Kardanova Elena, Ivanova Alina, Merrell Christine
Hawker David, Tymms Peter

THE ROLE OF THE IPIPS ASSESSMENT IN PROVIDING HIGH QUALITY VALUE ADDED INFORMATION ON SCHOOL AND SYSTEM EFFECTIVENESS WITHIN AND BETWEEN COUNTRIES

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: EDUCATION
WP BRP 20/EDU/2014

This Working Paper is an output of a research project implemented within NRU HSE’s Annual Thematic Plan for Basic and Applied Research. Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.
THE ROLE OF THE IPIPS ASSESSMENT IN PROVIDING HIGH QUALITY VALUE ADDED INFORMATION ON SCHOOL AND SYSTEM EFFECTIVENESS WITHIN AND BETWEEN COUNTRIES

Despite the growing influence of international surveys of student achievements such as PISA, TIMSS, and PIRLS on policy, there is currently no international baseline study of children’s development on starting school. As a result, nobody knows for sure the extent to which the differences in performance between countries which are observed in these later assessments are already present when children start school, and the impact of educational policy and school effectiveness.

This working paper describes the development of such an international study (iPIPS) which aims to help answer some key questions relating to early years and schooling. iPIPS takes as its basis, the Performance Indicators in Primary Schools (PIPS) On-entry baseline and follow-up assessments which were developed by the Centre for Evaluation and Monitoring (CEM) at Durham University (UK). PIPS has been used for twenty years in schools in a range of countries, providing a wealth of data for research purposes. iPIPS aims to collect assessment data from representative samples of children internationally. The paper will also report on the work undertaken in 2013-2014 to adapt and pilot the iPIPS instruments for use in Russia.

JEL Classification: I2.

Keywords: iPIPS, International comparative study, Rasch model, test adaptation.
1. Introduction - Why is a baseline assessment of children starting school and a follow-up assessment at the end of their first year of school needed?

Figure 1 shows the average results in mathematics of selected countries from the 2012 PISA assessment, including the Russian Federation. The graph is by now very familiar to education professionals across the world. One interesting but rather puzzling fact is that, even though these are results for 15 year old students, they are having a major influence on pre-school policies in many countries. In a 2010 survey, OECD found that, of around 35 countries which responded to a survey, over one third said that the PISA results had had a direct influence on their policies for early childhood education.

![Figure 1: selected countries’ 2012 PISA results: mathematics](image)

Clearly, it is not possible to tell anything directly from PISA about the relative effectiveness of different countries’ early years education policies. But the responses to the survey show that countries are attempting to do this nonetheless. So, how much better would it be to provide an

---

assessment which did actually focus on this age range? There are many key questions which such a study would set out to answer, and by way of illustration some are listed below:

1. What do children know and what can they do when they start school, and how does this differ within and between countries?
2. What is the value for money and relative effectiveness of different early years programs?
3. How much do children learn in their first year at school and how effective is one educational system compared with others?
4. How do different contextual factors influence children’s learning?
5. How can teachers and schools improve?
6. What are the most promising policies for long term effectiveness in children’s learning?

In many countries the national or other tests have become one of the principal means of school accountability. However, judging schools purely on the basis of their ‘raw’ results on the tests is unsatisfactory, because of the evidence that children from socially and economically deprived backgrounds tended to be less advanced cognitively and emotionally when they started school than children from more privileged backgrounds. Schools in areas of high deprivation are therefore bound to fare badly in such comparisons. So some kind of ‘value added’ system is needed which takes account of children’s starting points and their progress in school in order to compare the performance of schools more fairly. The value-added system can be developed by means of a baseline assessment when children start school and follow-up assessments at later time-points. iPIPS aims to administer a baseline assessment at the start of school and then a follow-up assessment at the end of the first school year in order to measure individual progress during that period.

The PIPS assessment has been demonstrated to be reliable and valid in a number of different countries and is therefore considered to be an ideal instrument for iPIPS. Over the years, from 1994, it has been used to assess more than two million children, and has provided thousands of schools in the UK and elsewhere with high quality information about children’s development and their own educational effectiveness.

A big international study is planned in 2015 aimed a comparison of children from different countries when they start primary school and after the first year of schooling. About 10 countries including Russia are going to join the study. Currently in Russia there are no standardized, valid, applicable for large-scale use assessments for evaluation of the initial level of a child starting
school and measuring the individual progress of the child during the first year of schooling. That’s why Russia was one of the first countries who decided to join the iPIPS study. It is necessary to note, that despite a wide usage of iPIPS in different countries, there were mainly English–speaking countries and there was no systematic work on translating and adaptation of iPIPS into foreign language and culture. Moreover, the very opportunity of comparison of children who (1) are from different countries, languages and cultures, and (2) are of more than two years of difference in age looked problematic.

Thus, the purpose of the paper is threefold. The first goal is to describe the iPIPS assessment. The second goal is to describe the adaptation process for creating a Russian language version. The third goal is to show the potential of iPIPS to compare children at the start and end of their first school year from three countries – Russia, England and Scotland, illustrating that this is valid, despite a big difference in the age of children starting primary school in different countries.

2. Overview of PIPS and iPIPS

Following the introduction of the first National Curriculum in England and Wales after the 1988 Education Reform Act, national tests were established for all pupils aged 7, 11 and 14. Over time the tests became one of the principal means of school accountability in the English education system. In 1995 the UK government proposed the introduction of some kind of baseline assessment when children started school at age 4/5 from which schools would measure children’s progress, rather than reporting just their raw attainment when they reached the first statutory assessments at age 7. A simple solution would have been a single baseline assessment measure, common to all schools across the country, to be used for this value added purpose. However, in the event, worries from teachers about another centrally imposed national test led the government to adopt a ‘licensing’ approach, where a number of different (and competing) baseline assessment schemes were allowed to operate, and schools would choose the one they liked best. One of these schemes was the PIPS On-entry Baseline assessment developed in 1994 by the Centre for Evaluation and Monitoring (CEM), now at Durham University.

The prime purpose of PIPS was to provide a baseline for later value added measurements of children’s progress. But it was also designed to give schools and teachers a multi-faceted analysis of their pupils’ cognitive and non-cognitive development when they entered school. As a result, teachers could (i) use the PIPS assessment diagnostically to assess children’s learning...
needs initially, then (ii) re-administer it at the end of the first year in order to assess progress during that crucial first year in school, and then (iii) use it as a baseline against which they could assess their children’s progress over the first few years of their school career. It also provided school principals with a view of their own pupils’ progress in relation to a larger national cohort. (For a detailed description of the PIPS assessment, its rationale and development, see Tymms, 1999).

PIPS begins with an assessment of children at the start of school. The same assessment is administered at the end of the first year of school. The assessment can be efficiently administered on computer or with a paper manual accompanied by an App for data collection. It takes between 15 and 20 minutes per child, has high reliability and good correlations to later attainment (Tymms, 1999; Tymms et al., 2012). It is possible to collect a very reliable yet comprehensive measure of children at the start of school because PIPS is adaptive, using sequences of items with stopping rules. The items are arranged into sections in order of increasing difficulty. Each child begins with easy items and move on to progressively more difficult ones. When they make a number of errors, the administrator (or the software if the assessment is computer-delivered) moves onto the next section and so the assessment continues. The assessment is repeated at the end of the school year, taking off from the point where the child began to falter on their first assessment. Thus, they do not repeat items which were clearly very easy for them at the beginning of the year.

Schools carry out the assessments with each child, using a computer program, and upload the results to a database held at Durham University. They receive back (within a week) a summary of the results, expressed mainly as standardised scores, in both tabular and graphical form, both of which allow them to compare their children’s performance with the full national sample. At the end of the year, the feedback also includes progress charts and regression plots which allow schools to compare the progress their children are making by comparison with children elsewhere.

The system is straightforward to use and very popular with schools. Over the years the assessments have proved to be very reliable, with a test-retest reliability of 0.98 and internal consistency (Cronbach’s alpha) of around 0.92 on the test as a whole. They have also proved to have extremely good predictive validity, with correlations of around 0.68 to later national assessments at age 7 and 11, and of around 0.5 to the national examinations at age 16.
The PIPS assessment has been used successfully in a number of countries including UAE (Abu Dhabi), Australia, England, Germany, New Zealand, Scotland and South Africa (See, for example, Archer et al., 2010; Tymms and Wylde, 2013; Wildy and Styles, 2008a and b). Comparisons between children starting school in English-speaking countries using PIPS have been published (Merrell and Tymms, 2007; Tymms and Merrell, 2009) and building upon this research, CEM has proposed a large-scale international comparative study of children starting school, called iPIPS. iPIPS is intended to provide comparative, system-level information to policy makers and researchers.

3. PIPS and international studies: first results

PIPS has been used in some previously-published studies to evaluate the effectiveness of systems in addition to providing information to teachers about their individual pupils. Below are some examples.

A study commissioned by the Scottish Government in 2004 looked at data from England, Scotland, Australia and New Zealand, to see what differences there were, if any, in the average attainment and progress of children in the four countries.

The study found that it was possible to use the same assessment for children starting school at different ages – in Australia, for example, children start school closer to the international norm of age 6. It also found that, while there was a clear relationship between children’s scores and their age when they started school, it was still possible to compare the progress of all four national cohorts in a meaningful way (Merrell and Tymms, 2007).

Data collected a number of years later (in 2011/12) from the same four countries produced some interesting results, which are shown below (Tymms, Merrell and Wildy, 2014).
Figure 2: children’s reading development at the end of their first year in four countries, relative to age

Two things are particularly striking from this picture. First, the amount of progress which children in all four countries make relative to their age is enormous. The difference in age between the oldest children starting school and the youngest children at the end of the first year in the same country is one month; but the difference in their reading development is the equivalent of between two and three years – this is the scale of the progress which children typically make in their first year of school. Second, the Australian children appear to have made rather less progress than their counterparts in the other three countries. There are a number of possible reasons, which would need to be explored through another study. It is also the case that the Australian sample was not fully representative, so some of the results may be skewed. Nevertheless, the fact that the PIPS assessment has the power to reveal such differences was an indication of its potential to act as the basis of a larger scale international comparative study.

All countries of this study are English-speaking countries. Thus it is especially important to check the potential of iPIPS to compare Russian children and children from other countries, because of different language and different age of Russian children. In average Russian kids start school at the age of 7 and half. It is obviously that for children from Russia and other countries who are of so different age, the tests can not be the same. They should be partly different
meaning that we need to use a special technique of measurement to equate the results from different tests.

4. Adaptation of iPIPS in Russia

4.1. Test adaptation process

It is important to note that test adaptation is not just test translation. It includes many activities from deciding whether or not a test could measure the same construct in a different language and culture, to checking equivalence of the initial and adapted test versions (Hambleton, 2005; Ercikan, 2013).

When developing the Russian version the main task was to ensure, as much as possible, the equivalence of the tests in both languages. Translation can affect the meaning of words and sentences, the content of the items, and the skills measured by the items. The degree and manner in which item features are changed during translation will determine whether the equivalence of items is maintained. Changes in any of these item features may alter its difficulty or even what is being measured.

The procedure of adaptation included several stages. Firstly, the baseline and follow-up booklets of the English version were translated into Russian by two independent translators. After editing and discussing the results, the final translation of booklets was presented to two Russian experts in literacy and numeracy for evaluation. They examined the translated booklets (taking into account the original version) and offered to merge items from two booklets into one, as many items in the original baseline booklet were clearly too easy for older Russian children. The final version of the first booklet consisted of the following sections:

- Handwriting, where the child is asked to write his/her own name
- Vocabulary, where the child is asked to identify objects within two pictures (the remaining in Russian version pictures were related to countryside and toy-shop)
- Ideas about reading, where concepts about printed text are assessed
- Phonological awareness: rhymes and repeats (This section differs mostly from the original version, but essentially is based on the same principles)
- Letter identification (only 15 the most difficult letters in the Russian version)
- Word recognition and reading.
- Counting
• Sums: addition and subtraction problems presented with pictures and without symbols
• Digit identification (five of the first ten digits, as well as some two- and three-digit numbers)
• Mathematics problems including sums with symbols.

For the verification of equivalence of different language versions of items, backward translation method was used. All items were translated back into the English language and compared with original test items by experts (both English and Russian) and iPIPS developers. The purpose of this was to ensure comparability of items in two language test versions. Criteria for evaluation included (1) differences in the actual meaning of the item; (2) differences in the item format; (3) differences in the item presentation; (4) difference in cultural relevance; (5) exclusion or inappropriate translation of key words; (6) differences in length or complexity of sentences; etc. (Ercikan, 2004). All translation errors were documented and discussed, and items were revised.

The next step was to guarantee the equivalence of test conditions that refers to whether different language versions of tests are administered in identical fashion. To provide it we used the same procedure of testing that was used in England and Scotland.

The first piloting in October 2013 in Russia revealed a ceiling effect on some items. So the significant revision of Russian booklet was made. Vocabulary and reading sections became more difficult and appropriate for older children. For example, reading passages were developed to check not only technical reading, but rather reading with understanding. The mathematics part was also changed quite remarkably. Counting and simple sums sections were shortened, some items were deleted and some others were changed. Finally, an extra section was created, which included real life-oriented math problems, connected with graphs, understanding and operating numbers in shopping situations.

Equivalence of measurement using adapting tests should be confirmed by empirical evidence that implies psychometric analysis. It included different kinds of analysis that can be grouped into two main categories: (a) IRT analysis of test items and tests (comparison of item characteristics, item maps, item hierarchy, dimensionality, etc., for two language versions); and (b) identification of DIF (differential item functioning) in favour of participants from one country and its sources. The results of this analysis will be presented below in the paper.
It is important that original English and Russian versions of the booklet have a sufficient number of common items to make it possible to equate them and put the results from baseline and follow-up assessments from different countries on a common scale. Different linking methods can be used, but the most appropriate in our case is separate monological group design (Sireci, 1997). It employs the use of a set of items determined to be equivalent in the two languages as anchor items in IRT-based calibration. In this case some items could be recognized as non-equivalent and considered as unique items for the language version of a test while others serve as anchors.

It is especially difficult to develop equivalent versions of verbal items where cultural and language differences are having the most impact on the test items. So in this study we consider only math items for comparison between countries.

4.2. iPIPS piloting in Russia

Participants

The Russian sample originally consisted of 310 children recruited from 21 classes of 21 schools in one of the Russian regions, located in the central part of Russia where the majority of the population is ethnic Russians. This region was selected because its socio-economic characteristics (e.g., average salary, unemployment, educational level, urban-to-rural ratio) were similar to those in the entire country (based on 2012 census, REF). For example, the distribution of the region’s population by educational level (62% college and above, 30% high school, 8% below high school) was parallel to that in the country (65% college and above, 29% high school, 6% below high school). Also, the ratio of urban to rural students in the region (72% urban, 28% rural) was similar to that in the country (see Table 1 below).

The target population was children enrolled to the 1st grade of schools on the 1st of September 2014. The sample represents about 5% of the general population of all the grade 1 students of this region. It is representative and stratified on two parameters. Firstly, we have taken into account the place of a school location (rural or urban area), and, secondly, different status of schools\(^8\), so that these parameters cover a wide range of environmental conditions in Russia. On these parameters the area types of schools were randomly selected from the total number of region’s areas. And then the subsequent random selections were made through schools within the

\(^8\)There are 3 main types of schools in Russia: comprehensive (general regular) schools, schools specializing in a certain subject, and gymnasiums (some of them are fee-paying).
selected areas and classes within the selected schools. All the chosen schools agreed to take part in the assessment. After parental agreements for children’s participation in the assessment were received, we randomly selected pupils within the selected classes.

The first stage of the assessment cycle was executed during the second week of October 2013. The second stage proceeded during the third week of April 2014. A significant number of pupils were sick with flu (according to teachers) during this period, so we lost about 10% of the original sample. Tables 1 and 2 give details of the sample for respective stages of the assessment cycle.

Table 1. The Russian sample, October 2013

<table>
<thead>
<tr>
<th>Gender, %</th>
<th>Place of living, %</th>
<th>Type of school, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>49</td>
<td>Urban</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>51</td>
<td>Rural</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In total: 310 pupils</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The Russian sample, April 2014

<table>
<thead>
<tr>
<th>Gender, %</th>
<th>Place of living, %</th>
<th>Type of school, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>49.8</td>
<td>Urban</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>50.2</td>
<td>Rural</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In total: 277 pupils</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Procedure description

The regional department of education, in consultation with school principals, provided the permission to conduct the study.

The assessment procedure in Russia took, as in other countries, approximately 20 minutes per child. Following the procedure designed by CEM, specially trained interviewers worked with individual pupils using a booklet and App on a tablet with the pre-installed software based on a semi-adaptive algorithm (a series of short sections with stopping rules). Following the instructions on the tab an interviewer presented a child with a question in the booklet, then recorded the child's response on-screen and the program selected the next question, which the interviewer would see on his or her tab screen. The adaptive algorithm allowed children to avoid inappropriately difficult questions. Each section of the program presented items of increasing
difficulty until the child had answered three or four questions wrong and then it moved on to the next appropriate section.

4.3. **Item analysis and scale construction for the Russian version**

First of all, two Russian data sets from the baseline and follow-up assessment cycles were treated separately to ensure the quality of the tests and possibility for students’ estimation. Because of space limitation, only the results of baseline assessment math scale analysis are presented here.

The one-parameter dichotomous Rasch model (Wright B.D.& Stone M.N., 1979) was implemented to transform raw scores of students into measures on an objective interval scale. As per this model, each test item is characterized by one parameter, which is difficulty, and each test participant is also characterized by one parameter, which is ability level. Rasch analysis places students and items on the same measurement scale with the logit as the unit of measurement. The reasons for choosing the Rasch model are both psychometrical and practical. Firstly, the Rasch model has optimal metric properties, and secondly, from practical point of view, it is useful to empirically determine the quality of test items, construct scales and carry out test equating (Bond & Fox, 2001).

Winsteps software (Linacre J. M., 2011) was used for parameters estimation and data analysis under Rasch model.

There were 43 items in this scale which included items such as recognition of numbers, ideas about mathematics, and use of arithmetical operations. Several items were removed from the scale. An item was removed if at least one condition was satisfied: (a) the item was extremely easy for Russian children, (b) the item demonstrated poor fit to the model, (c) the item was close to the second dimension, (d) the item demonstrated DIF (differential item functioning), for example by sex or whether the assessment was carried out at the start or at the end of the school year. The remaining items (19 in total) formed a good scale: it was unidimensional, all items demonstrated fit to the model and good psychometrics characteristics. The person reliability was 0.81, that means that the proportion of observed student variance considered true was 81%. (This index is close on value and interpretation to classical reliability: $\alpha = 0.82$ for the math baseline scale).

Thus, the analysis of Russian data revealed good psychometric qualities of the iPIPS Russian version test items and the whole test. It is reliable and valid measure of pupils’ math ability both at the beginning and end of the first grade. The next task was to compare the results of children
from Russia with those from England and Scotland by equating the assessments to place all students onto the same scale.

5. Equating iPIPS versions for England, Scotland and Russia

The Russian sample differs quite significantly from the English and Scottish ones on two main parameters – the age of children and the sample size. Children in Russia are older than in England and Scotland at the start of school, as illustrated in Table 3.

Table 3. Average age of children at the time of the first assessment in samples of 3 countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Mean, age in years</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>4.56</td>
</tr>
<tr>
<td>Scotland</td>
<td>5.09</td>
</tr>
<tr>
<td>Russia</td>
<td>7.33</td>
</tr>
</tbody>
</table>

Another challenge for this research was the significant difference in sample sizes from the three countries. To date, only a small-scale pilot has been conducted in Russia, in which the sample was limited to just above 300 pupils. Conversely, PIPS has been used in England and Scotland on a large scale for many years. Details of the sample sizes are given in Table 4.

Table 4. Sample size for 3 countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of participants in the baseline assessment</th>
<th>Number of participants in the follow up assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>6985</td>
<td>5837</td>
</tr>
<tr>
<td>Scotland</td>
<td>6627</td>
<td>6627</td>
</tr>
<tr>
<td>Russia</td>
<td>310</td>
<td>277</td>
</tr>
</tbody>
</table>

So the question is how to make valid comparisons of children’s achievement from different countries taking into account that they start school at different age? The Rasch measurement provides a method of test equating that allows the creation of a single continuous scale of item difficulty and children’s developed ability, enabling children’s scores to be calculated for both baseline and follow-up cycles of assessment and for all countries.

We have six data sets in total that need to be equated; two assessment cycles (baseline and follow-up), from England, Scotland and Russia. The results from all data sets were equated in a single model using Rasch measurement (Wolfe, 2004). This allowed us to perform both equating between countries and equating between baseline and follow-up stages of the assessment. It
should be noted that the most items were the common ones for all countries because actually it was the same test. But as we described earlier, the follow-up assessment test in Russia was changed significantly. Thus some items between countries were common while others were not.

To conduct the analysis of joint scale functioning we decided to use random subsamples of comparable size from English and Scottish samples. From baseline assessment samples for both countries three random subsamples were constructed with sample size of 320 students each. The same students were chosen from follow-up assessment samples. Thus we had three single matrices for equating. They were for Russia, England and Scotland. Each had data on students who were assessed at the start and end of the year. These matrices were analyzed separately to ensure the stability of the results.

Each student was included in the single matrix twice in accordance with the assessment cycle (baseline or follow-up). Thus, the total sample size was of 1867 students. The total number of items was 81 including common and unique items. The data analysis was performed in several steps as follows:

*Step 1. Analysis of model fit.* Starting with 81 items three items were deleted because of poor psychometric quality - low discrimination and/or misfitting the model (for statistics out of range) (Wright & Stone, 1979).

*Step 2. DIF analysis relating country.* Twelve items demonstrated DIF relating the country. An item demonstrates DIF (Differential Item Functioning) if test participants with the same ability level who belong to different groups have varying chances to answer the item correctly. In other words, the item functions in a different manner for different groups of test takers, and representatives of one of the groups can be evaluated unfairly. Two statistics – Student’s t-test and Mantel-Haenzel statistics – were used, according to circumstances, to check DIF in this study (Smith. R. M., 2004; Dorans N.J., 1989). Generally, the focus fell on differences between Russia and England as Scotland and England were so similar. An item was considered as an item with DIF if at least one condition was satisfied: (1) the Mantel-Haenzel statistics had the significance level p<.01; (2) t-statistics had value out of range (-2, +2), and (3) difference in the item difficulty for different groups of students was more than .5 logits.
Figure 3 shows difficulties of items separately for children from different countries. We see that most of items demonstrate stable estimates of the difficulty, so they are DIF free. The 12 items with DIF were analyzed.

These items included items such as recognition of numbers, use of arithmetical operations and logic sequencing. All 12 items tended to operate differently for the Russian group and other countries groups.

Figure 3: item difficulties for different countries

The first 3 items, as well as items # 8-10 showed Russian students performing better than English and Scottish students, while the other six items showed the reverse. Table 5 lists these items and the direction of DIF.

Table 5. Items showing DIF

<table>
<thead>
<tr>
<th>#</th>
<th>List of items</th>
<th>Direction of DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number teen (10)</td>
<td>Ru&gt;En,SC</td>
</tr>
<tr>
<td>2</td>
<td>Number 2dig (25)</td>
<td>Ru&gt;En,SC</td>
</tr>
<tr>
<td>3</td>
<td>Number 3dig (100)</td>
<td>Ru&gt;En,SC</td>
</tr>
<tr>
<td>4</td>
<td>What is 1 more than 5?</td>
<td>En,Sc&gt;Ru</td>
</tr>
<tr>
<td>5</td>
<td>What is 3 less than 7?</td>
<td>En,Sc&gt;Ru</td>
</tr>
<tr>
<td>6</td>
<td>What is 2 more than 6?</td>
<td>En,Sc&gt;Ru</td>
</tr>
<tr>
<td>7</td>
<td>What is 3 more than 8?</td>
<td>En,Sc&gt;Ru</td>
</tr>
</tbody>
</table>
A possible reason why items on number recognition (items 1-3) and arithmetical operations (items 8-10) tend to be relatively easy for Russian children is that these children are significantly older than English and Scottish children.

One way to deal with items which show DIF is just to delete them. But in this case the total number of items would be essentially less that could impact the precision of estimation. So we decided to keep them in the analysis, but to use as the country unique items. In other words, these items were not considered more as common items for all countries, but rather as unique items for Russia and unique items for England and Scotland. We created 12 such dummy-items and repeated the Rasch analysis. Even though the overall results improved, some items showed poor fit to the model. Finally, we retained country-specific items # 7-12 for the Russian sample and items # 9-12 for English and Scottish samples (see Table 5). Other items were deleted. This new data set was then analyzed.

**Step 3. Analysis of items.** We repeated analysis of model fit with the new data set. Six items were deleted due to poor fit.

**Step 4. Dimensionality study.** We examined the dimensionality of the scale by conducting a principal component analysis (PCA) of the standardized residuals (Linacre, 1998; Smith, 2002). The test could be recognized as essentially unidimensional with one strongly dominant dimension so all items were kept in the analysis after this step.

**Step 5. DIF analysis relating to assessment cycles.** None of the items demonstrated DIF relating to the assessment cycle based on the same criteria for DIF analysis relating to countries. It means that all items functioned in a similar manner for test participants with the same ability level in the baseline and follow-up assessments.

After item analysis we had 59 items left that could be used for equating.
**Step 6. Analysis of the whole scale.** The next part of analysis was devoted to the properties of the whole scale. Rasch analysis provides the reliability index and alternative statistics, separation index, that characterize the ability of the scale to reliably distinguish between individual participants (Stone, 2004). This index compares the distribution of person measures (estimates of ability) with their measurement errors and indicates the spread of person measures in standard error units. The index is used to estimate the number of distinct levels, or strata (separated by at least three errors of measurement), in the distributions (Wright & Stone,1979). The number of strata are calculated as: \( \text{Strata} = \frac{4G+1}{3} \), where \( G \) is the separation index. Our analysis produced the person reliability of 0.94, meaning that the proportion of observed student variance considered true was 94%. The classical reliability: \( \alpha = 0.79 \) for the scale, but it should be noted that classical reliability is not an appropriate measure in our case because of a big amount of missing data in our data design for equating. The person separation index was 4.01, indicating at least four statistically distinct groups of students along the iPIPS-Math continuum.

Figure 4 presents the Rasch variable map, which shows the relative distribution of all items and test takers from all countries for both cycles of assessment in a common metric. Specifically, the variable map depicts the joint distribution of items operationally defining the mathematics variable and the locations of children, based on their total correct scores, along this variable. The left column is the “logit” unit of measurement scale. On the map students are represented on the left side and the items are on the right. More difficult items and higher-performing students are located in the upper part of the map (positive logits), while easier items and lower-performing students are placed in the lower part of the map (negative logits).

The letter M on the right side of the dividing line shows the position of the mean of the item difficulties and the letter M on the left side shows mean of pupil abilities. The letters S and T mean one and two standard deviations correspondingly relating to the means. When a child’s name is opposite an item, s/he has a 50:50 chance of getting the item right. The further a child above an item the greater the chance s/he has of being able to answer it correctly and vice versa.
Figure 4: the iPIPS math variable map for the common scale
The distribution of students is wide and represents, for measurement purposes, good differentiation between higher and lower scoring students. The distribution of item locations, too, is also good because the span includes very easy items appropriate for less able students and very difficult items appropriate for advanced students. Furthermore, the progression of items from easier-to-more difficult represents a smooth, uniform, progressive continuum of increasing difficulty. The student sample is relatively well-located against the mathematics variables, which means that the test was appropriately targeted for the sample.

Step 7. Analysis of stability of item measures across samples. Six steps of scale construction were conducted for all three subsamples. The relative difficulties of the items for the three samples were then compared. The purpose was to check whether an item difficulty varies between samples, or in other words whether it depends on the sample of English and Scottish children chosen for analysis. The results showed no differences. Correlation of the difficulties of the items in three samples were all higher than 0.99. Additionally we calculated t-statistics that took into account the error of measurement for estimation of the item difficulty. None item had the value of t-statistics out of range (-2, +2) that means that all measures of item difficulty are stable across the samples. So the first sample was chosen for item difficulty estimation.

Step 8. Children estimation. Estimation of children’s math measures was conducted for each country separately with the whole samples for England and Scotland. During students’ estimation, the items difficulties were fixed and were not re-estimated. As a result we have measures of all students’ math ability for both baseline and follow-up cycles of assessment and for all countries on the same metric scale. It enabled valid comparisons to be made of children’s achievement from different countries at different time points.

6. How do children in different countries vary in what they know and can do starting school and after the first grade

Table 6 shows mean scores (in logits) and standard deviations of the math section of iPIPS for each country and cycle of assessment. We see a large difference between countries when children start school as well as when they finish the first grade. One factor which may explain this is the difference in the age of children starting school in different countries. The pupils in England tend to be the youngest, pupils in Scotland are six months older than their English counterparts on average, while pupils in Russia are about two years older. Even at the end of the
first grade, the English and Scottish pupils are about one year younger than Russian pupils when they start school.

*Table 6.* Average math-level of pupils across 3 countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Baseline assessment</th>
<th>Follow-up assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Deviation</td>
</tr>
<tr>
<td>England</td>
<td>-3.85</td>
<td>2.24</td>
</tr>
<tr>
<td>Scotland</td>
<td>-2.30</td>
<td>2.08</td>
</tr>
<tr>
<td>Russia</td>
<td>1.40</td>
<td>2.13</td>
</tr>
</tbody>
</table>

Dispersion analysis with one way ANOVA showed significant differences in average math level of pupils throughout three countries both at the start and at the end of the first school year.

Additionally Figure 5 also shows a very strong and clear relationship between mathematics experience and countries taking into account distribution of pupils’ measures, rather than only their means. The values on the y-axis are mean scores in logits while the box plots show the interquartile range of each country’s sample with the horizontal line inside showing the median and whiskers showing the minimum and maximum pupils’ scores. Two colored boxes for each country show the results for baseline and follow-up assessment cycles.
Figure 5: comparing countries in mathematics experience

The common scale for all countries and both cycles of assessment provides us with an opportunity to compare directly pupils’ performance at the start and at the end of the first grade, in other words, to measure individual progress for each child. Table 7 shows the average progress in math for the first school grade in different countries.

Table 7. Average relative progress (in logits) of pupils, a cross-country comparison

<table>
<thead>
<tr>
<th>Country</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>4.14</td>
<td>1.87</td>
</tr>
<tr>
<td>Scotland</td>
<td>3.88</td>
<td>1.78</td>
</tr>
<tr>
<td>Russia</td>
<td>1.99</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Dispersion analysis revealed significant differences in average relative progress of pupils throughout the three countries. The learning gain over the year, from baseline assessment to follow-up, was found to be slightly larger in England than in Scotland and much larger than in Russia.

The next analysis of the results relates to the comparison of pupils’ achievement depending on age. The children were put into 17 age categories corresponding to increments of 3 months each. The average scaled scores were then plotted against age to produce Figures 6 and 7 below.

The values on the y-axis in Figure 6 are mean scores in logits with error bars denoting the 95 percent confidence interval. The confidence intervals for Russia are wider than for England and Scotland, very likely because of the smaller sample.
The values on the y-axis in Figure 7 are just mean scores in logits of children depending on their age.

The first thing to notice from the figures is that math scores rise steadily with age for all countries, and it is true for both cycles of assessment. But the level of increase is different. It is larger for England and Scotland than for Russia. Russian children starting school later look more homogeneous in their math knowledge compared with English and Scottish children.
The second thing is that the patterns for England and Scotland are very much in line with one another. But the math scores of children in Scotland are slightly higher than of similar age children from England. It is true both for baseline and follow-up assessments where the difference between the countries is maintained.

The third thing is that the math scores of Russian children starting school are similar to those of English and Scottish children in the end of the first year of schooling, despite the fact that Russian children when they start school are one year older in average than English and Scottish children when they finish the first grade.

The fourth thing is that progress from starting to finishing the first grade is strong for all countries. It supports the claim that the first year of schooling is crucial for children’s development in all countries.

The fifth thing is that the progress is different between the countries. The Russian sample has rather less progress in math than in the other two countries. The possible reason for it can be the following. The time between the two assessments was much less in Russia than in England and Scotland: 6 months in Russia comparing to 10 months in other countries. Therefore, it is
expected that less progress would be seen in children in Russia. So, another study should be conducted in order to find out the possible reasons of such differences in the countries results. Given that rapid progress is made in the first year of school, as clearly demonstrated in England and Scotland, in the next study it is important for the time between the two assessments to be similar for Russia so as to make the results more comparable.

It should be emphasized that all findings are based on a very small Russian sample and need to be checked with a big sample.

7. Conclusion

iPIPS is an international monitoring survey of children starting school. It is being developed by the Centre of evaluation and monitoring at Durham University and has been used for twenty years in schools in England and Scotland. Also there were several international studies in English-speaking countries. These studies found that it was possible to use the same assessment for children starting school at different ages – in Australia, for example, children start school at age 6, while in England at age 4 and half in average. It also found that, while there was a clear relationship between children’s scores and their age when they started school, it was still possible to compare the progress of all national samples in a meaningful way. But all countries of this study are English-speaking countries. So it was especially important to check the potential of iPIPS to compare Russian children and children from other countries, because of different language and different age of Russian children.

The first findings of international comparative study of children from three countries starting school are presented in the paper. English children start school at the age of 4 and half in average. Scottish children start school five-six months later than in England. And Russian children start school at the age older than 7 on average. The attempt to compare the math achievements of children from different countries when they start school and when they finish the first grade has been made in the paper.

The study has shown that it is indeed possible to equate iPIPS scores in mathematics from the start and end of the first year at school across Scotland, England and Russia.

The findings of the comparison between countries are limited by the small Russian sample and the results should be checked with a larger representative Russian sample. The second limitation is connected with different time between the two assessments in Russia and the other countries.
The next study is planned during 2015 with a large Russian sample and the time between the two assessments to be longer, so as to make the results more comparable.

References


27


Authors:

1. Kardanova Elena, Higher School of Economics (Moscow, Russia). Center for Monitoring the Quality in Education. PhD. The Head;  
   e-mail: ekardanova@hse.ru  
   tel. +7 916 393 8649 (mob.)
2. Ivanova Alina, Higher School of Economics (Moscow, Russia). Center for Monitoring the Quality in Education. PhD. The researcher;  
   e-mail: aeiivanova@hse.ru  
   tel. +7 925 834 8120 (mob.)
3. Christine Merrell, Durham University (Durham, UK). Centre for Evaluation & Monitoring (CEM) . PhD. Director of Research  
   e-mail: christine.merrell@cem.dur.ac.uk
   e-mail: davidandjudyhawker@gmail.com
5. Peter Tymms, Durham University (Durham, UK). Centre for Evaluation & Monitoring (CEM). PhD. Director of iPIPS and international study of children starting school
e-mail: Peter.Tymms@cem.dur.ac.uk

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

© Kardanova, Ivanova, Merrell, Hawker, Tymms, 2014