. The tests for different grades had 20 common items to make it possible to equate them.
Starting from the 20 items that were common across the grade 1 and 3 tests, 7 items were selected as good candidates for linking between grades. Items that were not considered good for linking were deleted at earlier stages, either because they did not have good psychometric qualities when they were analysed for inclusion in a particular test, or because they exhibited DIF for at least one test. This number of common items is about 18% of the total number of items in each test and is close to the 20% recommended by Angoff (1971). The seven common items were checked for DIF between grades by using the same approach was used for DIF investigation between the countries, and no items showed significant DIF.

The difficulty of selected anchor items was fixed with values for the grade 1 test when calibrating the grade 3 test. As a result, the parameter estimates for the grade 3 test were placed onto the scale of the grade 1 test. To evaluate the quality of the link between the grade 1 and grade 3 tests, we calculated the item-within-link statistic. Its value of 0.95 indicated a reasonable fit within the link.

Constructing the common physics scale between different grades

Eleven common items across the grade 1 and 3 physics tests were selected as good candidates for linking between the grades. As with the mathematics test, we used these common items for linking between the two grades and constructing the single scale for both grades. The gains (differences between first and third year students) in physics and mathematics scores are consistent with expectations (Loyalka et al. 2015). Specific results are omitted for the sake of brevity but are available upon request.

Conclusion

The last two decades have witnessed a rapid growth in the number of engineering graduates worldwide. As such, the quality of engineering education across a wide range of countries and institutions has become a topic of discussion among researchers and policy-makers (e.g. NAS 2007; Gereffi et al. 2008). However, assessing the quality of engineering education – in particular, the skill gains of engineering students – in such a way that allows for intra-national and international comparisons, necessitates the creation of special assessment instruments. This paper took a step in that direction by developing instruments that can assess skill gains in mathematics and physics among engineering students in China and Russia. We demonstrate that it is possible to construct reliable and valid instruments for these purposes.

The creation of such instruments includes numerous methodological and empirical challenges. The instruments must be verified for cross-national equivalence and provide comparable results. In addition to taking careful steps to ensure the content validity, construct validity and psychometric quality of the instruments (dimensionality, reliability, cross-national comparability), it was important to establish the possibility of constructing a common scale between grades and countries. Our procedure used anchor items to link the tests and included DIF analysis to ensure that the items were functioning in the same way across countries. We also used simultaneous and separate calibration procedures for creating a common scale between grades and countries. These choices ensure that these instruments can be used to compare test scores directly and to estimate student progress between grades in different countries.
As one of the first attempts to develop assessment instruments intended to measure and compare skill gains among engineering students within and across countries, we acknowledge that our study is subject to limitations. In particular, our evaluation of construct validity thus far has relied entirely upon the judgment of experts. While our selection of experts was careful, the process of expert validation was rigorous, and our analyses found no evidence of any bias in the expert feedback, more work is needed to substantiate construct validity of the assessments. As this is a promising and wide field for future research, we look forward to pursuing this type of work in the near future.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

Notes on contributors
Elena Kardanova is a researcher and senior lecturer at the Institute of Education, National Research University Higher School of Economics, Moscow, Russian Federation. Her research focuses on psychometrics, Rasch measurement, test development, assessing cross cultural comparability of test results.

Prashant Loyalka is an assistant professor (Research) at the Graduate School of Education and a centre research fellow at REAP/FSI, Stanford University as well as a senior visiting researcher at the Institute of Education National Research University Higher School of Economics, Moscow, Russian Federation. His research areas include the economics of education, international comparative education, higher education and impact evaluation.

Igor Chirikov is a senior research fellow at the Institute of Education National Research University Higher School of Economics, Moscow, Russian Federation. His research interests include student experience, comparative higher education and organisational sociology.

Lydia Liu is Director of Research in Higher Education at the Educational Testing Service. Her research interests include applied measurement in educational and psychological settings.

Guirong Li is a professor in the School of Education at Henan University. Her research interests include education policy, impact evaluation and higher education.

Huan Wang is a researcher at the Center for Experimental Economics in Education, Shaanxi Normal University. Her research interests include impact evaluation, social networks and higher education.

Ekaterina Enchikova is a researcher at the Center for Monitoring the Quality in Education, Institute of Education, National Research University Higher School of Economics, Moscow, Russian Federation. Her research focuses on data analysis, psychometrics, DIF identification.

Henry Shi is a doctoral student in the Economics of Education program at Stanford University. He is also a researcher at the Rural Education Action Program at Stanford’s Freeman Spogli Institute. His research interests include higher education, vocational education and international development.

Natalie Johnson is a project manager for the Rural Education Action Program at Stanford University.
References


