Constructing a Russian-Language Version of the International Early Reading Assessment Tool

Ivanova A., Kardanova-Biryukova K.

Alina Ivanova
Research Fellow, Center for Monitoring the Quality in Education, Institute of Education, National Research University Higher School of Economics. Address: 20 Myasnitskaya St, 101000 Moscow, Russian Federation. Email: aeivanova @hse.ru

Ksenia Kardanova-Biryukova
Candidate of Sciences in Philology, Associate Professor, Institute of Foreign Languages, Moscow City University. Address: 5B Maly Kazenny Lane, 105064 Moscow, Russian Federation. Email: Kardanova-BirukovaKS@mgpu.ru

Abstract. Successful adaptation of first-graders to school largely determines their subsequent educational attainment. In Russia as well as across the globe, there are few high-quality standardized assessment instruments providing a comprehensive picture of what children know and what they can do when they start school. Large-scale evaluation of reading literacy is particularly challenging due to age-specific characteristics and the assessment format. This article outlines a step-by-step procedure of localizing a part of the international instrument iPIPS designed to measure early reading skills at the start of school and the progress made during the first school year, within the Russian educational paradigm. Localization is understood as transformation of an instrument originally developed in another language (English in this case) so that it makes account of the cultural and linguistic characteristics of the target audience. The procedure included development of a Russian-language version of iPIPS and a series of studies to verify its construct validity. The process involved analyzing the linguistic characteristics of the original tasks, finding equivalent linguistic means in the Russian language, and designing Russian-language tasks identical to the original ones in terms of functionality. To verify construct validity of the localized instrument, we evaluated the psychometric properties of the scale, tested its reliability, and studied compliance of the task structure and hierarchy with the theoretical framework. The findings have revealed that large-scale local or regional tests administered using this localized assessment instrument may yield valuable data which can be further used for analysis of the current situation and informed decision-making in educational policy.

Keywords: reading assessment instrument, iPIPS, elementary school, localization, validation.

Early elementary grades are crucial for children's cognitive and personal development. Numerous studies have shown that successful adaptation of first-graders to school largely determines their subsequent educational attainment [Bezrukikh, Filippova, Baydina 2012; Tymms et al. 2009]. Politicians, researchers, and educators are concerned about using high-quality assessment instruments to measure children's knowledge and skills at the start of school and the progress they make in order to provide for evidence-based learning organization and early academic interventions, where necessary.

Education assessment systems use both national instruments and the results of international comparative studies, which improves the overall efficiency of educational management [Bolotov et al. 2013]. However, planning and administration of large-scale international assessments are fraught with a number of methodological challenges, of which measurement pitfalls are the greatest ones, as both analysis and interpretation depend on the instrument's ability to measure the intended construct similarly across all the participating countries. When it comes to reading assessment, researchers and instrument developers from different countries face a daunting task of elaborating national versions of the instrument that will overcome the effects of a specific language on reading in that language. These days, reading literacy is measured by two international comparative assessments, the Programme for International Student Assessment (PISA) for 15-year-old students and the Progress in International Reading Literacy Study (PIRLS) for fourth-graders. However, no international comparative studies of similar scale exist to measure reading literacy at the start of school objectively and reliably. Such studies are incredibly hard to implement, since the influence of language is the strongest at the early development stage [Ercikan, Roth, Asil 2015]. Nevertheless, researchers attempt to compare early reading competencies at the start of school across different countries, English-speaking for instance [Tymms et al. 2014].

This article describes the step-by-step procedure of localizing a part of an international instrument originally designed in the English language to measure early reading skills, by the example of iPIPS[1]. The procedure included development of a Russian-language version of iPIPS and a series of studies to verify its construct validity.

1. Approaches to Early Childhood Assessment

Knowledge and skills of preschool and early school-age children can be assessed either by asking them directly what they know and can do or by using indirect evaluation methods, such as observation or teacher (parent) surveys. There are some proprietary methodologies in Russia designed to assess specific skills in children and often re-

---

[1] http://ioe.hse.ru/ipips

quiring psychological assistance in administration and interpretation, such as the voluntary behavior assessment instrument Graphic Dictation or the phonemic awareness study The First Letter [Kovaleva et al. 2012]. There is a number of well-known integrated measures of a broad range of children's competencies at school entry abroad. For example, the Early Development Instrument (EDI), designed in Canada and applied by many other countries as well, is a teacher-completed measure of physical, social, emotional, communicative, and language and cognitive domains of child development in the last year of preschool [Janus, Brinkman, Duku 2011]. Other countries use preschool and school environment assessment instruments, such as Early Childhood Environment Rating Scales (ECERS) and School-Age Care Environment Rating Scales (SACERS). Those are based on observation and structured expert assessment of child's environment [Harms 2013].

Additional difficulties in child assessment emerge when it comes to large-scale early literacy studies which require heavy resource investments, elaborated design, and standardized procedures [Mislevy, Steinberg, Almond 2003]. In case of international or cross-cultural assessments, there are also challenges associated with the need to provide uniformity of measurements across countries, cultures, and languages [Rutkowski et al. 2010].

No projects similar to PISA or PIRLS exist to measure first-year performance, and few of the existing instruments meet the criteria of objectivity, integrity, and quality to be used for international comparisons. One of the few examples is International Performance Indicators in Primary Schools (iPIPS), designed to measure early reading skills at the start of school and the progress made during the first school year.

The iPIPS instrument, originating from the University of Durham in Great Britain, exists today in a few versions localized for different countries [Tymms 1999; Archer et al. 2010; Niklas, Schneider 2013; Wildy, Styles 2011]. Children are asked to do computer-adaptive tests with the assistance of an examiner (school psychologist, counselor, or pre-briefed teacher). Each assessment cycle consists of two stages: at the beginning and at the end of the first school year. The total of iPIPS tests allows assessing child development in four domains: early reading and math skills, phonemic awareness, and vocabulary.

The iPIPS characteristics and the opportunities it offers make it quite a promising objective to elaborate a Russian version of the instrument and join the growing iPIPS project. First, even though the direct assessment format requires some additional resources (adult supervision), it allows making inferences about learning abilities and learning difficulties of every child. Individualized approach eliminates bias, which is rarely achievable with indirect assessments. Second, computer-adaptive testing allows assessing children's knowledge and skills in a friendly and comfortable environment without giving them tasks that are overly complicated for their current level. Third,

the instrument also measures the progress that children make in their first year at school. Fourth, proven reliability and validity of the original version and its adaptations for other English-speaking countries [Demetriou, Merrell, Tymms 2017; Tymms et al. 2009; Wildy, Styles 2011] offer prospects for international comparative first-grader studies that will include non-English orthographies, too.

2. The iPIPS Reading Model

Early literacy research findings show that learners acquire early reading skills in phases, the language of teaching affects how long students need to acquire those skills, and learners of all alphabet-based languages pass through the same phases [Seymour, Aro, Erskine 2003; Rayner et al. 2012; Gove, Wetterberg 2011].

The most basic reading model, which the iPIPS authors constructed using the results of many years' research and which is consistent with the theoretical ideas of Russian educators [Egorov 2006; Kornev 1995], postulates that children pass through some important phases when learning to read: they develop a general idea of how the language works, learn to recognize letters and graphic representations of words, gradually develop decoding skills, and finally achieve the level of reading comprehension [Merrell, Tymms 2007].

Knowing how the language works implies understanding the fundamental organization of language and its forms, which includes knowing how to hold a book and where to start reading, being aware of left-to-right progression, being able to distinguish between letters and words, etc. [Clay 1985]. Letter-name knowledge is another important longitudinal predictor of learning to read, as letters serve fundamental functions in alphabetic writing systems [Foulin 2005]. Word recognition often comes to children as they learn letters. Research findings indicate that a lot of children are capable of recognizing and understanding simple familiar words even before they acquire reading comprehension skills [Harrison 2004].

Children who already know how the language works will require some specific teaching methods to take their reading further to embrace decoding and comprehension skills [Harrison 2004; Merrell, Tymms 2007; Zhurova, Elkonin 1963]. In the course of practicing in reading with the help of one or two strategies, children develop decoding skills, gain experience, and learn to read more and more words automatically—this is when their information processing abilities can be redirected to text comprehension [Merrell, Tymms 2007].

The reading model offered by the iPIPS authors, describing step by step the development of reading skills in children, includes the following phases:

1) Ideas about reading (the concept and structure of text)
2) Letter recognition
3) Sight word recognition

4) Decoding (mechanical reading)
5) Reading comprehension

3. Adaptation and Localization

Adaptation of an English-language instrument is an extremely complex and resource-intensive process [Leong et al. 2016]. Besides, even if all the procedures are followed accurately, the outcome is not always suitable for cross-country comparative assessments [Ivanova 2018].

Adaptation of an instrument seeks to ensure validity of interpreting the results obtained with country-, language-, or culture-specific versions of that instrument [Sireci, Patsula, Hambleton 2005; Leong et al. 2016]. Research institutions involved in educational assessment offer a variety of guides and recommendations designed to provide a high quality of instrument adaptation in international studies [AERA, APA, NCME 2014; Leong et al. 2016]. Those procedures are aimed at achieving maximum result comparability as the main prerequisite for further use of international study findings, which is only possible if measurements obtained with the instrument versions developed for different languages and cultures are equivalent. Measurement equivalence, in its turn, implies three critical components: construct equivalence, equivalence of method, and equivalence of items [Ercikan, Roth, Asil 2015].

When developing the Russian-language version of iPIPS, we proceeded from the firm belief that localization is the only possible solution for the iPIPS reading test. Localization involves taking a product and making it linguistically and culturally appropriate to the target locale (country/region and language) where it will be used [Esselink 2000]. In this article, localization of an assessment instrument is understood as the process of developing a version of the instrument in another language that lies within the original theoretical framework but makes account of the target country's cultural characteristics. The main difference between localization and adaptation is that localization does not imply direct comparison of individual students' test scores across countries.

Elaboration of the Russian-language version of the early-grade reading assessment instrument took a significant amount of time and effort and involved multiple adjustments and improvements in the process. Age difference of more than two years between the English-speaking target audience (four- to five-year-olds) and Russian first-graders (six- to seven-year-olds) along with a number of substantial differences between the Russian and English languages were the greatest challenges faced by the developers.

4. Methodology
4.1. iPIPS reading test localization

Translation and expert evaluation of the baseline reading assessment tests targeting British children came to be the first step towards cre-

ating the Russian-language version of iPIPS. Translation (direct and reverse) was performed in compliance with the recommendations of the International Testing Commission (ITC) [Leong et al. 2016]. The original iPIPS reading test was designed to evaluate language development of elementary school students who were native speakers of English. This part of the iPIPS instrument included a few modules corresponding to phases of the reading development model proposed by the iPIPS authors. Understanding of text structure was tested by a module of tasks asking students to indicate the beginning and end of a given text, etc. Letter recognition was assessed using tasks that asked children to name a letter or the corresponding sound. Word recognition skills were tested by asking children to match spoken words (pronounced by examiners) to written ones presented in the test (e.g. find the pronounced word in a row of four or five words offered by the task). Decoding skills were measured by having children read a short story aloud and examiners scoring the number of words pronounced correctly.

When developing the Russian-language version of the iPIPS reading tasks, the latter were preserved nearly unchanged (translation was performed, and equivalent letters, words and texts in Russian were selected with due regard to usage and baseline literacy rates). Meanwhile, the reading comprehension module was much more difficult to work with, as the texts offered to Russian- and English-speaking children had to be comparable in the level of complexity, and the tasks hidden within them (the so-called "traps") had to evaluate the same competencies. Eventually, this part of the iPIPS instrument was localized in stages. First, linguistic characteristics of the original text were analyzed; next, equivalent "traps" in Russian were found; finally, a Russian-language text with "traps" and content close to the English-language original was produced. The sections below will dwell on every stage of the work done.

4.1.1. Linguistic characteristics of the original reading comprehension test

The student is asked to read a text and, in some parts of it, choose one of three response options (expert evaluation was performed using the texts Underground and Walking to School). While making their choice, children have to deal with the so-called "traps", which represent some of the biggest challenges faced by learners of their native language (English), such as distinction between words and different grammatical forms (temporal and aspectual verb forms; subjective, objective, and possessive personal pronouns; singular and plural nouns, etc.), articles and how to use them, distinction between the meanings of language units, valency and concurrence characteristics of words, simple and double prepositions, etc.

The units in a "trap" are related pairwise, one of the three options being part of both pairs. For example, in the cluster creatures–annuals–animal, the first pair creatures–animal contrasts its components by noun number (the former being plural and the latter, singular), and

the annuals–animal pair is based on the principle of phonological similarity. While animal is the unit common for both pairs, it is creatures that should be selected as the correct answer.

Some "traps" use a more intricate mechanism of concealing difficulty. In the sentence They can run quickly and are very good at leaping upon to fences, trees and quit / other / offer high places, the pair other–offer is based on phonological/orthographic similarity, but there is no direct relationship between quit and other. The "trap" consists in quit being spelled similar to quite which could be used instead of other in this context: other high places–quite high places. Therefore, children are expected to choose quit in case they are unable to distinguish visually between quite and quit (otherwise speaking, failure will be caused by orthographic similarity between the words).

To explicate the logic behind selecting "traps" for each of the two texts, a somewhat deeper insight into the English language is required.

4.1.1.1. Structural characteristics of English as compared to Russian

Modern English is a flectional language of the analytic type. Throughout its evolution, synthetic forms blending the semantic and grammatical meaning within the same phonological and orthographic units (s-del-a-l; de-motiv-ate-d) have been gradually replaced with analytic ones representing sequences of independent phonological and orthographic units loaded with discrete semantic and grammatical meanings (have been asked). Although synthetic forms still exist in English, they are becoming ever more simplified (consisting of ever fewer components which are relatively simple and often monosemantic, as in re-do, where the only meaning conveyed by re- is that of "repetition").

A relatively high incidence of homonymy—homophones (of–off; night–knight, etc.), homographs (to wind [ai]–wind [i], to tear [ɛə]–tear [iə]), and homoforms (heard (past tense of hear)–herd, left (past tense of leave)–left (opposite of right))—is another distinctive feature of the English language. Along with homonyms, there are a lot of words in English that are similar in sound or spelling but do not make homonymic pairs. Phonological and/or orthographic similarity is represented in the following examples extracted from the texts analyzed: wake–work–walk, buy–boy, carried–cared, leaf–leave, etc.

Verbs are central to the English language system [Fillmore 1981]. Apparently, use of personal and impersonal verbs is mastered at the earliest stages of first-language acquisition. This is reflected in the texts offered by the English-language version of the test, which contain "traps" testing children's ability to recognize the spelling and the semantic meaning of various verb forms as well as their knowledge of verb form components (auxiliary verb + form of main verb).

Another important characteristic of the English language is fixed word order, which manifests itself, in particular, in a large number of stable syntactic constructions (complex object, the for-to-infinitive construction, etc.). Native English speakers begin to learn those

syntactic patterns in a Gestalt-like manner at the earliest stages of first-language acquisition—that is why "traps" testing children's ability to use them are abundant in the English-language texts.

Difficulty of a text is determined by the amount and quality of "traps" in it. Orthographic/phonological similarity "traps" prevail in Walking to School (eight "traps"[2]), while Underground contains only six traps of this type, yet a large amount of grammar/syntax "traps".

4.1.1.2. Structural characteristics of Russian as compared to English, and making account of them when developing the Russian-language version of the test

Development of a comparable test for Russian-speaking children is possible provided that the distinctive characteristics of the Russian language are taken into account. The main difference between the two languages is that Russian has been an inflected language of the synthetic type throughout its history of development, which implies prevalence of synthetic forms and coexistence of different semantic and grammatical meanings within the same phonological and orthographic units. For instance, the form pri-dorozh-n-ogo represents the semantic meanings of "near" and "road" and the grammatical ones of gender, number, and case. Unlike English, Russian has few analytic forms consisting of different independent components, each with a semantic or grammatical meaning of its own (compare, for example, have been working to budu gotovit'). Isolated cases of analytism in Russian are represented by comparative and superlative adjectives (bolee udachny / samy vazhny) and some forms of the future verb tense (budem zanimat'sya). This divergence between the languages must be taken into account when devising "traps" testing distinction between grammatical forms. Whereas a number of "traps" in the easiest English-language text Walking to School ask children to choose the right answer from three continuous tense forms (is / was / were shining or was wearing / wear / wore), "traps" in the Russian language may suggest choosing the verb form that agrees with the subject in person and/or number (svetilsya / svetilas' / svetilos' or bylo / byli / byla).

Another distinctive feature of the Russian language is being centered around nouns and noun groups in terms of syntactic organization. In other words, the noun in Russian is a carrier of sense that is critical to understanding the meaning at the level of sentences. By contrast, English assigns this paramount role to the verb (and thus is said to be verb-based, or verb-centered). This difference, in particular, is obvious when we compare nominal parts of speech in Russian, with their extensive system of grammatical categories (gender, number, case, etc.), to English noun phrases which only have a number, a rather reduced category of case, and that of definiteness, and some have one category only (e.g. adjectives are only varied by degrees of

---

[2] Statistical data analysis is somewhat impaired by different types of logic underlying the "traps", so approximate figures are given.

comparison). "Traps" testing first-graders' ability to distinguish between the grammatical forms of nominal parts of speech are extremely few in the English-language texts, but they must be included in the Russian-language version as being highly relevant to language development (e.g. zhivotnye / zhivotnoe / zhivotnogo or vysokomu / vysokogo / vysokiy or rebyata / rebyatam / rebyat).

When adapting the texts to test native Russian speakers, the numerous orthographic and phonological identity/similarity "traps" in the original version should be abandoned, as Russian has much fewer words like that than English (the high degree of homonymity in English is explained by the prevalence of one-syllable words, which increases the likelihood of concurrence). Besides, "traps" testing the ability to distinguish between different articles and auxiliary verbs are also impossible to transplant into Russian (the few auxiliary verbs in Russian used to build future tense forms (budet rabotat') are easy to recognize for native speakers even at the earliest stages of language development).

Finally, another distinctive feature of Russian is free word order. Synthetism of the Russian language (grammatical markers being realized within orthographic and phonological units) grants relative freedom in sentence building. Using an extensive system of grammatical markers, native speakers of Russian establish the necessary logical connections and formulate complete thoughts without being constrained by component arrangement (of course, again, this freedom is relative, since collocation rules for logically bound text elements postulate that an adjective cannot be separated from the element it modifies and that noun as an object should stay within the verb phrase). As a result, Russian has no stable syntactic constructions that are typical of English (such as complex subject or complex object), so syntax "traps" of this kind should also be left out in the process of localization.

Relying upon the distinctive characteristics of the Russian language, a typology of "traps" centered around the linguistic competencies of native Russian speakers is proposed.

One category includes "traps" based on phonological and orthographic similarities, in which three response options have similar sound and/or spelling: poka / pora / gora ... vstavat'; Vo dvore yeye uzhe ... zhdal / sdal / zhban ... Kostya, etc.

Another group consists of "traps" designed to test the ability to distinguish between the grammatical forms of words. These may include, for example, nouns differing in number / case: rebyata / rebyatam / rebyat ... poshli k shkole; vypey ... moloka / moloku / molokom; verbs in different tense-aspect forms, forms of person and number: Oni bystro ... begat' / begaet / begayut; Togda im ... nuzhen / nuzhna / nuzhno ... pomoshch'; personal and possessive pronouns in various forms: No ... ona / on / oni ... vse zhe lovyat myshey.

The third category is represented by "traps" testing students' ability to choose the right response depending on the context. In this case,

preceding context allows using any of the options available, while limitations are imposed by the context that follows, or the choice may be contingent on a broader context. For instance, in the sentence Ya budu s nim … uchit' / igrat' / gulyat' … any of the options will do, and the choice can only be done as the child has read the continuation of the phrase, komandu "Ryadom".

Lexical "traps" constitute a special subcategory of semantic "traps" and imply choosing from prefixed words (with the same stem). Not only are students expected to identify the right form with the context but they also should distinguish between the options offered. In Prezhde chem vyiti iz doma, Yulya sobralas' / nabralas' / zabralas', students need first of all to make sense of the meanings of each verb and then match the right meaning with the immediate context.

Taking into account the differences between Russian and English, we have managed to develop a typology of "traps" that makes allowance for the distinctive structural characteristics of Russian as the target language. Next, the resulting version of the instrument had to be tested for reliability and validity.
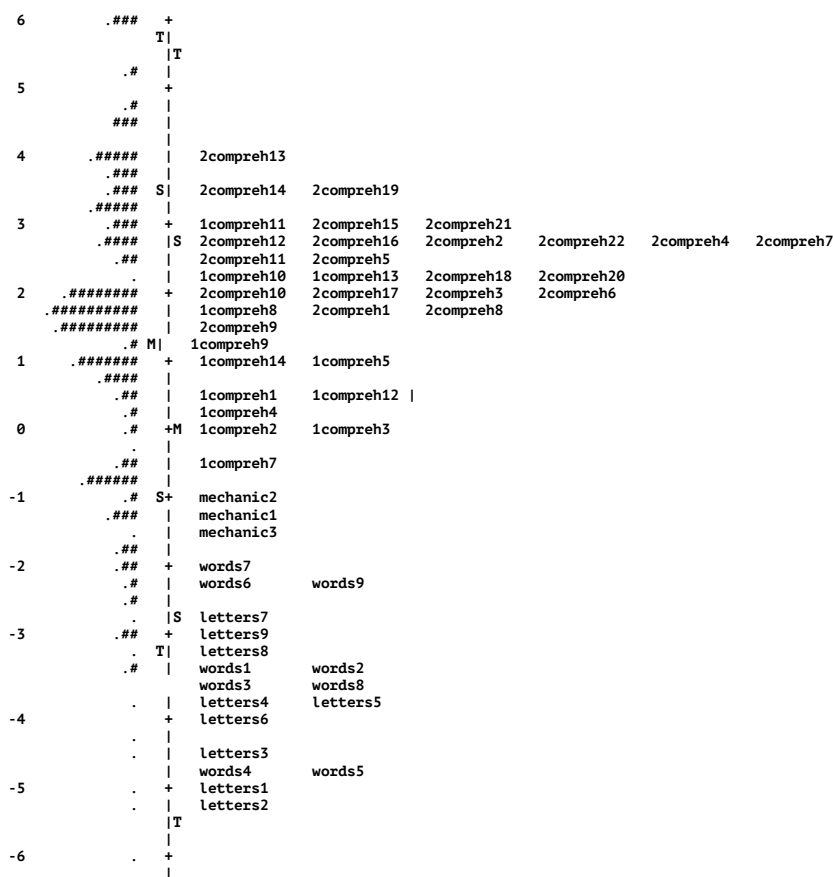
4.2. Collecting evidence of validity

The first iPIPS pilot study was carried out in Veliky Novgorod and Novgorod Oblast in October 2013 on a sample of 300 first-graders. It turned out that a number of tasks had a low ceiling that too many children could reach. Next year, the project participants met with elementary and preschool education experts to discuss the challenges in the study. The tasks in the Russian-language version of iPIPS were reviewed with due consideration of expert advice and in close collaboration with partners from the University of Durham, and a few more evaluation tests followed in 2014–2018. In particular, the tasks that had been too easy were replaced with more challenging ones. As a result, a localized version of the iPIPS instrument was created, for which a series of validity tests was performed.

Instrument validation is an indispensable part of proving the quality of an instrument, and a time-consuming process that requires a substantial amount of research. At the first stage of proving validity of data obtained with iPIPS, evidence of construct validity was collected. For that purpose, the internal structure of the reading scale was examined, its capacity and the psychometric characteristics—based on the assessment results obtained in 2015 on a sample of 1,822 first-graders (average age 7.4, 51% girls) in several schools of Moscow.

Conducting a psychometric analysis of the Russian-language reading scale, we used the basic Rasch model for dichotomous choice assessment instruments [Rasch 1960]. The same model was used to evaluate the original iPIPS version in English [Tymms 1999]. Test data was analyzed using WINSTEPS software [Linacre 2011].

Figure 1. **Reading scale variable map**

```
6        .###   +
                T|
                 |T
          .#     |
5                +
          .#     |
         ###     |
                 |
4        .####   |    2compreh13
         .###    |
         .###   S|    2compreh14    2compreh19
         .####   |
3        .###    +    1compreh11    2compreh15    2compreh21
         .####  |S    2compreh12    2compreh16    2compreh2     2compreh22    2compreh4     2compreh7
         .##     |    2compreh11    2compreh5
          .      |    1compreh10    1compreh13    2compreh18    2compreh20
2     .########  +    2compreh10    2compreh17    2compreh3     2compreh6
    .##########  |    1compreh8     2compreh1     2compreh8
    .#########   |    2compreh9
         .# M|        1compreh9
1     .#######   +    1compreh14    1compreh5
        .####    |
         .##     |    1compreh1     1compreh12 |
         .#      |    1compreh4
0        .#     +M    1compreh2     1compreh3
          .      |
         .##     |    1compreh7
     .######     |
-1       .#    S+    mechanic2
        .###     |    mechanic1
          .      |    mechanic3
         .##     |
-2       .##     +    words7
         .#      |    words6         words9
         .#      |
          .     |S    letters7
-3       .##     +    letters9
          .    T|     letters8
         .#      |    words1         words2
                      words3         words8
          .      |    letters4       letters5
-4               +    letters6
          .      |
          .      |    letters3
                 |    words4         words5
-5        .      +    letters1
          .      |    letters2
                 |T
                 |
-6        .      +
                 |
```

5. Results

5.1. General description of the reading scale

Figure 1 shows the map of variables, depicting the relative position of task difficulty and student ability on the reading scale. At the left end of the figure, the logit scale is shown. On the right-hand side of the vertical axis, the reading tasks are presented, ranging from the easiest (e.g. letter recognition—at the bottom) to the most difficult (e.g. reading comprehension—at the top). At the left side of the axis, the distribution of student ability in the scale is shown.

Empirical data obtained from tests and visualized as a map is consistent with the theoretical expectations of task distribution on the reading scale. In particular, we can see that the letter-naming task is the easiest and the reading comprehension one with semantic "traps" is the most difficult for children. Finally, the map clearly demonstrates a wide distribution of both task difficulty and student ability. Such distribution corresponds to the original English-language version, which outlines prospects for indirect cross-country comparison of student performance using the modern test theory paradigm.

5.2. Data-model fit

Rasch goodness-of-fit test is based on response residuals, i.e. the difference between observed and expected response [Ludlow, Haley 1995]. Adequate fit of data to the model is found in every task except two letter-naming ones (which may be due to their low levels of difficulty).

5.3. Dimensional analysis

Measure unidimensionality, often defined as the existence of one latent trait underlying the data, is one of the principles of the Rasch model for measurement [Hattie 1984]. In this case, reading is measured as a unidimensional construct realized in the array of instrument tasks. Principal component analysis of Rasch model residuals proved the test to be substantially unidimensional [Linacre 2011].

5.4. Reliability analysis

Classical test theory reliability (Cronbach's alpha) of the test is 0.98, which is an extremely high degree of internal consistency. Item spread exceeds 9 logits for task difficulty and 15 logits for student ability.

Therefore, the iPIPS reading test can be considered a quality measure that may be used to assess early reading development in children at the start of school. This inference was reconfirmed in the course of psychometric analysis of the results of assessing first-grade students in Kazan in 2016 [Republican Center for Monitoring the Quality in Education 2016].

6. Discussion

The main purpose of this article was to outline a step-by-step procedure of localizing and validating the Russian-language version of the iPIPS reading test. The iPIPS instrument targets children at the start of school, hence at the start of learning to read.

Reading is an extremely complex skill that is fundamental for school education, shaping the child's overall ability to learn [Antipkina, Kuznetsova, Kardanova 2017; Stanovich 2000]. Reading comprehension is achieved through a series of cognitive processes allowing to analyze lexical (at the level of word recognition) and synthetic (at the level of sentences) information, make inferences, and use metacognitive strategies (self-directed learning, ability to concentrate on the reading process, etc.) [Magliano et al. 2007; Stanovich 2000]. Children master reading comprehension skills stage by stage, from the first acquaintance with a text in their native language, to letter and word recognition, to reading sequences of letters and combining them into words, and, finally, to understanding what they have read. Such stage-by-stage reading acquisition is exactly what the reading model proposed by the iPIPS developers implies [Merrell, Tymms 2004].

The purpose of this study—localization and validation of a Russian-language version of an instrument testing such a complex skill as reading in such a specific target audience as children starting school— is so complicated in its nature that it appears impossible to simply use the existing practices of cross-national and cross-cultural instrument

adaptation, such as The ITC International Handbook of Testing and Assessment [Leong et al. 2016], or the international experience of the PIRLS-participating countries [Mullis et al. 2009].

In a situation where children only begin to start learning to read, adaptation of an international assessment instrument implies using concrete units of Russian and, in the case of iPIPS, English language, which makes it impossible to ensure strict equivalence between the two versions [Ercikan, Roth, Asil 2015]. That is to say, the instrument should be localized with due regard to the distinctive characteristics of the target language, Russian in this case. Assessment of reading comprehension skills is a major challenge, as texts offered to children who are native speakers of Russian and English should be comparable by complexity and the text-related tasks should evaluate the same skills. That is why the iPIPS reading test was localized in stages and involved analyzing the linguistic characteristics of the original tasks in the first place, finding equivalent linguistic means in the Russian language, and, finally, designing Russian-language tasks identical to the original ones in terms of functionality.

Due to a range of substantial structural differences between English and Russian (most importantly, English being verb-centered and Russian noun-centered, 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.

In order 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.

This article describes step by step the procedure of localizing the iPIPS reading test and the process of gathering evidence of its validity within the framework of modern test theory. In particular, it presents the results of analyzing the scale structure and dimensionality, the functioning of individual tasks and the scale as a whole, and internal consistency of scale items. Those procedures have proved psychometric goodness and reliability of the scale and confirmed the compliance of task structure and hierarchy with the theoretical framework of the study.

The stage of collecting construct-related validity evidence described in this article is indispensable yet not the final one in the long process of instrument validation. Additional studies are needed—in particular, to collect content and predictive validity. Series of such studies have already been administered; however, they are left be-

yond the scope of this article and need to be analyzed specifically in the future.

Finally, research proving the fundamental possibility of conducting an indirect (given the non-equivalence of reading assessment instruments between Russian and English and, on the other hand, availability of uniform standardized procedures, theoretical frameworks, and constructs) comparative assessment of first-graders' reading skills in Russia and Great Britain could be regarded as another piece of evidence for instrument validity.

References

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

Antipkina I., Kuznetsova M., Kardanova E. (2017) Chto sposobstvuet i chto meshaet progressu detey v chtenii [What Factors Help and Hinder Children's Progress in Reading?]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 2, pp. 206–233. DOI: 10.17323/1814-9545-2017-2-206-233

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*, vol. 28, no 1, pp. 77–88.

Bezrukikh M., Filippova T., Bajdina V. (2012) Diagnostika razvitiya detey starshego doshkolnogo vozrasta kak sposob rannego vyyavleniya riskov dezadaptatsii [Developmental Diagnostics of Elder Pre-School Children as the Means of Early Detection of Dysadaptation Risks]. *Novye issledovaniya*, no 1 (30), pp. 145–157.

Bolotov V., Valdman I., Kovaleva G., Pinskaya M. (2013) Rossiyskaya Sistema otsenki kachestva obrazovaniya: glavnye uroki [Russian Quality Assessment System in Education: Key Lessons]. *Education Quality in Eurazia*, no 1, pp. 85–121.

Clay M.W. (1985) *The Early Detection of Reading Difficulties*. Auckland, New Zealand: Heinemann.

Demetriou A., Merrell C., Tymms P. (2017) Mapping and Predicting Literacy and Reasoning Skills from Early to Later Primary School. *Learning and Individual Differences*, vol. 54, February 2017, pp. 217–225.

Egorov T. (2006) *Psikhologiya ovladeniya navykom chteniya* [The Psychology of Learning to Read], St. Petersburg: KARO.

Ercikan K., Roth W.M., Asil M. (2015) Cautions about Inferences from International Assessments: The Case of PISA 2009. *Teachers College Record*, vol. 117, no 1, pp. 1–28.

Esselink B. (2000) *A Practical Guide to Localization*. *Vol. 4.* Amsterdam, Philadelphia: John Benjamins.

Fillmore Ch. (1981) Delo o padezhe [The Case for Case]. *New in Foreign Linguistics* (ed. V. Zvegintsev), Moscow: Progress, iss. 10, pp. 369–495.

Foulin J.N. (2005) Why Is Letter-Name Knowledge Such a Good Predictor of Learning to Read? *Reading and Writing*, vol. 18, no 2, pp. 129–155.

Gove A., Wetterberg A. (2011) *The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy.* Research Triangle Park, NC: RTI International.

Harms T. (2013) *School-Age Care Environment Rating Scale (SACERS)*. New York, NY: Teachers College.

Harrison C. (2004) *Understanding Reading Development*. London: Sage.

Hattie J. (1984) An Empirical Study of the Various Indices for Determining Unidimensionality. *Multivariate Behavioral Research*, vol. 19, no 1, pp. 49–78.

Ivanova A. (2018) Problema sopostavimosti rezultatov v mezhdunarodnykh sravnitelnykh issledovaniyakh obrazovatelnykh dostizheniy [Problem of Comparability of Results in International Comparative Studies of Educational Achievements]. *Otechestvennaya i zarubezhnaya pedagogika*, vol. 1, no 2, pp. 68–81.

Janus M., Brinkman S.A., Duku E. K. (2011) Validity and Psychometric Properties of the Early Development Instrument in Canada, Australia, United States, and Jamaica. *Social Indicators Research*, vol.103, no 2, pp. 283–297.

Kornev A. (1995) *Disleksiya i disgrafiya u detey* [Dyslexia and Dysgraphia in Children], St. Petersburg: Gippokrat.

Kovaleva G., Danilenko O., Ermakova I., Nurminskaya N., Gaponova N., Davydova E. (2012) O pervoklassnikakh: po rezultatam issledovaniy gotovnosti pervoklassnikov k obucheniyu v shkole [On First-Graders Based on the Results of First-Grade Readiness Assessments]. *Munitsipalnoe obrazovanie: innovatsii i eksperiment*, no 5, pp. 30–37.

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*, vol. 45, no 3, pp. 640–661.

Leong F.T., Bartram D., Cheung F., Geisinger K.F., Iliescu D. (2016) *The ITC International Handbook of Testing and Assessment*. New York, NY: Oxford University.

Linacre J.M. (2011) *Winsteps (Version 3.73) (computer software)*. Chicago, IL: Winsteps. com.

Ludlow L.H, Haley S.M. (1995) Rasch Model Logits: Interpretation, Use, and Transformation. *Educational and Psychological Measurement*, vol. 55, no 6, pp. 967–975.

Magliano J.P., Millis K., Ozuru Y., McNamara D.S. (2007) A Multidimensional Framework to Evaluate Reading Assessment Tools. *Reading Comprehension Strategies: Theories, Interventions, and Technologies* (ed. D.S. McNamara), New York: Lawrence Erlbaum Associates, pp. 107–136.

Merrell C., Tymms P. (2004) *Diagnosing and Remediating Literacy Problems Using INCAS Software: Identifying Reading and Spelling Difficulties and Providing Help*. Available at: http://www.pipsproject.org/Documents/CEM/publications/downloads/CEMWeb039%20Incas.pdf (accessed 1 November 2019).

Merrell C., Tymms P. (2007) Identifying Reading Problems with Computer-Adaptive Assessments. *Journal of Computer Assisted Learning*, vol. 23, no1, pp. 27–35.

Mislevy R.J., Steinberg L.S., Almond R.G. (2003) Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives*, no 1, pp. 3–62.

Mullis I.V., Martin M.O., Kennedy A.M., Trong K.L., Sainsbury M. (2009) *PIRLS 2011 Assessment Framework.* Amsterdam, Netherlands: International Association for the Evaluation of Educational Achievement.

Niklas F., Schneider W. (2013) Home Literacy Environment and the Beginning of Reading and Spelling. *Contemporary Educational Psychology*, vol. 38, no 1, pp. 40–50.

Rasch G. (1960) *Studies in Mathematical Psychology: I. Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.

Rayner K., Pollatsek A., Ashby J., Clifton Jr C. (2012) *Psychology of Reading*. New York, NY: Psychology Press.

Republican Center for Monitoring the Quality in Education (2016) *Materialy o monitoringe v 1-kh klassakh iPIPS. Otchet. Vesna 2016* [Materials on iPIPS First-Grade Monitoring. Spring 2016 Report]. Available at: http://rcmko.ru/ meropriyatiya/monitoringi/ipips/materialy-o-monitoringe-v-1-klassah-ipips/ (accessed 1 November 2019).

Rutkowski L., Gonzalez E., Joncas M., von Davier M. (2010) International Large-scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher*, vol. 39, no 2, pp. 142–151.

Seymour P.H., Aro M., Erskine J.M. (2003) Foundation Literacy Acquisition in European Orthographies. *British Journal of Psychology*, vol. 94, no 2, pp. 143–174.

Sireci S.G., Patsula L., Hambleton R.K. (2005) *Statistical Methods for Identifying Flaws in the Test Adaptation Process. Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. New Jersey: Lawrence Erlbaum Associates.

Stanovich K.E. (2000) *Progress in Understanding Reading: Scientific Foundations and New Frontiers*. New York: Guilford.

Tymms P. (1999) Baseline Assessment, Value-Added and the Prediction of Reading. *Journal of Research in Reading*, vol. 22, no 1, pp. 27–36.

Tymms P., Jones P., Albone S., Henderson B. (2009) The First Seven Years at School. *Educational Assessment, Evaluation and Accountability*, vol. 21, no 1, pp. 67–80.

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*. Available at: http://dro.dur.ac.uk/23562/1/23562.pdf (accessed 1 November 2019).

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*, vol. 42, pp. 11–20.

Wildy H., Styles I. (2011) Measuring What High-Achieving Students Know and Can Do on Entry to School: PIPS 2002–2008. *Australasian Journal of Early Childhood*, vol. 36, no 2, pp. 51–62.

Zhurova L., Elkonin D. (1963) K voprosu o formirovanii fonematicheskogo vospriyatiya u detey doshkolnogo vozrasta [On the Development of Phonemic Perception in Preschoolers]. *Sensornoe vospitanie doshkolnikov* [Sensory Development of Preschoolers] (eds A. Zaporozhets, A. Usova), Moscow: RSFSR Academy of Pedagogical Sciences, pp. 213–227.