NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Elena Y. Kardanova, Ekaterina S. Enchikova
Shi H., Johnson N.,Lydia O. Liu, Liyang Mao,
Prashant Loyalka

CONSTRUCTING TESTS THAT CAN MEASURE AND COMPARE THE MATHS AND PHYSICS SKILLS OF ENGINEERING STUDENTS IN RUSSIA AND CHINA

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: EDUCATION
WP BRP 28/EDU/2015

This Working Paper is an output of a research project implemented within NRU HSE's Annual Thematic Plan for Basic and Applied Research. Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE

Elena Y. Kardanova[1], *Ekaterina S. Enchikova* [2], *Shi H.*[3], *Johnson N.*[4], *Lydia O. Liu*[5], *Liyang Mao*[6], *Prashant Loyalka*[7]

CONSTRUCTING TESTS THAT CAN MEASURE AND COMPARE THE MATHS AND PHYSICS SKILLS OF ENGINEERING STUDENTS IN RUSSIA AND CHINA[8]

Although the number of engineering graduates has expanded rapidly in the last two decades, relatively little is known about the quality of engineering programs worldwide. In particular, few studies look at differences in the degree to which students are learning skills across different engineering programs within and between countries. There is particular interest in the investigation of the engineering education quality in the countries with the rapidly growing economy, such as BRICS countries. Until now, there was little research in this field and one of the main reasons for this is the difficulty in developing an assessment approach and the accompanying set of instruments, which would allow for measurement and international comparison. Our study describes a set of procedures for developing such an assessment framework of instruments, to measure and compare skill levels and gains across engineering programs.

We first describe a systematic approach for constructing cross-nationally comparable instruments in maths and physics for students in the first two years of their undergraduate engineering programs. The approach includes both a priori procedures (including expert assessments to avoid construct, method, and item bias), and a posteriori procedures (including the psychometric analysis of test quality, differential item functioning, and identifying and reducing bias in the data).

In addition to describing this set of procedures in theory, we also show how we systematically implemented these procedures. Drawing on data that we collected from over 24 engineering experts and 3,600 engineering students across Russia and China, we provide evidence that it is possible to create tests that are cross-culturally valid, equate-able, and free from bias.

Keywords: engineering education, BRIC countries, quality of education, cross-cultural measurement
JEL Classification: Z

---

[1] Higher School of Economics (Moscow, Russia). Center for Monitoring the Quality in Education. PhD. The Head; e-mail: ekardanova@hse.ru

[2] Higher School of Economics (Moscow, Russia). Center for Monitoring the Quality in Education. The researcher; e-mail: enchikova@hse.ru

[3] Stanford University, (USA) Graduate School of Education, PhD student; e-mail: zshi2@stanford.edu

[4] Stanford University (USA), Rural Education Action Program, Freeman Spogli Institute for International Studies, Project Manager; e-mail: nsydneyj@stanford.edu

[5] Educational Testing Service (USA), Director of Research, Higher Education, PhD; e-mail: LLiu@ETS.ORG

[6] Educational Testing Service (USA), Associate Research Scientist, PhD; e-mail: lmao@ets.org

[7] Prashant Loyalka, Stanford University (USA), Stanford Graduate School of Education. PhD. Affiliated Faculty; e-mail: aielman@stanford.edu

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

This Working Paper is an output of a research project implemented within HSE's Annual Thematic Plan for Basic and Applied Research. Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

1. Introduction

The number of engineering education graduates has expanded rapidly in the BRIC countries over the last two decades (Carnoy et al., 2013). Whereas twenty years ago, higher education systems in developed economies produced more engineering graduates, today the higher education systems of emerging economies produce the majority of the world's engineering graduates (Carnoy et al., 2013; Gereffi et al., 2008). In fact, the number of engineering graduates produced each year by China's institutions is more than twice the total annual number of engineering graduates produced by the United States (Loyalka et al., 2014).

As a result, the quality of engineering education in BRIC countries has become a focus of discussion between scholars and politicians. However, in contrast to the variety of efforts to measure and compare the quality of education in the engineering programs in the United States (see Arum and Roksa, 2011; Bok, 2006; Pascerella and Terenzini, 2005), there are no published studies that do so for emerging economies like the BRICs. A few studies have used indirect methods to determine the quality of engineering programs in BRIC countries. Such studies have, for example, examined employer feedback on the skill levels of engineering graduates (Blom and Saeki, 2011; Levin Institute, 2010; Klintsov et al., 2009; Gereffi et al., 2008; Mooney and Neelakantan, 2006; Bondarenko et al., 2005; Borsch, 2010). More recently, Loyalka et al. (2014) examined the quality of engineering programs in BRIC countries along a number of dimensions using an educational production function approach. Although Loyalka et al. (2014) arrived at tentative conclusions about the quality of elite and non-elite institutions within and across BRIC countries the authors were unable to directly measure the students' progress during the course of engineering programs. Other research focuses on the skills and competencies of students and graduates which indicate their ability to work as engineers (Gereffi et al., 2008; Zlatkin-Troitschanskaia et al., 2015; Amaral and Rosa, 2010; Pearce, 2015). Although this approach is very promising, it is difficult for large international assessment since it is very dependent on the particular engineering major. To make the research more comparable for students of different majors, this research is focused on the basic knowledge in maths and physics that is required for the successful development of professional skills and competences.

OECD's AHELO project is an example of an international study, which seeks to investigate the quality of engineering education (OECD, 2012). However, this project was developed for measuring

skill levels and not measuring skill gains (by "skill gains" we mean the degree to which skills have increased during university between two or more points of time). Since engineering universities differ within the countries, it is important to keep in mind the initial level of the students to estimate the contribution of the teaching program. In these terms, students' gains will be a more representative indicator for the quality of education in the university. Another point of criticism was the question about cross-cultural validity of AHELO instruments (Pearce, 2014; Wolf et al. 2015).

The quality of engineering education is of interest for international competition, and is important for education policies at home; for understanding the best ways to improve the quality of engineering programs. In particular, examining the quality of engineering programs in Russia is of direct policy interest. There are approximately 1 million engineering students in Russia today. The quality of education on those programs is the subject of on-going debate (Venig, 2011). It is also argued that the quality of engineering education in the majority of non-elite institutions is deteriorating over time (Pokholkov et al., 2012). Part of the concern is that many non-elite institutions have been forced to accept high school graduates that lack the basic technical (maths and physics) skills they need to succeed in engineering programs (Aleksandrov et al., 2013; Carnoy et al., 2013).

Although the number of engineering graduates has expanded rapidly in the BRIC countries, there is a lack of research on the quality of engineering programs; in particular how much engineering programs are leading to skill gains, which are more likely than skill levels to reflect the quality of engineering programs. In fact, there are no studies we know of which have attempted to measure and compare skill gains inside BRIC countries or across two or more countries. However, all quantitative research must be based on precise measures gained with the reliable instruments. Now, there is a lack of reliable and valid instruments for measuring the quality of engineering education. Such instruments must be developed taking into account the educational systems and the cultural backgrounds of different countries. Therefore, we focus on those countries and prove that measurement instruments are valid and fair for all countries. An examination of cross-cultural equivalence is a prerequisite for accurate and meaningful comparisons across different countries.

This study describes a set of procedures for creating assessment instruments for measuring and comparing the basic knowledge in maths and physics of engineering programs students in two countries: Russia and China. The study is a part of International Study of Higher Education Learning (ISHEL), which aims to measure and compare the quality of engineering programs both

within and across BRIC countries. We describe a set of procedures in theory and also show how we systematically implemented these procedures for two majors (electrical engineering and computer science) in these two countries. We develop an approach for constructing international test instruments and prove their reliability and validity, including cross-cultural validity. For this purpose, the procedure of test development was carefully controlled at every step to enable meaningful comparisons between countries. Overall, our results show that creating assessment instruments for comparing the quality of engineering programs both within and across countries is possible.

2. Approach for test constructing

As did other international studies, this research faces many methodological challenges, and the most crucial among them are the issues related to the cross-cultural equivalence of measurements. Without equivalence, it is unclear whether the observed differences across groups are due to true differences in maths or physics ability or to such differences as the understanding of items, or irrelevancy of item contents. The validity of measurements and comparisons depends on the degree to which the versions of tests in two languages indeed measure the intended constructs and provide comparable measurements.

There are different reasons why instruments can be incomparable across countries. These are construct differences, instrument differences, administration differences, sample differences, and response procedure differences (Ercikan, 2013). Our test development methodology takes into account all these differences and includes several stages, as described below, each of which contributes into the comparability of results.

We selected two undergraduate engineering majors for this study, electrical engineering (EE) and computer science (CS). These majors were selected because universities in China and Russia produce a large number of graduates each year in these two programs and because both of these majors teach skills that are important and highly valued in the modern economy.

During stage 1, a priori procedures were conducted to ensure the construct validity of the assessment instruments. We developed pilot assessments in maths and physics. Stage 1 was carried out in four steps. First, we selected comparable majors within the categories of EE and CS across China and Russia. Second, we selected content and sub-content areas by using expert evaluation of content

maps for each test (maths and physics) solicited from each field from experts at Chinese and Russian universities. Third, we collected items for the tests, which matched these content areas from official sources and verified the items based on another evaluation by local experts. Finally, in order to catch any final issues with question wording or test format, we conducted a small clinical pilot study. At the end of this stage, we developed tests, which were ready for a larger pilot study. To collect data on the quality of our test items and allow for a posteriori procedures to further improve the quality of our tests we conducted a study of over 3,600 year 1 and year 3 engineering students across Russia and China.

Stage 2 included a posteriori procedures to address remaining issues in the pilot tests. We analysed our pilot data to provide evidence that the tests are reliable, cross-culturally valid, equate-able, and free from bias. In particular, we used Item Response Theory (IRT) (Embretson and Reise, 2000) modelling to conduct item analysis as well as tests of dimensionality and reliability. We also paid particular attention to differential item functioning (DIF) analysis to provide evidence concerning the cross-cultural comparability of the test results and to ascertain the possibility of creating a common scale between the two grades and across the two countries.

As a result of these procedures, we were able to construct reliable and valid assessment instruments that measure and compare individual progress in the first two years of higher education engineering programs within and across the countries.

The next sections of the paper are organized as follows: first, we briefly describe the procedures that were conducted to establish cross-cultural validity; second, we provide a more detailed description of the procedures and analysis, which proved that it is possible to construct a common scale between the two grades and the two countries.

3. Stage 1: A priori procedures

A priori procedures were performed in several steps as follows:

Step 1. *Selection of comparable EE and CS majors across China and Russia*

We selected majors in China and Russia that had both consistent coursework and curricula across universities within each country, and substantial overlap in coursework and curricula across countries. For example, we selected the EE majors of Electrical Communications Engineering, and

Electronic Information Science and Technology in China and Electrical Energy and Electric Engineering, Radio Engineering, Information and Communication Technology and Communication Systems, and Design and Technology of Electronic Instrumentation in Russia on the basis of their shared core requirements—computer programming, circuits, analogue electronics, signal processing, and digital systems.

Step 2. *Selection of the content and sub-content areas for the tests*

Given that the curricular requirements for these majors in different countries differ in significant ways, it is crucial to verify the content validity of the tests for each country at the start of the test development process. In other words, the content of the tests must be invariant for all the countries tested, but should also be selected to match the curricular content for the sample majors in each country as closely as possible.

To achieve this goal, we first produced content maps in maths and physics for year 1 and year 3 students based on the national curriculum standards in Russia and China. These content maps contained:

- content and sub-content areas taught in high school (for the grade 1 test) and in university (for the grade 3 test) in each country,
- the relative weight of the content areas in each country's national curriculum—as measured by the number of units in popular textbooks devoted to the study of these content areas.

Table 1 shows the number of content and sub-content areas included in the content maps.

Table 1

The total number of content and sub-content areas

Test	Number of content areas	Number of sub-content areas
Maths, year 1	8	36
Maths, year 3	10	116
Physics, year 1	10	49
Physics, year 3	10	104

We then offered the content maps to twelve experts in each country for evaluation. The experts were professors in maths and physics from universities in Russia and China. We devoted special attention to balancing the number of experts who instructed EE students with those that instructed CS students to ensure that we obtained accurate assessments of the importance of each content area for our two sample majors. Since there is a big differentiation between elite and non-elite universities in both countries, we paid attention to select experts from different types of universities, both elite and non-elite.

The experts were presented with the list of content and sub-content areas and asked to rate the importance of each area for the academic progression of students in EE and CS majors. Based on this evaluation, we selected the content areas with the highest average ratings in each subject. Table 2 shows the list of content areas selected for each subject.

Table 2 *The content areas selected*

Test	Content areas selected based on expert evaluation
Maths, year 1	Functions and domains
	Equations
	Derivatives and their application
	Mathematical reasoning and logic
	Trigonometric functions and equations
Maths, year 3	Single variable differentiation
	Single variable integration
	Linear algebra
	Multivariate differentiation
	Series
Physics, year 1	Electromagnetic fields
	Electromagnetic induction
	Circuits
	Optics
	Oscillation and mechanical waves
Physics, year 3	Electricity and Electric Fields
	Electromagnetic Induction

Magnetism and Magnetic Fields
Optics
Waves and Oscillation

Step 3. *Collection and verification of test items*

Our item collection procedure consisted of two steps. First, we collected tests items that reflected the content areas selected. Second, we reenlisted the help of our team of experts to evaluate the items to make sure they were valid, relevant, clear, and of suitable difficulty.

We collected a pool of test items from the following sources:

- year 1 items from China's Gaokao (Higher Education Entrance Examination) and Russia's Unified State Exam;
- year 3 items from standardized exams and popular exercise books in China and Russia;
- year 1 and year 3 items from maths and physics tests created by the Educational Testing Service (ETS) for college level assessment in the United States.

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

- fall under the content areas in the content maps (Table 2);
- have short and simple sentence structure and simple grammatical form to make them translatable into other languages;
- have a multiple-choice (MC) format because this is the most familiar format for our target populations.

Additionally, a small number of items which reflected some of the content areas not included in the content maps, but that had high ratings from experts of only one country, were also included.

We collected item pools of 85–90 items for each test. Approximately one third of the items came from Russian sources, one third came from Chinese sources, and the rest from ETS.

After collecting the items, we translated them into Russian and Chinese and evaluated the quality of the translations. All items other than those from ETS 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. English speaking professors of maths and physics from each target country (China and Russia) were asked to compare back-translated items to their original English versions to rule out the possibility of any discrepancy in meaning. The few items that showed discrepancy were rectified or dropped.

Finally, to ensure the items were valid, relevant, clear, and of suitable difficulty, we interviewed the same 12 experts from each country. They were asked to rate the items according to four criteria: (1) comprehensibility of the item wording, (2) appropriateness of the item in measuring the content area of interest, (3) item difficulty, and (4) expected time required to solve the item.

The consistency of expert ratings was analysed using Chronbach's Alpha, and correlations between the ratings of each expert against ratings of the other experts. Additionally the behaviour of experts was analysed in order to reveal possible distortions in their scores which might have affected the final ratings. The detailed description of the expert analysis is beyond the focus of the paper, but the main results were: (1) consistency between experts was high (the Chronbach's Alpha coefficient was over 0.8) and (2) the experts demonstrated no effects and did not bring bias into the evaluation procedure.

Based on the ratings of experts and taking into account item difficulty, total testing time (no more than 55 min per subject) and the intention to keep the balance between the items from different countries, we selected the items for the year 1 and year 3 clinical pilot tests in maths and physics. The total number of items in each test was 55.

Step 4. *Conducting a clinical pilot*

We conducted a clinical pilot with a small number of students to check for such things as language ambiguity and formatting issues. Also our intention was to define empirically the item difficulty and get feedback from the students and their teachers about the items. We gave the clinical pilot tests to 40 year 1 students and 40 year 3 students in each country.

We found 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. The clinical pilot showed the items were understandable and had an acceptable difficulty level.

4. Stage 2: Pilot study and posteriori procedures

Method

Sample

The target population for this study was defined as all students in the first and third years of EE and CS undergraduate engineering programs 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 should be selected) and university location (both big and small cities across the country should be represented). Based on these criteria we selected 11 universities in China and 11 universities in Russia for the pilot study. In each university, classes of year 1 and year 3 students from their respective EE and CS departments were sampled until the number of students in each department was 60 students for each year (or when all of the students had been sampled). In each sampled class, two thirds of the students were randomly selected to take the maths and physics tests, with maths and physics being given in randomized order. Our final sample consisted of 1797 students in China and 1802 students in Russia.

Instrument

The tests in maths and physics for the first and third grades included 45 items each. As indicated earlier, the items in each test reflected the five main content areas presented in Table 2, and some additional content areas that were highly ranked in importance by the experts in one country (but not the other). All items were in multiple-choice format with one correct answer from 4–5 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 tests for each subject had around 20 common items between the year 1 test and the year 3 tests to make it possible to equate the test scores from different grades and place the results on a common scale.

Procedure and data collection

The pilot study was conducted at the end of October 2014 using a paper and pencil format. In Russia, the pilot was carried out by the staff members of the respective sample universities. Observers from the research team were present at some universities during the testing to ensure compliance with standardized research protocols. In China, the research team trained dedicated survey enumerators and accompanied them into the field to carry out the pilot in 11 universities.

The testing was conducted during two 55-minute sessions. Students were also asked to fill in a questionnaire about their background and schooling experience after the testing was complete.

Analysis

We used IRT modelling for item analysis and tests of dimensionality and reliability. We also conducted DIF analysis to provide evidence concerning the cross-cultural comparability of the test results.

One of the intentions of the pilot study and subsequent analysis was to shorten the test by selecting only the items with the best psychometric properties. In particular, while the pilot tests were 55 minutes each, the research team intends to cut the length of each subject test to 40 minutes for the main study. In the pilot study we included more items than we needed for the main study and gave more time so that we would be able to delete some of the items from the tests due to poor psychometric quality—not fitting the IRT model or threatening unidimensionality. The expected number of items for the main study is 35–40 for each subject test.

Because the tests in each subject include common items across the two years, we could also ensure simultaneous calibration and vertical scaling. The large number of common items allows for the selection of the items that are most appropriate for equating. These items should be of good psychometric quality and DIF free.

The one-parameter dichotomous Rasch model (Wright and Stone, 1979) was used for IRT analysis. Under this model, each test item is characterized by one parameter, which is difficulty, and each student is also characterized by one parameter, which is ability level. Rasch analysis places students and items on the same measurement scale with the logit as the unit of measurement. The reasons for choosing the Rasch model are both psychometrical and practical. First, the Rasch model has optimal metric properties, and second, from a practical point of view, it is useful for data analysis—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.

The data analysis was performed for each subject separately in two stages. During the first stage, we treated the data sets for each grade separately. The purpose of this stage was to discover whether it would be possible to construct a common scale for the two countries for 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 find out whether it would be possible to place all parameters onto a common scale between the two grades and across the two countries.

During the first stage the data analysis was performed for year 1 and year 3 separately in several steps as follows:

Step 1. *Analysis of model fit*

To measure the extent to which the data fit the Rasch model, we used the unweighted and weighted mean square statistics (in terms of Winsteps output: OUTFIT MNSQ and INFIT MNSQ, respectively). These statistics rely on standardized residuals, which represent the differences between the observed response and the response expected under the model (Wright and Stone, 1979). The OUTFIT MNSQ statistics represent the mean squared standardized residuals. They are known to be very sensitive to unexpected responses (Smith, 1991). INFIT MNSQ statistics that are weighted by the information function and take into account the variance of expected responses are more useful for the goodness-of-fit analysis. Generally, a criterion of +1.2 for these statistics is used to flag potential problems. A few items were deleted because they fit the model poorly.

Step 2. *Country-related DIF analysis*

An item demonstrates DIF if test participants with the same ability level who belong to different groups have markedly different chances of completing that item correctly. We test for DIF across countries because of possible differences in language, tradition of teaching, and curriculum. In this study we used the ETS approach for DIF classification (Zwick et al., 1999), which designates items as A (negligible or nonsignificant DIF), B (slight DIF), or C (large DIF) items depending on the magnitude of the Mantel-Haenszel statistic (Dorans, 1989) and its statistical significance. An item was considered a C item if two conditions were satisfied: (1) the difference in item difficulty between different groups of students was more than 0.64 logits, and (2) the Mantel-Haenzel statistic had the significance level $p<.05$ (Linacre, 2011).

We found a small number of items with country-related DIF. One possible way forward would be just to delete the C items, but as this could negatively impact the precision of our estimation. For our analysis, we decided to keep the C items and to treat them as country unique items. Thus, although some test items demonstrated DIF, a sufficient number of DIF-free items allowed us to construct a common scale for both countries.

Step 3. *Analysis of the whole data set*

For examination of the dimensionality of the scale we used a principal component analysis (PCA) of the standardized residuals (Linacre, 1998; Smith, 2002). Theoretically, if all the information in the data is explained by one latent variable, the residuals would represent random noise and would be independent of each other. As a consequence, correlations between the residuals would be near zero. If there is no second dimension within the data, then a PCA of the standardized residuals should generate eigenvalues all near one and the percentage of variance across the components should be uniform (Ludlow, 1985).

To analyse the reliability, we used the person reliability index (which is close in value and interpretation to classical reliability) and alternative statistics, the separation index provided by Rasch analysis (Stone, 2004). The separation index compares the distribution of student measures (the estimates of ability) with their measurement errors and indicates the spread of student measures in standard error units. The index can be used to calculate the number of distinct levels, or strata separated by at least three errors of measurement, in the distributions (Wright and Stone, 1979; Smith, 2001). The number of strata are calculated as: $\text{Strata}=(4G+1)/3$, where G is the separation index.

In order to show the relative distribution of item difficulty and students scores in a common metric, we constructed the variable map (Wright and Stone, 1979).

As a result of the first stage, we were able to construct final tests in maths and physics for year 1 and year 3. They were essentially unidimensional, reliable and valid instruments to measure and compare the maths and physics skills of engineering students in Russia and China.

During the second stage, the data analysis placed the parameter estimates for year 1 students and year 3 students onto the same scale. Linking data from two datasets used equating methodology. Wolfe (2004) describes commonly used designs and estimation procedures. The most widely-used approach is called common-item equating or anchor-test design. In this approach, linkage between forms is ensured by including common items on each form. Relative item locations can then be determined by holding the item calibrations for the common items constant across forms. If the equating is done in two separate calibrations, items can be anchored to estimate relative item locations.

The tests for different grades had 20 common items to make it possible to equate them. Recommendations relating the number of common items vary. In general, a larger number of common items results in more precise and stable item calibration, but Wright and Bell (1984) recommend 5 to 10 items to form the link. We had a larger number of common items which gave us an opportunity to select items that are appropriate for linking. Only items that exhibited adequate fit to the model within the tests and that were DIF free were used for equating purposes.

For equating, we used separate calibration design with anchoring items from one test when calibrating the other test (Wolfe, 2004). We first calibrated the year 1 test. The difficulties of the common items were anchored during the calibration of the year 3 test. That is, when we calibrated the year 3 test, we treated the common items as being fixed and their difficulties were not estimated. As a result, the remaining parameter estimates for the year 3 test were forced onto the same scale as for the year 1 test.

After equating, we evaluated the quality of the link between the tests by calculating the item-within-link statistic (Wright and Bell, 1984). Under the null hypothesis the items exhibit perfect fit within the link, this statistic has an expected value of 1. Link adequacy was also evaluated by determining

the stability of the item difficulty estimates across the year 3 test with and without anchoring. To do that, we calculated correlation between the item difficulty estimates.

Due to space limitations, for the purposes of this paper we present the results for the maths tests only. The results of the analysis of the physics tests are substantively similar.

5. Results

The year 1 maths test

Step 1. *Analysis of model fit*

Starting with our pool of 45 items, eight items were deleted because of poor psychometric quality (low discrimination and/or misfitting the model) or as threat to unidimensionality. For the rest of the analysis we consider the reduced set of 37 items.

Step 2. *Country-related DIF analysis*

13 items (from 37) demonstrated DIF across countries: 7 items in favour of China and 6 items in favour of Russia. Figure 1 shows the difficulties of items separately for students from different countries. We see that most items are DIF free as they demonstrate stable estimates of difficulty (the difference in item difficulty between countries is not significant).

The 13 items with DIF were analysed. For example, the item where students are asked to indicate the interval where the roots of the function $f(x) = 2^x + 3x$ are located, demonstrates DIF in favour of China. It means that Russian students with the same level of maths ability have markedly lower chances of completing the item correctly. It can be explained by the fact that Russian students don't have much experience with this type of tasks. On the contrary, Chinese students spend much more time in high school studying functions and their properties, and such an item is familiar for them.

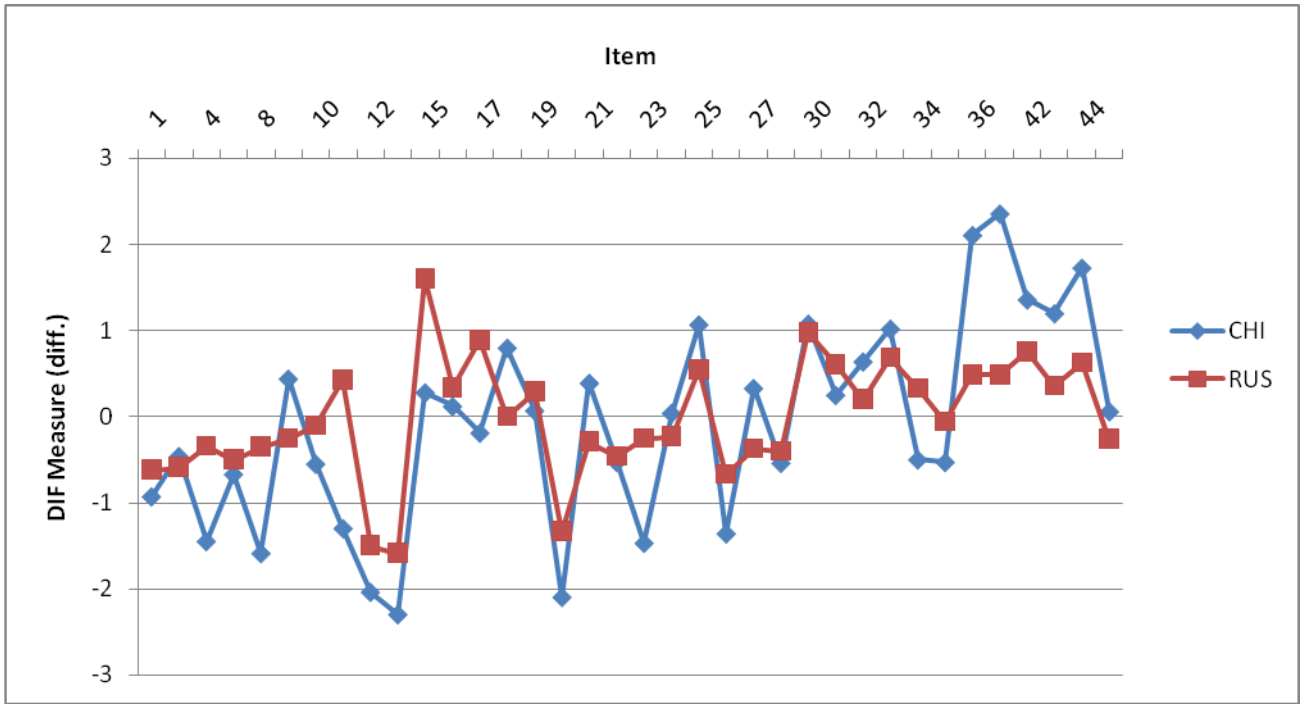

***Figure* 1**. Item difficulties for different countries

While the reasons for the observed DIF are not of immediate concern for this paper, in order to create comparable assessment instruments, we must decide how to handle the items that exhibit DIF. We decided to treat these 13 items as unique items for each country. Therefore, we have 24 common DIF free items for both countries and 13 items, which are specific for each country. Thus, the total number of items for each country is 37, but 13 from them are specific for the country, so the total number of items for further analysis is therefore 50.

Step 3. *Analysis of the new data set*

We repeated analysis of model fit with the new data set. Our analysis showed that the values of both INFIT and OUTFIT MNSQ for all items are in acceptable range with a mean 1.00 and SD = 0.08 for INFIT MNSQ, and mean 0.99 and SD=0.15 for OUTFIT MNSQ. This result indicates that all items in the test fit the model in accordance with the chosen criteria.

Then we examined the dimensionality of the test by conducting a PCA of the standardized residuals. The eigenvalues of the residual correlation matrix for the five primary components ranged from 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%. Based on these results, there is no evidence for a second dimension in the data.

The next portion of the analysis was devoted to the properties of the entire test. The person reliability is 0.85, which means that the proportion of observed student variance considered true is 85% (for comparison classical reliability $\alpha = 0.83$ for the test). In addition, our analysis produced a person separation index of 2.39, indicating at least three statistically distinct groups of students along the continuum.

Figure 2 presents the variable map, which shows the relative distribution of items and students in a common metric. The horizontal axis is the logit unit of measurement scale. On the map, students are represented in the upper part and the items are 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 placed on the left side of the map (negative logits). The distribution of students is wide and representing a good differentiation between higher and lower scoring students for measurement purposes. An analysis of the distribution of item locations shows that while the student sample is well distributed relative to the items, there is a lack of difficult items appropriate for very high-performing students although this is not a problem because of the small number of such students in the sample.

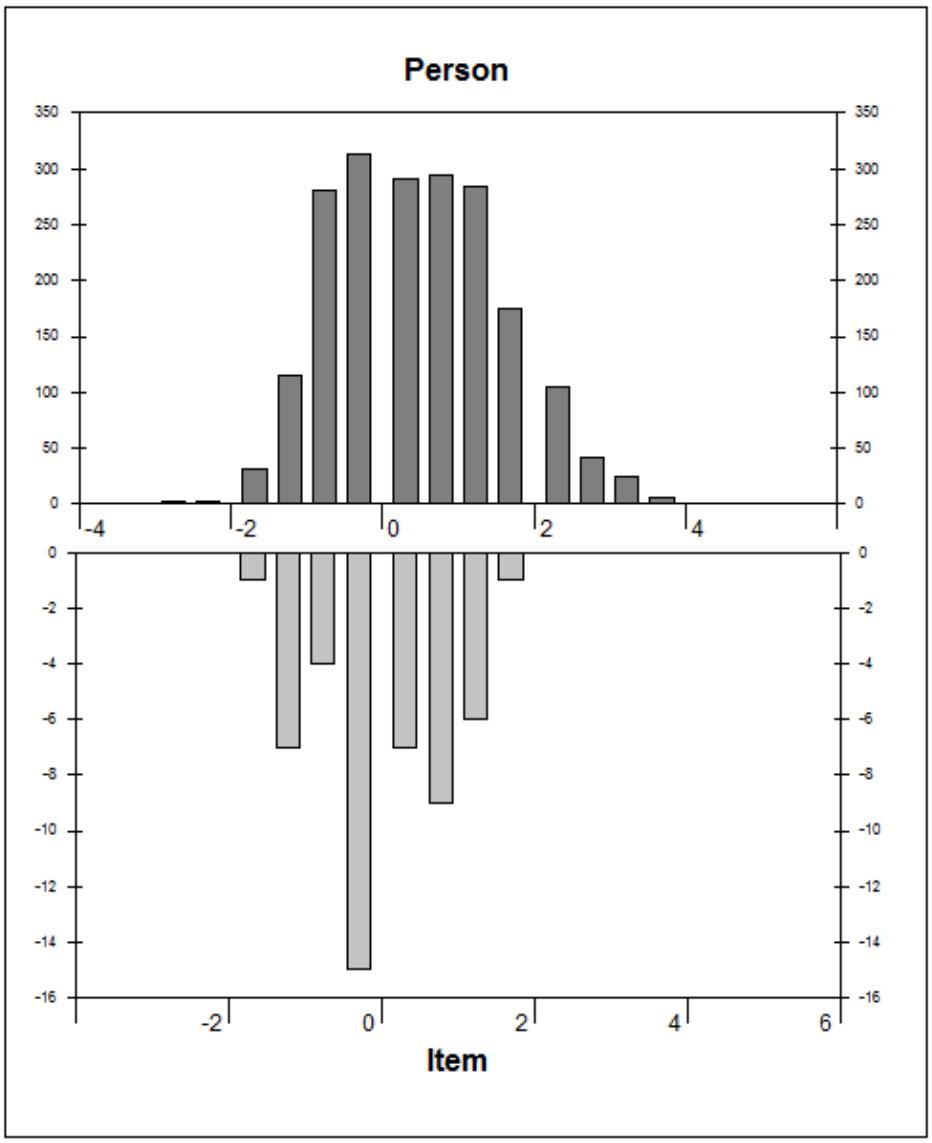

Figure 2. The year 1 maths variable map

The year 3 maths test

The results for the year 3 maths test are substantively similar. Six items were deleted because of poor psychometric quality or threats to unidimensionality. For further analysis, we considered 39 items. 11 of these demonstrated country-related DIF: 5 items in favour of China and 6 items in favour of Russia. Therefore, we have 28 common items for both countries (that are DIF free) and 11 items, specific for each country. The total number of items for each country is 39, but 11 from them

are specific for the country, so the total number of items for further analysis is 50. All items fit the model, and the test can be considered essentially 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 3 presents the variable map for the year 3 test. We see that the distribution of students is wide and the student sample is well located relative to the items. Similar to the year 1 test, there is a lack of difficult items appropriate for very high-performing students, but, again, this is not a significant issue because there are a very small number of such students in the sample.

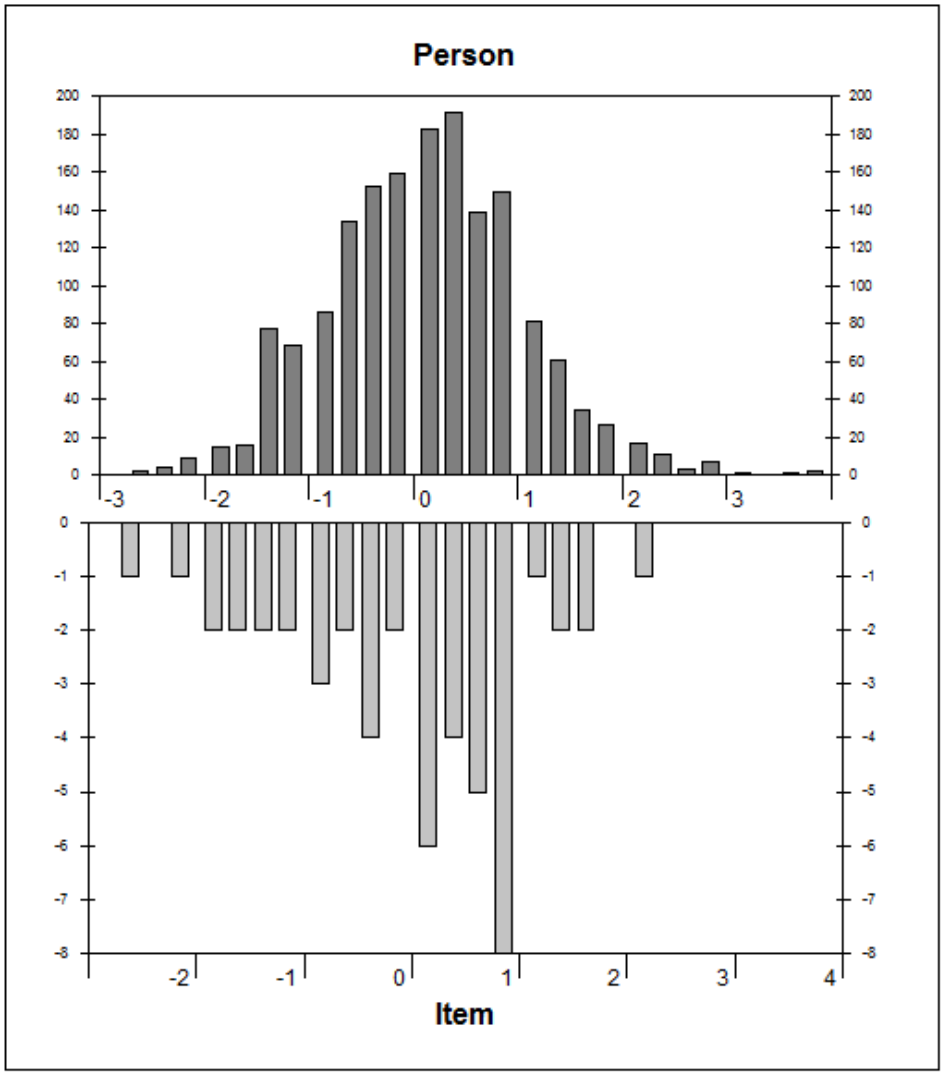

Figure 3. The year 3 maths variable map

Thus, we constructed the year 1 and year 3 maths tests, which are essentially unidimensional, reliable and valid instruments to measure and compare the maths skills of the year 1 and year 3 engineering students in Russia and China.

Tables 3 and 4 show the content of the final versions of the year 1 and year 3 tests.

Table 3. The content of the year 1 test

Number	Topic	Frequency	%
1	Derivatives and their application	7	18.9
2	Equations	7	18.9
3	Functions and domains	5	13.5
4	Inequalities	3	8.1
5	Mathematical reasoning and logic	5	13.5
6	Single Variable Differentiation	4	10.9
7	Trigonometric functions and equations	6	16.2
	Total	**37**	**100**

Table 4. The content of the year 3 test

Number	Topic	Frequency	%
1	Derivatives and their application	3	7.7
2	Equations	1	2.6
3	Functions and domains	1	2.6
4	Inequalities	1	2.6
5	Linear Algebra	5	12.8
6	Mathematical reasoning and logic	2	5.1
7	Multivariate Differentiation	6	15.4
8	Ordinary differential Equations	1	2.6
9	Probability and statistics	3	7.7
10	Series	2	5.1
11	Single Variable Differentiation	7	17.9

12	Single Variable Integration	5	12.8
13	Trigonometric functions and equations	2	5.1
	Total	**39**	**100**

The next step was to ascertain whether it would be possible to create a common scale between the two grades. This step required performing a data analysis for both grades together.

Equating the tests between different grades

Starting from 20 items that were common across the year 1 and 3 tests, only 7 items were selected as good candidates to be anchor items. This number is about 18% of the total number of items in each test and is close to the 20% recommended by Angoff (1971). Other common items either were deleted during the first stage of investigation or exhibited DIF for at least one test. The difficulties of selected anchor items were fixed with values for the year 1 test when calibrating the year 3 test. As a result, the parameter estimates for the year 3 test were placed onto the scale of the year 1 test.

To evaluate the quality of the link between the tests we calculated the item-within-link statistic. Its value of .95 indicated a reasonable fit within the link. Correlation between the item difficulty estimates across the year 3 test with and without anchoring was .99, which indicated stability of item order. In addition, we checked the quality of the year 3 test with anchoring by repeating analysis of model fit, dimensionality and reliability. All test characteristics were the same.

6. Conclusion and discussion

The quality of higher education is a subject of on-going discussion among scholars, politicians, teachers, students and employers. All sides seek to improve the quality, but suggest different methods and focus on different details. However, to improve the quality and to capture the progress there is a need for special measurement instruments which will indicate improvement. In addition, since the definition of "education quality" varies significantly, those instruments must be developed according to the context of the research. This paper focuses on the development of such instruments for the assessment of engineering programs of higher education in Russia and China.

Researchers faces many methodological challenges described in this paper. The instruments must be verified for cross-cultural equivalence and the sampling procedures must provide comparable results. In this study, we paid particular attention to the description of a systematic approach for constructing cross-nationally comparable instruments. The approach included both a priori procedures (including expert assessments to avoid construct, method, and item bias) and a posteriori procedures (including the psychometric analysis of test quality, differential item functioning, and identifying and reducing bias in the data).

During this study, we systematically develop test instruments for measuring and comparing the maths and physics skill levels and gains of year 1 and year 3 engineering students in Russia and China. Based on the results of pilot study we demonstrate that it is possible to construct essentially unidimensional, reliable and valid instruments for these purposes. We show that our test instruments have a good quality and can be used for further international research.

Since the study has a complex design and suggests testing students of different grades in different countries, it was important to establish the possibility of constructing a common scale between grades in different countries. We used anchor items to link the tests. The procedure includes DIF analysis to ensure that items are functioning in the same way in different tests. We used simultaneous and separate calibration for creating a common scale and built a common scale between grades and countries. This has great practical implications because it allows us to compare test scores directly and to estimate the student progress between grades in different countries. Therefore, it gives us a base for international comparisons and further statistical investigations.

This paper presents the results of a pilot study used to develop reliable and valid test instruments for the main wave of ISHEL study taking place in 2015. This study will allow the measurement of gains in academic skills for a representative population of higher education students (in a representative sample of higher education institutions) in two of the world's largest emerging economies (China and Russia) and benchmark those gains among selected populations of higher education students in the United States. This study utilises different research designs to examine which factors help higher education students develop academic and critical thinking skills. In particular, to examine whether (a) institutional characteristics (specifically attending higher education institutions of varying selectivity); (b) faculty characteristics (educational qualifications; faculty ranking; time faculty spend on research versus teaching); and (c) pre-existing, day-to-day curricular/instructional practices

(active, collaborative/cooperative, project-based, and small group learning) impact student skills. The results of the main study will be useful for policy makers for improving the quality of engineering education.

This paper enlightens the process of test development for cross-cultural research and demonstrates a methodology for the creation of valid and reliable tests. It can be useful for researchers who are looking for methodology for test development in international research. The approach described in the paper is effective for test construction as shown by the results of the pilot study.

References

Aleksandrov, A.A., Fedorov, I.B., Medvedev, V.E. (2013). Engineering Education Today: Problems and Solutions. *Higher Education in Russia*, 12, 3-8.

Altbach, P., Reisberg, L., Rumbley, L. (2009). «*Trends in Global Higher Education: Tracking an Academic Revolution»*. Chestnut Hill, MA: Boston College Center for International Higher Education.

Amaral, A., and Rosa, M. J. (2010). Recent trends in quality assurance. *Quality in Higher Education*, *16*(1), 59-61.

Angoff, W.H. (1971). Scales, norming, and equivalent scores. In R.L.Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Arum, R., and Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Blom, A., Saeki, H. (2011) *Employability and Skill Set of Newly Graduated Engineers in India*: Policy Research Working Paper, World Bank.

Bok, D. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton University Press.

Bond, T.G., Fox, C.M. (2001). *Applying the Rasch model*. Mahwah, NJ: Lawrence Erlbaum.

Bondarenko, N., Krasilnikova M., Kharlamov K. (2005). Demand for Labour Force – View of Employers. *Monitor Economics of Education,* 2005, 1. Moscow: State Research University Higher School of Economics.

Borsch, V. (2010). Mechanisms of independent assessment of educational quality by means of analysis of graduates' state of being relevant in the labor market, and recommendations for their practical use. *Inzhenernoe obrazovanie*, 6, 4-9

Carnoy M., Loyalka P., Dobryakova M. S., Dossani R., Froumin I., Kuhns K., Tilak J. B.,Rong W. (2013). *University Expansion in a Changing Global Economy: Triumph of the BRICs?* Stanford: Stanford University.

Dorans, N.J. (1989). Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel-Haenszel Method. *Applied Measurement in Education*, 2(3), 217-233.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ercikan, K., Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K.F.Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology: vol.3. Testing and Assessment in School Psychology and Education.*

Gereffi, G., Wadhwa, V., Rissing, B., and Ong, R. (2008). «Getting the Numbers Right: International Engineering Education in the United States», China, and India. *Journal of Engineering Education*, 97, 1: 13-25.

Klintsov, V., Shvakman I., and Solzhenitsyn Y. (2009). How Russia Could be more Productive. McKinsey Quarterly, September 2009. Available at: https://www.mckinseyquarterly.com/Europe/How_Russia_could_be_more_productive_2435

Levin Institute. (2010). "The Evolving Global Talent Pool: Lessons from the BRICS Countries." Report by the Levin Institute, State University of New York.

Linacre, J.M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome measurement*, 2, 266-283.

Linacre J. M. (2011). A User's Guide to WINSTEPS. Program Manual 3.71.0. Available at: http://www.winsteps.com/a/winsteps.pdf

Loyalka, P., Carnoy, M., Froumin, I., Dossani, R., and Tilak, J.B. (2012). "Getting the Quality Right: Engineering Education in the BRIC Countries." Report: 01-42.

Loyalka, P., Carnoy, M., Froumin, I., Dossani, R., Tilak, J.B., Dobryakova, M. (2014). "Factors Affecting the Quality of Engineering Education in the Four Largest Emerging Economies." *Higher Education* 68(6): 977-1004.

Ludlow, L.H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement*, 45(4), 851-859.

Mooney, Paul and Shailaja Neelakantan. (2006). Foreign Academics Question the Quality of their Countries' Engineering Programs. *Chronicle of Higher Education*, September 8, 2006.

OECD (2012) AHELO Feasibility Study Report. Volume 1. Design and Implementation http://www.oecd.org/education/highereducationandadultlearning/AHELOFSReportVolume1.pdf

Pascarella, E. T., and Terenzini, P. T. (2005). *How college affects students* (Vol. 2). K. A. Feldman (Ed.). San Francisco, CA: Jossey-Bass.

Pascarella, E. T., Blaich, C., Martin, G. L., and Hanson, J. M. (2011). How robust are the findings of Academically Adrift?. *Change: The Magazine of Higher Learning*, *43*(3), 20-24.

Pearce, J. (2014). Ensuring quality in AHELO item development and scoring processes. In: Musekamp F, Spöttle G, editors. Vocational Education and Training: Research and Practice: Vol. 12. Kompetenz im Studium und in der Arbeitswelt. Nationale und internationale Ansätze zur Erfassung von Ingenieurkompetenzen. Competence in Higher Education and the Working Environment. National and International Approaches for Assessing Engineering Competence. (1st ed.). Frankfurt am Main: Peter Lang

Pearce, J. (2015). Assessing vocational competencies in civil engineering: lessons from AHELO for future practice. *Empirical Research in Vocational Education and Training*, *7*(1), 1-15.

Pokholkov,Yu. P., Rozhkova, S.V.,Tolkacheva, K.K. (2012). The Level of Preparation of Engineers in Russia. Assessment, porblems and ways for their solution. Problems of Management in Social Systems, 4, 7, 6-15.

Smith, R.M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541-565.

Smith, E.V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2, 281-311.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3:2, 205-231.

Stone, M.H. (2004). Substantive scale construction. In E.V.Smith, R.M.Smith (Eds.), *Introduction to Rasch measurement* (pp.201-225). Maple Grove, MN: JAM Press.

Venig, S.B. (2011). Contribution of head classical universities in the development of engineering education. *Inzhenernoe obrazovanie*, 8, 88-90.

Wolfe, E.W. (2004). Equating and Item Banking with the Rasch Model. In E.V.Smith, R.M.Smith (Eds.), *Introduction to Rasch measurement* (pp.366-390). Maple Grove, MN: JAM Press.

Wolf, R., Zahner, D., and Benjamin, R. (2015). Methodological challenges in international comparative post-secondary assessment programs: lessons learned and the road ahead. *Studies in Higher Education*, *40*(3), 471-481.

Wright, B.D., Stone, M.N. (1979). *Best Test Design. Rasch Measurement*.– Chicago: Mesa Press.

Wright, B.D., Bell, S.R. (1984). Item banks: What, why, how. *Journal of Educational measurement*, 21, 331-345.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., and Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393-411.

Zwick, R., Thayer, D.T., Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, 36, 1, 1-28.

Authors:

Kardanova Elena, Higher School of Economics (Moscow, Russia). Center for Monitoring the Quality in Education. PhD. The Head;
e-mail: ekardanova@hse.ru

Enchikova Ekaterina, Higher School of Economics (Moscow, Russia). Center for Monitoring the Quality in Education. PhD. The researcher;
e-mail: enchikova@hse.ru

Henry Shi, Stanford University, (USA) Graduate School of Education, PhD student;
e-mail: zshi2@stanford.edu

Natalie Johnson, Stanford University (USA), Rural Education Action Program, Freeman Spogli Institute for International Studies, Project Manager;
e-mail: nsydneyj@stanford.edu

Lydia Liu, Educational Testing Service (USA). Director of Research, Higher Education. PhD;
e-mail: LLiu@ETS.ORG

Liyang Mao, Educational Testing Service (USA). Associate Research Scientist. PhD;
e-mail: lmao@ets.org

Prashant Loyalka, Stanford University (USA). Stanford Graduate School of Education. PhD. Affiliated Faculty;
e-mail: aielman@stanford.edu

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of National Research University Higher School of Economics.

© Kardanova, Enchikova, Shi, Johnson, Liu, Liyang, Loyalka, 2015