Cross-National Comparability of Assessment in Higher Education

D. Federiakin

Received in October 2019

Denis Federiakin
Intern Researcher, Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Address: 20 Myasnitskaya St, 101000 Moscow, Russian Federation. Email: dafederiakin@hse.ru

Abstract. The last three decades have seen an increase in researchers' interest in international comparative assessments of educational outcomes, particularly at the level of secondary schools. Achieving cross-national comparability is the main methodological challenge in the design of such studies. Cross-national comparability of test scores implies that the measure operates similarly across all the participating countries, regardless of their linguistic and cultural differences. The process of achieving cross-national comparability in higher education is more complicated due to specific features of higher education. This article explores the modern understanding of cross-national comparability of student assessment results and the possible ways of achieving it. It analyzes the specific aspects of higher education that complicate standardized measurement of educational outcomes and trivial achievement of cross-national comparability. The process of designing and conducting the Study of Undergraduate Performance—an international comparative research project aimed to assess and compare higher engineering education across nations—is described as an example of overcoming those challenges.
Keywords: quality of higher education, international comparative assessments, cross-national comparability of test scores, Study of Undergraduate Performance.

DOI: 10.17323/1814-9545-2020-2-37-59

Approaches to education system development drawing on human capital theory substantiate economic interest in education. In an effort to explain the economic success of developed countries, researchers tend to treat human competencies as public and private investment increasingly often [Marginson 2019]. The logic of investment at the core of "human capital" as a concept propelled the development of education policies around the world [Kuzminov, Sorokin, Froumin 2019]. The 20th century witnessed an unprecedented increase in the number of educational institutions and enrollment, education economists say. Remarkably, while a substantial proportion of national expenditures was channeled to primary and secondary education dur-

Translated from Russian by I. Zhuchkova.

ing the first half of the 20th century [Meyer, Ramirez, Soysal 1992], the second half saw an increase in the number of higher education institutions (HEI) [Cantwell, Marginson, Smolentseva 2018].

Indicators used to evaluate the impact of education on human capital have been constantly growing in number and improving in accuracy. The earliest studies conducted by the founders of human capital theory to measure human competences used such variables as years of school attainment (e. g. [Schultz 1961]). The resulting findings were helpful for substantiating the role of budgetary decision making in education and the economic approach to this sphere of social relations in general. Later research showed, however, that years of schooling alone were not enough to measure educational outcomes (e. g. [Hanushek, Woessmann 2008]). This resulted in a boom of international comparisons of educational achievements, starting with the 1980s. Stakeholders' desire to verify investment feasibility and return, along with the intention to borrow best practices, drove the need for measuring educational outcomes and comparing the results across countries.

Using cross-national measures finely tuned to assess subject-specific competencies is fraught with a number of challenges, comparability being the most critical one, especially in higher education.

This study has two main goals, (i) to explore the methodological issues of achieving cross-national comparability of test scores and (ii) to devise a methodology of doing cross-national assessments in higher education that will minimize the risk of non-comparability and control for the specific features of higher education. The first part of the article explores the modern understanding of comparability in student assessments. Next, the existing methods of achieving cross-national comparability in International Comparative Studies (ICS) for quality of education are compared, which is followed by a description of specific methodological issues associated with ICS in higher education. Finally, ways of solving those issues are analyzed using the example of the Study of Undergraduate Performance, one of the few international student assessments in higher engineering education.

Comparability of Test Results

The problem of comparability in assessment comes up every time research findings are used to compare different groups, so challenges associated with ICS in education represent a special case of a broader psychometric problem. It was in 1984, after the case of Golden Rule Life Insurance Company v. John E. Washburn[1], that comparability in educational assessment became a subject of public discussion for the first time. The suit was initiated by the insurance company to seek damages and fees from the Illinois department of *Educational*

[1] Golden Rule Insurance Company et al. v. Washburn et al., 419–76 (stipulation for dismissal and order dismissing case, filed in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1984).

Testing Service (ETS) on the grounds that the ETS examination was discriminated against black Americans. The plaintiff and the defendant reached an out-of-court settlement in 1984, with ETS assuming an obligation to revise all of its standardized tests to eliminate the biased items.

Cross-group comparability of test results implies that the measure functions similarly across all the subsamples, whether participants are grouped by age, gender, native language, or ethnicity [Meredith 1993]. Otherwise speaking, representatives of all the groups should equally interpret the theoretical construct of interest and its relation to the items. If no measurement invariance is established, it will be impossible to say whether the differences observed between the groups (or the absence of such) result from unequal functioning of the measure across the samples or from objectively existing differences in the level of ability or trait [Borsboom 2006; Schmitt, Kuljanin 2008]. Achieving comparability is especially challenging in ICS, as specific cultural and linguistic features get in the way.

There is a direct relationship between measurement validity and measurement invariance in ICS. For instance, a popular approach to validity defines this property of assessment as a sum of evidence supporting the interpretation of test scores [American Educational Research Association, American Psychological Association, National Council on Measurement in Education 2014]. Therefore, since measurement instruments are used for comparative analysis, comparability of test results should be verified. In addition, it has been shown that low measurement invariance may lead to unsatisfactory psychometric properties of tests in cross-national assessments [Church et al. 2011].

Approaches to Achieving Cross-National Comparability

One of the most widespread approaches to establishing measurement invariance across countries was proposed by European researchers Fons van de Vijver and Norbert K. Tanzer, who identified three levels of equivalence [van de Vijver, Tanzer 2004]: construct, method, and item. Construct equivalence implies that the construct structure is the same across all the cultural groups involved. Method equivalence means equivalence of data collection procedures, samples, etc. across the countries. Item equivalence is obtained when tests in all the participating countries function equally at all levels of ability, i. e. there is no Differential Item Functioning (DIF) bias. Below, we will dwell on all the three levels of comparability, which correspond to three types of bias.

If no construct equivalence is not achieved, the construct will be interpreted in conceptually different ways by respondents in different countries. Construct comparability is established by obtaining theoretical and empirical evidence of the construct's structural similarity across all the cultural groups involved. At the preliminary stage, expert analysis of the construct components is a critical procedure, in which relevance of each component to every group of prospective

testees is assessed [Carmines, Zeller 1979; Wynd, Schmidt, Schaefer 2003]. A major challenge associated with this procedure is the choice of experts, who must be well-informed of how the construct interferes with country-specific cultural features and how it manifests itself within each of the national samples as a result of such interferences. The stage of item development begins as soon as experts have identified the construct aspects that are equally relevant across the samples, and the corresponding behavioral indicators.

Post-hoc procedures designed to achieve construct comparability usually represent analysis of how items (or other behavioral indicators) are grouped, i. e. they serve to detect structural differences in the latent variables used for differentiating among the respondents. If measure dimensionality differs across the groups—for instance, if supposedly a single latent variable breaks into two or more variables in only one group—the results obtained in different countries will be incomparable.

The most common method biases include differences in environmental administration conditions, incomparability of samples, ambiguous instructions for testees, differential familiarity with response procedures, differential response styles, etc. [van de Vijver, Tanzer 2004]. Such sources of non-equivalence may become critical for comparability when they are not controlled for [Davidov et al. 2014].

To achieve method comparability, all the measurement procedures should be completely standardized. For example, a large section of PISA technical reports is devoted to description of all the testing procedure requirements [OECD2015].

Post-hoc statistical analysis of method comparability remained unstudied for a long time of the history of ICSs, as it requires collection of data on the *process* of testing, not only the *results*—which were traditionally the focus of psychometric studies throughout the greater part of the 20th century. However, advancements in computer-based testing technology made it possible to collect data on respondents' behavior while test administrating, which soon gave rise to publications analyzing the process and strategies of task performance. (Such data is often referred to as collateral information in scientific literature [Mislevy 1988].) The article by Wim J. van der Linden, who uses modeling of response time on test items [van der Linden 2007], is one of the corner stones that gave rise to this movement in psychometrics. Later studies analyzed not only response times but also researchers' perceptions of the sequence of choices made by respondents [Jeon, de Boeck 2016] and changes in their cognitive strategies when answering different items (e. g. [Tijmstra, Bolsinova, Jeon 2018]). On the whole, this area of psychometric research is one of the most thriving today.

The article by Louis Roussos and William Stout [Roussos, Stout 1996] is a major work on measurement item comparability. Item bias normally implies that certain items contain additional latent dimen-

sions—secondary constructs which are also measured by the task and differ across the national groups of respondents. Possible sources of item bias usually include poor item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, and influence of cultural specifics such as nuisance factors or connotations associated with the item wording [van de Vijver, Tanzer 2004].

To avoid item bias, test developers usually resort to various formalized translation procedures, such as forward and backward translation, where the original item and its detailed description are first translated into a foreign language and then back, by two different translators. Originally, the procedure was believed to enable test developers to capture subtle shifts in the meaning of items [e.g., Hambleton 1993]. However, this method is also one of the most criticized in scientific literature [e.g. Brislin 1970], since it may result in numerous iterations without any significant improvement in the quality of adapted versions. Alternative procedures include:

- Double translation and reconciliation (two or more translators independently translate the source version, including the description; then, a domain expert reconciles these translations into a single national version [OECD2017; 2016]);
- Translation in groups (a group of translators meets face to face and translates every item together, one by one [Hambleton, Kanjee 1995]);
- Translation by bilinguals, who are not just qualified translators but also native speakers of two or more languages, living in bilingual environments since early childhood and thus having a strong "feel" for the language [Annette, Dannenburg, Janet 1994]; and
- Numerous combinations of the above techniques (e. g. [Lenz et al. 2017]).

Differential Item Functioning (DIF) analysis [Holland, Wainer 1993] is one of the most commonly applied techniques of post-hoc statistical analysis in measurement invariance evaluation. DIF analysis is used to find out whether items demonstrate comparable psychometric characteristics across the groups while controlling for the level of target ability or trait.

Methodological literature on achieving item comparability emphasizes the importance of interpreting the statistics after completing the phase of statistical analysis (e. g. [Wang, Shih, Sun 2012]). In particular, it is shown that if a test item demonstrates incomparable psychometric characteristics but domain and cross-cultural experts are unable to provide any contensive explanation for the differences, the item should not be regarded as biased.

In case some of the test items exhibit DIF, special procedures are required to neutralize that effect—such as scale purification, where

DIF items are removed from the tests [Magis, Facon 2013]. However, this may result in a content imbalance, which poses risks for content validity of inferences drawn from the comparison. In addition, item purification may increase measurement error, thus reducing test reliability. Item splitting is an alternative procedure that is only possible within the framework of item response theory [Brodersen et al. 2007]: an item functioning unequally in different groups is treated as a set of group-specific items, which may feature parameters differing across the groups. This approach allows balancing content validity of a test at the national scale and avoid increasing measurement error, while maintaining the psychometric characteristics at an acceptable level.

Another popular approach to achieving cross-national comparability in educational assessment was proposed by Kadriye Ercikan and Juliette Lyons-Thomas [Ercikan, Lyons-Thomas 2013], who identified several categories of potential differences that affect comparability of test scores and validity of inferences drawn from comparison:

1) Differences in the sample;
2) Differences in the construct (non-equivalence of psychological reality behind the construct assessed, which stems from cultural differences);
3) Differences in the measurement instrument (DIF in the first place, but also linguistic differences and associated differences in information presentation);
4) Differences in the instrument administration procedures;
5) Differences in the item response procedures (first of all, item processing strategies).

Obviously, the two theoretical frameworks described above have a lot of parallel features to them and consider similar sources of bias. Besides, both frameworks implicitly suggest that elimination of such sources of bias (i. e., reasons for incomparability) automatically leads to achieving cross-national comparability of test results. The approach proposed by Ercikan and Lyons-Thomas appears to be more convenient for adapting the already existing measures to new national samples, as it allows eliminating the major sources of bias at every stage of instrument design and measurement result analysis. Meanwhile, it would be reasonable to use the framework offered by van de Vijver and Tanzer when developing tests from scratch specifically for cross-national assessment, as it integrates various sources of bias and examines them at all levels of instrument development.

Specific Aspects of Achieving Cross-National Comparability in Higher Education

Modern studies describe challenges in achieving cross-national comparability of test results regardless of the stage of formal education at which CIS are performed [Kankaraš, Moors 2014]. In particular, methodological challenges include differences in cultural and eco-

nomic environments in which education systems are compared [Bray, Thomas 1995] and in the way those systems are organized [Schneider 2009].

However, higher education has some distinctive features affecting the procedures of measuring student achievement. Assessment in higher education is different from that at the secondary education, where advancement in ICS methodology is promoted by a number of large-scale international comparative studies, such as PISA and TIMSS [OECD2017; Martin, Mullis, Hooper 2016].

Assessment of Learning Outcomes in Higher Education (AHELO) was one of the first projects designed to compare higher education systems in different countries [Tremblay 2013] and the one that revealed the specific nature of cross-national assessment in higher education. Criticism faced by the project formed the basis for nearly all methodological developments in international comparative higher education research.

ICSs in higher education differ from those at other educational stages due to the following specific features:

- Great curriculum variance within countries (even within majors in the same country) and wide curriculum differences across the countries, as compared to largely standardized curricula in secondary education [Zlatkin-Troitschanskaia et al. 2017]. For this reason, it is hard to find suitable source material to ensure that achievement tests are not biased against any group. Incompliance of measurement instruments to this requirement was one of the main points of criticism against AHELO [Altbach 2015];
- High selectivity in higher education, which results in longitudinal studies being preferred over cross-sectional ones. Criticism of AHELO was largely founded on the lack of attention for longitudinal changes in the indicators. Not so much does the lack of a dynamic perspective make it difficult to achieve measurement invariance as it complicates ICS design and, consequently, the evaluation of comparability and the choice of source material:
  - If higher education is more selective in one country and less so in another, a cross-national comparison will be challenged by biased estimators of population parameters, as part of the of the differences observed will be explained by highly selective admission. In addition to cross-sectional comparison, such studies require measuring the institution's contribution to student success, i. e. a longitudinal design [Jamelske 2009];
  - Even within national samples, HEIs may differ in their selectivity. Top universities select the most talented candidates, which makes it difficult to measure the institution's contribution to student progress. However, estimation of this factor is extremely important for assessments in higher education, as the cohort

participation rate is far not as high as in secondary education [Jamelske 2009];
  - Students failing to meet educational standards risk being expelled. It was shown on a dataset covering 18 OECD countries that on average 31% of students did not complete the tertiary studies for which they enrolled. Besides, the indicator varied essentially from 10% in Japan to 54% in the United States. Therefore, test scores should be adjusted for student retention rate to avoid bias when measuring the institution's contribution [OECD2010];
- Student achievement-centered approach as a requisite for an educational assessment to be relevant to the education system. For instance, comparison of a newly-developed education program to the one that has been in place for a few years does not stand up to scrutiny. Fine-tuning of education programs may take decades even in signatory countries of the Bologna Process [Rauhvargers 2011]. A separate challenge consists in the evaluation of higher-order thinking skills [Zlatkin-Troitschanskaia, Pant, Greiff 2019]. Some countries and universities focus on fostering higher-order thinking, while others focus on domain-specific knowledge; this difference adds up to the difficulty of achieving measurement invariance;
- A high risk of test data misuse. Awareness of the potential impact of assessment on institutional autonomy and academic freedom may result in deliberate bias at various levels of test administration. This problem is emphasized in the AHELO project documentation. AHELO-like assessments are not envisaged as ranking tools, yet there have been documented attempts to misuse their results [Tremblay 2013]. In particular, HEIs may try to inflate their performance level to raise their rankings. Therefore, misuse of test data may lead to incorrect, unsubstantiated conclusions and bias in data collection [Ibid.];
- Students' motivation for participation as a prerequisite for reliable test results. Unless students are motivated to do their best, their performance cannot be used as an indicator of higher education quality. Developers of ICS in education try to minimize the risk of wrongful conclusions affecting the education system (e. g. by anonymizing respondent data, avoiding rankings, etc.). This may produce "low-stakes" testing situations where respondents are not motivated to give their best effort. Under conditions like that, students' answers cannot be expected to actually reflect the quality of education at their HEI, so additional motivational tools have to be used [Banta, Pike 2012]. Not only does students' motivation influence measurable learning outcomes but it can also vary across countries, which complicates achievement of cross-national comparability of test scores [Goldhammer et al. 2016].

To minimize the risk of cross-national incomparability in higher education assessments, allowance should be made for all of the factors listed above—which are by far less powerful at other educational stages. Therefore, ICS in higher education are essentially distinct from those in other subdivisions of formal learning due to the specific features of the higher education system, which affects research methodology.

Methods of Achieving Cross-National Comparability in the SUPER-test Project

The Study of Undergraduate PERformance (SUPER-test) is a project designed to assess the quality of computer science and electrical engineering skills in higher education across national representative samples from Russia, India, China, and the United States [Loyalka et al. 2019] and to identify institutional and individual factors that influence student achievement in computer science majors. Data was collected in two stages from two cohorts, allowing to measure students' individual skills, see how they changed in two years, and evaluate the HEI's contribution to student progress (baseline assessments in the 1st and 3rd years, and outcome assessments in the 2nd and 4th years).

The project toolset includes a combination of techniques to measure students' competencies in general and specialized disciplines and their higher-order thinking skills (relational reasoning, critical thinking, and creativity) and a series of questionnaires for students, faculty, and administrators to collect a large amount of contextual information. The survey was computer-assisted, which not only optimized the data collection procedure but also allowed obtaining data on respondent behavior in the survey.

SUPER-test instrument development procedures used van de Vijver and Tanzer's approach [van de Vijver, Tanzer 2004] to achieve maximum possible comparability across the countries. The three levels of equivalence assessment proposed by van de Vijver and Tanzer do not coincide with stages of test development. Moreover, a design process based on that approach requires integration of all the three levels at each stage of design. As the SUPER-test project was being developed, some stages of instrument design targeted more than one level of comparability. Sample comparability is not an issue in this case as the target audience is narrowly defined by the project objectives.

The first step involved analysis of test content and construct validity by national education experts, who evaluated content domains and subdomains in education as well as items measuring skills in specific disciplines. The purpose of this step was to ensure construct comparability across the countries. Experts were recruited from a number of highly-selective and regular HEIs in China, Russia, India, and the United States. National experts in tertiary computer science and electrical engineering education were invited to select elements of content to be measured. They singled out the content domains covered in every country in conformity to national curricula. Next, the selected

elements of content were translated into the national languages of the participating countries by a team of qualified translators and domain experts who were native speakers of the target language. After that, the national education experts ranked all the listed areas of potential test content in order to identify the most important topics for graduates' future career success. The resulting rankings were processed using the multi-facet Rasch model so as to determine the most relevant areas and avoid researcher effects and expert bias (e. g. [Zhu, Ennis, Chen 1998]).

As soon as the most important areas of professional competence had been objectively established, it was time for item selection. At that stage, item comparability had to be achieved. In a joint effort of all the experts, an extensive pool of items was generated, which included every single item that any of the experts felt it necessary to submit. Next, that pool was sifted step by step against a series of criteria and assessment procedures. First of all, expert evaluation was applied—same as with the elements of content. The item evaluation criteria included expected item difficulty, the amount of time that an average student would spend on the item, cognitive load required to answer the item correctly, etc. Such evaluation allowed to select items measuring the most significant, cross-nationally relevant elements of content for the pilot study.

The pilot study was performed to establish cross-national comparability of test scores at all levels at once. First, a series of think-aloud interviews and cognitive labs was carried out in the participating countries to find out how respondents from the target audience perceived and processed information contained in the test items. Feedback from every country was documented and translated into foreign languages, allowing to identify the most ambiguous and confusing items. In particular, nationally conditioned difficulties with task understanding were considered at that point. Then, another series of brief pre-tryouts were carried out, followed by focus groups in which the respondents were asked to discuss the test material and their perception of the items. That stage was designed to analyze not so much the test content as the methods of content presentation and organization as well as respondents' recommendations on improving the testing procedure. That portion of tryouts served to establish method comparability by analyzing item response strategies, measuring respondent familiarity with the particular types of tasks, and finding methods of testing standardization that would be acceptable for respondents in all the participating countries.

That done, the test administration procedures were fully standardized and agreed with representatives of all the participating countries in order to allow for objective evaluation of item characteristics.

The next phase of test development was that of large pilot studies, in which psychometric characteristics of the items were assessed to control for construct comparability (using psychometric evaluation of

the number of latent traits measured by the instrument) and item comparability (DIF analysis). From stage to stage, the pool of items reduced as the most troubling items were sorted out, such as those with the most ambiguous wording (including those displaying high translation ambiguity and terminology variation) and the ones that caused difficulty making sense of the graphic representation. In addition, the actual tryouts allowed elaborating the instructions given to testees and test administrators, which contributed as well to method comparability and student motivation [Liu, Rios, Borden 2015].

After that, nationally representative samples were drawn from all the participating countries. The sampling procedure controlled for respondent clustering to reduce the costs of research. Random sampling was used to provide method comparability of test scores.

Data collection was followed by post-hoc analysis to ensure cross-national measurement invariance. The statistical procedures used in SUPER-test lie within the framework of item response theory, so they are contingent on a particular measurement instrument. Cross-national construct comparability was verified using assessment of local item independence for unidimensional instruments [Kardanova et al. 2016] and bifactor structural modeling [Wang, Wilson 2005] for composite measures [Dumas, Alexander 2016]. Cross-national method comparability was evaluated using generalized item response tree models [Jeon, de Boeck 2016] controlling for response times [Molenaar, Tuerlinckx, van der Maas 2015]. Item comparability was tested using the most well-researched and widely applied methods of DIF analysis [Rogers, Swaminathan 1993]. Describing those procedures is beyond the scope of this paper as they represent implementation of some isolated statistical methods.

The methodology described above allowed developing measurement instruments that made it possible to compare engineering undergraduates' competencies across countries. Besides, subsequent statistical analyses used in the project assessed cross-national comparability of the data collected.

Conclusion

Methodology of international comparative educational research is largely driven by studies measuring performance of school students, such as PISA and TIMSS, which laid the foundation of ICS administration and shaped the traditional understanding of ICS design and goals. The recent years have seen a growing need for similar research methods in higher education. However, attempts to apply ICS methodology to higher education have shown little success so far.

Experience with the AHELO project prompted the development of other higher education assessment initiatives (e. g. [Zlatkin-Troitschanskaia et al. 2017; Shavelson, Zlatkin-Troitschanskaia, Mariño 2018; Aloisi, Callaghan 2018]), but it also demonstrated that using conventional approaches in measurement instrument design

was a bad strategy for ICSs in higher education. Given the great variety of education programs, it is extremely difficult to obtain interpretable test results even from different universities within a country, let alone cross-national assessment scales. Methodological challenges specific to higher education complicate implementation of such projects in higher education.

Subsequent projects, SUPER-test in particular, provide convincing evidence that ICSs in higher education are not impossible. However, essential design modifications are necessary, first of all to ensure cross-national comparability of test scores. Such modifications should be based on one of the approaches to cross-national comparability that systematize the sources of bias and provide a coherent theoretical framework for understanding them and minimizing their impact. The SUPER-test project uses the approach proposed by van de Vijver and Tanzer [van de Vijver, Tanzer 2004], which is optimal for designing a measure from scratch, naturally making developers control for all the three levels of comparability: construct comparability (equivalence of construct structure and meaning), method comparability (equivalence of data collection procedures), and item comparability (psychological meaning of each isolated indicator). Using this approach resulted in producing an ICS methodology that is inherently designed to develop instruments for cross-national assessment. The methodology of instrument development described in this article is highly universal and can be used in other ICSs of educational achievements.

References

Aloisi C., Callaghan A. (2018) Threats to the Validity of the Collegiate Learning Assessment (CLA+) as a Measure of Critical Thinking Skills and Implications for Learning Gain. *Higher Education Pedagogies*, vol. 3, no 1, pp. 57–82.

Altbach P. G. (2015) AHELO: The Myth of Measurement and Comparability. *International Higher Education*, no 82, pp. 2–3.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*. Washington, DC.

Annette D. G., Dannenburg L., Janet G. (1994) Forward and Backward Word Translation by Bilinguals. *Journal of Memory and Language*, vol. 33, no 5, pp. 600–629.

Banta T., Pike G. (2012) Making the Case Against–One More Time. *Occasional Paper of National Institute for Learning Outcomes Assessment*, vol. 15, pp. 24–30.

Borsboom D. (2006) When Does Measurement Invariance Matter? *Medical Care*, vol. 44, no 11, pp. S176–S181.

Bray M., Thomas R. M. (1995) Levels of Comparison in Educational Studies: Different Insights from Different Literatures and the Value of Multilevel Analyses. *Harvard Educational Review*, vol. 65, no 3, pp. 472–490.

Brislin R. W. (1970) Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, vol. 1, no 3, pp. 185–216.

Brodersen J., Meads D., Kreiner S., Thorsen H., Doward L., McKenna S. (2007) Methodological Aspects of Differential Item Functioning in the Rasch Model. *Journal of Medical Economics*, vol. 10, no 3, pp. 309–324.

Cantwell B., Marginson S., Smolentseva A. (eds) (2018) *High Participation Systems of Higher Education*. Oxford: Oxford University.

Carmines E. G., Zeller R. A. (1979) *Reliability and Validity Assessment*. Vol. 17. Thousand Oaks, CA: Sage.

Church A. T., Alvarez J. M., Mai N. T., French B. F., Katigbak M. S., Ortiz F. A. (2011) Are Cross-Cultural Comparisons of Personality Profiles Meaningful? Differential Item and Facet Functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology*, vol. 101, no 5, pp. 1068–1089.

Davidov E., Meuleman B., Cieciuch J., Schmidt P., Billiet J. (2014) Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, vol. 40, pp. 55–75.

Dumas D., Alexander P. A. (2016) Calibration of the Test of Relational Reasoning. *Psychological Assessment*, vol. 28, no 10, pp. 1303–1319.

Ercikan K., Lyons-Thomas J. (2013) Adapting Tests for Use in Other Languages and Cultures. APA Handbook of Testing and Assessment in Psychology (ed. K. F. Geisinger), Washington: American Psychological Association, vol. 3, pp. 545–569.

Goldhammer F., Martens T., Christoph G., Lüdtke, O. (2016) *Test-Taking Engagement in PIAAC.* Paris: OECD.

Hambleton R. K. (1993) Translating Achievement Tests for Use in Cross-National Studies. *European Journal of Psychological Assessment*, vol. 9, no 1, pp. 57–68.

Hambleton R. K., Kanjee A. (1995) Increasing the Validity of Cross-Cultural Assessments: Use of Improved Methods for Test Adaptations. *European Journal of Psychological Assessment*, vol. 11, no 3, pp. 147–157.

Hanushek E. A., Woessmann L. (2008) The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature*, vol. 46, no 3, pp. 607–668.

Holland P. W., Wainer H. (1993) *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jamelske E. (2009) Measuring the Impact of a University First-Year Experience Program on Student GPA and Retention. *Higher Education*, vol. 57, no 3, pp. 373–391.

Jeon M., De Boeck P. (2016) A Generalized Item Response Tree Model for Psychological Assessments. *Behavior Research Methods*, vol. 48, no 3, pp. 1070–1085.

Kankaraš M., Moors G. (2014) Analysis of Cross-Cultural Comparability of PISA 2009 Scores. *Journal of Cross-Cultural Psychology*, vol. 45, no 3, pp. 381–399.

Kardanova E., Loyalka P., Chirikov I. et al. (2016) Developing Instruments to Assess and Compare the Quality of Engineering Education: The Case of China and Russia. *Assessment & Evaluation in Higher Education,* vol. 41, no 5, pp. 770–786.

Kuzminov Ya., Sorokin P., Froumin I. (2019) Obshchie i spetsialnye navyki kak komponenty chelovecheskogo kapitala: novye vyzovy dlya teorii i praktiki obrazovaniya [Generic and Specific Skills as Components of Human Capital: New Challenges for Education Theory and Practice]. *Foresight and STI Governance*, no 13 (S2), pp. 19–41.

Lenz S. A., Soler I. G., Dell'Aquilla J., Uribe P. M. (2017) Translation and Cross-Cultural Adaptation of Assessments for Use in Counseling Research. *Measurement and Evaluation in Counseling and Development*, vol. 50, no 4, pp. 224–231.

Liu O. L., Rios J. A., Borden V. (2015) The Effects of Motivational Instruction on College Students' Performance on Low-Stakes Assessment. *Educational Assessment*, vol. 20, no 2, pp. 79–94.

Loyalka P., Liu O. L., Li G. et al. (2019) Computer Science Skills Across China, India, Russia, and the United States. *Proceedings of the National Academy of Sciences*, vol. 116, no 14, pp. 6732–6736.

Magis D., Facon B. (2013) Item Purification Does Not Always Improve DIF Detection: A Counterexample with Angoff's Delta Plot. *Educational and Psychological Measurement*, vol. 73, no 2, pp. 293–311.

Marginson S. (2019) Limitations of Human Capital Theory. *Studies in Higher Education*, vol. 44, no 2, pp. 287–301.

Martin M. O., Mullis I. V.S., Hooper M. (eds) (2016) *Methods and Procedures in TIMSS2015*. Available at: http://timssandpirls.bc.edu/publications/timss/2015-methods.html (accessed 10 April 2020).

Meredith W. (1993) Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika*, vol. 58, no 4, pp. 525–543.

Meyer J. W., Ramirez F. O., Soysal Y. N. (1992) World Expansion of Mass Education, 1870–1980. *Sociology of Education*, vol. 65, no 2, pp. 128–149.

Mislevy R. J. (1988) Exploiting Collateral Information in the Estimation of Item Parameters. *ETS Research Report Series*, vol. 2, pp. 1–31.

Molenaar D., Tuerlinckx F., van der Maas H. L. (2015) A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times. *Multivariate Behavioral Research*, vol. 50, no 1, pp. 56–74.

OECD (2010) Education at a Glance 2010: OECD Indicators. . Available at: www.oecd. org/edu/eag2010 (accessed 10 April 2020).

OECD (2015) *PISA 2018 Technical Standards*. Available at: https://www.oecd.org/pisa/pisaproducts/PISA-2018-Technical-Standards.pdf (accessed 10 April 2020).

OECD (2016) *PISA 2018 Translation and Adaptation Guidelines*. Available at: https://www.oecd.org/pisa/pisaproducts/PISA-2018-translation-and-adaptation-guidelines.pdf (accessed 10 April 2020).

OECD (2017) PISA 2015 Technical Report. Available at: https://www.oecd.org/pisa/data/2015-technical-report/ (accessed 10 April 2020).

Rauhvargers A. (2011) *Global University Rankings and Their Impact: EUA Report on Rankings 2011.* Brussels: European University Association.

Rogers H. J., Swaminathan H. (1993) A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, vol. 17, no 2, pp. 105–116.

Roussos L., Stout W. (1996) A Multidimensionality-Based DIF Analysis Paradigm. *Applied Psychological Measurement*, vol. 20, no 4, pp. 355–371.

Schmitt N., Kuljanin G. (2008) Measurement Invariance: Review of Practice and Implications. *Human Resource Management Review*, vol. 18, no 4, pp. 210–222.

Schneider S. L. (2009) *Confusing Credentials: The Cross-Nationally Comparable Measurement of Educational Attainment* (PhD Thesis), Oxford: Oxford University.

Schultz T. W. (1961) Investment in Human Capital. *The American Economic Review*, vol. 51, no 1, pp. 1–17.

Shavelson R. J., Zlatkin-Troitschanskaia O., Mariño J. P. (2018) International Performance Assessment of Learning in Higher Education (iPAL): Research and Development. *Assessment of Learning Outcomes in Higher Education* (eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, C. Kuhn), Springer, pp. 193–214.

Tijmstra J., Bolsinova M. A., Jeon M. (2018) Generalized Mixture IRT Models with Different Item-Response Structures: A Case Study Using Likert-Scale Data. *Behavior Research Methods*, vol. 50, no 6, pp. 2325–2344.

Tremblay K. (2013) OECD Assessment of Higher Education Learning Outcomes (AHELO): Rationale, Challenges and Initial Insights from the Feasibility Study. *Modeling and Measuring Competencies in Higher Education* (eds S. Blömeke, O. Zlatkin-Troitschanskaia, Ch. Kuhn, Ju. Fege), Rotterdam: Brill Sense, pp. 113–126.

Van de Vijver F., Tanzer N. K. (2004) Bias and Equivalence in Cross-Cultural Assessment: An Overview. *European Review of Applied Psychology*, vol. 54, no 2, pp. 119–135.

Van der Linden W. J. (2007) A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, vol. 72, no 3, pp. 287–308.

Wang W. C., Wilson M. (2005) The Rasch Testlet Model. *Applied Psychological Measurement*, vol. 29, no 2, pp. 126–149.

Wang W. C., Shih C. L., Sun G. W. (2012) The DIF-free-then-DIF Strategy for the Assessment of Differential Item Functioning. *Educational and Psychological Measurement*, vol. 72, no 4, pp. 687–708.

Wynd C. A., Schmidt B., Schaefer M. A. (2003) Two Quantitative Approaches for Estimating Content Validity. *Western Journal of Nursing Research*, vol. 25, no 5, pp. 508–518.

Zhu W., Ennis C. D., Chen A. (1998) Many-Faceted Rasch Modeling Expert Judgment in Test Development. *Measurement in Physical Education and Exercise Science*, vol. 2, no 1, pp. 21–39.

Zlatkin-Troitschanskaia O., Pant H. A., Greiff S. (2019) Assessing Generic and Domain-Specific Academic Competencies in Higher Education. *Zeitschrift für Pädagogische Psychologie*, vol. 33, no 2, pp. 91–93. ·

Zlatkin-Troitschanskaia O., Pant H. A., Lautenbach C., Molerov D., Toepper M., Brückner S. (2017) *Modeling and Measuring Competencies in Higher Education. Approaches to Challenges in Higher Education Policy and Practice*. Wiesbaden: Springer.