Checking the Possibility of an International Comparative Study of Reading Literacy Assessment for Children Starting School

A. Ivanova, E. Kardanova

Received in May 2020

Alina Ivanova
Research Fellow, Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Email: aeivanova@hse.ru
Elena Kardanova
Candidate of Sciences in Mathematical Physics, Associate Professor, Tenured Professor, Director of the Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Email: ekardanova@hse.ru

Address: Bld. 10, 16 Potapovsky Lane, 101000 Moscow, Russian Federation

Abstract. The early years of school, when a child is only learning to read, are critically important for later development and learning. Cross-cultural assessments of reading literacy provide a rich source of data for researchers, practitioners and policymakers on the opportunities and prospects of early childhood development in different countries, circumstances and contexts. There are few publications of this sort available, and none of them has involved Russian-speaking children on entry to school so far.

Data obtained using two language versions of the International Performance Indicators in Primary Schools (iPIPS) on representative samples of first-graders from the Republic of Tatarstan and Scotland is used to compare the early reading assessment results between children starting school in countries with linguistic, cultural, and school entry age differences.

Two studies are conducted to analyze the possible methods of comparing assessment results of children from different countries in the absence of a uniform measurement scale. Study 1 uses the rank-ordering method to establish a correspondence between the levels of reading development among Russian- and English-speaking children by expert judgment. In Study 2, the constructed model of literacy levels is used to set the benchmarks of student assessment results.

Keywords: cross-cultural assessment (CCA), elementary school, expert judgment, paired comparison, Rasch modelling.

DOI: 10.17323/1814-9545-2020-4-8-36

Translated from Russian by I. Zhuchkova.

The role of large-scale national and international assessments (ILSA) has been increasing in developmental research. Recent ILSA allow verifying, adjusting and improving the existing theories of human development [Peña 2007; Shuttleworth-Edwards et al. 2004], providing an important source of data on predictors of student achievement in different countries, circumstances and sociocultural contexts [Ainley, Ainley 2019; Carnoy et al. 2016; Caro, Cortés 2012].

The worldwide interest in ILSA is reflected in their number growing rapidly since the early 2000s. For instance, the number of countries participating in the Programme for International Student Assessment (PISA) increased from 43 in 2000 to 80 in 2018 [Liu, Steiner-Khamsi 2020]. Researchers point out that more and more governments internalize the logic of ILSA in their national education policies by attempting to make learning measurable, comparable and accountable [Espeland 2015; Liu, Steiner-Khamsi 2020].

The results of international assessments in preschool and early school education, in particular those measuring early reading literacy, are of major significance to educational researchers and policymakers. Indeed, the role of reading literacy in today's world keeps increasing, and the early years are critical for further reading development. Besides, there is always a demand for rational spending in education. Finally, researchers and policymakers seek to make evidence-based decisions by studying the experience and best practices of other countries [Suggate 2009]. Although each country develops and implements their own education goals and programs, it needs external, international benchmarks and information about new opportunities and prospects for early childhood development [Buzhardt et al. 2019].

The use of research and comparative analysis to enhance education policy is only possible under the condition of reliable and valid measurements in ILSA. Adaptation plays a key role in ensuring valid interpretation of assessment results obtained with instrument versions designed for different countries, languages or cultures. Research institutions administering assessments offer recommendations concerning the procedures intended to provide a high quality of adaptation in ILSA [American Educational Research Association, American Psychological Association, National Council on Measurement in Education 2014; Leong et al. 2016]. The goal of those procedures is to achieve the highest possible level of measurement comparability, which is indispensable for further use of the assessment results.

International comparability of ILSA results is only possible if measurements performed with instruments designed in different languages are equivalent. The concept of measurement equivalence suggests ensuring and empirically validating (i) construct equivalence, (ii) equivalence of instruments, and (iii) equivalence of procedures [Ercikan, Roth, Asil 2015]. Therefore, to minimize possible cultural and linguistic bias in results, the ILSA procedures and methods of instrument design and results validation should guarantee that assessment of relevant behavior (skill, competency or any other construct) is not affected by other variables (nationality or ethnicity, socioeconomic status, etc.).

Instruments designed to measure reading skills are especially difficult to adapt to languages of other countries and cultures. Even the most reputable assessments, such as the Progress in International Reading Literacy Study (PIRLS) and the Programme for International Student Assessment (PISA), face the issue of incomparability across language versions [Goodrich, Ercikan 2019; Kreiner, Christensen 2014]. Furthermore, very few instruments allow for cross-country comparison of emergent literacy.

PIRLS measures reading achievement in elementary school graduates who already know how to read. In the recently initiated first round of the OECD's International Early Learning and Child Well-Being Study (IELS) assessing children on entry to elementary school, which involves only three countries so far, emergent literacy is measured through listening comprehension, vocabulary and phonological awareness [OECD2020]. The Early Grade Reading Assessment (EGRA), another well-known international project measuring reading acquisition of elementary school students, is not designed to make any cross-country comparisons. Developed in English and adapted to other languages, the instrument is only used at national levels [Dubeck, Gove 2015]. Researchers believe that emergent literacy is hard to measure across cultures because the influence of language in assessment is too strong when children make their very first steps in learning to read [Ercikan, Roth, Asil 2015].

This study attempts to compare reading literacy in children starting school in Russia and the UK using the International Performance Indicators in Primary Schools (iPIPS) test [Tymms 1999]. Originally designed in English, iPIPS is currently applied in a variety of countries, including not only the English-speaking Australia and New Zealand but also, for example, Germany, Brazil and Russia [Bartholo et al. 2019; Kardanova et al. 2018, Tymms et al. 2014; Vidmar et al. 2017].

When developing the Russian-language version of iPIPS, it became obvious that some part of the instrument was inadaptable and had to be localized. Localization involves taking a product and making it linguistically and culturally appropriate to the target locale (country, region, etc.) [Esselink 2000]. The main difference between localization and adaptation is that the former does not imply cross-country comparison of assessment results.

The need for localization was dictated by the essential structural differences between the English and Russian languages, the most important being the verb-centered nature of English and the noun-centered nature of Russian, categorical and functional mismatches in parts of speech between the languages, fixed word order in English, and a number of others. Because of those gaps, the stages of language development do not coincide for English- and Russian-speaking children [Ivanova, Kardanova-Biryukova 2019], which undoubtedly affects the process of reading acquisition and assessment.

In an earlier study [2019], we described the procedures used for localizing the Russian-language version of the iPIPS reading test, demonstrated the means of ensuring equivalence of construct—emergent reading literacy on school entry—at the stage of localization design, substantiated the impossibility of achieving full measurement equivalence, and described the procedures of collecting evidence of construct validity.

The first step towards creating a Russian-language version of iPIPS consisted in translation and expert evaluation of items designed to assess the basic reading skills of British children. Translation (forward and backward) was performed in compliance with the guidelines of the International Test Commission [Leong et al. 2016]. The iPIPS reading test was comprised of a few modules corresponding to the stages of reading development in the iPIPS theoretical model: text structure awareness, letter identification, word recognition, and reading/decoding automaticity.

Items assessing reading skills were fairly easy to localize for Russian-speaking children. Meanwhile, localization of the reading comprehension module turned out to be a challenge. This module included large narrative texts with hidden "traps"—gaps to be filled in by choosing one of the three words suggested. The "traps" targeted different aspects of language—spelling, grammar, phonology and semantics. Since the texts offered to Russian- and English-speaking children had to be of comparable difficulty and the "traps" had to evaluate the same competencies, much more effort was spent localizing this module.

Specifically, the process involved first analyzing the linguistic characteristics of the original text, then finding equivalent "traps" in Russian and, finally, producing a Russian-language text with "traps" and content close to the English-language original.

Although localization of an international instrument into Russian does not imply direct comparison of student outcomes, there is a demand for comparing children's basic skills on school entry, which can be satisfied by indirect comparison, understood here as comparison of assessment results in a cross-cultural context at the level of groups, not individuals.

This article aims at exploring the possibility of a cross-country assessment of emergent literacy in children starting school in Russia and the UK. Two studies are conducted for this purpose. Study 1 uses the rank-ordering method to establish a correspondence between the levels of reading development among Russian- and English-speaking children by expert judgment. In Study 2, the constructed model of literacy levels is used to establish the benchmarks of student assessment results in the two jurisdictions. This series of studies will render possible comparison of first-graders' reading test results between Russia and the UK for the first time.

1. Study 1. The Rank-Ordering Method as a Basis for Comparison

Expert evaluation of the construct—the model of literacy levels in Russian- and English-speaking children—was performed as part of analysis of comparative assessment opportunities using the rank-ordering method as a methodological framework.

This method was applied in a study comparing the raw marks of two tests for 14-year-old students in England [Bramley 2005] on different cohorts and different panels of experts. It represents a combination of expert judgment and mathematical modelling of judgmental data and allows comparing the test results between different versions of the instrument where no full measurement equivalence can be achieved.

The goal of Study 1 is to show how expert judgments can be used for building a reading literacy scale, identifying the item-difficulty hierarchy for the construct measured, comparing item hierarchies between two language versions and establishing the benchmarks.

1.1. Methodology of Study 1

1.1.1. Item Selection

The study uses the original English-language version of the 44-item iPIPS reading test [Merrell, Tymms 2007] with 18 “traps” in the reading comprehension module and the localized 40-item Russian-language version of the instrument [Ivanova, Kardanova-Biryukova 2019] containing 14 reading comprehension “traps”.

1.1.2. Participants

The study was assisted by twelve experts speaking fluent English, including teachers and professors of English, linguists and philologists, all with at least a Master’s degree and with two to over ten years of professional experience, of whom one was male and eleven were females.

1.1.3. Procedure

Prior consent to participate in the study was obtained from all experts. Each of them was given two packs with versions of the test in Russian and English, judgment instructions, and a short questionnaire.

Every item was represented as a picture on a separate sheet of paper with instructions that were given to children starting school in Russia and the UK and a short notice of what the item measured. Instructions for experts also contained information on the testing procedure. In addition, experts were provided a link to the video demonstrating the testing process.

The experts were notified that the items were randomly intermixed within their language-specific packs. They were asked to rank the items from the easiest to the most difficult within each pack by giving their personal holistic assessment of item difficulty based on their expert knowledge and experience. They were also asked to allow a minimum of two days between evaluations of the Russian- and English-language versions.

1.1.4. Analytical Approach

Rank-ordered data have characteristics that are in line with the family of Rasch measurement models. Ranks are observations of elements implying qualitatively more, ordered along an implicit or explicit variable [Linacre 2006]. A single set of ranks, called a "ranking", contains enough information to order the elements. If there are two or more rankings of the same elements, then there may be enough information to construct interval measures of the distances between the elements. Interval measures allow assessing the difficulty of every item by fixing a zero point (e. g. at the mean difficulty of all items, as in Rasch modelling), which means that they support inferences about measurement results and investigation into the consistency of particular rankings [Linacre 2006].

John M. Linacre [Linacre 1989; 2006] developed two approaches to modelling ranked-ordered data based on the method of paired comparison proposed by Louis L. Thurstone [Thurstone 1927]:

1) Decomposing the rank orderings into paired comparisons, e. g. a rank-ordering of 10 objects yields 45 paired comparisons for analysis: 1st against 2nd, 1st against 3rd, etc.;

2) Modelling each ordering as a partial-credit item.

The dataset for subsequent analysis includes 1,008 observations: 12 experts evaluated a pack of 44 items, plus 12 experts evaluated a pack of 40 items. For each expert, ranking of the items within each language version of the instrument represents a set of ranks. Both approaches proposed by Linacre are used to build an early reading development scale based on expert judgments.

Let us dwell on the method of paired comparison first. In the simplest case where items are ranked by paired comparisons, items are compared in pairs and ordered according to their ranking. In each ordering, any particular item is ranked higher or lower than any other particular item. What is decisive is the number of times one item is ranked higher than another [Linacre 1989]. Otherwise speaking, item n with measure B_n might be ranked HIGHER than item m with measure B_m a total of H times across the orderings made by the different experts. In contrast, item n might be ranked LOWER than m a total of L times. The ratio H/L is the essential data for the estimation of a distance between items n and m as in $(B_n - B_m)$.

A measurement model for rank orders is

$$Ln\ (P_{nm} / P_{mn}) = B_n - B_m\,,$$

where P_{nm} is the probability that n is ranked higher than m, P_{mn} is the probability that m is ranked higher than n, and $P_{mn} + P_{nm} = 1$.

The ratio P_{nm} / P_{mn} becomes the empirical data for estimating the parameters. For rankings of more than two items, there are added constraints because items are not compared independently, but are reported in a composite rank-order.

In the model proposed by Linacre for Thurstone's method of paired comparison (hereinafter TM—for Thurstone Model), rank orderings are decomposed into paired comparisons. A measurement model for this conceptualization is

$$Ln\ (P_{nrk}\ /\ P_{nrk+1}) = B_n - B_r - F_{rk},$$

where P_{nrk} is the probability that, in ordering r, item n will be ranked k, P_{nrk+1} is the probability that, in ordering r, item n will be ranked $k + 1$, B_n is the difficulty of item n, B_r is the mean difficulty of the items included in ordering r, F_{rk} is the step difficulty up from a ranking of k+1 to a ranking of k within ordering r.

A delight of this measurement model is that it doesn't matter, in general, how many experts include each item in their rankings, or how many items each expert ranks. The estimates of the measures are derived merely from counting each item's location in each ordering [Linacre 1989].

The other approach suggests modelling each ordering as a polytomously-scored response, where the number of response options corresponds to the number of ranks assigned in an ordering. Analytically, this is implemented as follows: 12 expert rankings (by the number of orderings made by 12 experts) will represent "items", and actual items in the English- and Russian-language versions of the instrument will be treated as "persons". This is where the Partial Credit Model (PCM) [Masters 1982] can be applied.

This approach was used by Tom Bramley [Bramley 2005], who fitted the PCM model specifically for equating tests by expert judgment:

$$Ln\ (P_{nrk}\ /\ P_{nr(k+1)}) = B_n - D_{rk},$$

where P_{nrk} is the probability that item n is ranked at position k in ranking r; $P_{nr(k+1)}$ is the probability that item n is ranked at position $k + 1$ in ranking r; B_n is the difficulty of item *n;* and D_{rk} is the difficulty of reaching the scale category k relative to category $k + 1$ in ranking *r.*

Analysis of our data was performed in the logic of both approaches using FACETS [Linacre, Wright 1994] and Winsteps [Linacre 2011] software, respectively.

1.2. Results of Study 1

Agreement among the judgments of all experts was analyzed to ensure sufficient reliability of the data collected. Kendall's coefficient of concordance, a classic measure of agreement among raters [Field 2014], was 0.84 for the Russian-language version and 0.87 for the English-language one. Consequently, there is a high agreement among the estimates of item difficulty made by the experts.

Next, the results of rank-order judgments were presented separately for each language version of the instrument within the two models described above, TM and PCM. Data analysis yielded similar results with both approaches. A summary of item analysis results is given in Table 1. The standard error for item difficulty (Model S. E.)

Table 1. **General scale parameters**

Model	Measures of quality			
	Goodness-of-fit statistics			
	INFIT MNSQ	OUTFIT MNSQ	Model S.E	Range of Measures
Russian-language version				
TM	1.0	0.6	0.16	9.86
PCM	1.3	1.3	0.07	2.99
English-language version				
TM	1.0	0.9	0.16	11.24
PCM	1.04	1.04	0.07	3.41

is rather small (especially for PCM). At the same time, the range of item difficulties (Range of measures), according to expert judgments, is much wider for TM. Goodness-of-fit statistics, designated as INFIT and OUTFIT MNSQ, serve as indicators of how well the chosen measurement model predicts the dataset and represent root-mean-square deviations of empirical values from those predicted by the model for each rank. As can be seen from Table 1, mean-squares of the goodness-of-fit statistics fall within the range [0.6; 1.4] recommended by psychometricians [Linacre 2011].

Table 2 shows data broken by items, namely the level of item difficulty in both language versions according to expert judgments, the standard error of item difficulty estimate, and goodness-of-fit statistics showing how well the data fits the measurement model.

It can be inferred from Table 2 that the data basically fits both models for both language versions. There is a very high correlation between the levels of item difficulty in each language version analyzed with different approaches, Pearson's correlation coefficient between the TM and PCM judgments being 0.95 ($p < 0.05$) for the Russian-language version and 0.96 ($p < 0.05$) for the English-language one.

It is convenient to illustrate the item-difficulty hierarchy using variable maps for the two versions of the instrument shown in Figures 1 and 2. The maps are built using the PCM approach (the ones based on TM have a similar appearance). The easiest items (letter identification tasks) can be found at the bottom; the items assessing text structure awareness and word recognition skills are just above; the middle part of the scale (around 0 logits) displays the items measuring reading/decoding automaticity; finally, the most difficult tasks for reading comprehension are at the top.

Both maps feature item clustering at the top, middle and bottom of the scale. Moreover, the distances on the continuum between the boundary items of the top and middle clusters as well as the middle

Table 2. **Expert judgments of item difficulty for two language versions**

		Russian-language version						English-language version					
		PCM			TM			PCM			TM		
Item	Task description	MEASURE	S.E	INFIT	MEASURE	S.E	INFIT	MEASURE	S.E	INFIT	MEASURE	S.E	INFIT
Task01	Text structure 1	–1.5	0.2	2.0	–2.6	0.1	1.4	–1.8	0.2	1.7	–3.4	0.2	1.5
Task02	Text structure 2	–1.0	0.1	0.7	–1.9	0.1	1.3	–1.5	0.1	0.6	–2.5	0.2	1.4
Task03	Text structure 3	–1.2	0.1	1.5	–2.6	0.1	1.5	–1.1	0.1	3.1	–2.6	0.1	1.5
Task04	Text structure 4	–1.0	0.1	0.3	–2.2	0.1	1.3	–1.4	0.1	0.4	–3.1	0.1	1.4
Task05	Text structure 5	–1.0	0.1	0.7	–1.9	0.1	1.3	–1.4	0.1	0.8	–2.6	0.1	1.5
Task06	Letters 1	–1.2	0.1	0.5	–4.4	0.2	0.9	–1.5	0.1	0.6	–6.2	0.2	1.2
Task07	Letters 12	–1.0	0.1	0.3	–4.2	0.2	1.1	–1.3	0.1	0.2	–4.7	0.2	0.9
Task08	Letters 13	–0.9	0.1	0.3	–5.2	0.2	1.1	–1.3	0.1	0.6	–5.3	0.2	1.0
Task09	Letters 14	–0.9	0.1	0.7	–3.4	0.2	0.9	–1.2	0.1	0.6	–5.1	0.2	1.0
Task10	Letters 15	–0.8	0.1	0.5	–3.5	0.2	0.8	–1.2	0.1	0.4	–4.3	0.2	0.8
Task11	Letters 16	–0.6	0.1	1.1	–3.5	0.2	0.9	–1.1	0.1	0.2	–4.7	0.2	0.9
Task12	Letters 17	–0.8	0.1	0.4	–3.2	0.2	0.7	–1.1	0.1	0.2	–4.3	0.2	0.8
Task13	Letters 18	–0.7	0.1	0.5	–3.0	0.1	0.7	–1.0	0.1	0.4	–3.9	0.2	0.7
Task14	Letters 19	–0.7	0.1	0.3	–3.0	0.1	0.8	–1.1	0.1	2.3	–4.0	0.2	0.7
Task15	Words 1	–0.6	0.1	0.8	–1.8	0.1	0.8	–1.0	0.1	2.8	–3.0	0.1	0.6
Task16	Words 2	–0.6	0.1	0.8	–1.2	0.2	0.9	–1.0	0.1	1.9	–2.0	0.2	0.8
Task17	Words 3	–0.6	0.1	0.2	–1.0	0.2	0.9	–1.0	0.1	2.7	–1.8	0.2	0.8
Task18	Words 4	–0.5	0.1	1.5	–2.0	0.1	0.7	–0.8	0.1	0.5	–3.0	0.1	0.6
Task19	Words 5	–0.8	0.1	1.4	–1.5	0.1	0.9	–0.9	0.1	0.3	–2.1	0.2	0.8
Task20	Words 6	–0.9	0.1	3.1	–1.0	0.2	0.9	–0.8	0.1	0.6	–2.5	0.2	0.7
Task21	Words 7	–0.9	0.1	2.4	–0.5	0.2	0.9	–0.8	0.1	0.5	–1.1	0.2	0.9
Task22	Words 8	–0.8	0.1	2.3	–2.0	0.1	0.8	–0.9	0.1	0.5	–1.5	0.2	0.9
Task23	Words 9	–0.8	0.1	2.5	–0.9	0.2	0.9	–0.7	0.1	1.1	–1.7	0.2	0.9
Task24	Story, Part 1	0.4	0.1	6.3	1.0	0.2	1.1	0.1	0.2	0.8	0.7	0.2	0.6
Task25	Story, Part 2	0.7	0.1	5.1	1.5	0.2	1.2	0.2	0.2	0.5	0.7	0.2	0.5
Task26	Story, Part 3	0.8	0.1	3.4	1.9	0.2	1.1	0.2	0.2	1.4	0.5	0.2	0.6
Task27	"Trap" text 1	1.1	0.1	1.1	2.7	0.2	1.0	1.3	0.1	1.5	4.2	0.2	1.0
Task28	"Trap" text 2	1.2	0.1	0.7	4.0	0.2	1.0	1.0	0.1	1.4	3.4	0.2	1.4
Task29	"Trap" text 3	1.3	0.1	1.1	3.9	0.2	0.9	1.3	0.1	3.0	4.6	0.2	0.9
Task30	"Trap" text 4	1.4	0.1	0.9	3.6	0.2	0.9	1.4	0.1	0.7	4.2	0.2	1.0
Task31	"Trap" text 5	1.2	0.1	1.1	3.3	0.2	1.0	1.4	0.1	1.3	4.3	0.2	1.0
Task32	"Trap" text 6	1.4	0.1	0.8	4.1	0.2	1.0	1.2	0.1	0.6	4.4	0.2	1.0
Task33	"Trap" text 7	1.3	0.1	1.0	3.4	0.2	0.9	1.5	0.1	0.5	4.1	0.2	0.9
Task34	"Trap" text 8	1.3	0.1	1.1	3.1	0.2	0.9	1.5	0.1	0.8	5.0	0.2	1.0
Task35	"Trap" text 9	1.5	0.1	0.6	3.8	0.2	1.0	1.3	0.1	1.4	3.0	0.2	0.8
Task36	"Trap" text 10	1.4	0.1	0.9	3.6	0.2	0.9	1.5	0.1	1.0	4.4	0.2	1.0
Task37	"Trap" text 11	1.3	0.1	0.8	4.2	0.2	1.0	1.2	0.1	2.2	4.5	0.2	1.0
Task38	"Trap" text 12	1.1	0.1	1.5	3.3	0.2	1.0	1.5	0.1	0.7	2.4	0.2	0.8
Task39	"Trap" text 13	1.2	0.1	0.8	4.6	0.2	1.0	1.4	0.1	0.9	4.7	0.2	1.0
Task40	"Trap" text 14	1.3	0.1	0.7	4.1	0.2	1.0	1.5	0.1	0.8	4.8	0.2	1.0
Task41	"Trap" text 15							1.4	0.1	0.9	3.6	0.2	1.0
Task42	"Trap" text 16							1.6	0.1	0.4	3.1	0.2	1.3
Task43	"Trap" text 17							1.6	0.1	0.9	5.0	0.2	1.0
Task44	"Trap" text 18							1.5	0.1	1.1	3.5	0.2	1.0

Note: In FACETS, estimates of item difficulty are presented as those of item "easiness".
For presentation, they were converted into estimates of item difficulty.
Item difficulty is measured in logits—specific units of measurement on a log-odds scale adopted in Item Response Theory.

Table 3. **Differences in item difficulty for grouping**

Difference in difficulty between boundary items in logits	Russian-language version	English-language version
Task24—Task23	1.9	2.5
Task26—Task25	0.8	3.7

and lower ones are large enough (Table 3), falling significantly outside the range of two standard errors, which means that the clusters represent three major groups of tasks corresponding to different levels of reading development.

Therefore, expert judgments of item-difficulty hierarchy for two language versions of the iPIPS instrument were obtained in the course of this study. A psychometric analysis of expert judgments using two measurement approaches allowed distinguishing among three clusters of items differing within the range of two standard errors by difficulty for both language versions of the iPIPS reading assessment test, represented by the same groups of items in both Russian and English.

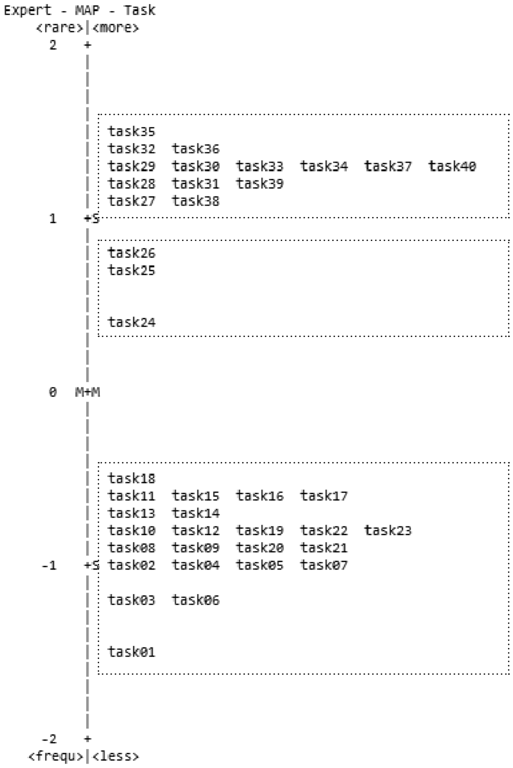
The first cluster (at the bottom of the map in Figures 1 and 2) includes the easiest test items measuring text structure awareness as well as letter identification and word recognition skills. This cluster corresponds to the earliest stage of reading development. The second cluster (middle part of the map) includes items of medium difficulty, which assess reading/decoding automaticity in both language versions. Finally, the third cluster (top of the map) is comprised of reading comprehension tasks.

The identified clusters of items constitute empirical evidence on applicability of the theoretical model of reading development underlying the original version of iPIPS to the Russian-language version. The clusters can be used as a uniform basis for setting comparable benchmarks on the reading scales of the two language versions of the instrument.

2. Study 2. Establishing the Benchmarks for Comparing Samples of Russian- and English-Speaking Students by Reading Development

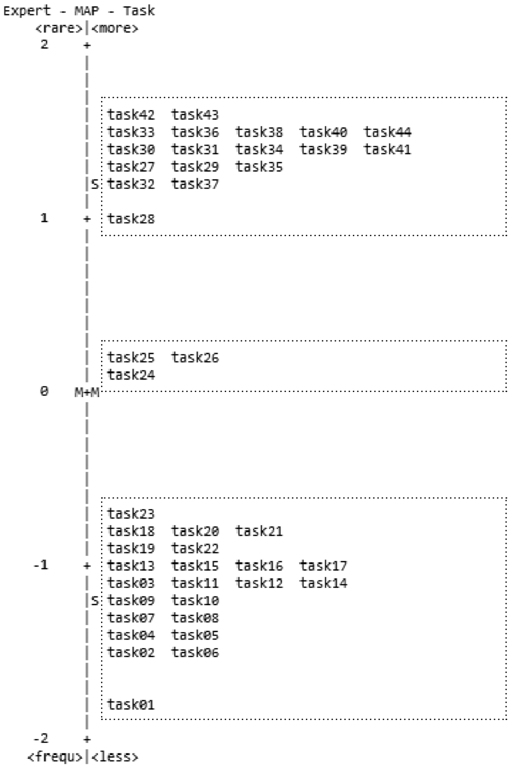
It was impossible to achieve full reading measurement equivalence in the course of adapting/localizing the iPIPS instrument into Russian [Ivanova, Kardanova-Biryukova 2019], which makes it impossible to compare student performance between the two countries directly. Nevertheless, using the actual results of tests administered under equivalent procedures in the two countries and having confirmed the possibility of using a uniform model of early reading literacy levels (Study 1), one may attempt to carry out an indirect cross-country comparison of student achievements. In particular, the principles of

Figure 1. **Variable map for the Russian-language version.** *Partial Credit Model. Winsteps*



```
Expert - MAP - Task
    <rare>|<more>
  2    +
       |
       |
       |
       |
       | task35
       | task32  task36
       | task29  task30  task33  task34  task37  task40
       | task28  task31  task39
       | task27  task38
  1   +S
       |
       | task26
       | task25
       |
       |
       | task24
       |
       |
       |
  0   M+M
       |
       |
       |
       |
       | task18
       | task11  task15  task16  task17
       | task13  task14
       | task10  task12  task19  task22  task23
       | task08  task09  task20  task21
 -1   +S task02  task04  task05  task07
       |
       | task03  task06
       |
       |
       | task01
       |
       |
       |
       |
 -2    +
  <frequ>|<less>
```

Figure 2. **Variable map for the English-language version.** *Partial Credit Model. Winsteps*



```
Expert - MAP - Task
    <rare>|<more>
  2    +
       |
       |
       |
       | task42  task43
       | task33  task36  task38  task40  task44
       | task30  task31  task34  task39  task41
       | task27  task29  task35
       |S task32  task37
       |
  1    + task28
       |
       |
       |
       |
       |
       |
       | task25  task26
       | task24
  0   M+M
       |
       |
       |
       |
       |
       |
       | task23
       | task18  task20  task21
       | task19  task22
 -1    + task13  task15  task16  task17
       | task03  task11  task12  task14
       |S task09  task10
       | task07  task08
       | task04  task05
       | task02  task06
       |
       |
       | task01
       |
 -2    +
  <frequ>|<less>
```

Rasch measurement models allow comparing samples of students in different countries according to the level of reading development.

The study uses empirical data obtained from samples of first-graders on entry to school in the Republic of Tatarstan and Scotland as reference data for the Russian- and English-language versions, respectively. The items used below in this article have been described in Study 1.

2.1. Methodology of Study 2

2.1.1 Sampling

In Russia, all the necessary data for sampling were collected in cooperation with the Republican Center of Education Quality Monitoring of the Republic of Tatarstan in 2017. A representative sample of over 5,000 children (44% of total population) was produced. For this sample, total population is understood as all first-graders in the selected regions of Tatarstan. The sample was stratified by school type and location. Classes of students selected randomly from the cohort of first-graders of a particular participating school served as sampling units. The study only involved children whose parents had given their consent to participation. The total sample of first-graders whose results will be used for analysis below was comprised of 4,940 students aged on average 7.4 on school entry.

The sample of children from the UK in this study consisted of 6,627 Scottish students, which was a representative sample for Scotland (data for the 2014/15 academic year) [Tymms, Merrell, Buckley 2015]. The average age of children on entry to school was 5 years, although a certain percentage of older children was observed.

2.1.2. Analytical Approach to Comparative Assessment

Based on Georg Rasch's principles of measurement, we suggest treating the test results as a continuum of reading development, some kind of a "path" that leads toward reading comprehension, while drawing on the iPIPS theory of reading development and the model of literacy levels constructed as a result of analyzing expert estimates of item difficulty in Study 1.

Using the scale of expert judgments of early reading assessment item difficulty, transformed in the course of modelling into a logit scale, we identified three clusters of items, or three theoretically interpretable stages of reading development confirmed by expert judgments. The model proposed here allows distinguishing among the levels of reading development, from zero level where children make their very first steps in learning to read, to the advanced level of reading comprehension.

Next, we can check how well the empirical item hierarchy fits the expert rankings and the constructed model of literacy levels. If psychometric analysis reveals item clustering that agrees with the model of literacy levels, it will be possible to set the benchmarks of transition between the levels of reading acquisition.

All items that fall into each cluster identified in Study 1 can be regarded as a separate subtest representing a certain level of literacy. To establish the benchmark scores, it is important to determine the criterion for transition from one level to another. Drawing on theoretical findings of Russian researchers [Bespalko 1989], we assume that a hypothetical level of skill development can be considered as achieved if at least 70% of items at this level are answered correctly (probabilistic estimation). According to this concept, acquisition of learning material by 70% indicates that a student is ready to learn new material and that a skill has been acquired.

The benchmarks for transition from one level of reading development to another were established using the methods of Item Response Theory. Three hypothetical items were formed to represent each level. Difficulty of each of the three items was estimated as the mean difficulty of all items at the relevant level. Next, a benchmark score was set for achieving each level on the literacy scale as a level of skill at which the probability of answering the hypothetical "average" item correctly is 0.7 (the 70% cutoff score for skill acquisition adopted above). All participants performing below this score should

Table 4. **Reliability analysis of empirical data**

Properties	Cronbach's α	Person Reliability	Person Separation
Russian-language version	0.97	0.87	2.56
English-language version	0.75	0.71	1.58

be considered to have not acquired the respective level as well as all the subsequent ones.

2.2. Results of Study 2

Applying the analytical approach described above, let us analyze the results of a preliminary psychometric analysis of available empirical data and graphically compare the item-difficulty hierarchies obtained for the Russian- and English-language versions of the test by expert judgments vs. actual testing. The one-parameter Rasch model for dichotomous data was used to convert raw scores into measures of reading literacy [Wright, Stone 1979]. Psychometric analysis of the items, analysis of scale dimensionality and reliability, and a goodness-of-fit test were carried out. Winsteps software [Linacre 2011] was used for psychometric analysis and assessment of item and person parameters.

Table 4 presents general psychometric properties of literacy scale quality for empirical data obtained on student samples in Russia and the UK.

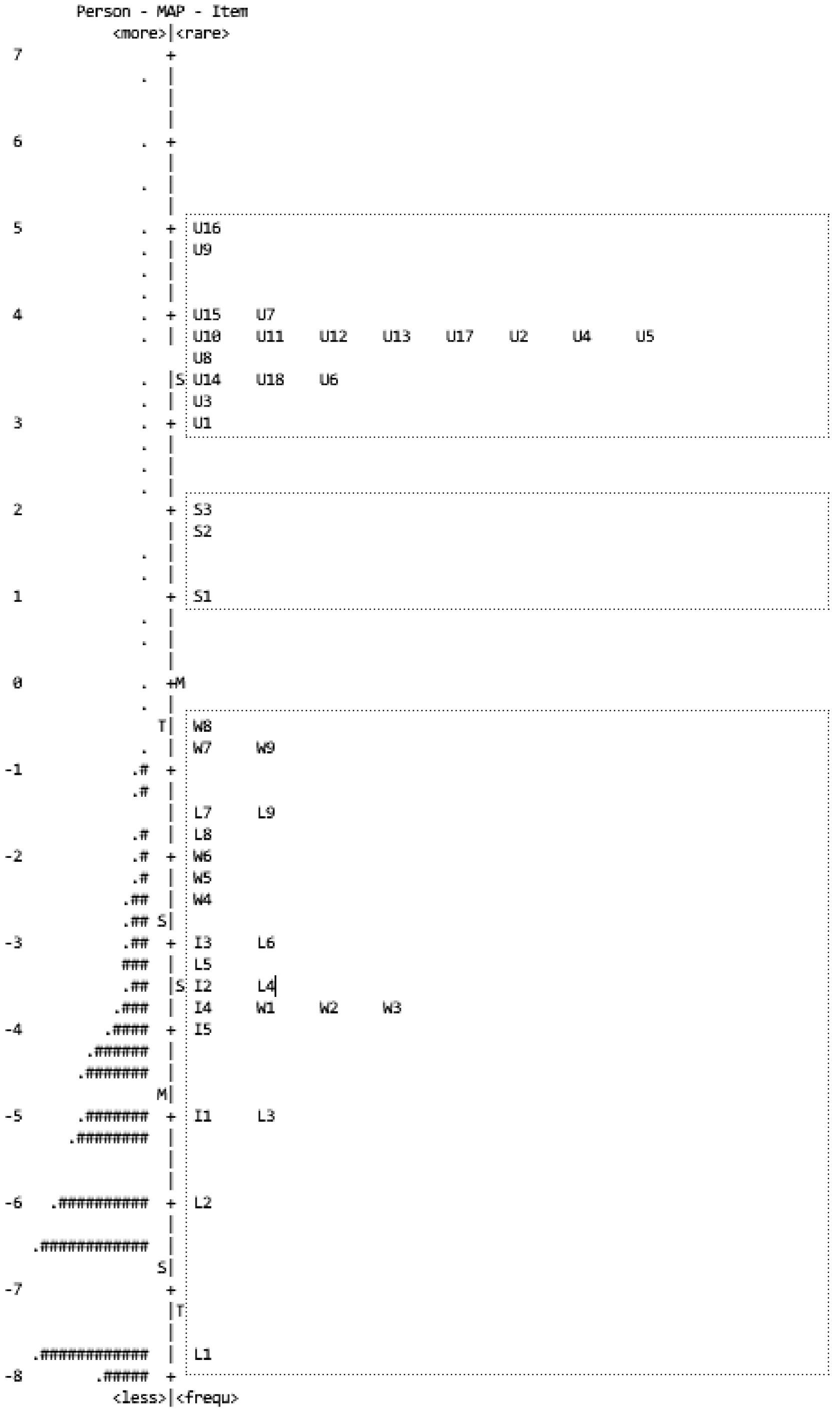
Both versions of the instrument yield highly reliable scores as indicated by both the classical and Rasch (person) reliability statistics and feature quite a high level of scale sensitivity (person separation) that allows grouping examinees into a minimum of three clusters[1] according to their level of reading literacy.

Having tested the psychometric quality of the literacy scale on a sample of students from Russia and the UK, we can examine the item hierarchy on variable maps and compare it to the one in Study 1 (Figures 3 and 4). In Figures 3 and 4, the items are shown on the right and the distribution of persons is shown on the left.

Although demarcations between the item clusters along the axis on the variable maps is less clear than in the case of expert judgments, the general structures of the clusters are identical. The easiest items measuring text structure awareness (I1–I5), letter identification (L1–L8) and word recognition (W1–W9) are at the bottom of the map, which corresponds to the items of the first cluster in Study 1. Items meas-

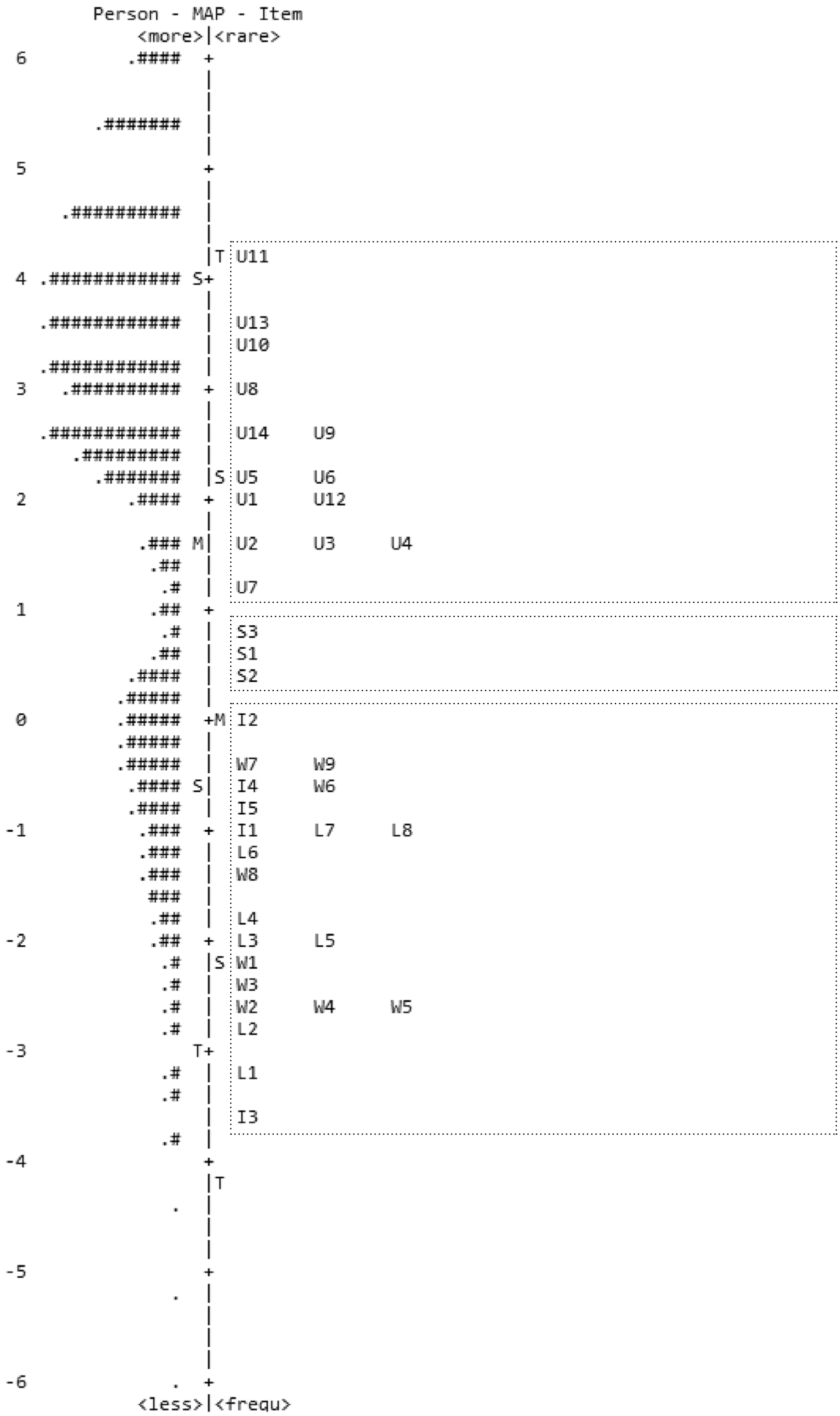
[1] To determine how many measurement strata could be statistically distinguishable among the examinees, we use the separation index formula and the procedure described in Winsteps tutorial [Linacre 2011].

Figure 3. **Variable map. Test results of English-speaking students**



```
                 Person - MAP - Item
                     <more>|<rare>
  7                        +
                       .   |
                           |
                           |
  6                    .   +
                           |
                       .   |
                           |
  5                    .   +  U16
                       .   |  U9
                       .   |
                       .   |
  4                    .   +  U15    U7
                       .   |  U10    U11    U12    U13    U17    U2     U4     U5
                              U8
                       .   |S U14    U18    U6
                       .   |  U3
  3                    .   +  U1
                       .   |
                       .   |
                       .   |
  2                        +  S3
                           |  S2
                       .   |
                       .   |
  1                        +  S1
                       .   |
                       .   |
                           |
  0                    .   +M
                       .   |
                          T|  W8
                       .   |  W7     W9
 -1                   .#   +
                      .#   |
                           |  L7     L9
                      .#   |  L8
 -2                   .#   +  W6
                      .#   |  W5
                     .##   |  W4
                     .## S|
 -3                  .##   +  I3     L6
                     ###   |  L5
                     .##   |S I2     L4
                    .###   |  I4     W1     W2     W3
 -4                .####   +  I5
                 .######   |
                .#######   |
                          M|
 -5             .#######   +  I1     L3
               .########   |
                           |
                           |
 -6          .##########   +  L2
                           |
           .############   |
                          S|
 -7                        +
                           |T
                           |
           .############   |  L1
 -8               .#####   +
                     <less>|<frequ>
```

Figure 4. **Variable map. Test results of Russian-speaking students**



```
        Person - MAP - Item
            <more>|<rare>
 6          .####  +
                   |
                   |
         .#######  |
                   |
 5                 +
                   |
      .##########  |
                   |
                   |T U11
 4 .############ S+
                   |
   .############   |  U13
                   |  U10
   .############   |
 3    .##########  +  U8
                   |
   .############   |  U14    U9
      .#########   |
        .#######   |S U5     U6
 2          .####  +  U1     U12
                   |
            .### M|  U2     U3     U4
             .##  |
              .#  |  U7
 1           .##  +
              .#  |  S3
             .##  |  S1
           .####  |  S2
          .#####  |
 0        .#####  +M I2
          .#####  |
          .#####  |  W7     W9
           .#### S|  I4     W6
           .####  |  I5
-1          .###  +  I1     L7     L8
            .###  |  L6
            .###  |  W8
             ###  |
             .##  |  L4
-2           .##  +  L3     L5
              .#  |S W1
              .#  |  W3
              .#  |  W2     W4     W5
              .#  |  L2
-3               T+
              .#  |  L1
              .#  |
                  |  I3
              .#  |
-4                +
                  |T
               .  |
                  |
                  |
-5                +
               .  |
                  |
                  |
                  |
-6             .  +
            <less>|<frequ>
```

Table 5. **Setting the benchmarks**

	Sample			
	Russia		UK	
Level benchmarks	Mean item difficulty	Benchmark score	Mean item difficulty	Benchmark score
Level 3 benchmark Reading comprehension	2.11	2.96	3.78	4.63
Level 2 benchmark Reading/decoding automaticity	0.61	1.46	1.53	2.38
Level 1 benchmark Text structure awareness, letter identification, sight word recognition	−1.63	−0.78	−3.16	−2.31

Note: Benchmark scores in the table are presented in logits, but they can easily be converted to any scale for presenting test results, for example, a 100-point scale.

uring reading/decoding automaticity (S1–S3) are in the middle of the map, and those assessing reading comprehension (U1–U14) are at the top, corresponding to the third cluster.

To assess how well expert judgments of item difficulty in each language version reflect the item hierarchy on the literacy scale, we conducted a correlation analysis of item difficulty estimates obtained by expert judgment and actual testing. Pearson's correlation coefficients were found to be high enough, 0.81 for the Russian-language version and 0.88 for the English-language one.

Therefore, analysis of the available empirical and expert data confirms the possibility of using the constructed model of literacy levels as a basis for setting benchmarks and conducting an indirect cross-country comparison.

As a result of the procedure described in the analytical approach section, benchmarks on the literacy scale were determined, setting the boundaries of reading acquisition levels (Table 5), which can be applied uniformly to children's test results in the two language versions of the instrument.

Setting of benchmarks for the Russian-language version of the test is illustrated in Figure 5. The level of examinees' literacy in logits is plotted on the abscissa axis, while the ordinate axis displays the probability of answering the item correctly. The three curves correspond to the characteristic curves[2] of the three hypothetical items reflecting mean item difficulty at the respective level. The horizontal line reflects the accepted 70% cutoff score of level acquisition. Benchmark

[2] An item characteristic curve reflects the probability of answering the item correctly depending on the level of literacy.

Figure 5. **Setting the benchmark scores for levels of reading development among the examinees**
(Russian-language version of the test)

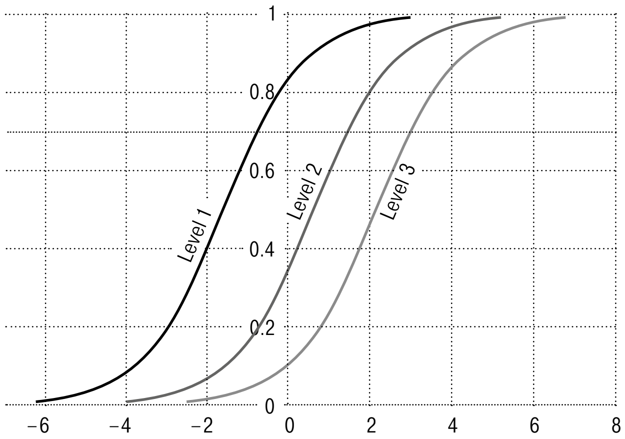
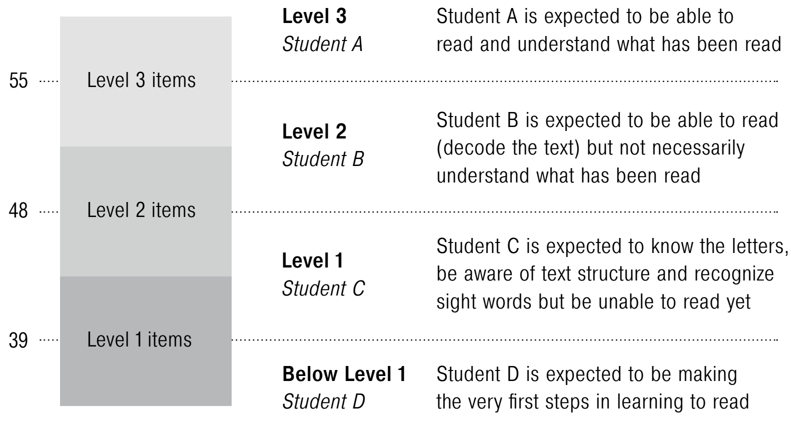


Figure 6. **Categorization of benchmark scores**
(example for the Russian-language version of the test)



Level 3
Student A
Student A is expected to be able to read and understand what has been read

Level 2
Student B
Student B is expected to be able to read (decode the text) but not necessarily understand what has been read

Level 1
Student C
Student C is expected to know the letters, be aware of text structure and recognize sight words but be unable to read yet

Below Level 1
Student D
Student D is expected to be making the very first steps in learning to read

scores are defined as abscissae of the points of intersection between the horizontal line and the item characteristic curve for each level of reading development.

Figure 6 shows the literacy scale with the benchmark scores converted into a 100-point scale which is used to present test results (in *z*-scores with the mean of 50 and standard error of 10). The resulting benchmarks are as follows: 39 for acquiring Level 1, 48 for Level 2, and 55 for Level 3.

Table 6. **Comparing the distribution of students by the levels of literacy on school entry between Russia and the UK**

Level of reading development	Level description	Share of the sample, %	
		Russia	UK
Level 3	Reading comprehension	32.7	0.3
Level 2	Reading/decoding automaticity	27.7	0.6
Level 1	Letter identification, sight word recognition, text structure awareness	23.8	9.9
Below Level 1	The very first steps in learning to read	15.9	89.2

2.2.1. Using the Benchmark Scores for Comparative Assessment

Therefore, Study 2 established the benchmarks setting the boundaries between the levels of reading development in each of the two language versions of the instrument. Taken together, the results of the two studies confirm the possibility of conducting a cross-cultural assessment of reading skills in groups of children starting school. Table 6 shows the distribution of students in Russia and the UK by the levels of reading development.

The distribution of students by the levels of reading development differs greatly between the two countries. Interpretation of the results obtained is beyond the scope of this article, but it is worth noting that the samples analyzed here differed significantly in examinee age, which may be the reason for considerable disparities in student achievement. The most important finding of Study 2, meanwhile, is that the methodology described in it can be used for indirect cross-country comparisons in the absence of a uniform metric scale.

3. Discussion

To make data from cross-country assessments reliable for educational researchers, policymakers and practitioners, it should be verified for credibility, validity and meaningfulness, and evidence should be provided that it adequately represents the construct measured and can serve the basis for informed decision making. The ITC Guidelines for Translating and Adapting Tests [Leong 2016] were applied when creating the Russian-language version of the iPIPS test that had been originally designed in English.

The iPIPS instrument is widely used by schools in the UK as well as in several other countries, including Australia, Brazil, Germany and South Africa [Archer et al. 2010; Bartholo et al. 2019; Tymms et al. 2014; Vidmar et al. 2017]. In some earlier publications, it was used to measure student performance across cultures, e. g. in the UK, Australia and New Zealand [Tymms et al. 2014], where the authors tried to assess possible differences in academic achievement and progress as well as the effectiveness of education systems.

Another study tested the potential of iPIPS for comparing math test scores between children in the UK (England, Scotland) and Russia, countries differing in school entry age, curricula, language and culture [Ivanova et al. 2018]. It is shown that, despite the obvious challenges, direct cross-country comparison of iPIPS test results in mathematics is not impossible.

In yet another study [Vidmar et al. 2017], reading progress of first-graders in Serbia and Germany was compared using the iPIPS instrument. However, this article only deals with sample means and provides no evidence of cross-national comparability.

Up to this point, there have been no studies using the iPIPS instrument to compare the reading test results of students starting school in different countries.

This study attempts to solve the challenging research problem of international assessment of reading ability on school entry using the test results of first-graders in Russia and the UK. Study 1 analyzed the expert judgments on the construct, i. e. the model of literacy levels in Russian- and English-speaking children starting school, using the rank-ordering method.

A cross-country comparison of item-difficulty hierarchy obtained as a result of data calibration using two Rasch modelling frameworks shows that three item clusters can be distinguished in both language versions. These clusters are represented by the same items in both Russian and English.

Study 1 demonstrates that expert judgments 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.

Study 2, using test results obtained from the samples of Russian- and English-speaking students in the two countries, sets the benchmarks and determines the levels of reading development. Those levels are 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.

We assume 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 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 hypothesis should be tested for other language versions of the iPIPS instrument.

The practical significance of this study is that it introduces educational researchers to the problems that can be encountered when assessing a particular construct (reading skills) on a specific sample of participants (children on school entry) from different countries. The methods used here can be applied to tackle research problems in studies that involve similar constructs and target elementary school students.

In the long term, this methodology can be used both for cross-country comparisons as well as other purposes, such as year-over-year comparison of test results obtained on different samples of students using different versions of the same assessment instrument. In the UK [Bramley 2005], this practice has been implemented for several years to compare scores on some written examinations.

References

Ainley M., Ainley J. (2019) Non-Cognitive Attributes: Measurement and Meaning. *The SAGE Handbook of Comparative Studies in Education* (eds L. E. Suter et al.), London: Sage, pp. 103–125.

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.

Archer E., Howie S. J., Scherman V., Coe R. (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.

Bartholo T. L., Koslinski M. C., Costa M. D., Barcellos T. (2019) *What Do Children Know upon Entry to Pre-School in Rio de Janeiro?* Ensaio: Avaliação e Políticas Públicas em Educação.

Bespalko V. (1989) *Slagaemye pedagogicheskoy tekhnologii* [Components of Pedagogical Technology]. Moscow: Pedagogika.

Bramley T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgment. *Journal of Applied Measurement*, vol. 6, no 2, pp. 202–223.

Buzhardt J., Greenwood C. R., Hackworth N. J., Jia F., Bennetts S. K., Walker D., Matthews J. M. (2019) Cross-Cultural Exploration of Growth in Expressive Communication of English-Speaking Infants and Toddlers. *Early Childhood Research Quarterly*, vol. 48, 3rd Quarter 2019, pp. 284–294.

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*, vol. 53, no 4, pp. 1054–1085.

Caro D. H., Cortés D. (2012) Measuring Family Socioeconomic Status: An Illustration Using Data from PIRLS2006. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments*, no 5, pp. 9–33.

Dubeck M. M., Gove A. (2015) The Early Grade Reading Assessment (EGRA): Its Theoretical Foundation, Purpose, and Limitations. *International Journal of Educational Development*, vol. 40, January, pp. 315–322.

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.

Espeland W. (2015) Narrating Numbers. *The World of Indicators: The Making of Governmental Knowledge through Quantification*. *Cambridge Studies in Law* *and Society* (eds R. Rottenburg, S. Merry, S. Park, J. Mugler), Cambridge: Cambridge University, pp. 56–75.

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

Field A. P. (2014) *Kendall's Coefficient of Concordance* // Wiley StatsRef: Statistics Reference Online.

Goodrich S., Ercikan K. (2019) Measurement Comparability of Reading in English and French Canadian Population: Special Case of the 2011 Progress in International Reading Literacy Study. *Frontiers in Education*, vol. 4. Available at: https://www.frontiersin.org/articles/10.3389/feduc.2019.00120/full (accessed 27 September 2020).

Ivanova A., Kardanova-Biryukova K. (2019) Sozdanie russkoyazychnoy versii mezhdunarodnogo instrumenta otsenivaniya rannikh navykov chteniya [Constructing a Russian-Language Version of the International Early Reading Assessment Tool]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 4, pp. 93–115. DOI: 10.17323/1814–9545–2019–4–93–115

Ivanova A., Kardanova E., Merrell C., Tymms 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 & Practice*, vol. 25, no 2, pp. 141–159.

Kardanova E., Ivanova A., Sergomanov P., Kanonire T., Antipkina I., Kayky D. (2018) Obobshchennye tipy razvitiya pervoklassnikov na vkhode v shkolu. Po materialam issledovaniya iPIPS [Patterns of First-Graders' Development at the Start of Schooling: Cluster Approach Based on the Results of iPIPS Project]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 1, pp. 8–37. DOI: 10.17323/1814–9545–2018–1–8–37

Kreiner S., Christensen K. B. (2014) Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, vol. 79, no 2, pp. 210–231.

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 Rasch Measurement. Version 3.71. Winsteps. com.*

Linacre J. M. (2006) Rasch Analisis of Rank-Ordered Data. *Journal of Applied Measurement*, vol. 7, no 1, pp. 129–139.

Linacre J. M. (1989) Rank Ordering and Rasch Measurement. *Rasch Measurement Transactions*, vol. 2, no 4, pp. 41–42.

Linacre J. M., Wright B. D. (1994) *A User's Guide to FACETS: Rasch Measurement Computer Program*. Chicago: MESA.

Liu J., Steiner-Khamsi G. (2020) Human Capital Index and the Hidden Penalty for Non-Participation in ILSAs. *International Journal of Educational Development*, vol. 73, iss. C, pp. 1–9.

Masters G. N. (1982) A Rasch Model for Partial Credit Scoring. *Psychometrika*, vol. 47, no 2, pp.149–174.

Merrell C., Tymms P. (2007) Identifying Reading Problems with Computer-Adaptive Assessments. *Journal of Computer Assisted Learning*, vol. 23, no 1, pp. 27–35.

OECD (2020) *Early Learning and Child Well-Being: A Study of Five-Year-Olds in England, Estonia, and the United States*. Paris: OECD.

Peña E. D. (2007) Lost in Translation: Methodological Considerations in Cross-Cultural Research. *Child Development*, vol. 78, no 4, pp. 1255–1264.

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*, vol. 26, no 7, pp. 903–920.

Suggate S. P. (2009) School Entry Age and Reading Achievement in the 2006 Programme for International Student Assessment (PISA). *International Journal of Educational Research*, vol. 48, no 3, pp. 151–161.

Thurstone L. L. (1927) A Law of Comparative Judgment. *Psychological Review*, vol. 34, no 4, pp. 273–286.

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., Merrell C., Buckley H. (2015) *Children's Development at the Start of School in Scotland and the Progress Made during their First School Year: An Analysis of PIPS Baseline and Follow-Up Assessment Data*. Edinburgh, UK: The Scottish Government. Available at: http://dro.dur.ac.uk/17417/ (accessed 27 September 2020).

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 27 September 2020).

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*, vol. 32, no 2, pp. 311–331.

Wright B. D., Stone M. H. (1979) *Best Test Design*. Chicago, IL: MESA.