CHAPTER 9

ON-ENTRY BASELINE ASSESSMENT ACROSS CULTURES

Peter Tymms and Christine Merrell

Introduction

This chapter describes the development of an on-entry baseline assessment that has been used extensively in England for several years and has also been translated and modified for use internationally. The baseline was originally devised to provide a fixed point from which progress could be assessed. Its purpose has, however, evolved and it is now seen as providing practitioners with a detailed profile of their pupils from which to plan an appropriate curriculum and against which progress can be measured. Additionally, collecting data internationally from a single assessment gives an opportunity for the investigation of the developmental pathways, levels and skills of children from different countries and cultures from the point when they start full-time education.

The PIPS on-entry baseline assessment

Development of the assessment

The original assessment was created for use in England to provide teachers with information about what children could do when they started school so that there would be reliable data from which later assessments in schools could be put into context. More specifically, the idea was to look at the relative progress of children in their own situation compared with the relative progress of children with similar baseline profiles in

other, similar contexts. This 'value-added' approach has been widely adopted within England and, increasingly, in the rest of the world (see, for example, Fitz-Gibbon, 1996; Fitz-Gibbon and Tymms, 2002; Sanders and Horn, 1995; Tymms, 1999a). The work described in this chapter was carried out within the Curriculum, Evaluation and Management (CEM) Centre at Durham University and forms part of the broad family of professional monitoring projects run from the Centre. It is distinctive in that it involved working directly with the profession (teachers, heads and advisers) as well as building upon theoretical research findings in order to construct a monitoring system that would be of use to the profession. This is in stark contrast to the government-run testing system of England, which was created in order to hold teachers and schools to account. The statutory End of Key Stage assessments currently used in England are part of this official system designed to drive up standards (see, for example, Shorrocks-Taylor, 1999) though their use in primary schools is currently under revision.

The baseline assessment, which is part of the Performance Indicators in Primary Schools (PIPS) project (see, for example, Tymms and Albone, 2002), was created in 1994 with the support of the National Association of Head Teachers and two local education authorities with working parties advising on the development work. It was refined over subsequent years in response to the statistical analysis of the data collected and comments from teachers largely concerning the practical issues of administration. Most changes were made in the first couple of years and, as time has passed, fewer and fewer changes have been made. This makes it possible to monitor trends over time with a stable assessment. The basis of the assessment was a structure that took the stance that children start at school having come from a very great variety of different backgrounds but, regardless of that, the assessment should establish what they know and can do at the point of entry. The choice of material for the assessment focused on information that would best inform us about how children would be likely to progress in school. In other words, the idea was to create a baseline that would allow for fair comparisons of progress later. The literature was consulted to see what were good predictors of later success or difficulty of young children. There is extensive literature on which to draw, particularly for progress in literacy, to a lesser extent for progress in mathematics and to a much smaller extent for progress in other areas. Later we were able to confirm that the subunits of the baseline, chosen by reading the literature, were acting well as predictors (Tymms, 1999b), although, as expected, there was variation in their efficacy. Teachers have traditionally carried out a baseline assessment, either of their own making or published, with new pupils when they started school but using that information for value-added purposes was rare at the time when the PIPS baseline assessment was first developed (see, for example, Wolfendale, 1993), and existing published baseline assessments were not designed for that purpose. There was no published baseline that included phonological awareness as well as vocabulary and digit identification, all of which have been shown to be good predictors of later achievement. Nor was there any computer-adaptive baseline.

The PIPS on-entry baseline assessment is now used in several thousand schools. These are largely in England but it is widely used in Scotland, Australia and New Zealand, with growing use in the Netherlands, Germany, Hong Kong and South Africa

as well as several international schools around the world. It is conducted on an individual basis and takes approximately 20 minutes per child.

Although the initial purpose of the assessment was to act as a base for value-added, its use has evolved in practice and teachers routinely use it to get to know the pupils when they first arrive at the school and to help to inform their practice. Further, many teachers have found the action of assessing to be useful in itself. They often comment that spending 20 minutes with each child helps build a good relationship and that it is not just the child's reaction to assessment items that matters but the way in which the child responds which gives valuable information. In response to requests from teachers, the assessment has now been extended to include personal, social and emotional development. There has been an additional extension of it down into the nursery years, where motor development is also monitored, in a project known as the Assessment Profile on Entry for Children and Toddlers (ASPECTS).

The manner in which the assessment was constructed has already been noted but there were other basic principles employed during the development phase and they were that it should be something that children enjoy doing, that teachers see as valuable and that it involves as little work and time as possible. Further, it would not simply record what the teachers knew already but it would develop new knowledge and would be as objective as possible, so that whoever carried out the assessment would get a similar result. Taken together, these requirements define an exacting task, especially with young children who are very variable in their attention spans and are often slow to respond. Getting a reliable assessment typically requires responses to many different items and this presented a problem. By careful work and refinement over the years, however, and crucially by constructing an assessment that was adaptive to the pupil's responses, the major problems have been overcome (further details can be found in Tymms, 2001).

How it works in practice

The assessment now comes in two formats – text and computer-delivered, although only the latter will be discussed here. Data from both formats of the assessment are returned to the CEM (Curriculum, Evaluation and Management) Centre for processing. Schools receive feedback in the form of standardised scores for each pupil and, at the same time, the CEM Centre has built up a large dataset over several years.

The following areas are assessed:

- handwriting – the child is asked to write his/her own name;
- vocabulary – the child is asked to identify objects embedded within a complex picture;
- ideas about reading – assesses concepts about print;
- phonological awareness – rhymes and repeats;
- letter identification – a fixed order of mixed upper and lower case letters;
- word recognition and reading;

- ideas about mathematics – assessment of understanding of mathematical concepts;
- counting;
- sums – addition and subtraction problems presented without symbols;
- shape identification;
- digit identification;
- maths problems – including sums with symbols.

The teacher works with individual pupils. The computer program presents the child with questions (orally) and, depending on the nature of the question, the child responds either by pointing to the answer from the choice of options on the screen or by saying the answer. The teacher records the child's response on-screen and the program selects the next question.

The way that the assessment works can be well illustrated by referring to the section relating to vocabulary. A child is shown a picture and asked to point out where a certain item is on the picture. The picture is of a kitchen and the first item for the English version is 'carrots'. We now know that practically every child starting school at the age of 4 in England whose first language is English can point to carrots on the picture, and this is, incidentally, also true in New Zealand, Scotland and Australia. The program then moves on to another item and another. Each time the item becomes harder.

There are three pictures to assess vocabulary, each with progressively more difficult items until finally it becomes too difficult for almost all children at the start of school, with items such as yacht and microscope. However, being adaptive, children are not faced with items that are inappropriately difficult. The computer continues until the child has got a few wrong and then moves to a different section. In each section the plan is the same: to start at an easy point and move through to harder items.

When we reassess pupils after the teachers have assessed them, with the pupils picked at random around England, we find almost coincident results. The latest exercise produced an exceptionally high reliability figure of 0.98. We also find that this assessment does exactly what it is intended to do: it predicts later success or difficulties well and the correlation up to reading three years later is about 0.7, and for mathematics about 0.7. The correlation with general academic success at the end of primary education seven years later when the children are 11 years old is also almost 0.7 (Tymms et al., 2007). In psychometric terms this is an exceptional assessment.

Adoption of the baseline assessment in England

The PIPS baseline assessment was used in England for several years before there was any statutory requirement for assessment in the Foundation Stage. It started off in a small way but expanded rapidly. Just three years after its introduction approximately 42,000 children were assessed. Its use in English schools has continued to become more widespread despite the introduction of statutory assessments, including the Foundation Stage Profile, with 2500 schools using it in the 2006–7 academic year.

International experiences of PIPS

The content, adaptive nature of administration and high correlation with later achievement of the PIPS baseline assessment has attracted the attention of researchers in other countries and it is now used in the ways described earlier and also as an instrument to evaluate the impact of particular educational initiatives.

In Scotland (Curriculum, Evaluation and Management (CEM) Centre, 2007) there has been quite a rapid growth of the project and a third of education authorities are now using the PIPS on-entry assessment. In New Zealand (CEM Centre, 2007) it is used in about 80 schools and in Australia (CEM Centre, 2007), in about 800. The adoption of the assessment in each of these three countries has followed a different pattern, although the foresight of individuals has been important each time and it has been individual schools or districts that have decided to implement the assessment. It has not been something that schools have been required to do in the same way as the statutory requirements, such as the Foundation Stage Profile in England. The administration of the PIPS baseline was modified for use in Scotland, New Zealand and Australia by using a voice with a local accent for the sound files. Although the majority of items have remained the same, some small changes have been inevitable. For example, the picture of a traditional English windmill was thought to be inappropriate in Australia and the picture was changed.

PIPS has been adapted for use with deaf pupils who use British Sign Language and also for first language speakers of Bengali, Cantonese and Urdu in England. Additionally there are versions in Dutch, French, German, South African, Chinese, Slovenian and Thai. Researchers who understood the assessment and the area of development that each question was probing carried out each translation/adaptation. Any alterations to the nature and difficulty of the questions were kept to an absolute minimum.

International comparisons of assessment data

The data generated from the assessment administered in different countries allow a number of different questions to be explored. The starting points of young children in different countries might be expected to reflect the impact of home, culture, nutrition, preschool provision and the mother's health during pregnancy, among other factors (see, for example, Bellamy, 2001). Detailed information about children's development before the formal state school system allows hypotheses about preschool developmental influences to be more clearly formulated. It also sets down markers, which can be used to assess the impact of schooling in different countries. Some of the most valuable educational research comes from longitudinal data and, by setting up an international project that looks at the starting point of children, the groundwork can be laid for follow-up studies in the years to come. The purpose of international comparative data must ultimately be to assist policy-making and key issues about the impact of schooling across countries that can be addressed only by knowing more about the starting points of young children.

But is it feasible to use the same assessment across different countries? Of course, Scotland, Australia and New Zealand have a very similar heritage to England when

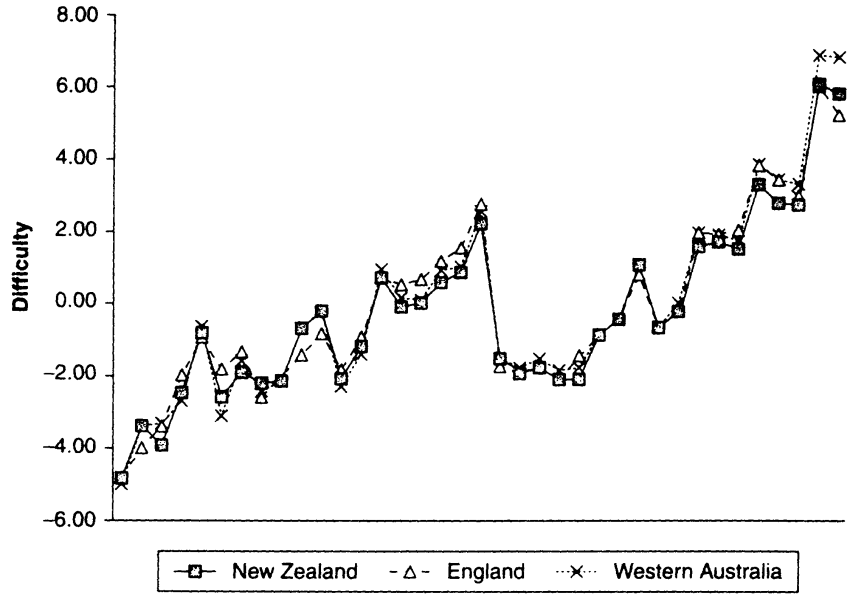

Figure 9.1 Maths (first part)

compared with most Asian, African, European or American countries. They are however, different in important ways in terms of their culture. The next logical step is to find out whether the questions in the assessment retain the same characteristics in different cultures. One way to explore this is to ask whether the relative difficulties of the items in those different countries remain stable.

Cross-cultural differences explored

Cultural influences are likely to be greater in some parts of the assessment. For example, while we might expect that the development in early mathematics (arithmetic) might be fairly consistent across different cultures, and indeed there is now evidence for a universal starting point of newborn infants in arithmetic (Wynn, 1992), nevertheless it might be that different cultures follow different developmental pathways after birth. It is clear, for example, that different counting systems have evolved in different cultures (Butterworth, 1999).

Figure 9.1 relates to the first part of the early maths section of the PIPS baseline and shows the relative difficulties of the items from England, Western Australia and New Zealand. An almost identical picture appears in each case. The data were based on a representative sample of 1000 cases from England and all of the data available in 2001 from New Zealand (1680 cases) and Western Australia (3390 cases). From each sample a Rasch model (see, for example, Bond and Fox, 2001) was used to estimate the difficulty levels of the items. The correlations between the difficulties were extremely high, as shown in Table 9.1.

At the other extreme, language clearly depends on the culture into which one is born. One would not expect a strong relationship between the difficulties of different

Table 9.1 Correlations between maths item difficulties

	New Zeland	England
England	0.987	
Australia	0.997	0.984

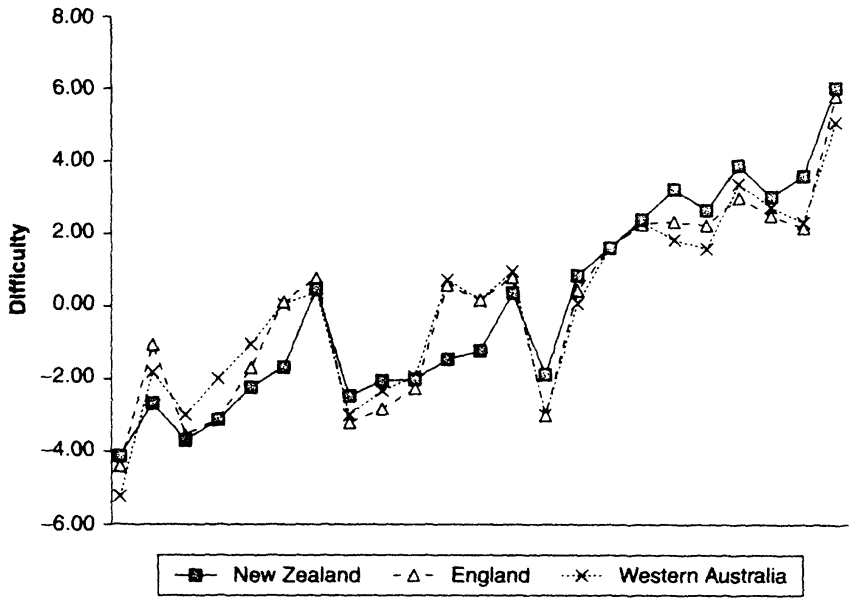

Figure 9.2 Vocabulary

words such as 'carrots' in different languages. Nevertheless there was a very strong relationship between the results from New Zealand, England and Australia as Figure 9.2 and the correlations in Table 9.2 show.

Lower correlations might be expected once the assessment is translated into another language, and that is indeed the case. The correlation between the English and Dutch maths item difficulties was 0.899. The correlation between the difficulties of vocabulary items (the Dutch items were simple direct translations of the English items) was 0.797. Despite being lower, the figures were still high.

Pursuing the theme of differences across cultures, it is instructive to ask if the assessment's predictive power lies in any one particular part of the assessment. Before this issue can be properly addressed, it is important to see how the assessment coped when predicting first reading and then mathematics of English pupils. This was done with a series of multiple regression analyses.

The results of these analyses are shown in Figure 9.3, which looks at the prediction of a maths assessment carried out three years later. The proportion of variance explained is plotted as each new variable is added into the equation. The line called 'full' shows that

Table 9.2 Correlations between vocabulary item difficulties

	New Zealand	England
England	0.934	
Australia	0.983	0.946

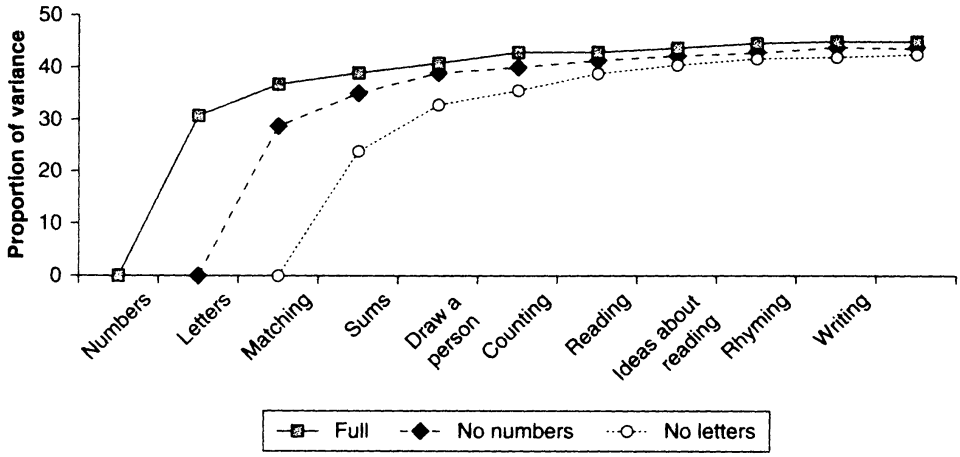

Figure 9.3 Prediction of maths assessment

30% is 'explained' by the best single predictor variable, called 'numbers' (a.k.a. digit identification). In fact digit identification remains the best single predictor of reading and mathematics achievement at age 11 years (Tymms et al., 2007). As each successive variable is added to the equation more variance is 'explained' but by less and less each time. The chart suggests the line asymptotically approaches about 45%. The second line was constructed by omitting the best predictor (numbers) entirely. Now the second best predictor comes into play immediately. This is 'Letters' (letter identification). Again the line continues to approach 45%. The third line omits both letter identification and numbers and the pattern repeats itself.

Is a Universal baseline assessment possible?

The analysis encouraged the view that it would be possible to construct a universal baseline that could act as an efficient predictor of later performance whatever the country and whatever the culture. This is because even if the most predictive parts of the assessment turned out to be heavily culturally related it would still be possible to rely on the more culturally independent sections of the assessment, of being able to do simple sums, count, understand basic concepts of print and write their own name.

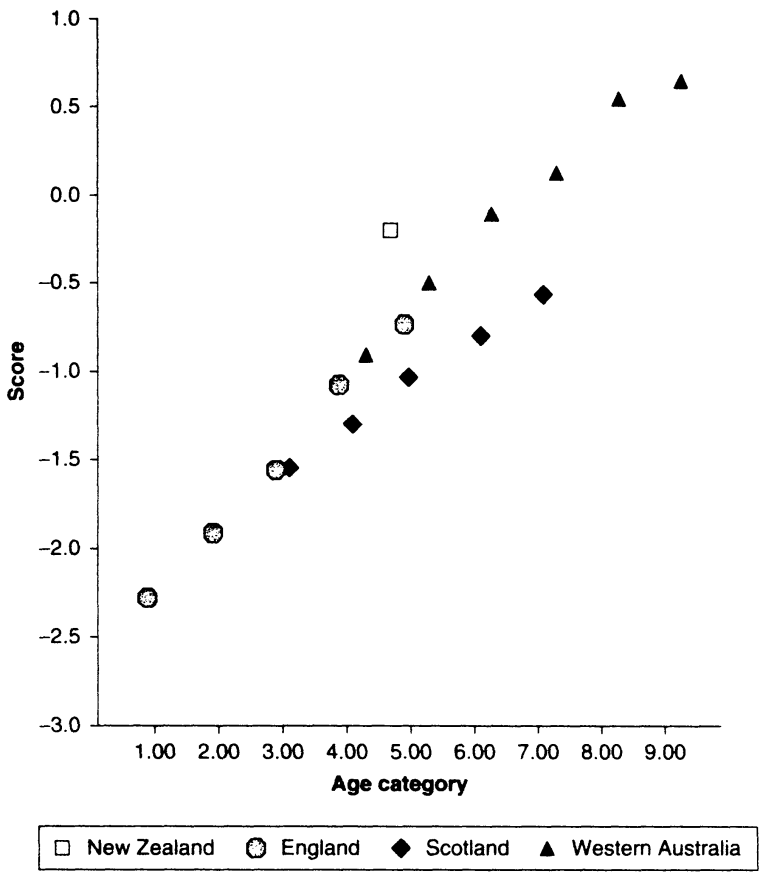

Figure 9.4 Reading scores of children starting school in England, New Zealand, Scotland and Western Australia

This should give a clear idea of the developmental level of children free of culture-dominated aspects. This is a strong claim and needs further investigation. But as more data have been collected the view has been strengthened. For example, when we compare the relative difficulties of the items from Indigenous populations in Australia against the non-Indigenous group we find very close similarities – the correlation between item difficulties for the whole assessment on entry to school for the two groups was 0.97 (but see page 101 for a different perspective). Within England we have found that the prediction of success in reading and mathematics follows almost identical patterns for children with various levels of hearing loss including those children with a severe level of impairment (Tymms et al., 2003). Data from children of different cultural backgrounds within the Netherlands show very similar patterns. So far, when comparing the assessment data of children from different cultures and very different situations, we have found very similar results for the pupils that have been assessed. Figure 9.4 compares the starting points of children in England, New Zealand, Scotland

and Western Australia. The age range of children on entry to school in these countries is different, although there is some overlap. The ages have been categorised on the bottom axis of the figure, with category 1 being the youngest at 4–4.25 years. The categories increase in quarter-year increments. Age for age, the reading scores of children in England and Western Australia are the same. The scores of the children in New Zealand, who all start school close to their fifth birthday, are slightly higher. The scores of children in Scotland show an interesting pattern, with the younger children being in line with those in England and Western Australia, but the older children starting school with relatively lower scores. It may be that the groups of older children include a significant proportion of children whose entry to school has been delayed because of special educational needs. The mathematics scores were very similar to those for reading. For a full description of the analysis, see Merrell and Tymms (2007).

One of the major areas of difference between countries is the writing system that is employed. Few difficulties arise when translating PIPS into European languages such as Dutch or German; they are essentially the same as English so far as the writing system is concerned and it is quite possible to include a section of letter identification for different countries. Similarly, a Thai translation can include a section assessing letter recognition since Thai writing is alphabetic. However, if one wants to translate the same section into Cantonese, there are considerable difficulties. Within the Cantonese system one is not dealing with an alphabetic system, which changes the entire nature of the section. It also, of course, has major implications for the child's education and the culture in the child lives. PIPS has been translated into Cantonese and is being refined but we have yet to test the hypothesis that prediction of later success at school will follow similar patterns to those in England.

The assessment climate in different countries

There are many uses for an international baseline assessment, some of which have been described in the previous section. It has also been shown that the format and content of the PIPS baseline assessment are appropriate for use in different countries and that the data collected are reliable with good predictive validity. However, an assessment cannot be successfully implemented without support from the teaching profession and possibly the government of a country. The assessment climate varies enormously from country to country and over time within countries, and in terms of both the methods of assessment and the uses of assessment data.

Significant changes in England during recent years provide a case in point. When the National Curriculum was starting, the PIPS project was also starting. The first author visited a primary school in the North-East of England and administered a half-hour test of mathematics with children aged 11. In doing the test, the author asked the children not to talk while they worked. The teacher was astounded and thought that the children would not be able to sit quietly for half an hour and that the test simply would not operate as was intended. She was genuinely surprised to see that the children managed to get on with what was required for 30 minutes. That was before

the National Curriculum appeared and national testing got under way on a large scale. Since then the atmosphere has changed significantly. Government policy relating to the increased accountability of schools led to testing becoming a major feature across English state primary schools. Within the early years there are groups of psychologists and other researchers who would regard objective testing of young children as absolutely essential in order to make progress. Conversely, there are others who would regard objective assessments to be unreliable and inappropriate with young children, partly because of behavioural fluctuations, and propose observational data as the best way forward. The differing views are strongly held and are backed by heartfelt arguments. The influence of opinion has led the English government to make the latest statutory assessment for 5-year-olds (the Foundation Stage Profile) observational (see Chapter 1). With this assessment there are concerns about moderation and its reliability. If an assessment is unreliable, this inevitably means that the uses of the data are severely limited.

Outside England, there is also a diversity of practices and views, and the take-up of PIPS within the Anglophone countries shows that the assessment has made sense to and has been seen to be useful by many professionals. For example, Cowie (2002) described how the PIPS on-entry baseline assessment was introduced into the city of Aberdeen very successfully without a period of consultation and at a time of uncertainty over assessment in Scotland. He concluded that the successful introduction was dependent on 'the quality of the material; the ongoing tension between managerial and professional accountability and on the integrity and commitment of staff in schools and the education authority'. In Australia Wildy et al. (2001) reported that a strong assessment culture had not developed in primary schools but that 'there is evidence of support among practitioners for the use of an entry-level assessment program in both government and non-government primary schools'. By building on that feeling, respecting, consulting and responding to practitioners as well as creating an efficient administrative arrangement, the project has grown throughout Australia. There is evidence that classroom teachers are beginning to use the PIPS data to plan their teaching programmes and to group students. School leaders are also using the data to allocate resources, particularly to support those with potential learning difficulties. Moving to Germany, the situation is different again. As with Australia the system is structured along state lines. There is a strong research tradition in the early years and in assessment but no universally available on-entry assessment such as PIPS. For this reason Wylde (2002), while working at the German School in London, saw the need for an assessment like PIPS and set about translating and adapting it for German children. This resulted in FIPS (Frühindikatoren zur Leistungsfähigkeit in der Primarstufe).

Harries (2002) conducted a small-scale trial of the PIPS baseline with young children in Lesotho, which is a country that currently suffers from a severe shortage of trained teachers. There were typically at least 50 pupils of a mixed age-range per class. He suggested that 'the country would benefit from a training programme for potential teachers which was clearly focused on the early learning foundation needed by pupils in order to give them the best opportunity of progressing and developing'.

One aspect of that early learning foundation is a baseline assessment and awareness of PIPS could be a useful addition in such a training programme.

Van der Hoeven-van Doornum (2002) conducted a longitudinal research project in 11 schools in the Netherlands using a translated version of the PIPS baseline assessment (OBIS – Onderbouw Informatiesysteem) to investigate 'the effects of regular assessment on pupils' progress and teachers' professionalism'. Teachers conducted OBIS and were then given feedback about the baseline assessment scores of their pupils from which they set individual learning targets. This intervention was found to have a positive impact and has been reported nationally and internationally. Since then, OBIS has been used for other research projects. Meanwhile, in the Netherlands, the government's attitude towards baseline assessment has undergone a period of change. From the mid-1980s the funds allocated to Dutch schools were partially based on the socioeconomic and ethnic composition of their pupil population. The underlying theory for the policy was that those children were considered to be educationally disadvantaged and the government hoped to combat the perceived disadvantages by providing extra funding. More recently, the Dutch government has proposed that rather than relying on background factors alone to determine additional funding, a baseline assessment for 4-year-old children should be compulsory and is exploring different possibilities.

Within Hong Kong the climate is different again. The idea that there might be an assessment from which the results were not made generally available and discussed, and pressure put on children to get higher scores, is anathema. In fact, it seems to be almost impossible to have an assessment within schools without creating pressure on the children, which would be seen as quite inappropriate in the West. A similar situation is found across much of South-East Asia.

Summary and conclusion

There is a vast literature which marks the key variables that can indicate the likely success of children as they advance through primary school. This literature has been used to create a computerised adaptive multimedia assessment known as the PIPS on-entry baseline. This assessment has proved to be attractive to teachers and is used widely within the UK and increasingly in other countries. It is used to help teachers in their professional roles and to start the vital process of monitoring children's progress.

With the wider use of the assessment, issues concerning its cultural appropriateness in different contexts have had to be addressed. The evidence so far is that with small adjustments the baseline is entirely appropriate in Anglophone countries. It is also encouraging to note that data from the Netherlands and statistical analysis of English data indicate that it can be usefully adapted to other European cultures and perhaps more widely. In other words the assessment can be translated and adapted for use in different countries, and yet maintain the nature and difficulty of the content, thus enabling reliable and meaningful international comparisons to be made of the cognitive development of children when they start school. This opens up the

possibility of broader work looking at the development of children across many different contexts with a threefold purpose. First, there is its established use by teachers. Second, there is the possibility of relating the starting points of children to the preschool facilities in various countries. Third, there is the possibility of interpreting later school-based data from international studies such as the Trends in International Mathematics and Science Study (TIMSS) (see, for example, Howie, 2002) in the light of PIPS data.