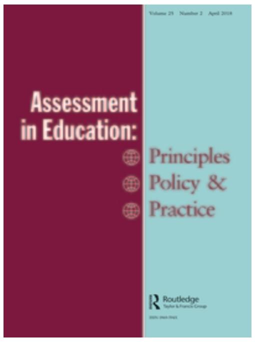

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

Checking the possibility of equating a mathematics assessment between Russia, Scotland and England for children starting school

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

To cite this article: Alina Ivanova, Elena Kardanova, Christine Merrell, Peter Tymms & David Hawker (2018) Checking the possibility of equating a mathematics assessment between Russia, Scotland and England for children starting school, Assessment in Education: Principles, Policy & Practice, 25:2, 141-159, DOI: 10.1080/0969594X.2016.1231110

To link to this article: https://doi.org/10.1080/0969594X.2016.1231110

Published online: 19 Sep 2016.

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE, 2018
VOL. 25, NO. 2, 141–159
https://doi.org/10.1080/0969594X.2016.1231110

Checking the possibility of equating a mathematics assessment between Russia, Scotland and England for children starting school

Alina Ivanova[a], Elena Kardanova[a], Christine Merrell[b], Peter Tymms[b] and David Hawker[b]

[a]Center of Education Quality Monitoring, Institute of Education, National Research University Higher School of Economics, Moscow, Russia; [b]Centre for Evaluation & Monitoring, Durham University, Durham, UK

ABSTRACT

Is it possible to compare the results in assessments of mathematics across countries with different curricula, traditions and age of starting school? As part of the iPIPS project, a Russian version of the iPIPS baseline assessment was developed and trial data were available from about 300 Russian children at the start and end of their first year at school. These were matched with parallel data from representative samples of equal numbers of children from England and Scotland. The equating of the scales was explored using Rasch measurement. A unified scale was easiest to create for England and Scotland at the start and end of their first year at school when children only differ by a half a year in age, and live in adjacent countries with a common language. Although fewer items showed invariance across the three countries, it was possible to link iPIPS scores in mathematics from the start and end of the first year at school across Scotland, England and Russia. The findings of this study suggest that, despite the apparent difficulties, meaningful comparisons of mathematics attainment and development can be made. These will allow for substantive interpretations with policy implications.

ARTICLE HISTORY
Received 27 April 2015
Accepted 24 August 2016

KEYWORDS
International; mathematics; baseline; primary school

Introduction

Despite the growing influence of international surveys of student achievement such as Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), there is currently no international baseline study of children's development on starting school. As a result, it is not possible to know 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 how far they are the result of differences in the effectiveness of schools, although a recent paper by Merry (2013) showed that the magnitude of PISA reading differences between Canada and the USA were paralleled in early childhood; this opens up possibilities on a wider scale.

CONTACT Alina Ivanova aeivanova@hse.ru

The results from PISA and TIMSS have a major influence on preschool policies in many countries, despite these assessments being of much older children. For example, the OECD (2012) reported 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.

Clearly, it is hard to conclude anything directly from PISA or TIMSS about the relative effectiveness of different countries' early years education policies, but countries are attempting to do this nonetheless. Additionally, the information gleaned from an assessment administered at a single time point at the start of school is limited. The first year of school is a time of rapid change for children's development and an assessment at the start and end of that important period not only provides valuable information about the effectiveness of schools at that time but also gives a more stable measurement basis from which to monitor progress up through the education system.

The Performance Indicators in Primary Schools (PIPS) baseline assessment (Tymms, 1999) was created by Tymms in 1994 and subsequently developed with Merrell. Over the years, it has been used to assess more than three 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. It is generally repeated at the end of the first year of school to provide a measure of children's progress during that year.

It has, for example, been used successfully in a number of countries for self-evaluation including Abu Dhabi, Australia, England, Germany, New Zealand, Scotland and South Africa. (Archer, Scherman, Coe, & Howie, 2010; Bäuerlein, Niklas, & Schneider, 2014; Tymms & Wylde, 2004; Wildy & Styles, 2008a, 2008b). As a result of the widespread use of PIPS, it has been possible to make comparisons between children starting school at different ages in English-speaking countries using PIPS (Merrell & Tymms, 2007; Tymms & Merrell, 2009; Tymms, Merrell, Hawker, & Nicholson, 2014). Building upon these studies, a new international comparative study of children starting school has been proposed called iPIPS. This project is intended to provide comparative, system-level information to policy-makers and researchers. It used the PIPS assessment, adapted and extended for the comparative work.

Previously published comparisons of children starting school using the PIPS assessment involved mainly English-speaking countries. The involvement of a sample from Russia with its different language and where children are, on average, 7 years old at the start of school presented an opportunity to explore the challenges of equating these data with samples from England and Scotland where the children are much younger at the start of school. This study focused on the iPIPS baseline assessment and follow-up, extending our understanding of the challenges and possibilities of making comparisons across countries of young children's development, which is an important contribution to the debate if meaningful conclusions are to be drawn about the effectiveness of countries' educational policies in future.

Early years education and care in Russia, England and Scotland

Russia, England and Scotland each have their own policies with regard to early education and care, which influence the type and amount of provision that children receive. They also have their own arrangements for the first year at school. A brief description of each is given in Appendix A.

The three educational systems – Russian, English and Scottish – have a number of features in common. First, all three countries place great importance on early childhood education and development. Second, preschool education is not compulsory in any of the countries, but the majority of children do attend. Thirdly, in all three countries, there is an understanding of the importance of baseline assessment.

The three schooling systems also differ significantly. First, they differ in the age of children starting school. Secondly, there are different country-specific traditions and cultures of assessment. For example, at present, in Russia, there are no standardised, valid assessments applicable to large-scale surveys for evaluating the initial level of a child starting school. Thirdly, the three countries have different curricula at the start of school, although all three include literacy and mathematics albeit in differing contexts with differing traditions and different foci.

The iPIPS baseline and follow-up assessment

The PIPS assessment was developed with the aim of providing teachers with a comprehensive profile of children's early reading and mathematics skills, and of their personal and social development at the start of school. This evolved over the years and now the iPIPS assessment can be efficiently administered on computer or with a paper manual accompanied by an app running on a smart phone or tablet. The app records responses and guides the administrator through the choice of items. The early reading and mathematics part takes between 15 and 20 min per child working on a one-to-one basis with the administrator. With the computer version, the software presents items to the child on screen with sound files. With the booklet and app version, the child sees the same pictures as for the computer version and the administrator asks the questions using the same script as the sound files. It is possible to collect a very reliable yet comprehensive measure of children at the start of school because iPIPS is adaptive, using sequences of items with stopping rules. The items are arranged into sections in order of increasing difficulty. The sections are described in the 'Instrument' section later in the paper. Each child begins with easy items and moves on to progressively more difficult ones. When they make a number of errors, the assessment progresses to 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.

The system is straightforward to use and very popular with schools. Over the years, the assessment has 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 for children starting school (Tymms et al., 2014). It has also proved to have extremely good predictive validity, with correlations of around 0.68 to later national assessments at ages 7 and 11, and of around 0.5 to the national examinations at age 16 (Tymms, 1999; Tymms, Merrell, Henderson, Albone, & Jones, 2012).

Adaptation of the PIPS assessment for use in Russia

Adaptation is not just translation. It includes many activities ranging from decisions about whether or not the same construct can be assessed in a different language and culture to checking equivalence of the initial and adapted assessment versions (Hambleton, 2005). The validity of comparisons using an adapted assessment critically depends on the degree to which the adapted versions do indeed measure the intended constructs and provide comparable measurements (Ercikan & Lyons-Thomas, 2013).

Several different assessment adaptation processes exist including parallel, successive, simultaneous and concurrent development of different language versions of assessments (Ercikan & Lyons-Thomas, 2013). To develop the Russian version of the PIPS baseline assessment, the method of successive assessment adaptation was used whereby assessments that are developed for one language and culture are subsequently adapted to other cultures. Therefore, the conceptualisation of the construct being assessed is based on one culture, the wording of assessment items, the actual items included in the assessment, how they should be evaluated and how they relate to the construct. These items are all based on the culture for which the assessment was originally developed.

In developing the Russian version, the main task was to ensure, so far as was possible, the equivalence of the assessments 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. The process of assessment adaptation involved input from specialists of differing perspectives, translators, cultural and linguistic reviewers and teachers. Back-translation was used to check the equivalence of the different language versions of the assessments. All the Russian items were translated back into English and compared with the original items by experts (both English and Russian) and with the iPIPS developers. Criteria for evaluation included (1) differences in the 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, Gierl, McCreith, Puhan, & Koh, 2004). All translation errors were documented and discussed, and items were revised.

Thus, firstly, the items from the English version were translated into Russian by two independent translators. After editing and further discussing the final translation, Russian subject specialists verified the suitability of the content. Subsequently, the Russian booklet was back-translated into English and the items compared with the original version.

Secondly, the administration procedure was standardised. To do this, the team which adapted the version for Russia discussed the procedure with the original authors of the assessment and then produced guidance to ensure that it was being administered in an equivalent way in both countries.

It has already been noted that the ages of the target populations in the three countries differed significantly. Additional items were added to the Russian version to try to avoid the assessment reaching a ceiling, particularly on the second, follow-up assessment later in the school year. Some of the very easy items that all children in Russia were able to answer correctly were omitted from the Russian version.

For the study, it was necessary to confirm the equivalence of the adapted assessments in measurement terms. Two approaches were used: (a) Rasch measurement theory analysis of assessment items and assessments (comparisons of item characteristics, item maps, item hierarchy, dimensionality, etc. for two language versions); and (b) identification of differential item functioning (DIF) across countries and within country variables.

The dichotomous Rasch model (Wright & Stone, 1979; Andrich, 1988) was used for data analysis. It transforms children's raw scores into measures on an equal interval scale. In this model, each assessment item is characterised by one parameter, (difficulty), and each assessment participant is also characterised by one parameter (ability). Rasch analysis places participants and items on the same log-odds measurement scale (logit) with an arbitrary unit. The reasons for choosing the Rasch model are both psychometrical and practical. Firstly, the Rasch model has optimal metric properties, and secondly, from a practical point of view, it is useful for parameter estimation and data analysis – empirically determining the quality of assessment items, constructing scales and carrying out assessment equating (Bond & Fox, 2001). Winsteps software (Linacre, 2011) was used for this process.

An item demonstrates differential item functioning (DIF) if assessment participants with the same ability level who belong to different groups have markedly different chances of completing that item correctly. Two methods – Mantel–Haenzel (MH) and Logistic Regression (LR) – were used, according to circumstances, to check DIF in this study (Dorans, 1989; Zumbo, 1999).

The Mantel–Haenszel DIF detection method is one of the most commonly used tests for detecting differential item functioning. It consists of comparing the item performance of two groups of participants, whose members were previously matched on the ability scale. The matching is carried out using the observed total test score as the criterion or matching variable. To test for DIF (across countries and across assessment cycles) with MH method, we used the Educational Testing Service (ETS) approach for DIF classification (Zwick, Thayer, & Lewis, 1999), which designates items as A (negligible or non-significant DIF), B (slight DIF) or C (large DIF) items depending on the magnitude of the difference and the statistical significance as found using the Mantel–Haenszel statistic (Dorans, 1989). An item was considered a C item if two conditions were satisfied: (1) the difference in item relative difficulty between different groups of students was more than 0.64 logits and (2) the Mantel–Haenzel statistic had a significance level of $p < .05$ (Linacre, 2011).

The LR method is also commonly used for detecting DIF. It is based on statistical modelling of the probability of responding correctly to an item as a logistic function of at least one or more predictor variables. Predictors include the total score as the ability measure, a grouping variable and the interaction between ability and group. An item is identified as DIF item when the latter two variables show a significant improvement in the data–model fit beyond a model that includes only ability (Zumbo, 1999). The variables are entered into the model in this order: (step #1) total score, (step #2) group and (step #3) the interaction term of ability and group. Such modelling allows to identify the presence of DIF (comparisons between the models at step 3 versus step 1), as well as the type of DIF, non-uniform and uniform. To identify the type of DIF, comparisons between the models at step 3 versus step 2 and step 2 versus step 1, respectively, should be made. In the framework of Rasch measurement, the non-uniform DIF is not a specific target of DIF analysis and it is considered rather as violation of model assumptions. But we included the identification of DIF type because it can give additional information.

Thus, DIF was identified by comparing models from step 3 (the full model) compared to step 1 (the ability only model). As Zumbo (1999) suggested, for an item to be classified as displaying DIF, the two degree-of-freedom chi-squared test in LR had to have a p-value less than or equal to 0.01 and the Zumbo–Thomas effect size measure had to be at least an R-squared of 0.13. To measure the magnitude of DIF, we used the Zumbo and Thomas

(1996) approach for DIF classification, which designates items in three categories: items which exhibited negligible DIF (*R*-squared values below 0.13), moderate DIF (*R*-squared values between 0.13 and 0.26) and large DIF (*R*-squared values above 0.26). Both the moderate and large categories also required the item to be flagged as statistically significant with the two degree-of-freedom chi-square test. After this process, to identify the type of DIF, comparisons were made between the models at step 3 versus step 2 and step 2 versus step 1 to determine the presence of non-uniform and uniform DIF.

The reasons to use these two methods for DIF analysis were the following. Firstly, MH and LR methods are the most often used. Second, although the Russian sample size was relatively small, it is sufficiently large to use MH and LR methods (Narayanan & Swaminathan, 1994; Zumbo, 1999). Third, taking into account the different ages of the target populations in the three countries, we assumed that ability distribution differences between the groups of participants would exist. It is known that the differences in ability mean and variance increase the Type I error rate for both DIF detection methods, but especially for MH (Narayanan & Swaminathan, 1994; Pei & Li, 2010).

In conducting DIF analysis, an item was considered as an item with DIF if two conditions were satisfied: (1) the MH method designated the item as C item (large DIF) and (2) the LR method designated the item as moderate or large DIF item.

After DIF detection, items that were identified as DIF were omitted, and the total score was recalculated. This recalculated total score was used as the matching criterion for a second DIF analysis to ensure the matching of groups was appropriate. Additionally, to investigate the sources of DIF, all items identified as DIF were analysed for content and cultural relevance.

To confirm the measurement equivalence of two assessments, it is necessary to establish a measurement unit and scalar equivalence. Scores from different adaptations of the same assessments cannot be considered comparable without a score linking exercise. Different methods can be used, but the most appropriate for this study was thought to be separate monological group design (Sireci, 1997). This employs a set of items found to be equivalent in the two versions as anchor items in Rasch-based calibration. It is especially challenging to develop equivalent versions of verbal items where culture and language have potentially large differential impact. In the present study, we considered only mathematics items for comparison between countries.

Method

Participants

The Russian sample consisted of 310 children recruited from 21 classes of 21 schools in the Novgorod region, 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 were similar to those in the country as a whole, based on the 2010 census (Social and demographic portrait of Russia, 2010). For example, the distribution of the region's population by educational level (62% college and above; 30% high school; 8% below high school) was similar to the national figure (65% college and above; 29% high school; 6% below high school), as was the ratio of urban to rural students in the region (72% urban; 28% rural).

The target population was children enrolled in first grade on the 1 September 2013. The sample represented about 5% of all the grade 1 students of this region. The sample was randomly selected after stratification on two parameters: (i) the school location (rural or urban area) and (ii) the different status of schools (there are three main types of schools in Russia: comprehensive (general regular) schools, schools specialising in a certain subject and gymnasia (some of them fee-paying)). All the chosen schools consented to participate. After parental consent was obtained (the majority of parents gave permission for their children to participate in the study), children were randomly selected within the selected classes.

The first cycle of assessment was administered in mid-October 2013. The second follow-up assessment was administered during the fourth week of April 2014. Ten per cent of pupils were absent during the second cycle. Tables 1 and 2 give details of the achieved sample for the two assessment cycles.

The Russian sample differed from both the English and the Scottish samples by the age of children and the sample size. Table 3 shows these differences.

The origin of the samples for England and Scotland and how their representativeness was established can be found in Tymms et al. (2014) and are based on PIPS data which were collected already.

Instrument

The final version of the Russian PIPS assessment was structured in the same sections as the original English version and used the same algorithms. Table 4 shows the content of the English and Russian assessments for the mathematics part.

Table 1. The Russian sample, October 2013.

Gender (%)		Place of living (%)		Type of school (%)	
Female	49	Urban	71.6	Gymnasium	16.1
Male	51	Rural	28.4	Specialised school	21.9
				Comprehensive school	61.9
In total: 310 pupils					

Table 2. The Russian sample, April 2014.

Gender (%)		Place of living (%)		Type of school (%)	
Female	49.8	Urban	70.8	Gymnasium	16.6
Male	50.2	Rural	29.2	Specialised school	20.9
				Comprehensive school	62.5
In total: 277 pupils					

Table 3. Average age of children at the time of the first assessment and numbers.

Country	Mean age in years	Number of participants in the baseline assessment	Number of participants in the follow-up assessment
England	4.56	6985	5837
Scotland	5.09	6627	6627
Russia	7.33	310	277

Table 4. Content of booklets in two versions.

English version	Russian version
Understanding of mathematical concepts (bigger, smaller, etc.)	Not included
Counting and numerosity of four and seven objects	Not included
Simple sums presented informally using pictures	The same
Recognition of single-digit numbers and then teens followed by two and three digits	Very similar starting with teens and including four- and five-digit numbers
Recognition of shapes and patterns	Not included
Counting on with dots as an aide	The same
More advanced calculations, some presented with formal notation	The same
Simple applied math problems	The same plus more difficult items

The first piloting in October 2013 in Russia suggested a ceiling effect on some sections. For the second cycle of the assessment, these sections were extended with items that were intended to be more difficult and some items were omitted.

All items in the baseline and follow-up assessments for the three countries were of the same type: they were short questions asked by the assessor requiring a short answer.

Data collection

The Russian children were assessed by specially trained assessors using the booklet and app.

In England and Scotland, the children were assessed by the staff in the school which they attended using the computer-delivered version.

Results: linking the English, Scottish and Russian data

There were six data-sets in total, baseline and follow-up for the three countries. Simultaneous Rasch equating was used to link and compare the results from all six data-sets (Wolfe, 2004). During this procedure, each item is either treated as common to at least two countries or as unique. Thus, the overlap between subsets of data allows us to simultaneously estimate parameters for the Rasch model.

To conduct the analysis, random subsamples of comparable size to the Russian data were created from the available English and Scottish baseline assessment samples. The same children were chosen from the follow-up assessment samples. Thus, we had a single matrix for equating, with data on children from three countries who had been assessed both at the start and at the end of the year. The total sample size was 1867 students. The total number of items was 81, including both common and unique items. There were 37 common items between all countries, 25 items were unique for Russia and 19 unique for England and Scotland. The data analysis was performed in several steps as follows[1]:

Step 1. Analysis of model fit. Items with low discrimination and/or those that did not fit the model were deleted. This applied to 3 of the 81 items (two common items and one Russian item). Two England and Scotland items were dropped from the analysis because of extreme difficulty. No further substantial or technical problems were identified. Thus, 76 items were left in the analysis after this step, with 35 common items between the three countries).

Table 5. DIF items across country (LR method).

Item	*R*-squared values at each step in the sequential hierarchical regression			DIF χ^2 (df=2) test	DIF *R* squared		
	Step #1	Step #2	Step #3		ΔR^2 (step 3–1)	ΔR^2 (step 3–2)	ΔR^2 (step 2–1)
I255	,348	,547	,547	309,457 *p=.000*	,199	,000	,199
I258	,293	,497	,504	293,901 *p=.000*	,211	,007	,204
I261	,024	,657	,657	684,044 *p=.000*	,633	,000	,633
I305	,351	,528	,533	224,112 *p=.000*	,182	,005	,177
I308	,175	,412	,420	163,042 *p=.000*	,245	,008	,237
I311	,016	,408	,422	145,541 *p=.000*	,406	,013	,392

Table 6. Items showing DIF.

Item ID	List of items	Direction of DIF
I255	Number identification: teen 1	Ru>En,SC
I258	Number identification: two digit	Ru>En,SC
I261	Number identification: three digit	Ru>En,SC
I305	Look at this set of numbers. What should be there instead of the asterisk? 10 20 30 40 *	En,Sc>Ru
I308	Can you do this sum? 4+11=	Ru>En,SC
I311	Can you do this sum? 15–4=	Ru>En,SC

Step 2. Country-related DIF analysis. Firstly, DIF analysis was conducted across England and Scotland. No items exhibited DIF in accordance with the chosen criteria. This is understandable because children in England and Scotland only differ by a half a year in age, and live in adjacent countries with a common language. For further country-related DIF analysis, Russian sample and joint English and Scottish sample were considered.

LR analysis revealed that six items exhibited moderate or large DIF. Table 5 lists the results from the DIF analysis of the detected items.

Although the exact type of DIF was not of concern, the analysis was conducted to understand what appeared to be occurring. As the last two columns in Table 5 display, all items were uniform DIF items: the difference in *R*-squared from Step #2 to Step #3 was quite small compared to the difference from Step #1 to Step #2.

The MH method revealed that eight items exhibited large DIF (C items), and six of them exhibited DIF according to the LR method. Thus, our analysis revealed that six items exhibited DIF in accordance with the two methods. The six items with DIF appeared in several different sections, including recognition of numbers, use of arithmetical operations and logic sequencing.

Table 6 lists these items and the direction of DIF. In the table, we use the following notations for DIF direction: Ru>En,SC, that means DIF in favour of Russia, that is to say the items were relatively easy for Russian children compared to children from England and Scotland of similar math attainment. We see that five items demonstrate DIF in favour of Russia and one item in favour of England and Scotland.

After reviewing the DIF items, we explored possible causes of DIF for the 6six items. Just why the items should vary in relative difficulty across countries is not clear but it is doubtless due, in general terms, to differences in age, the practices of pre-schools and the upbringing at home. Interestingly though, this 'why' question is it is not of concern for this paper; rather, we need to delete the items that exhibit DIF from the linking procedure.

Seventy items remained at this stage. Among them, there are 29 common items, 24 items unique to Russia and 17 items unique to England and Scotland. After the DIF items were removed, all the remaining items were assessed again for DIF across countries. Based on LR method, no items exhibited DIF now.

Step 3. Dimensionality study. We examined the dimensionality of the scale by conducting a principal component analysis (PCA) of the standardised residuals, which are the differences between the observed response and the response expected under the model (Linacre, 1998; Smith, 2002). The scale was essentially unidimensional with one strongly dominant dimension and no further items were dropped.

Step 4. DIF analysis relating to assessment cycles. DIF analysis across cycles was conducted with the same approach as across countries. Fifty-five items were used for both cycles, baseline and follow-up. Figure 1 shows item relative difficulties separately from different cycles of assessment – baseline and follow-up. The majority of items demonstrate stable estimates of their relative difficulty, which means that the items function in a similar manner at baseline and follow-up, so they are DIF free. Only three items were detected as DIF items, which included recognition of three-digit numbers (two items) and applied math problem (one item). Taking into account the small size of DIF for these items, we decided to keep them in the analysis.

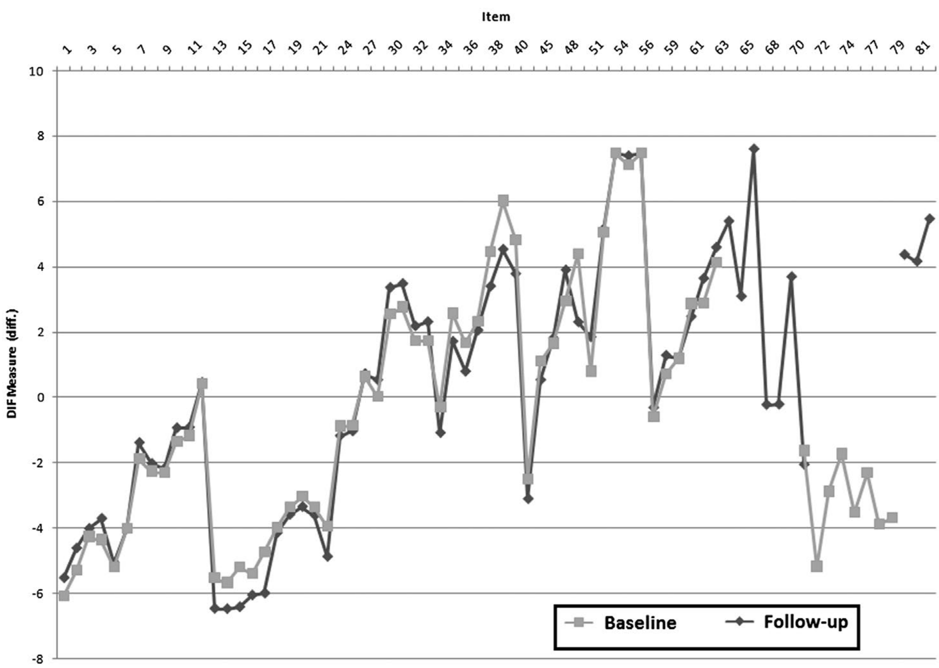

Figure 1. Item relative difficulties for different countries.

Step 5. Analysis of the whole scale. The next part of analysis was devoted to the properties of the whole scale. Our analysis produced a person reliability of 0.95, meaning that the proportion of observed person variance considered true was 95%.

Figure 2 presents the Rasch variable map, which shows the relative distribution of all items and assessment takers from all countries for both cycles of assessment in a common metric.

The distribution of students is wide and, for measurement purposes, clearly differentiates between higher and lower scoring students. The distribution of item locations 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 continuum of increasing difficulty. The student sample is well located relative to the mathematics items, which means that the assessment was targeted for the sample.

To conclude, although only 29 common items showed invariance across the three countries, it was possible to equate iPIPS scores in mathematics from the start and end of the first year at school across Scotland, England and Russia. However, it is acknowledged that deleting items can reorient the variance.

Children estimation. Estimation of children's math measures was conducted using the model outlined above. As a result, we have measures of the whole samples in terms of math

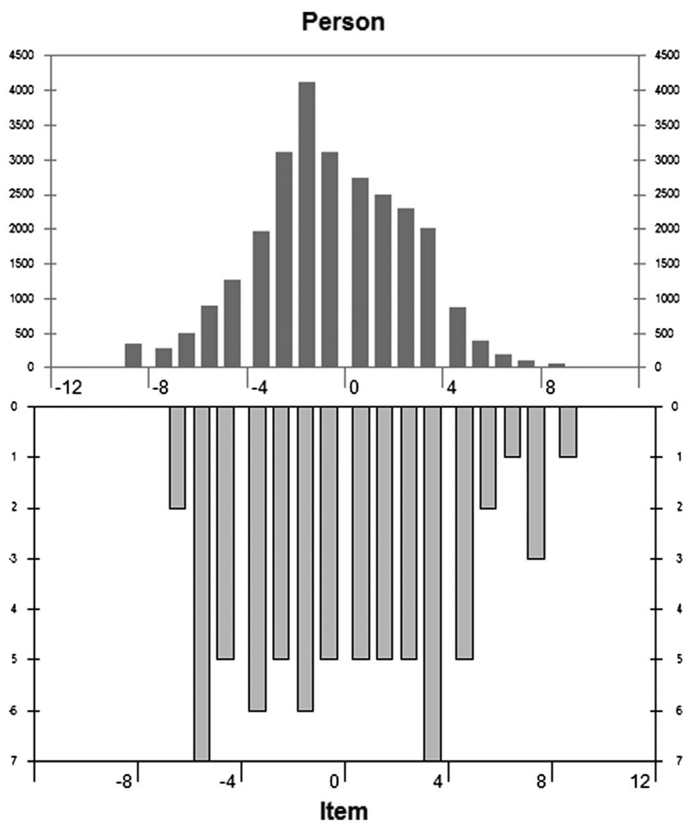

Figure 2. The iPIPS math variable map for the common scale.

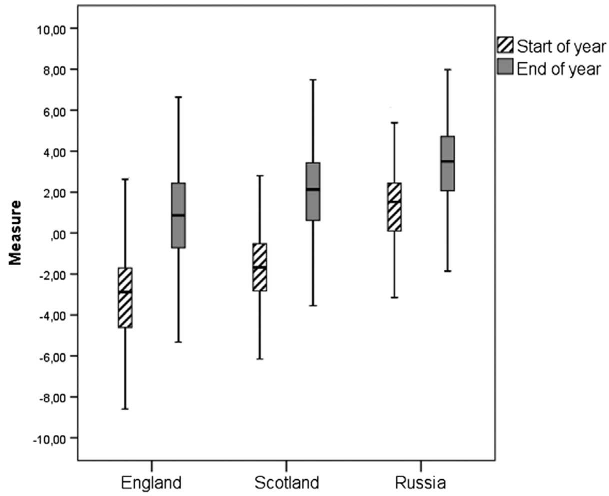

Figure 3. Box-and-whisker plots of math attainment in the three countries on the two occasions.

Table 7. Average math level of children and progress across three countries.

Country	*Start of year* Mean	SD	*Follow-up* Mean	SD	Mean difference	SD of difference	Progress per month
England	−320	221	086	235	408	180	045
Scotland	−173	198	207	219	384	169	043
Russia	149	185	344	195	197	108	032

ability for both baseline and follow-up cycles of assessment and for all countries on the same metric scale. This allowed us to make valid comparisons of children's achievement from different countries at different time points.

Results: variation across countries

Figure 3 shows box-and-whisker plots of the math attainment of the children in the samples for the three countries at the start and end of the year.

The chart shows a considerable range of math performance from the weakest children starting school in England with some who were not able to count four objects to the strongest children in Russia at the time of the second assessment who were able do formal sums such as 42–17.

The chart shows the very clear progress made by each country's cohort between the start and end of the year. And despite the differences, there is a considerable overlap between all the cohorts.

The chart also shows that the median score for Scotland was higher than for England on both occasions and that medians for Russia were higher still.

One-way ANOVA showed significant differences ($p < 0.01$) between the average math levels of children in the three countries both at the start and at the end of the first school year. Table 7 illustrates this final point.

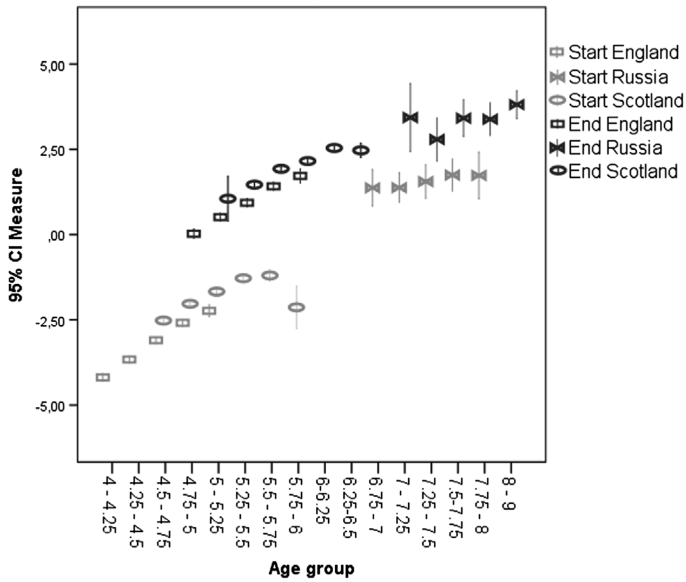

Figure 4. Three country age-related comparisons.

The table also shows that the learning gain from baseline assessment to follow-up was found to be larger in England than Scotland (slightly) or Russia (markedly). This difference is partially explained by shorter time between the two assessments in Russia: 6 months as against between 8 and 9 months in the other countries. To provide a fairer comparison, we computed the progress per month. This is presented in the last column of the table. The average progress per month is still less for Russia than in the other countries. Possible reasons for this are picked up in the discussion section later.

The next analysis of the results relates to comparisons of children's achievement to age. The children were put into 17 age categories corresponding to increments of 3 months. The average scaled scores were then plotted against age to produce Figure 4 below.

The values on the y-axis in Figure 4 are mean scores in logits with error bars denoting the 95% confidence interval. The confidence intervals for Russia are wider than for England and Scotland because of the smaller sample of children.

Figure 4 shows that, within confidence intervals, the math scores tend to rise steadily with age, and this holds true for both cycles of assessment and for all three countries. The strength of this relationship is stronger for the younger the cohort, which coincides with differences between countries.

Second, the patterns for England and Scotland are very much in line with one another, although the scores of children in Scotland are slightly higher than for children of a similar age in England at baseline and follow-up assessments.

Third, 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 at this point in time they are considerably older. Nevertheless, their scores more or less coincide with an extrapolated line from the English and Scottish children starting school.

Fourth, progress from starting school to the end of the first year is strong for all countries, although less so for in Russia. This supports the claim that the first year of schooling is crucial for children's development.

Conclusion

The primary focus of this paper is methodological. Our research set out to see if Rasch measurement procedures could be applied to mathematics attainment measures so that they could reasonably be compared across very different situations. It has shown that it is possible to equate attainment in mathematics at different ages (4 to 7) in different countries (England, Scotland and Russia) at the start of school and at the end of the first year. A small Russian sample from only one region of Russia is a limitation of the study; so to confirm the conclusion, it is necessary to repeat the study with a big sample. The present research has shown the potential possibility of equating, which provides a proof of concept.

It follows that an international study of children starting school with a one-year follow-up is possible and we hypothesise that the more fundamental the measure and less culturally tied, the more it will be possible to equate measures across countries. We expect, for example, that short-term memory measures will be easier to equate than mathematics which will in turn be easier than reading. A highly language-specific construct, such as rhyming, will be close to impossible to equate across different languages.

In designing an international study of children around the start of their school career, an important question arises as to whether the study should be age or stage based. Figure 4 makes it clear that a purely aged-based study could produce data which are very difficult to interpret because of the major impact of schooling. Consider a survey conducted with children who had finished their first year at school in England and Scotland but had yet to start in Russia; the surveyors would conclude that the English and Scottish children were, on age-corrected scores, ahead of children from Russia. But, if the survey focused on a time before all children had started school, extrapolation of the data in Figure 4 suggests that the researchers would reach a very different conclusion. It therefore makes sense to collect data at the start and end of the first year of school in each country and estimates can then be made of attainment at different ages with and without a year at school, and the link between age and attainment can be established. Slopes can in themselves be seen as measures worthy of study (Burstein, Kim, & Delandshere, 1989).

The Russian data available for this paper, although widely based, were from a small sample from one region and, although the region was chosen to reflect the wider Russian demography, it cannot be said to be truly representative of the country as a whole because of the huge variations between the different regions. Therefore, no conclusions can be made about Russia's educational system as a whole. However, it is possible to set out a number of questions which could be tackled if, or when, a larger representative sample becomes available from Russia and other countries.

(a) To what extent does the on-entry and follow-up data predict PISA performance?
(b) To what extent do preschool policies relate to on-entry developmental levels, progress measures and the age-/developmental-level gradients?
(c) How do developmental levels vary across schools and to what extent is this related to social segregation?

(d) To what extent do relative progress (value-added) measures vary from school to school?
(e) How do (a) and (b) compare to other countries?
(f) If the data can be linked to performance at the end of elementary school across countries, do they suggest an optimum age for starting school?These are the key policy questions which have inspired the proposal to establish an international study of children starting school. This paper has demonstrated the technical feasibility of using the PIPS assessment to compile the data needed to start on this journey.

Note

1. The data and syntax are available from the authors by request.

Acknowledgements

Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.
Support from Durham University is gratefully acknowledged.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Alina Ivanova is a junior researcher, Centre of Education Quality Monitoring, National Research University Higher School of Economics, Russia. She graduated the master's programme 'Measurements in psychology and Education'. Her research interests include psychometric, primary education and large-scale educational assessments.

Elena Kardanova is an associate professor and director of Centre of Education Quality Monitoring, National Research University Higher School of Economics, Russia. She has been an expert in the assessment of the quality in education at CICED (Centre for International Cooperation in Education Development) in Moscow. Her main research interests include assessment, psychometric and test development.

Christine Merrell is the director of Research, Centre for Evaluation and Monitoring (CEM) and reader in School of Education, Durham University. Christine has extensive experience in the development of assessments for children aged between 3 and 11 years. Her other research interests include the academic attainment and progress of severely inattentive, hyperactive and impulsive young children and ways to help them succeed in the classroom.

Peter Tymms is the director of iPIPS and professor of Education , School of Education, Durham University. He is also an adviser to the German NEPS project, led the start of the Online Educational Research Journal and started the PIPS project. His main research interests include monitoring, assessment, performance indicators, ADHD, reading and research methodology.

David Hawker is a professor of the College of Teachers, an honorary professor at the University of Durham and a visiting professorial fellow at the Institute of Education in London. He runs his own international education consultancy. He has been an adviser to several international organisations, including the OECD, the World Bank and the Open Society Foundations.

ORCiD

Alina Ivanova http://orcid.org/0000-0003-3340-7651
Peter Tymms http://orcid.org/0000-0002-7170-2566

References

Andrich, D. (1988). *Rasch models for measurement*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-068. Beverly Hills and London: Sage.

Archer, E., Scherman, V., Coe, R., & Howie, S. J. (2010). Finding the best fit: The adaptation and translation of the Performance Indicators for Primary Schools (PIPS) for the South African context. *Perspectives in Education, 28*, 77–88.

Bäuerlein, K., Niklas, F., & Schneider, W. (2014) Fähigkeitsindikatoren Primarschule (FIPS) – Überprüfung des Lernerfolgs in der ersten Klasse [Performance indicators in primary school (FIPS) – Verification of first grade learning success]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Jahrbuch der pädagogisch-psychologischen Diagnostik. Tests und Trends, Bd. 12 Lernverlaufsdiagnostik* (pp. 127–144). Göttingen: Hogrefe.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: King's College London School of Education.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model*. Mahwah: Lawrence Erlbaum.

Burstein, L., Kim, K. S., & Delandshere, G. (1989). Multilevel investigations of systematically varying slopes: Issues, alternatives, and consequences. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 194–211). New York, NY: Academic Press.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: standardization and the mantel-haenszel method. *Applied Measurement in Education, 2*, 217–233.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 3, pp. 545–569). Washington: American Psychological Association.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301–321.

Federal State Statistics Service. (2010) Social and demographic portrait of Russia. Retrieved September 1, 2014, from http://www.gks.ru/free_doc/new_site/perepis2010/croc/Documents/portret-russia.pdf

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah: Lawrence Erlbaum.

Kolchanova, S. S. (2012). Startovaya diagnostika pervoklassnikov kak osnova planirovaniya individual'nykh obrazovatel'nykh trayektoriy. *Regional Education in XXI century: Problems and prospects, 1*, 11–14.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266–283.

Linacre, J. M. (2011). *A user's guide to WINSTEPS. Program manual 3.71.0*. Retrieved September 1, 2014, from http://www.winsteps.com/a/winsteps.pdf

Merrell, C., & Tymms, P. (2007). What children know and can do when they start school and how this varies between countries. *Journal of Early Childhood Research, 5*, 115–134.

Merry, J. J. (2013). Tracing the US deficit in PISA reading skills to early childhood: Evidence from the United States and Canada. *Sociology of Education, 86*, 234–252.

Narayanan, P., & Swaminathan, H. (1994). Performance of the mantel-haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315–328.

Novoselova, Y. M. (2012). O pervoklassnikakh goroda Tyumeni. *Regional Education in XXI century: Problems and prospects, 1*, 14–17.

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, mantel-haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement, 34*, 453–456.

PISA, OECD. (2012). Results in focus. What 15-year-olds know and what they can do with what they know. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf

Scottish Government. (2010). *The building the curriculum 5: A framework for assessment*. Edinburgh: The Scottish Government.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*, 12–19.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205–231.

Tymms, P. (1999). *Baseline assessment and monitoring in primary schools: Achievements, attitudes and value-added indicators*. London: David Fulton.

Tymms, P., & Merrell, C. (2009). On-entry baseline assessment across cultures. In A. Anning, J. Cullen, & M. Fleer (Eds.), *Early childhood education: Society & culture* (2nd ed., pp. 117–129). London: Sage Publications.

Tymms, P., & Wylde, M. (2004). Basis pruefverfahren und Dauerbeobachtung in der Grundschule [Basic testing procedures and observation in primary school]. In G. Faust, M. Götz, H. Hacker, & H.-G. Roßbach (Eds.), *Anschlussfaehige Bildungsprozesse im Elementar- und Primarbereich* (pp. 190–203). Bad Heilbrunn: University of Bamberg, Verlag Julius Klinkhardt.

Tymms, P., Merrell, C., Henderson, B., Albone, S., & Jones, P. (2012) Learning difficulties in the primary school years: Predictability from on-entry baseline assessment. *Online Educational Research Journal, June 2012.* Retrieved October 6, 2014, from www.oerj.org

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: Research report*. London: Department for Education. Retrieved October 6, 2014, from https://www.gov.uk/government/publications/performance-indicators-in-primary-schools

Wildy, H., & Styles, I. (2008a). Measuring what students entering school know and can do: PIPS Australia 2006–2007. *Australian Journal of Early Childhood, 33*, 43–52.

Wildy, H., & Styles, I. (2008b). What Australian students entering primary school know and can do. *Journal of Australian Research in Early Childhood Education, 15*, 75–85.

Wolfe, E. W. (2004). Equating and item banking with the Rasch model. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 366–390). Maple Grove: JAM Press.

Wright, B. D., & Stone, M. N. (1979). *Best test design. Rasch measurement*. Chicago: Mesa Press.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential Item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1996). A measure of DIF effect size using logistic regression procedures. Paper presented at the National Board of Medical Examiners, Philadelphia, PA. Retrieved from http://scholar.google.com/scholar_lookup?title=A%20measure%20of%20DIF%20effect%20size%20using%20logistic%20procedures.&author=B.%20D.%20Zumbo&author=D.%20R

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical bayes approach to mantel-haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28.

Appendix A. Early years education and care in Russia, England and Scotland: a brief description of educational systems

The Russian system

The political, social and economic transformations that took place in Russia at the end of the 1980s and the beginning of the 1990s significantly influenced the Russian education system, which was experiencing difficulties during those years.

On 1 January 2014, a new Federal Standard for preschool education was established. This states that the aim of preschool educational programmes must be the diversified development of preschool age children including their physical, artistic and aesthetic, social and communicative, cognitive and speech development. Preschool educational programmes should be directed towards 'purpose orientation', which is the range of social and psychological characteristics of a child's achievements. These purpose orientations suppose that a preschool age child at the stage of finishing preschool education should already have prerequisites for educational activity fully formed. The Standard does not prescribe any form of pupil assessment, but it does prohibit its use for selection.

Attending a kindergarten is not obligatory, but about 90% of preschool children attend for at least one year, just before school. Children can start elementary school at any time from the age of 6 years and 6 months if there are no contraindications connected to their state of health, and they must have started before their eighth birthday. A school year usually starts on the 1 September and lasts 34 weeks (33 weeks in the first grade).

Since 1 September 2011, all educational institutions in Russia have been required to adhere to the new Federal educational standard for elementary general education. There is no national curriculum, but instead all schools are required to develop their own basic educational programme. The educational programmes vary depending on different educational and methodological complexes (the complex includes a set of course books, guidelines for teachers, workbooks, etc.). Schools select from about 10-12 complexes, based on which they can form their curriculum. Different classes in one school can use different complexes.

The assessment of first-grade pupils is accomplished mainly with the help of techniques that do not always have evidence of reliability and validity (Kolchanova, 2012). Teachers usually summarise the results of the children's diagnostics in free form, divide children into groups according to their preparedness levels, from low to high, or create individual profiles. The results of such assessments are used by teachers to plan lessons; by head teachers to prepare public reports; and by parents. They also can be used by educational managers and education quality control services.

At present, in Russia, there are no standardised, valid assessments applicable to large-scale surveys for evaluating the initial level of a child starting school. Some small-scale initiatives do exist at local levels (in particular regions or municipal districts) to organise small measurements of first-year pupils' preparedness to school (Novoselova, 2012).

The English system

Although not compulsory, most children start school in the Reception year when they are aged 4, prior to the statutory school starting age of 5 when the National Curriculum begins.

Preschool education is provided by a mixture of state, private and voluntary sector organisations, but the government funds all children aged 3 and 4 on an equal basis for 15 h per week, and makes similar funding available to two-year olds from deprived backgrounds. Parents are allowed to pay for additional hours if they wish. Early years providers are expected to operate according to the published 'Early Years Foundation Stage' standards, and the quality of early years provision is inspected and regulated by the national school inspectorate, Ofsted. The Early Years Foundation Stage extends to the end of the child's first year of school, the Reception year. During the Reception year, teachers will begin to teach reading and more formal methods of calculation will be introduced as appropriate to the stage of development of each child. There is currently a requirement for teachers to assess children's development at the end of the Early Years Foundation Stage by means of the Early Years Foundation Stage Profile and the scores from these profiles are collated centrally.

From 2014, the accountability system has been strengthened by introducing a government expectation for schools to demonstrate good progress. To support this policy, the government has proposed that a 'baseline assessment' should be administered to children on entry to school. A number of baseline assessments produced by assessment development organisations will be accredited for this purpose, from which schools can choose one that is most suitable for their context. The policy will be fully implemented by 2016.

The Scottish system

In Scotland, the statutory school starting age is 5 as in England, and there is a similar pattern and funding of preschool provision. The early years curriculum was introduced in 2010 and is set out in the 'Curriculum for Excellence' document issued by Education Scotland on behalf of the Scottish Government, which covers ages 3–18. It provides high-level guidance from which local education authorities, in collaboration with schools, are expected to develop the detail of the curriculum locally rather than imposing a prescriptive approach. A framework for assessment was also provided (Scottish Government, 2010). This built on a previous document: 'Assessment is For Learning (AiFL)', which in turn was underpinned by the research of Black and Wiliam (1998). Black and Wiliam proposed that the wealth of information about pupils' learning, progress and difficulties could be used by both teachers and the pupils themselves to inform subsequent learning, i.e. for formative purposes. Exemplars were made available via the National Assessment Resource to enable teachers to benchmark their own judgements against agreed standards.

Since the curriculum provides high-level guidance, teachers can decide when it is appropriate to begin to introduce new material in mathematics and to teach children to read. The local education authorities provide supportive guidance.

The Scottish Government does not currently collect information on all pupils through national assessments to monitor progress and standards at a system level. However, it does expect schools to be able to report information about improvements in their practices that have led to improvements in pupils' outcomes. Education authorities are expected to have moderated their schools' assessment outcomes against national benchmarks and to be able to feed information into the National Performance Framework.