NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS
INSTITUTE OF EDUCATION

As a manuscript

Ivanova Alina Evgenievna

Development and validation of the Russian version of an instrument for an international comparative study of what students know and can do at school entrances and their progress during the first year: problems of adaptation and localization

Summary of the thesis
for the purpose of obtaining the academic degree
Doctor of Philosophy in Education

Academic supervisor:
Elena Kardanova, PhD

Moscow 2021

List of publications	Main publications Ivanova A. (2018). Problema sopostavimosti rezul'tatov v mezhdunarodnykh sravnitel'nykh issledovaniyakh obrazovatel'nykh dostizheniy (The problem of ILSA' results comparisons). Otechestvennaya i zarubezhnaya pedagogika, 2(1), 68-81 (in Rus.). Ivanova A., Kardanova E., Merrell C., Tymss P., Hawker D. (2018). Checking the possibility of equating a mathematics assessment between Russia, Scotland and England for children starting school. Assessment in Education: Principles, Policy and Practice, 2(25), 141-159. Ivanova A., Kardanova-Biryukova K. (2019). Sozdaniye russkoyazychnoy versii mezhdunarodnogo instrumenta otsenivaniya rannikh navykov chteniya (Constructing a Russian-Language Version of an International Early Reading Assessment Tool). Educational studies Moscow, 4, 93-115 (in Rus.). Ivanova A., Kardanova E. (2020). Izucheniye vozmozhnosti provedeniya mezhstranovogo sravnitel'nogo issledovaniya navyka chteniya uchashchikhsya na vkhode v shkolu v Rossii i Velikobritanii (Checking the Possibility of an International Comparative Study of Reading Literacy Assessment for Children Starting School). Educational studies Moscow, 4, 8-36 (in Rus.). Additional publications: Ivanova A. (2018). Validizatsiya oprosnika povedencheskikh kharakteristik mladshikh shkol'nikov. Journal of Modern Foreign Psychology, 7(3), 86-95 (in Rus.). Kardanova E., Ivanova A., Sergomanov P., Kanonir T., Antipkina I., Kayky D. (2018). Obobshchennyye tipy razvitiya pervoklassnikov na vkhode v shkolu. Po materialam issledovaniya iPIPS (Patterns of First-Graders' Development at the Start of Schooling: Cluster Approach). Educational studies Moscow, 1, 8-37 (in Rus.). Kuzmina, Y., Ivanova, A., & Kaiky, D. (2019). The effect of phonological processing on mathematics performance in elementary school varies for boys and girls: Fixed-effects longitudinal analysis. British Educational Research Journal, 45(3), 640-661. Vasilyeva, M., Dearing, E., Ivanova, A., Shen, C., & Kardanova, E. (2018). Testing the family investment model in Russia: Estimating indirect effects of SES and parental

	beliefs on the literacy skills of first-graders. Early Childhood Research Quarterly, 42, 11-20. Kuzmina, Y., & Ivanova, A. (2018). The effects of academic class composition on academic progress in elementary school for students with different levels of initial academic abilities. Learning and Individual Differences, 64, 43-53.
List of conferences	AERA 2018 Annual Meeting «The Dreams, Possibilities, and Necessity of Public Education» (USA). Presentation: «The ability to read numbers: A universal measure?» Rasch Measurement Conference 2018. Australia. Presentation «Setting benchmarks of children reading development for potential comparisons across different countries and cultures». AERA 2016 Annual Meeting «Public Scholarship to Educate Diverse Democracies» (USA). Presentation: «Equating iPIPS measures across different countries and cultures». ITC 2016 Conference (Canada). Presentation: Development and Validation of the Russian Version of the iPIPS Study.

Table of content

Introduction

Relevance of the topic

Large-scale national and international studies are playing an increasingly large role in child development research. Modern international large-scale assessments (ILSA) allow researchers, among other things, to verify, refine and improve existing development theories (Shuttleworth-Edwards et al., 2004; Peña, 2007). ILSAs are an important source of data on predictors of child learning success in different countries, settings, and social and cultural contexts (Ainley, & Ainley, 2019; Carnoy et al., 2016; Caro, & Cortés, 2012).

Interest in ILSAs worldwide has been confirmed by the rapid growth of their number since the beginning of the 2000s. For example, the number of PISA (Program for International Student Assessment) participants increased from 43 in 2000 to 80 in 2018 (Liu, & Steiner-Khamsi, 2020). Researchers note that an increasing number of governments are trying to follow the logic of ILSAs in their domestic education policies in an effort to achieve reliable, predictable, quantifiable results (Espeland, 2015; Liu, & Steiner-Khamsi, 2020).

It is also worth noting the particular interest of researchers and politicians in the results of international studies in the fields of preschool and early school education. This interest is associated with the increasing role of literacy in modern society as a whole (for example, according to the UN Sustainable Development Goals adopted by 193 countries, the world community needs to achieve universal access to high-quality preschool and primary school education (UN, 2015)). The overall interest in the topic is also associated with the fact that the first years of schooling are critically important for later development. Additionally, there is a need for high-quality data for the reasonable use of educational resources. Finally, researchers and policymakers are interested in the data produced by ILSAs to make evidence-based decisions and to take into account the experience and best practices of other countries (Suggate, 2009). Despite the fact that each country develops and implements its own educational goals and programs, other external international guidelines and information about other opportunities and prospects for the development of primary school-age children are welcomed (Buzhardt et al., 2019). An example of an international comparative study focusing on the beginning of schooling is the international project iPIPS (international Performance Indicators in Primary Schools), which involves the baseline assessment of children at the time of entrance into school and an assessment of their progress during the first year of schooling. The iPIPS tool can provide data for a wide range of secondary studies on what children know and can do when they start schooling.

However, as with any other international comparative study, the comparability of the data obtained using this instrument must be thought out in advance and subsequently proven if the purpose of the study is to comparatively interpret the students' scores and generalize the results to different countries and cultures. The development and implementation of an international comparative study is always an extremely difficult task. An assessment instrument developed in one culture to evaluate a particular construct based on certain values and knowledge will not always be equivalent to measuring the same construct in another culture. This is a common methodological challenge of all international studies in the field of education, and therefore special studies are needed to prove the equivalence of measurements obtained with the help of the instruments used.

There are a number of examples showing that there are problems with the comparability of results for individual countries and constructs even for the largest renowned international studies (Ercikan, Roth, Asil, 2015; Oliveri, & von Davier, 2011). However, in the case of an international study focused on primary school students, at the start of education, the researchers have to face additional difficulties associated with the developmental level of children, their age, and all the limitations behind these factors.

The implementation of the international comparative study iPIPS required solving a number of problems that international studies traditionally face, as well as solving additional problems. The latter were associated with ensuring the validity of the results based on early school age students' assessments in reading and mathematics for the purpose of international comparison in conditions of different student ages and partially differing assessment instruments created in different countries.

This work describes the process of developing the Russian-language version of the international iPIPS instrument and searches for validity arguments. The work shows the problems that have arisen in adapting the instrument and their solutions. In addition, this work discusses the possibilities of an exploratory comparative study of first-graders' baseline assessment results and their first-year progress in reading and mathematics, obtained using the original English version and, for the first time for the iPIPS project, a non-English language version of the instrument.

Literature review

International large-scale assessment studies (ILSAs) in education, such as the Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS), or the Programme for International Student Assessment (PISA), produce significant amounts of objective data for researchers and policymakers. Some researchers are convinced that ILSAs often shape the way education is understood and what value it has in participating countries (Sellar, Lingard, 2014). In Russia, which has been taking part in international comparative educational studies conducted by the Organization for Economic Cooperation and Development (OECD) and the International Association for the Evaluation of Educational Achievements (IAE) since 1988, significant attention has been given to the results of ILSAs (Bolotov et al., 2013). In our country, for several decades in educational policy, among researchers and practitioners, data on the educational results of schoolchildren, as well as data obtained from contextual questionnaires of parents, teachers and school principals, have been actively studied and used. Educational researchers claim the ILSA to be a part of the Russian system for assessing the quality of education in the country (Bolotov, 2018; Kovaleva, 2017).

The largest ILSAs, such as PIRLS, TIMSS or PISA, are today the most illustrative examples of effective design and implementation of the study of educational achievements in the languages of different countries and cultures. However, on an annual basis worldwide, other international comparative studies are carried out involving a smaller number of countries, resources and media attention. These include various joint educational projects and initiatives of several countries (Ellefson, Zachariou, Ng, Wang, & Hughes, 2020) or educational organizations that assess the educational achievements of students. Even with internal national monitoring, versions of assessment instruments in languages different from the main state language are quite common (Ercikan, Oliveri, & Sandilands, 2013). Alternatively, countries and institutions need educational exams, which give students the right to choose the language in which to take the exam (Sears, Othman & Mahoney, 2015). These examples suggest the need to compare the educational outcomes of participants using culturally or linguistically different versions of the instruments. However, regardless of the scale of the international comparative study being conducted, they will all face similar methodological problems and challenges and should be guided by similar quality standards.

Numerous scientific papers show that international comparative studies in the field of education are faced with a large number of methodological challenges associated with the development of high-quality measurement instruments and adaptation procedures for participating countries. These challenges are explained by the fact that the linguistic versions of any international assessment instrument that involves making comparisons across different countries and cultures inevitably contain culturally dependent components that are in excess of the construct of the interest (Braun, 2013).

If the methodological problems of ILSAs are not resolved, the likelihood of errors in conclusions from the results of such studies increases significantly. For example, the study's findings for differences in educational test scores among students from different countries may be due to artefacts of measurement rather than actual differences in student abilities (Allalouf, 1999; Sears, Othman & Mahoney, 2015).

Despite the existence of international standards and guidelines containing recommendations for conducting international comparative studies, even strict adherence to all the rules and procedures for adapting instruments to the languages of the participating countries does not guarantee that the students' assessment results will be comparable for all participating countries (Laschke, Blömeke, 2016; Grisay et al., 2009; Stubbe, 2011). Thus, a fundamental problem in international comparative studies or ILSAs is ensuring comparability of assessment results (Rutkowski et al., 2010).

A separate research interest and, at the same time, a methodological challenge are international comparative studies of the educational achievements of children at the start of schooling. In addition to the challenges outlined above, which all international studies face, ILSAs targeting primary school children must take into account the age-specific developmental characteristics of children, significantly limiting existing assessment opportunities. Most children of preschool and early school ages cannot read, have a limited vocabulary, cannot focus on a specific task for a long time, and do not always have sufficient psychological maturity to participate in testing (Castro, Swauger, Harger, 2017; McClelland et al., 2007; Merrell, Tymms, 2016; Weigel, Martin, & Lowman 2007).

If we address international experience, we can single out a number of well-known studies focused indirectly (through a survey of parents or teachers, through observation of the child's behaviour in the learning environment, etc.) on the assessment of the child's development at the time of entrance into school. For example, the Early Development Instrument (EDI), developed in Canada and currently used in several other countries, assesses the physical, social, emotional, communicational and cognitive-linguistic spheres of children's development in the form of a survey of kindergarten teachers a year before school (Janus et al, 2007). Early Childhood Environment Rating Scales (ECERS) and School-Age Care Environment Rating Scales (SACERS) are used in many countries. These assessment instruments involve observation and structured peer review of the child's environment (Harms, Clifford, Cryer 2015; Harms 2013).

It is extremely difficult to implement an international comparative study of children's skills at the time of school entrance based on the direct assessment of children (i.e., working with a child individually and recording how children demonstrate what they know and can do). Nevertheless, attempts are being made to conduct such studies, although the number of participating countries is still small. An example of such a study is, for example, the recently launched OECD's International Early Learning Child Well-being Study, in which three countries - the United Kingdom, the United States and Estonia – have taken part (OECD, 2018). The project is aimed at children aged 5-6 who attend schools or kindergartens. Among other things, the study involves a direct assessment of children's vocabulary and phonemic literacy, as well as basic math skills. However, such an important component of a child's basic skills as early reading was not assessed in this study. In addition, like all the studies mentioned above, this project assumes cross-sectional data collection.

Another example of research focused on the beginning of schooling is the international project iPIPS (international Performance indicators in Primary Schools), which involves the baseline assessment of children at the time of entrance into school and an assessment of their individual progress during the first year of schooling. The instrument includes, among other things, an assessment of children's early reading and math skills. The possibility of measuring the progress of children, which means the potential for studying the dynamics of children's development from an intercountry perspective, is a unique feature of this international study.

The main purpose and objectives

Participation in the iPIPS study requires researchers from the acceding countries to solve a number of methodological problems that are not typical for ILSAs related to the age of the students and the possibility of assessing their progress. Despite the fact that the iPIPS instrument is widely used in a number of countries outside the UK, including Australia, New Zealand, Brazil, Germany, and South Africa (Archer et al., 2010; Bartholo et al., 2019; Howie et al., 2016; Tymms et al., 2014; Vidmar et al., 2017), there is a lack of studies devoted to the problems of quality assurance of international comparisons based on these data.

The main goals of this work are to develop the Russian-language version of an international study for assessing students' skills at the time of school entrance and their progress in reading and mathematics during the first school years, to find evidence of validity for the use of these study results, and to develop a mechanism to ensure the comparability of the results of at least two language versions of the instrument.

Terminology used in the dissertation. Adaptation of an assessment instrument is a necessary step if the instruments will be used in at least two versions that differ significantly from each other in terms of language and/or culture. The adaptation of the instrument involves its transfer to the language of another culture and the minimum necessary change of the instrument, associated with the peculiarities of the language, content and use of terminology, as well as making an informed decision that the final version of the instrument in another language and/or in another culture reflects the measurement of the same construct as in the original version (Hambleton and Patsula, 1999; ITC, 2016). The adapted version will inevitably differ slightly from the original version in terms of its language and culture. It is important to note that these differences are minimal, while the equivalence of measurements is maintained. The equivalence of measurements can be ensured through 1) the equivalence of the construct; 2) the equivalence of the instrument; and 3) procedure equivalence (Ercikan, 2013).

In this work, the concept of instrument localization will also be used. The concepts of adaptation and localization reflect the same process - the development and modification of an assessment instrument for use in a different cultural or linguistic environment. However, they look at the process in terms of different uses and interpretations of the final assessment results.

Localization is a term used in various fields of social sciences that reflects the process of transforming a product in such a way that it takes into account the cultural and linguistic specifics of the target audience (country, region, etc.) where it will be used (Esselink, 2000). It seems reasonable to introduce this concept into the sphere of international comparative studies in cases where even strict adherence to adaptation procedures in accordance with international standards and guidelines will not allow achieving complete equivalence of measurements. Then, localization can be defined as the process of creating a measurement tool in the language of another culture, which is based on the same theoretical model as in the original version, but the cultural characteristics of the country of localization are more fully taken into account. At the same time, for localization, the impossibility of achieving the equivalence of measurements is fixed, which means that it becomes impossible to conduct a direct comparison of the assessment results using different versions of the instrument at the individual level.

The research design implies solving several research tasks

- First, within the framework of the dissertation research, a profound analysis of existing research on the problems of international comparative studies of educational achievements is carried out, and approaches to their solution are discussed.
- Second, the features of adaptation of an international comparative study of children's basic knowledge and skills at school entrances and assessment of their progress in the first year of schooling are analysed. The process of adaptation will be illustrated based on the development of the Russian version of the mathematical part of the iPIPS instrument. A set of validation studies

is being carried out. Additionally, an approach is proposed to ensure the comparability of the Russian-language and original versions of the instrument at the individual level.

- Third, the features of the localization of an international comparative study of children's basic knowledge and skills at school entrances and assessment of their progress in the first year of schooling are analysed. The process of localization will be illustrated based on the development of the Russian version of the reading part of the iPIPS instrument. A set of validation studies is being carried out. Additionally, an approach is proposed to ensure the comparability of the Russian-language and original versions of the instrument at the group level.

Research questions

In a typical situation, when conducting a cross-sectional ILSA focused on students (or adults) who can already read, the comparability of assessments for international comparison is ensured using specially developed procedures for adapting assessment tools to the languages of countries and cultures participating in the ILSA. International organizations operating in the assessment sphere offer a variety of guidelines and recommendations to ensure the quality of adaptation procedures for ILSAs (for example, AERA, APA, NCME; ITC). The researchers also suggest their own solutions to specialized or narrow ILSA issues, such as translation quality control (Sperber, Devellis, Boehlecke, 1994), or measures to ensure the uniformity and consistency of field research administration mechanisms (Jowell, 2007) or methods of empirical analysis of the comparability of the results obtained (Oliveri, 2012).

Nevertheless, as numerous studies show, even these measures do not guarantee strong data comparability (Laschke, Blömeke, 2016; Grisay et al., 2009, Stubbe, 2011). Special studies are needed to prove that in the course of the developmental procedures, the separate national versions of the international study assessment instruments do produce truly comparable results. However, in the case of a study such as iPIPS, additional efforts are required in the processes of adaptation and validation and when ensuring the comparability of the results of the participating countries obtained from samples of students who just started formal schooling.

Considering all of the above, we can assume that this study will answer the following *research questions*:

1. What are the challenges in conducting an international comparative study on student skills in elementary school?
2. How can the validity of the use and interpretation of assessment results obtained using the ILSA be ensured?
3. How can the primary testing data obtained by the ILSA be organized to further use them for scientific research?
   a. How can international comparability of students' assessment results be ensured and validated at the start of school and their progress in mathematics?
   b. How can international comparability be ensured and validated between early-school assessment and reading progress?

Methods

Measures. The iPIPS instrument, originally developed in the UK (Tymms, 1999), is now widely used in various countries, particularly in Australia, New Zealand, Germany, South Africa and several others (Archer et al., 2010; Niklas, & Schneider, 2013; Wildy, & Styles, 2008). The assessment of children is carried out in the format of computer adaptive testing and involves the direct interaction of the child with the tasks with the help of a specially trained assessor (interviewer). The iPIPS assessment is carried out in two stages: when children just start school and when they finish their first year of school, which allows us to assess their baseline level and progress in learning. The instrument assesses a child's development in four areas: vocabulary,

phonemic literacy, early reading and mathematics. In the course of the dissertation, we discuss the problems aroused in adapting the two parts of the iPIPS instrument: mathematics and reading.

Participants. To solve the problems of the empirical part of the study, the iPIPS data – an international comparative study of what children know and can do when starting schooling – were used. The data came from different countries, i.e., Great Britain and Russia. In particular, the study used Russian data collected during the iPIPS approbation study in Veliky Novgorod in 2013 on a representative stratified randomized sample of 310 first-grade students, data from a representative stratified randomized sample of approximately 1489 first-graders in Krasnoyarsk in 2014, and data of 1289 first-graders in Kazan for 2016, as well as data from large-scale iPIPS testing in 2017, collected on a sample of approximately 5000 (also representative stratified randomized) first-graders of the Republic of Tatarstan[1]. The sample of students from the UK is represented by data from large-scale testing in England and Scotland (approximately 16,000 students) in 2012 and 2013.

Theoretical framework

This study builds on advances in the validity theory represented by the Michael Kane model (Kane, 1992, 2006, 2013).

In modern measurement practice in education, the concept of validity is a fundamental concept for the design and use of assessment tools. Over the past 15 years, researchers (in particular, in the United States and Canada) have conceptualized validity as a single concept rather than a specific set of validities as characteristics of tests that previously included, for example, face, content, criterial, and other types of validity. Some of the most authoritative standards in testing today, the unified Standards of the American Association for Educational Research, the American Psychological Association, and the National Council for Educational Measurement (AERA, APA, NCME, 2014), define validity as "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests". The standards provide an indication of what kind of evidence it is – the content, the process of performing the test (including receiving responses from respondents to the test items), the internal structure of the test, relationships with other variables, and the consequences of using the test results.

Michael Kane's model of validity is consistent with the modern vision of the validity concept and further develops it, including methodologically, using an approach based on formal argumentation (Cook et al., 2015). Kane's validation process can be built on the basis of four categories of inferences about the use and interpretation of test results, which include scoring, generalization, extrapolation, and implications.

Raw test results in and of themselves are not of interest to researchers, policymakers, and practitioners, but they have to support some conclusions that are much broader than the goals set out in the test specification. The test results, for example, are used to show that the trainee (person, examinee) has achieved certain educational results in a certain area, has a high probability of successfully completing a certain educational program, or is recognized as ready to use his knowledge and skills in real life. However, none of these conclusions are obvious and require proof (Kane, 2013). Validating a particular interpretation and use of test scores means evaluating the feasibility and acceptability of the conclusions that can be drawn from the test results. Kane's

[1] Note that a large-scale study in the Republic of Tatarstan was carried out on a sample of approximately 5000 people as part of a dissertation project. The research results were published in an article by Ivanova and Kardanova (2020). The presence of the another official language of the Republic in addition to Russian may raise the question of the appropriateness of conducting an iPIPS study in a given region for the purposes of a dissertation project. However, the sample size of this particular region was commensurate with the sample of Scotland. According to a survey of parents of pupils in the sample, approximately 80% of them speak Russian with their children at home (Regional Report, 2017). Additionally, according to one of the studies by the author of the thesis (Ivanova et al. 2016), no statistically significant differences were found in the iPIPS data on a representative sample of first-graders from Tatarstan and the Krasnoyarsk Territory.

argument-based approach provides a framework for assessing the feasibility and acceptability of such conclusions. The key idea of the approach is to build a network of claims and arguments for the intended use and a specific interpretation of the test results.

According to Kane scoring, inferences imply the linking of the observed behaviour of the test person and the observed score. The evidence begins to be collected from the moment of planning the test design and beyond and includes a profound description of the instrument's development process, justification for the choice of the task format, test procedures and formats, scoring procedures, and scaling procedures (Cook et al., 2015). For generalization inferences, we need to know how well the test represents the entire possible general population of test items that measure a given construct and corresponds to the theoretical structure of the construct. Evidence on generalization includes the reliability of measurements, the stability and reproducibility of estimates, the consistency of expert estimates, and evidence of the absence or insignificance of variance irrelevant to the construct.

The extrapolation inferences advance the necessity to determine the relationship of the test results with other variables from the real world. This may include theoretical work to link the test goals and educational goals set out in the educational standards and the test content and the curriculum content. In certain cases, this may include expert review of the content and objectives of the instrument. It also includes evidence linking the test scores with other relevant indicators, including results from the testing with other conceptually relevant test results. Finally, to draw inferences, evidence is needed regarding the long-term impact of assessment on learners and other stakeholders, including evidence about how decisions made on the basis of test instruments led to objective improvements in educational practice. Kane's work notes that the latter category of findings is rarely reported in today's scientific publications about validation due to its high resource intensity (Kane, 2006, 2013).

Thus, within the framework of Kane's concept, the researcher begins by identifying those decisions that can be made based on the results of a particular test and the purposes of its use (for example, to identify risk groups in the student environment and propose an action plan to prevent the onset of risks of learning failure). The interpretation of test results that might support such a decision would include, for example, the statement that high test scores reflect a high probability of success in learning, while low test scores reflect a low probability of success in learning. Guided by this argument about the use/interpretation of test results, it is necessary to collect supporting (or refuting) evidence.

Within the framework of Kane's concept used in this study, the following central claim can be formulated: test results using the iPIPS in reading and math can be used to conduct an international comparative study of what children from different countries know and can do at the time of entrance into school. The results of this study, published in four scientific articles, provide evidence to support this claim.

Methodology

The methodology of our work is rooted in item response theory and the Rasch approach to measurement (Rasch, 1966; Hambleton, 2002; Ercikan, Lyons-Thomas, 2013; Oliveri, von Davier, 2014). Item response theory and the Rasch approach allow planning, developing and evaluating the quality of the measurements. The choice of a mathematical model describing the relationship of the elements of the measured construct, the characteristics of the instrument and the primary results of the sample, psychometric analysis of the quality of items and the instrument as a whole, analysis of the structure of the measured construct, evaluating the error with which it was measured, assessment of the reliability of measurements, creation of scales and final assessments – in modern ILSAs in education, all of these metrics are based on IRT (Kastberg, Roey, Lemanski, Chan, & Murray, 2014; Mullis, & Martin, 2019).

Thus, when using the Rasch approach to measurements, a student's test score can be represented as a mark on a scale reflecting the progress of this student in mastering a certain educational competence (Rasch, 1966). For the test instrument to be built on the basis of this model, the set of items must satisfy the following basic requirements: unidimensionality, variation in the indicators of the item difficulty and the abilities of students, as well as the hierarchical distribution of items across the entire scale, from easy to difficult, in accordance with the theoretical expectations of this instrument. Thus, the principles of Rasch measurement lay the foundation for the conceptual construction of the assessment instruments. As an example, one of the models of the Rasch family is embedded in the constructs measured by PISA (Turner, Adams, 2007).

Additionally, in international studies, the means of item response theory are traditionally used for a posteriori confirmation of the equivalence of measurements. First, the adequacy of the functioning of the items and the instrument as a whole within each country is assessed. Then, the structure (dimension) of the measured construct and its similarity in all countries are assessed. Finally, an analysis of possible distortions in items in different language versions of the toolkit is carried out (Oliveri, 2012).

As applied to our work, the principles of item response theory and the Rasch approach to measurements underlie both the theoretical model of the iPIPS instrument (for example, in accordance with the theoretical principles of the iPIPS model for reading, a set of tasks evaluating reading should be a one-dimensional construct that represents a child's assimilation reading from basic ideas of reading to reading with comprehension) and in the basis of adaptation and localization procedures, including collection of the proofs of equivalence of the measurements carried out by the instrument and the possibility of international comparisons.

The Rasch family of models is used as the basis for constructing an argument about using test results and searching for evidence in favour of this argument. Verification of the quality of adaptation of the iPIPS in mathematics and the possibility of conducting an international comparative study of children's mathematics skills at the time of school entrance and their progress in the first year of schooling included several types of analysis: psychometric analysis of the mathematical scale, analysis of the possibility to link the results of assessment of the participating countries by the method of simultaneous calibration, and analysis of the items' fair functioning in relation to groups of children from different countries. In addition, in the work (in the presented articles), the collection of validity evidence of the results of the Russian-language version of the iPIPS is considered.

The localization procedure for the iPIPS instrument in reading, as well as the analysis of the possibilities of conducting an international comparative study of children's reading skills at school entrances, was also carried out based on the Rasch approach to measurements. In addition, the work with the reading part of the iPIPS included an analysis of the instrument in terms of content and language and justification of the need for localization and refusal to carry out adaptation. The work also describes the collection of validity evidence of test results.

It is important to note that localization requires following the same rigorous procedures and methodology as recommended in the international standards and guidelines for adaptation. However, in the case of localization, it is necessary to justify the impossibility of achieving full comparability of future data produced by the instrument versions at the individual level and take into account the consequences of such a decision for their use. As a mechanism to ensure the comparability of the reading assessment results at the group level for the Russian language and the original version of the instrument in the absence of a common metric scale between the assessments of children from different countries, a two-stage approach is used. The first stage involves the use of the expert ranking method to establish correspondence between the models of assessed reading skills in the Russian and English languages. In the second stage, a common reading-level model was built and used to establish threshold scores (benchmarks) for student

assessment results in two countries. Using this procedure, it was possible to compare the basic reading skills of groups of students from two countries.

Results

In the first article within the framework of this dissertation research (Ivanova, 2018), the challenges faced by researchers conducting international comparative studies are discussed in detail. In particular, using the examples in a number of empirical studies, it has been shown that measurement results, including those in the field of education, being part of cross-cultural research, will always contain measurement errors. Any bias in a version of a study conducted in different languages or cultures within countries potentially jeopardizes the ability to draw fair conclusions from such research. The paper discusses in detail the sources of data incompatibility, which can be summarized as differences in the construct measured by different language versions of the instrument; differences in design, format, elements of the measurement instrument; and finally, differences in the administration of the instrument in different countries or cultures. The article states that the comparability of cross-cultural comparative studies should not be taken for granted; it should be ensured at the stage of developing versions of an instrument and subsequently proven in the process of validation, including studies on the equivalence of measurements. The article also introduces a distinction between the concepts of adaptation and localization of international comparative research instruments.

The results presented in this paper provide an overview of the challenges faced in conducting an international comparative study, regardless of its scale; they also lay the foundation for building an ILSA design, taking into account the recommendations set out in international guidelines and standards for conducting ILSAs, as well as the experience in this area of the considered individual researchers and research teams.

In the second article within the framework of this dissertation research (Ivanova et al., 2018), the process of adaptation of an international comparative study of the basic skills of primary school students in mathematics is considered using the example of the mathematical part of the iPIPS instrument. The work describes in detail the process of consequential adaptation, which assumes that the operationalization of the measured constructs, the style, form and way of formulating the tasks, and the methodology for assessing the tasks of the adapted version laid down in the tool will be based on the content and cultural basis of the original version of this tool. The adaptation resulted in the Russian-language version of the iPIPS instrument (math) potentially applicable for international comparison. A necessary step in substantiating the quality of the created assessment tool was its validation. To collect evidence of the validity of the results of the mathematical part of the iPIPS instrument for its use in the Russian language, a detailed psychometric analysis of the scale was carried out, the psychometric characteristics of individual tasks and the entire test were determined, and the reliability of measurement with this tool was established. Separate evidence of the argument presented in the theoretical part of this summary in favour of the validity of using iPIPS results for conducting an international comparative study was an exploratory cross-country comparison of the results of mathematics students from three countries – Russia, England and Scotland – including a series of studies proving the equivalence of measurements using two language versions of the instruments.

Using the Rasch dichotomous model, a linking procedure was implemented both between the assessment stages (at the beginning of the first grade and at the end of the first grade to measure student progress) and between countries. To place the students' test results from two assessment stages and three countries, the method of simultaneous calibration of items and persons was used (Wolfe, 2004). This method allows for alignment between countries and between initial and subsequent assessments, even with partially different instruments (i.e., several additional tasks that are more difficult than the original version were added to the Russian version). In our case, data

from all countries were analysed simultaneously, and each task was assessed either as common for all or some countries or as unique.

The psychometric analysis of the equated data was carried out in several stages: 1) analysis of the conformity of the data to the model (analysis of fit statistics) to provide conclusions about the quality of tasks and test forms, 2) analysis of the differential item functioning (DIF analysis) by country and by assessment stage to provide conclusions about the fairness of assessments for groups of students from different countries, 3) dimensionality analysis to provide inferences about the structure of the measured construct in both instrument versions; 4) analysis of the stability of items parameters for different samples to ensure conclusions about the reliability of measurements, and finally, 5) analysis of the general scale, which made it possible to assess the levels of preparedness of children in mathematics in three countries.

Thus, as a result of a series of studies described in the article by Ivanova et al. (2018), a reliable Russian-language version of the iPIPS instrument was created in terms of starting diagnostics of children's skills in mathematics and assessing their progress in the first year of school. The results of this article allow us to draw inferences about the scoring and generalization procedures within the framework of Kane's concept and to provide the collected evidence in favour of the validation argument that the results of iPIPS testing can be used for the purpose of conducting an international comparative study.

In the third article (Ivanova, Kardanova-Biryukova, 2019), a set of two studies was carried out on the localization of the international iPIPS instrument’s reading portion. The work carried out included the planning of the design and the localization of the instrument, the psychometric analysis of its quality, and the collection of evidence of the validity of the test results. The set of implemented procedures with a detailed discussion of the problems encountered during their implementation can be considered an example of a methodology for localizing an early reading assessment instrument.

The article discusses how, in the process of developing the Russian-language version of the reading assessment based on linguistic expertise, it was decided to localize this scale and to accept the impossibility of results in adaptation of the English version of the iPIPS test for assessing reading skills in Russian. The work on the localization of this portion of the iPIPS test took place in several stages: the linguistic characteristics of the original test items were studied, similar means were selected in Russian, and finally, Russian-language items were simulated, meaningfully close to the English-language original.

Due to a range of substantial structural differences between English and Russian (most importantly, English being verb-centred and Russian noun-centred, different sets of parts of speech and their functions, fixed word order in English resulting in a high incidence of stable syntactic constructions vs. free word order in Russian that agrees well with a developed system of grammatical markers), the stages of language development are not the same for English- and Russian-speaking children, which surely affects the process of reading acquisition.

To make the instruments testing reading development in British and Russian elementary school students as identical as possible, it was necessary to carry out linguistic analysis of the original iPIPS version, identify the functionally comparable linguistic means in both languages, and create tasks in Russian that would test equivalent reading skills.

In addition to describing the procedures for localizing this iPIPS instrument in the reading portion, the article describes the process of collecting evidence of the validity of the results obtained using item response theory methods. In particular, the analysis of the structure (dimension) of the scale, the analysis of the functioning of individual items and the scale as a whole, and the analysis of internal consistency were carried out. Thus, in the course of a series of performed procedures, the psychometric quality of the scale was substantiated, the reliability was assessed, and the internal structure and hierarchy of items were confirmed, corresponding to the

theoretical structure of the construct set out in the work. As part of collecting evidence of validity according to Kane's concept, the results of this article allow inferences to be drawn about scoring and generalization.

In the fourth article within the framework of the dissertation research (Ivanova, Kardanova 2020), in a series of two studies, an exploratory analysis of the possibilities of conducting an international comparative study of the iPIPS reading assessment results was carried out for first-grade students from different countries with different languages, cultures, and ages. Data obtained using two language versions of the iPIPS on representative samples of first-graders from the Republic of Tatarstan, Russia, and Scotland, UK, were used to compare the early reading assessment results. As part of the study of the possibilities of cross-country comparisons in the first of a series of studies, an examination of the construct was carried out – the models of children's emergent literacy levels at the time of entrance into school in Russian and English were compared. The methodological basis of the examination involved the expert judgment method and the rank-ordering method. Comparison of the hierarchy of item difficulties obtained as a result of data calibration using two approaches of Rasch modelling showed that three item clusters can be distinguished in both language versions of the instrument. These clusters are presented by the same items in both Russian and English. Thus, in the first study of this article, it was shown that expert judgements of the difficulty of items measuring emergent literacy on school entry can be used to build an item hierarchy along the construct continuum, to compare item hierarchies between the two language versions, and, finally, to form the basis for setting benchmarks between the levels of reading development in two languages, Russian and English.

In the course of the second in a series of studies using empirical data from samples of Russian- and English-speaking students from two countries, benchmarks were established, and the levels of reading development were determined. Those levels were applied in a uniform manner to the reading test results in both language versions of the instrument to group children in both countries into categories according to their level of reading development. In the conclusion of the article, it is assumed that if the structure of the proposed iPIPS theoretical model of reading development is confirmed for any two countries compared (i.e., if the test item clusters identified by experts and confirmed by psychometric analysis measure the same construct), this can serve as the basis for setting the international benchmarks that will allow comparing cumulative percentages of children at a particular level of reading development across countries. This assumption could be tested in the future for other language versions of the iPIPS projects. Thus, the series of studies presented in this article provided support for the argument that the results of iPIPS testing can be used for the purpose of conducting an international comparative study.

Scientific and practical significance

The importance with which participants of international large-scale assessment studies relate to their results determines the need to ensure high quality at each stage of the ILSA life cycle – from planning such a study to collecting evidence of the validity of its results and interpretation in general and in each language version separately.

Extensive ILSA data allow us to study the effects of a variety of factors on educational outcomes, assess the interrelationships of these factors, and gain a deeper understanding of the mechanisms that make up and drive educational systems. At the same time, the generalized form of the results of the ILSA is most often some kind of rating that allows us to identify a country or a group of countries with the highest level of educational achievement based on some, as a rule, cross-sectional study. Based on such ratings, it is often assumed that educational practices existing in a given country are the most effective and therefore deserve to be adapted by other participants (Steiner-Khamsi, Waldow, 2012). These assumptions, of course, require theoretical understanding and empirical testing. There are many works that urge researchers and policymakers to treat these

assumptions carefully and thoughtfully, including in view of the cross-sectional nature of the design of most ILSAs (Ercikan, Roth, 2015).

The longitudinal nature of the international iPIPS study, which in the long term makes it possible to obtain internationally comparable data on the progress of students for the first year, which is key for their subsequent education at school, is of significant scientific and practical interest. Despite the number of objective difficulties associated with the development of national versions of the iPIPS instrument, it can provide data for conducting a wide range of secondary studies devoted to the study of what children know and can do when entering school, what are the dynamics of their progression during the first year at school, what factors predict the success of children during this important period, and how these processes proceed in different countries.

The implementation of the iPIPS international comparative study required solving a number of problems that international studies traditionally face, as well as solving additional problems related to ensuring the validity of a potential comparison of the results of assessment of primary school students in emergent reading and math skills in conditions of different students' ages and the partially differing assessment instruments. In this work, for the first time for the participants of the international iPIPS project, an exploratory comparative study was conducted to assess the results of first-graders taking the non-English language version of the instrument. Within the framework of the general field of scientific works devoted to the problem of international comparative studies, the possibilities of international comparison of the progress of participants from different countries were presented for the first time.

The analysis of the scientific literature and empirical analysis and the proposed solutions to the problems that arose in the course of the research allow us to expand the existing scientific knowledge about international comparative studies of children of early school age in the fields of reading and mathematics.

The practical significance of this work is that it acquaints the interested community with the challenges and problems faced by the ILSA, focuses on a specific sample of participants (children at the time of entrance into school), and allows them to assess their academic progress during the first year of schooling. The results of this work can be generalized to solve research problems that emerge while dealing with similar constructs and are aimed at early-age students.

Adaptation and localization procedures, as well as studies of the comparability of assessment results presented within this dissertation, can serve as a set of methodological recommendations when conducting international studies on the assessment of educational achievements at the time of school entrance and students' academic progress.

It is also worth noting that in the long term, the methodology presented in the article by Ivanova, Kardanova (2020) can be applied for both the purposes of international comparisons of data and other purposes, such as for year-to-year comparisons of test results from different samples of students and different versions of the same assessment instruments.

Finally, in the course of the work done, a high-quality assessment instrument was created for the initial diagnostics children at the time of school entrance and assessing his or her progress during the first year of schooling. This instrument can be used not only for international comparisons but also within the country as a standardized monitoring instrument applicable for large-scale use, as well as a source of reliable and valid data for a wide range of scientific research in the field of education.

Points for the PhD thesis defence

International comparative studies, in particular, focused on the beginning of children's education in school face methodological challenges associated with ensuring cross-cultural comparability in the context of restrictions imposed by the age, cultural and linguistic specifics of its target audience. In this regard, it is necessary to distinguish between the concepts of adaptation and localization. The following definitions are taken in the work: adaptation is the process of transferring an instrument into the language of another culture with the minimum necessary changes associated with the peculiarities of the language, content and use of terminology, preserving the equivalence of measurements; localization is the process of creating another language or culture version of an instrument that is based on the same theoretical model, but does not imply the equivalence of measurements.

The methodology based on international standards for ILSAs makes it possible to adapt the iPIPS instrument in mathematics for use in the Russian language. The methodology proposed within the current research allows localizing the iPIPS instrument in the reading portion for use in the Russian language.

The set of activities carried out in the course of the dissertation allows validating the use and interpretation of the assessment results obtained using the Russian-language version of the iPIPS. Evidence of the validity of the assessment results of the Russian-language version of the iPIPS is based on the description of the procedures for its development (adaptation and localization), on the construction of empirically substantiated inferences about the quality of individual items and scales, the reliability and stability of the measurements, and the structure of the scales used.

The Russian version of the iPIPS instrument's development and validation allowed the expansion of the range of high-quality standardized tools used in the Russian language to research what students know and can do when they start schooling.

The application of the methodology outlined in the work allows us to ensure the comparability of the assessment results obtained using the non-English language version of the iPIPS instrument. Thus, the results of the Russian version of the iPIPS reading and math tests can be used for international comparative studies of what children from different countries know and can do when entering school and to measure their progress in the first year of school.

References

1. AERA, APA, NCME. Standards for educational and psychological testing. – Amer Educational Research Assn, 2014.
2. Ainley, M., & Ainley, J. (2019). Non-Cognitive Attributes: Measurement and Meaning. In: Suter, L.E., et al. (Eds.). The SAGE Handbook of Comparative Studies in Education (pp. 103-125). SAGE Publications Limited.
3. Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. Journal of educational measurement, 36(3), 185-198.
4. Archer, E., Scherman, V., Coe, R., & Howie, S. J. (2010). Finding the best fit: the adaptation and translation of the Performance Indicators for Primary Schools for the South African context. Perspectives in Education, 28(1), 77-88.
5. Bartholo, T. L., Koslinski, M. C., Costa, M. D., & Barcellos, T. (2020). What do children know upon entry to pre-school in Rio de Janeiro?. Ensaio: Avaliação e Políticas Públicas em Educação, 28(107), 292-313.
6. Bolotov V.A., Val'dman, I. A., Kovalova, G. S., & Pinskaya, M. A. (2013). Rossiyskaya sistema otsenki kachestva obrazovaniya: glavnyye uroki (Russian system for assessing the quality of education: main lessons). Educational quality in Eurasia, 1, 85-121 (in Rus.).
7. Bolotov, V. A. (2018). Proshloye, nastoyashcheye i vozmozhnoye budushcheye rossiyskoy sistemy otsenki kachestva obrazovaniya (The past, present and possible future of the Russian system for assessing the quality of education). Educational studies Moscow, 3, 287-297 (in Rus.).
8. Braun, H. (2013). Prospects for the future: A framework and discussion of directions for the next generation of international large-scale assessments. In The role of international large-scale assessments: Perspectives from technology, economy, and educational research (pp. 149-160). Springer, Dordrecht.
9. Carnoy, M., Khavenson, T., Loyalka, P., Schmidt, W. H., & Zakharov, A. (2016). Revisiting the relationship between international assessment outcomes and educational production: Evidence from a longitudinal PISA-TIMSS sample. American Educational Research Journal, 4(53), 1054-1085.
10. Caro, D. H., & Cortés, D. (2012). Measuring family socioeconomic status: An illustration using data from PIRLS 2006. IERI Monograph Series Issues and Methodologies in Large-Scale Assessments, 5, 9-33.
11. Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to K ane's framework. Medical education, 49(6), 560-575.
12. Ellefson, M. R., Zachariou, A., Ng, F. F. Y., Wang, Q., & Hughes, C. (2020). Do executive functions mediate the link between socioeconomic status and numeracy skills? A cross-site comparison of Hong Kong and the United Kingdom. Journal of Experimental Child Psychology, 194, 1-19.
13. 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. Three. Washington, American Psychological Association.
14. Ercikan, K., Oliveri, M. E., & Sandilands, D. (2013). Large-scale assessments of achievement in Canada. International guide to student achievement, 456-459.

15. Ercikan, K., Roth, W. M., & Asil, M. (2015). Cautions about Inferences from International Assessments: The Case of PISA 2009. Teachers College Record, 117(1), n1.
16. Esselink B. (2000). A Practical Guide to Localization. Vol. 4. Amsterdam, Philadelphia: John Benjamins.
17. Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. In M. von Davier & D. Hastedt (Eds.), IERI monograph series: Issues and methodologies in large scale assessments, Vol. 2
18. Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. Methodological advances in cross-national surveys of educational achievement, 58-79.
19. Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. Journal of Applied Testing Technology, 1(1), 1-13.
20. Harms, T., Clifford, R. M., & Cryer, D. (2014). Early childhood environment rating scale. Teachers College Press.
21. Howie, S., Combrinck, C., Tymms, P. & Merrell, C. (2016). What children know and can do when they start school in the Western Cape. URL: https://docs.google.com/a/ipips.org/viewer?a=v&pid=sites&srcid=aXBpcHMub3JnfGlwaXBzfGd4OjZmNGE5YjIwMDNkODliOTk
22. International Test Commission. (2017). International Guidelines for Test Use. Int. J. Test., 2(1), 93–114.
23. Ivanova A. (2018). Problema sopostavimosti rezul'tatov v mezhdunarodnykh sravnitel'nykh issledovaniyakh obrazovatel'nykh dostizheniy (The problem of ILSA' results comparisons). Otechestvennaya i zarubezhnaya pedagogika, 2(1), 68-81 (in Rus.)
24. Ivanova, A., Kuznetsova, M., Semenov, S., & Fedorova, T. (2016). Faktory, opredelyayushchiye gotovnost' pervoklassnikov k shkole: vyyavleniye regional'nykh funktsiy (School Readiness of First-Graders and Its Predictors: Identifying Region-Specific Characteristics). Educational studies Moscow, (4), 84-105 (In Rus.).
25. Ivanova A., Kardanova E., Merrell C., Tymss P., Hawker D. (2018). Checking the possibility of equating a mathematics assessment between Russia, Scotland and England for children starting school. Assessment in Education: Principles, Policy and Practice, 2(25), 141-159.
26. Ivanova A., Kardanova-Biryukova K. (2019). Sozdaniye russkoyazychnoy versii mezhdunarodnogo instrumenta otsenivaniya rannikh navykov chteniya (Constructing a Russian-Language Version of the International Early Reading Assessment Tool). Educational studies Moscow, 4, 93-115 (in Rus.).
27. Ivanova A., Kardanova E. (2020). Izucheniye vozmozhnosti provedeniya mezhstranovogo sravnitel'nogo issledovaniya navyka chteniya uchashchikhsya na vkhode v shkolu v Rossii i Velikobritanii (Checking the Possibility of an International Comparative Study of Reading Literacy Assessment for Children Starting School). Educational studies Moscow, 4, 8-36 (in Rus.).
28. Ivanova, E., & Vinogradova, I. (2018). Scales SACERS: Results of the Study of the Educational Environment of Moscow Schools. European Journal of Contemporary Education, 7(3), 498-510.

29. Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 39(1), 1.
30. Jowell R. et al. Measuring attitudes cross-nationally: Lessons from the European Social Survey. – Sage, 2007.
31. Kane, M. (2006). Content-related validity evidence in test development. Handbook of test development, 1, 131-153.
32. Kane, M. (2013). Validity and fairness in the testing of individuals. Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity, Emerald Group Publishing, Bingley, 17-53.
33. Kane, M. T. (1992). An argument-based approach to validity. Psychological bulletin, 112(3), 527-535.
34. Kastberg, D., Roey, S., Lemanski, N., Chan, J. Y., & Murray, G. (2014). Technical report and user guide for the Program for International Student Assessment (PISA). NCES 2014-025.
35. Kautz T. et al. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. National Bureau of Economic Research, w20749.
36. Kovaleva G. S. (2017). Finansovaya gramotnost' kak sostavlyayushchaya funktsional'noy gramotnosti: mezhdunarodnyy kontekst (Financial literacy as part of functional literacy: an international context). Domestic and foreign pedagogy, 2 (37), 31-43 (in Rus.).
37. Laschke C., Blömeke S. (2016). Measurement of job motivation in TEDS-M: testing for invariance across countries and culturesıo Large-scale Assessments in Education, 1(4), 16.
38. Mullis, I. V., & Martin, M. O. (2019). PIRLS 2021 Assessment Frameworks. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.
39. Niklas, F., & Schneider, W. (2013). Home literacy environment and the beginning of reading and spelling. Contemporary Educational Psychology, 38(1), 40-50.
40. OECD (2018). International Early Learning and Child Well-being Study. URL: http://www.oecd.org/education/school/international-early-learning-and-child-well-being-study.htm
41. Oliveri M. E. et al. (2012). Methodologies for investigating item-and test-level measurement equivalence in international large-scale assessments. International Journal of Testing, 12(3), 203-223.
42. Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. Psychological Test and Assessment Modeling, 53(3), 315-333.
43. Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. International Journal of Testing, 14(1), 1-21.
44. Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. International Journal of Testing, 13(3), 272-293.

45. Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. Child development. 78(4), 1255-1264.
46. Rasch, G. (1966). An item analysis which takes individual differences into account. British journal of mathematical and statistical psychology, 19(1), 49-57.
47. Regional report. (2017). Starting diagnostics of students at the entrance to school: Republic of Tatarstan. 2017. URL: http://rcmko.ru/meropriyatiya/monitoringi/ipips/startovaya-diagnostika-uchashhihsya-na-vhode-v-shkolu-respublika-tatarstan/
48. Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. Educational Researcher, 39(2), 142-151.
49. Sears, N. A., Othman, M., & Mahoney, K. (2015). Examining the relationships between NCLEX-RN performance and nursing student factors, including undergraduate nursing program performance: A systematic review. Journal of Nursing Education and Practice, 5(11), 10-15.
50. Sellar, S., & Lingard, B. (2014). The OECD and the expansion of PISA: New global modes of governance in education. British Educational Research Journal, 40(6), 917-936.
51. Shuttleworth-Edwards, A. B., Kemp, R. D., Rust, A. L., Muirhead, J. G., Hartman, N. P., & Radloff, S. E. (2004). Cross-cultural effects on IQ test performance: A review and preliminary normative indications on WAIS-III test performance. Journal of clinical and experimental neuropsychology, 26(7), 903-920.
52. Sperber A. D., Devellis R. F., Boehlecke B. (1994). Cross-cultural translation: methodology and validation //Journal of cross-cultural psychology, 25(4), 501-524.
53. Steiner-Khamsi G., Waldow F. (ed.). World yearbook of education 2012: Policy borrowing and lending in education. Routledge.
54. Stubbe, T. C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. Educational Research and Evaluation, 17(6), 465-481.
55. Turner, R., & Adams, R. J. (2007). The programme for international student assessment: An overview. Journal of Applied Measurement, 8(3), 237-248.
56. Tymms P., Merrell C., Hawker D., Nicholson F. (2014) Performance Indicators in Primary Schools: A Comparison of Performance on Entry to School and the Progress Made in the First Year in England and Four Other Jurisdictions. URL: http://dro.dur.ac.uk/23562/1/23562.pdf
57. Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. Journal of Research in Reading, 22(1), 27-36.
58. Tymms, P. (2013). Baseline assessment and monitoring in primary schools. David Fulton Publishers.
59. Tymms, P., Jones, P., Albone, S., & Henderson, B. (2009). The first seven years at school. Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21(1), 67-80.
60. UN. 2015. The 2030 Agenda for Sustainable Development. URL: https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=R
61. Vidmar, M., Niklas, F., Schneider, W., & Hasselhorn, M. (2017). On-entry assessment of school competencies and academic achievement: a comparison between Slovenia and Germany. European journal of psychology of education, 32(2), 311-331.

62. Wildy, H., & Styles, I. (2008). Measuring what students entering school know and can do: PIPS Australia 2006-2007. Australian Journal of Early Childhood, 33(4), 43-52.
63. Wolfe, Edward 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.
64. Yudina, Ye. G. (2015). Shkaly ECERS kak metod otsenki kachestva i razvitiya rossiyskoy sistemy doshkol'nogo obrazovaniya (ECERS scales as a method for assessing the quality and development of the Russian preschool education system). Preschool education today. Theory and Practice, 7(59), 22-26 (in Rus.).