Finding the best fit: the adaptation and translation of the Performance Indicators for Primary Schools for the South African context

ELIZABETH ARCHER AND VANESSA SCHERMAN
Centre for Evaluation and Assessment, University of Pretoria, South Africa
elizabeth.archer@up.ac.za

ROBERT COE
Curriculum Evaluation and Management Centre, University of Durham, United Kingdom

SARAH J. HOWIE
Centre for Evaluation and Assessment, University of Pretoria

Reform and improvement are imperative in the current South African education system. Monitoring of school and learner achievement is an essential for establishing praxis for school improvement. Diversity of culture and South Africa's 11 official languages make it difficult to develop valid monitoring systems. Limited resources, time constraints and the need to redress neglect of large portions of the education infrastructure from the apartheid era make it problematic to develop new monitoring systems for all official languages. Adaptation and translation of existing international monitoring instruments provide alternative solutions to developing new monitoring systems. Adaptation and translation of existing instruments is a daunting process, which balances statistical analysis, translation processes and user and expert evaluations. We investigate how to balance these different processes in order to create an instrument that provides valid data for educational decisions. The processes utilised in the adaptation and translation of the vocabulary subtest of the Performance Indicators for Primary Schools (PIPS) test for the South African context are used to illustrate the complex interplay between user and expert input as well as psychometric rigour. It is hoped this paper will contribute to the development of the necessary instrument adaptation skills in South Africa.

Keywords: expert and user collaboration; instrument adaptation and contextualisation; Rasch analysis; translation of assessment instruments

Introduction

South African learners consistently achieve poorly on international assessments of learner performance, such as the Trends in Mathematics and Science Study (TIMSS) 1995 (Howie, 1997), 1999 (Howie, 2001), 2003 (Martin, Mullis, Gonzalez & Chrostowski, 2004) and the Progress in International Reading Literacy Study (PIRLS), 2006 (Howie, Venter, Van Staden, Zimmerman, Long, Scherman & Archer, 2008). These poor performances have been mirrored by national studies such as the Grade 3 and Grade 6 National Systemic Evaluation Report (National Department of Education, 2006a; Department of Education, 2006b), despite a relatively high proportion of the Gross Domestic Product being invested in education (National Treasury Republic South Africa, 2005).

Monitoring of school achievement is essential to stimulate informed action to improve education in South Africa. The adaptation of existing and proven international monitoring systems provides a cost-effective alternative to developing new instruments for the South African context. Taking this into account, the Centre for Evaluation and Assessment (CEA) at the University of Pretoria approached the Curriculum Evaluation and Management centre (CEM centre) at the University of Durham in the United Kingdom (UK) in 2003 to adapt their monitoring systems for

use in South African schools. The process of adaptation, translation and contextualisation has been ongoing and has delivered valid, reliable and contextually appropriate South African instruments at primary and secondary school level (Scherman, 2007).

The complex process of adaptation and translation incorporates statistical data analysis, translation, as well as user and expert information. This paper documents the process of adapting the vocabulary subtest, which is one of the subtests in the original Performance Indicators for Primary Schools (PIPS) assessment for the South African context.

Bringing PIPS to South African soil

The CEM centre has developed a number of monitoring systems that are used by about 7,000 schools in the UK. These assess the progress made by over a million learners every year (Tymms & Coe, 2003). For the South African adaptation, two points in schooling were identified as crucial areas where monitoring was necessary, namely, entry into primary and into secondary school (Howie, 2002). South African schools often receive learners with highly diverse backgrounds and levels of skills from a wide feeder area. At the time that this project was initiated, there were no monitoring systems in place that focused specifically on these transitional points.

The CEA therefore decided to put PIPS into practice, adapting it for the South African context and implementing it through funding from the National Research Foundation (Howie, 2002). The PIPS assessment fulfilled the CEA's criteria for an assessment measure in that it provides an indication of a child's readiness for academic learning as scores on the test administered at the start of schooling and also correlates well with subsequent academic achievement (Tymms & Coe, 2003).[1] PIPS is administered twice a year to provide a measure of progress. The test also provides information on a child's profile of performance in a number of domains which can be used to identify particular learning difficulties or strengths.

As the South African context differs widely from that in the UK, the unique learning environment of South Africa was expected to influence how children perform on the CEM centre instruments. Therefore it was necessary to adapt aspects of the monitoring system.

Monitoring quality in education, reliability and validity

An important aspect of establishing the 'validity argument' (Kane, 2006) is to demonstrate that test scores measure the same thing across all groups for whom the instrument is intended. Validity addresses the question regarding the extent to which the interpretation of results is appropriate as well as meaningful (Gronlund, 1998). However, the impact of culture on the assessment process is important. The core validity issue in adapting an assessment for the South African context is therefore determining which adaptations and accommodations would preserve the meaningfulness of the scores (Fuchs, Fuchs, Eaton, Hamlett & Karns, 2000). The removal of the irrelevant construct variance — created by the differences in culture, context, language, and social practices — results in validity. Validity is a unitary concept that is based on various forms of evidence, with construct-related validity being the central concept. Ultimately validity is concerned with the consequences of using the assessment (Gronlund, 1998; Linn & Gronlund, 2000; Killen, 2003).

One strategy for identifying bias in an assessment instrument is to look for differential item functioning (DIF) (Smith, 2004). If the relative difficulty of an item differs significantly across various groups, it indicates that scores that include that item are not measuring a unidimensional construct. Meaning that performance on the item is being influenced by some characteristic of that group other than the underlying construct being assessed (Smith, 2004).

Reliability refers to the consistency of scores that are obtained by the same individuals when these individuals are requested to complete the assessment on different occasions (Anastasi & Urbina, 1997). Reliability indicates not only how much confidence can be placed in a particular score, but also how constant the scores will be in different administrations (Owen & Taljaard, 1996). Internal consistency is employed in this study and is a prerequisite for construct validity, since one

would expect high inter-item correlations among items measuring the same construct (Kline, 1993).

Research design

Sampling

Multi-phase sampling took place (Cohen, Manion & Morrison, 2007). Initially, schools were stratified according to language of instruction. Seven schools were then selected (two Afrikaans-, three English- and three Sepedi-medium schools). Geographic representation of the Tshwane area was found to be satisfactory. Between one and four classes were assessed per school, depending on the size of the school and classes identified for assessment.

A sample of 417 learners participated in the 2005 assessment, to be tracked from the baseline to follow-up assessment. The average age for the sample was 7 years and 54% were male. While it would have been desirable for sampling to occur from a population that could allow for generalisability across South Africa, practical constraints such as funding did not allow for this. This population of Grade 1 learners in the Tshwane region therefore represents the accessible population from which the sample was drawn (Best & Kahn, 2006).

The instrument

The original PIPS for South Africa assessment was implemented in its computer-based format. Trained fieldworkers administered the assessment to learners via laptop computers at the participating schools. The PIPS instrument consists of 17 subtests, which are combined into three scales:

1. The Early Phonics Scale: It focuses on phonic awareness as an important basis for the development of reading ability.
2. The Early Reading Scale: It focuses on the prerequisite skills for reading development and includes the vocabulary subtest.
3. The Early Mathematics Scale: It aims to establish the learners' abilities in early mathematics skills.

In this paper, only the adaptation process of the vocabulary subtest of the Early Reading Scale is discussed. Vocabulary assessments have been noted to be a good indicators of language proficiency as well as academic performance (Cooper & Van Dyk, 2003). The vocabulary subtest is suited to illustrating the adaptation and contextualisation process as it encompasses both translation issues and aspects of the contextualisation of graphical elements. This subtest is aimed at evaluating the receptive vocabulary of learners and consists of 23 items. Learners are asked to point out objects in three pictures, graded according to difficulty. The first picture depicts a kitchen scene, with questions such as 'Can you point to some carrots?'; the second shows an outdoor scene with questions such as 'Can you point to a windmill?'; and the third a toy shop with more advanced items such as 'Can you point to a yacht?', which tap into learners' exposure to literature. A termination rule requires that the subtest be discontinued when a candidate supplies three consecutive incorrect answers.

During the adaptation process, a shift was made from the computer-based assessment to a paper-based equivalent to facilitate the ease of adaptation and piloting. The instrument was also translated into Afrikaans and Sepedi by registered translators and submitted to a process of back-translation and checking to establish appropriate translation. These translations were then recorded as part of the computer-based assessment.

The administration of the assessment

A team of people were trained to operate the software and conduct the assessment. Each fieldworker assessed one learner at a time. The use of the computer-based assessment meant that standardised procedures could easily be followed as the assessment was guided by the programme itself. Assess-

ments took place in English, Afrikaans and Sepedi, depending on the medium of instruction at each school.

Contextualisation and adaptation process

The contextual adaptation process utilised the data from classical test theory and Rasch analyses[2] of the Differential Item Functioning (DIF) and reliability of the subtest. This was supplemented with data from teacher evaluations regarding the face validity of the vocabulary subtests. These data formed the basis for an expert panel evaluation of the vocabulary subtest. The panel consisted of two research psychologists, two educational psychologists, three educators, two educational researchers and two subject experts involved in teacher education at a tertiary institution. These experts were invited to discuss the data on the instrument and make suggestions for the adaptation process. The suggestions were further investigated through discussions with translators and the CEM centre before adaptations were implemented. In the following sections each set of data or information presented to the expert panel is discussed and the process followed in the adaptation is illustrated.

Statistical analysis of instrument

Both classical test theory and the Rasch measurement model were employed to examine the functioning of the vocabulary subtest. The results were then discussed by the expert evaluation panel, during which item discrimination, item difficulty and subtest reliabilities were explored. The DIF analysis from the Rasch analysis was employed mainly by the researchers to further elucidate item functioning across languages.

Reliability analysis and item statistics

Internal consistency reliability (Cronbach's alpha) was used in the analysis. The level of acceptable reliability is influenced by whether the data are used for decision making on groups or individual learners and whether the data are used in isolation or in conjunction with other data (Frisbee, 1988). For the PIPS assessments, reliability values of above 0.8 were sought, though creating reliable assessments for very young children is notoriously difficult. Besides indicating the stability of measures over time, a high reliability figure strengthens the inferences made about the content-related validity of the assessment (Suen, 1990). Comparison between language groups on the PIPS is not encouraged as equivalence has not been established. Data for each medium of instruction is thus represented separately. The reliability scores for the vocabulary subtest for the 2005 assessment are indicated in Table 1.

Table 1. Reliability for the vocabulary subtest across the three languages

English (n = 211)	Afrikaans (n = 62)	Sepedi (n = 144)
0.85	0.92	0.63

The reliability for the English and Afrikaans language groups indicates that the South African English and Afrikaans vocabulary subtests seemed to function well. The reliability figure of 0.63 for the Sepedi learners, however, is a concern.

As the test was administered through standardised, computer-based procedures, the poorer Sepedi reliability could possibly be ascribed to a characteristic of the translated test, rather than the testing conditions or administrator characteristics. The items were further examined using the expert evaluation, item statistics and DIF analysis to determine the basis for the poor reliability figure.

Item facility and discrimination values

The item facility values (also referred to as item difficulty or difficulty values), along with the item

discrimination values, were used as indicators of items that needed closer examination. Item discrimination of 0.25 or higher was aimed for when examining the item-total correlation values (Barnard, 2009). Item discrimination points to the ability of an item to differentiate between high and low achievers. Facility values that show the percentage of learners that answered the items correctly are presented separately for each of the three pictures used in the vocabulary subtest (in Table 2). A termination rule is applied with the vocabulary subtest, so most candidates were not even presented with the most difficult items. For the purposes of calculating item facilities, these missing items were treated as incorrect. This means that discrimination and facility values of items that were completed by only a few learners should be interpreted with caution.

Table 2. Facility and discrimination values for items across the three languages

Item	English			Afrikaans			Sepedi		
	n	**Facility**	**Discr**	***n***	**Facility**	**Discr**	***n***	**Facility**	**Discr**
				Kitchen scene					
1 - carrots#	211	92.9	0.23#	62	98.4	0.21#	144	82.6	0.39
2 - the knife#	211	79.6	0.39	62	98.4	0.21#	144	93.8	0.44
3 - a fork#	211	90.0	0.31	62	98.4	0.21#	144	97.2	0.44
4 - a cupboard#	207	72.0	0.35	61	90.3	0.18#	141	63.2	0.33
5 - some cherries#	198	59.2	0.36	61	61.3	–0.19#	141	59.7	0.25
6 - a pan	197	63.0	0.46	61	93.5	0.28	141	91.7	0.31
7 - a bowl#	183	48.8	0.32	59	58.1	–0.14#	139	75.7	0.33
				Outdoor scene					
8 - the butterfly	159	12.8	0.57	57	17.7	0.89	132	2.1	0.33
9 - the kite	140	11.8	0.58	54	17.7	0.78	127	4.9	0.29
10 - the castle	112	7.1	0.67	44	19.4	0.85	110	0	0
11 - the wasp	39	4.3	0.62	14	19.4	0.85	10	-	-
12 - the pigeon	26	5.7	0.61	13	14.5	0.82	5	-	-
13 - the windmill	17	3.3	0.66	12	14.5	0.91	3	-	-
14 - the turtle	14	5.2	0.70	12	17.7	0.89	3	-	-
15 - the violin	13	2.8	0.64	10	12.9	0.81	3	-	-
16 - the padlock	11	2.8	0.62	11	12.9	0.85	1	-	-
17 - the toadstool	11	0.9	0.28	9	9.7	-	1	-	-
				Toy store					
18 - the yacht	8	-	-	7	-	-	1	-	-
19 - some cash	7	-	-	7	-	-	1	-	-
20 - the microscope	6	-	-	7	-	-	0	-	-
21 - some jewellery	5	-	-	7	-	-	0	-	-
22 - the saxophone	5	-	-	6	-	-	0	-	-
23 - the cosmetics	5	-	-	6	-	-	0	-	-

: items to be investigated further as indicated by discrimination values

For very easy items such as item 1, the discrimination is lower than the parameter of 0.25. However, this is to be expected as both low- and high-achieving learners typically answered this item correctly. The items from the toy store scene were challenging to learners from all language groups, which is appropriate as these are the more advanced items.

Two items have a negative item discrimination value. Item 5 asks, 'Can you show me some

cherries?' The Afrikaans translation of cherries is 'kersies', which is a homonym for birthday cake candles and cherries, both of which appear in the picture. This meant that learners indicated the birthday candles instead of the more difficult item of cherries. The approved translation of bowl as 'papbakkie' for 7 of the Afrikaans assessment was also problematic, and an alternative translation was suggested to address the negative discrimination value. Fieldworkers reported similar problems with many Sepedi learners — words are often borrowed from Afrikaans.

The fact that most of the items were significantly more difficult for the Sepedi learners than the English or Afrikaans learners is of concern. This may be due to the way in which items have been translated or graphically represented. Possibly the translations are accurate, but these words are used less frequently or are more advanced, which decreases the item facility. If the graphic representations are found alien or distracting by the Sepedi learner, it may well act as a confounding variable in measuring receptive vocabulary. Alternatively this particular sample of Sepedi learners may have a poor vocabulary. It was necessary to explore this phenomenon further through techniques such as Rasch analysis.

Rasch analysis

The Rasch model locates the difficulty of items and the ability of persons on a single latent trait continuum. The probability that a person of ability, β, will correctly answer an item of difficulty, δ, is determined entirely by the difference $\beta - \delta$. The relative difficulty of two items is independent of the abilities of the sample of persons that have attempted them (Baker, 2001). This is a particular strength in the South African context, where the aim is to examine DIF across groups whose average scores are quite different.

DIF analysis looks at the relative difficulties of items for persons in different groups, in this case the three language groups, English, Afrikaans and Sepedi. The aim is to establish whether the relative difficulty of items to one another is similar across language groups. The equal-interval property is crucial here, since the groups differ appreciably in their overall performance.

All 23 items were included for the analysis pertaining to the vocabulary subtest. Rasch analysis copes well with the missing data, which is beneficial in assessments that employ a termination rule. Missing data were thus not recoded as incorrect. This also meant that no persons were deleted from the analyses.

The person separation reliability was 0.67, indicating that the scale discriminates between persons. The items also created a well-defined variable (as indicated by the item separation reliability of 0.98). The OUTFIT mean-square for both persons and items were slightly more than 1, indicating underfit (1.04 and 1.57, respectively). Conversely the INFIT mean-square for both persons and items was below 1, indicating overfit or rather that responses are too predictable (0.83 and 0.80, respectively).

Upon inspection, several items were identified by misfit statistics (namely, Item 4, 'Can you point to a cupboard?'; Item 5, 'Can you point to some cherries?' and Item 7, 'Can you point to a bowl?'). These three items, although falling within the criteria of 0.5–1.5 for productive items with regard to the INFIT mean-square, did not fall within the prescribed range for the OUTFIT mean-square, indicating that outliers are present in the data.

Figure 1 represents the DIF for the three language groups, where the Y axis represents difficulty and the X axis the items included in the analysis. Similar ability levels can be observed between the three language groups for some of the items of the assessment. However, differences are noted. The vocabulary items at the beginning of the assessment were very easy for Afrikaans learners. Possibly, the kitchen as represented in the picture is similar to the kitchen in learners' own homes. Item 10 'Can you point to the castle?' was very difficult for Sepedi learners to identify in comparison with English and Afrikaans learners, although these learners also found this item challenging. Item 12, 'Can you point to a pigeon?', was easier for the Afrikaans learners than the English and Sepedi learners. This item taps into exposure to literature in the South African context

where in the UK, castles and the types of pigeons depicted in the vocabulary subtest are more common. Furthermore, Item 17, 'Can you point to a toadstool?', and Item 18, 'Can you point to a yacht?', were much more difficult for English learners than the other two language groups. Of all the items, Item 22, 'Can you point to the saxophone?', seemed to be the most difficult for the Afrikaans learners. For the Sepedi learners, from Item 20 onwards there was not enough information to obtain difficulty measures.

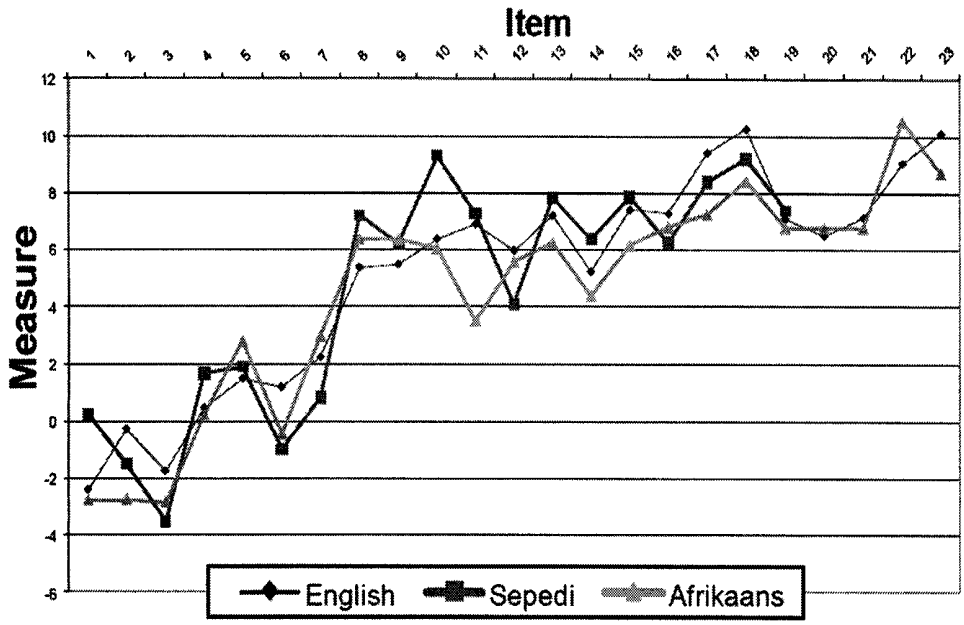

Figure 1. Differential item functioning for the vocabulary subtest for the three different language groups

The DIF analysis indicated that some items functioned differentially across the language groups. These items are to be examined further to establish whether the differences are due to lack of exposure to the stimulus or to the translation and graphical challenges experienced throughout the adaptation process.

Teacher evaluations

Six Grade 1 educators (two from each language group) were asked to evaluate the vocabulary subtest, taking particular note of whether they found the items and graphical representations of the items fair or not in terms of exposure and culture. They were also asked to rate the difficulty of each item. The results, according to language group, are indicated below. (Item 23 is not part of the paper-based assessment.)

The teacher evaluations of the vocabulary subtest raised issues about the fairness of several items. The more difficult items, such as Item 20, 'Can you point to a microscope?', and Item 22, 'Can you point to the saxophone?', were questioned, but maintained as these items are specifically to be difficult in order to tap into learners' exposure to literature and level of stimulation. The cherries item (Item 5) was questioned by one of the Afrikaans teachers, reaffirming the difficulty identified for this item through the item analysis. The validity of the graphical representations of Items 12, 14, 16, and 17 (the pigeon, the turtle, the padlock and the toadstool) was also questioned by the

Table 3. Difficulty and values indicated by educators and assessment of fairness of items

Items	Difficulty			Fairness			
	Easy	Average	Difficult	Culture		Exposure	
				Yes	No	Yes	No
1 – carrots	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	✪✪ ◎◎ ❖❖	.
2 - the knife	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎◎❖❖	.
3 - a fork	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎◎❖❖	.
4 - a cupboard	✪✪ ◎◎ ❖❖	.	.	◎◎❖❖	.	✪✪ ◎◎ ❖❖	.
5 - some cherries#	◎ ❖	✪✪ ◎◎	❖	◎◎ ❖	❖	✪✪ ◎◎❖	❖
6 - a pan	✪✪ ◎◎ ❖	❖	.	✪✪ ◎◎ ❖❖	.	◎◎ ❖❖	.
7 - a bowl	✪✪ ◎◎ ❖	❖	.	✪✪ ◎◎ ❖❖	.	◎◎ ❖❖	.
8 - the butterfly	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎◎ ❖	.
9 - the kite	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎◎ ❖	.
10 - the castle	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎◎ ❖	.
11 - the wasp	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎◎ ❖	.
12 - the pigeon#	◎ ❖	✪✪ ◎	❖	✪✪ ◎◎ ❖❖	.	◎◎ ❖	❖
13 - the windmill	✪✪ ◎ ❖	◎	❖	✪✪ ◎◎ ❖❖	.	◎◎ ❖❖	.
14 - the turtle#	✪✪ ◎	◎ ❖	❖	✪✪ ◎	◎❖❖	◎	◎❖❖
15 - the violin	◎◎ ❖❖	.	✪✪	✪✪◎❖❖	.	◎❖❖	.
16 - the padlock#	✪✪ ❖	◎◎	❖	✪✪ ◎ ❖	◎	❖	◎◎❖
17 - the toadstool#	✪✪ ◎❖	◎	❖	✪✪ ◎◎ ❖	.	◎◎ ❖	❖
18 - the yacht	✪ ◎❖❖	✪✪ ◎	.	✪✪ ◎◎ ❖❖	.	◎ ❖❖	.
19 - some cash	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎ ❖❖	.
20 - the microscope#	◎◎	✪✪	❖❖	✪✪ ◎	❖❖	◎	❖❖
21 - some jewellery	✪✪ ◎◎ ❖❖	.	.	✪✪ ◎◎ ❖❖	.	◎ ❖❖	.
22 - the saxophone#	✪✪	◎◎	❖❖	✪✪ ◎	◎❖❖	.	❖❖

◎ = English, ✪ = Sepedi and ❖ = Afrikaans

educators. In some cases there was a tendency to confuse the difficulty of the item with fairness, although there was some agreement between the DIF analyses and teacher judgements.

Expert evaluation panel

The evaluation panel was presented with the information above on the vocabulary subtest. Particular attention was paid to items flagged by the analyses and teacher evaluations Concerns were raised by the panel that some aspects in the vocabulary were too eurocentric, which may act as extraneous distracters to South African learners. This includes basic aspects such as the colouring used in the kitchen, which may have caused confusion for learners from lower socio-economic circumstances. The outside country scene in the second picture is also depicted as a view through a window; this may cause confusion with many learners as the window did not have burglar bars, as is the norm in South Africa. Based on the evaluation, changes to the items were suggested by the panel (Table 4).

Further examination of translation and graphical elements

Based on the statistical analyses, and teacher and expert panel appraisals, the identified items were explored in order to determine how to address the concerns. There was some correspondence between the conceptual and empirical processes, with some of the same items being highlighted (Items 4, 5, 12, 17, 20 and 22). There were also differences, for example, the turtle, which functioned well across the groups, according to the Rasch analysis and classical test theory.

Table 4. Changes that were proposed by expert evaluation panel to specific items

Item	Proposed changes
4 - a cupboard	This may be unfamiliar to learners from very rural areas or with very low socio-economic status
5 - some cherries	This should perhaps be replaced with a South African fruit (translation into Afrikaans leads to confusion with the candles)
9 - the kite	This is a very culturally specific pastime and should possibly be replaced with a more South African item
10 - the castle	None. This may be a very European concept, but learners should have exposure to this through literature; it is also a very well-known South African brand name
11 - the wasp	The drawing of the wasp is inaccurate for the South African wasp species and should be adapted. The translation of 'wasp' into Sepedi is very complex
12 - the pigeon	The colouring of the pigeon may have to be changed
13 - the windmill	The item can be maintained by changing the graphic representation of the windmill to the South African windmill
14 - the turtle	It would be more appropriate to the South African context if this item were changed to tortoise
15 - the violin	None. This item demands a certain level of educational stimulation and exposure
16 - the padlock	The graphic presentation of this item should be changed to a grey lock with a square shape which is more familiar in the South African context
17 - the toadstool	Toadstools are relatively unfamiliar in South Africa. Mushroom would be an appropriate replacement for this item with a concurrent change in the graphic representation
18 - the yacht	None. This item requires some exposure and previous educational stimulation from learners who are not located in a coastal area
19 - some cash	It should possible be replaced with the word 'money', which is more commonly used in South Africa

The first phase was a re-examination of translation issues. Although a strict protocol of translation and back-translation was followed, it seemed that there were still some difficulties. While the translations were correct, they were sometimes more complex, which increased the difficulty of the item. This was the case with the word for Item 7 - bowl in Afrikaans was originally translated as 'papbakkie', but this was changed to the shorter 'bakkie'. The translations for Sepedi had to be examined in depth. Since group names are often employed in Sepedi instead of specific differentiated words, some of the translations from Sepedi are academically correct, but not often used in the spoken language. Regional dialects of Sepedi are also quite prolific and complicated translation further. Careful re-evaluation of translations in terms of the difficulty of items was undertaken with a number of translators. In most cases, the translations could be rectified, but no appropriate translation with a similar difficulty value could be identified for the word 'wasp' and a completely new item of a similar difficulty value had to be incorporated for the Sepedi subtest.

After consultation with the CEM centre, it was determined that some of the more difficult items in the South African test (such as the cherries, saxophone and microscope) were quite advanced in the UK context as well and should not be altered purely because they were more difficult. The difficulty with cherries for Afrikaans learners was addressed by removing the candles on the birthday cake and introducing an extra distracter in the form of a box in the kitchen picture. Alterations to the colouring in the pictures made the items more accessible to Sepedi learners.

Discussion of the adaptation process

We report on an adaptation of an existing assessment from one context (the UK) to become more appropriate in another context (South Africa). This approach has the benefit of building on existing

effective instruments, being less time consuming than developing an instrument from scratch and creating an opportunity to compare the performance of learners across countries on similar instruments. On the other hand, it is hard to determine whether cultural biases may be inherent in the instrument, how difficult these will be to identify, or how an approach that is radical enough to eliminate the differences will be at odds with the desire to compare results with those from other contexts.

The adaptation and contextualisation process may sometimes generate somewhat contradictory information which forms the basis of adaptation. Expert appraisals may sometimes question items, which seem to be statistically sound and vice versa. Translations may be professionally conducted and monitored through back-translations, but correct translations may influence the difficulty value of an item. Inputs by users such as educators and learners are invaluable in ascertaining why certain items are not functioning as expected. The consultation process is essential not only to adapting and contextualising the instruments, but also to establishing buy-in by users and build trust in the quality of the instrument. Often seemingly insignificant changes to the graphical representation provide subtle cues which can make an item more accessible to the learners. The adaptation and contextualisation process inherently places a large amount of power in the hands of the researcher, who needs to weigh all the information in order to determine how to achieve the most appropriate adaptation. The guiding principle in these decisions, however, should be maintaining the intentions of the instrument.

The desire to create different language versions of the same instrument also brings opportunities and challenges. On the one hand, trying to keep the items similar across different languages potentially allows comparison between different language groups in terms of their levels of common vocabulary at the start of school and the progress they make during the first year of school. These are important advantages. On the other hand, the need to limit the assessment to items that work well and work similarly across all languages may present a significant constraint. Only items that can be translated satisfactorily and whose relative difficulty is the same in all languages can be used.

One limitation is that the scores on PIPS have not yet been related to any other measures. Establishing appropriate levels of internal consistency, item facility, item discrimination, and differential item functioning are important prerequisites for the validity of an instrument. However, it is also necessary to establish convergent and discriminant validity with other measures in the future. In particular, comparison of results on the vocabulary test with performance in high quality assessments of academic achievement taken later on in the learners' school career would be valuable.

Conclusion

This paper provides an example of one approach that has been effective in producing valid and reliable instruments based upon an existing international monitoring system. The adaptation, translation and contextualisation of existing international monitoring systems does provide a viable alternative for producing educational monitoring systems in the South African context for all 11 official languages. The process, however, is complex and calls for collaboration with experts and users, as is required in the development of an instrument from scratch. Adaptation and contextualisation can be achieved through a number of approaches, but usually calls for the incorporation of a number of techniques to interpret the difficulties with specific items and to determine how to address these issues most appropriately. Piloting and testing of various changes during the development process is essential an expected improvement may adversely influence other items in an unpredictable manner.

Translation issues are particularly complex in the South African context as the skills of professional registered translators are required for languages that are not yet standardised and where agreement has not yet been reached on terminology. Furthermore the researchers need to work closely with the translators as the researchers themselves often do not speak all the languages. While these translators are versed in translation, they need strict guidelines in terms of the requirement for

translation of assessments. It is recommended that the researchers take the time to brief translators about the necessity of maintaining the difficulties of items. Notwithstanding all efforts to keep various language versions equivalent, it would be wise not to compare data across language groups unless equivalence has been established. It may also be prudent to consider the motivation for comparison across language groups. Adaptation, translation and contextualisation of instruments does not constitute a hard and fast science, but a matter of finding the best fit or adaptation, based on the information gathered on the instrument.

Notes

1. Lekgogo and Winskel (2008) have illustrated the importance of phonemic awareness and letter knowledge as a predictor of future reading performance and skill transference for learners who switch from one medium of instruction to another, particularly Tswana to English.
2. Classical test theory refers to traditional statistical methods that including reliability, validity and item analysis. Classical test theory is dependent on the sample used for the statistical analysis and only refers to that sample. Rasch analysis falls into the category of modern test theory where the results are not dependent on the sample as item difficulty and person skill level is mapped on the same continuum.

References

Anastasi A & Urbina S 1997. *Psychological testing,* 7th edn. New Jersey: Prentice Hall.

Baker FB 2001. *The basic of item response theory,* 2nd edn. Retrieved on 15 November 2004 from http://ericae.net.

Barnard JJ 2009. *Measurement theory workshop*. Workshop held at Unisa, Pretoria, 13-15 October.

Best JW & Kahn JV 2006. *Research in education,* 10th edn. Boston: Pearson Education.

Cohen L, Manion L & Morrison K 2007. *Research methods in education*, 6th edn. London: Routledge Falmer.

Cooper T & Van Dyk T 2003. Vocabulary assessment: A look at different methods of vocabulary testing. *Perspectives in Education*, 2:67–79.

Frisbee DA 1988. Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, 7:25-35.

Fuchs LS, Fuchs S, Eaton SB, Hamlett CL & Karns KM 2000. Supplementing teacher judgements of mathematics test accommodations with objective data sources. *School Psychology Review*, 29:65-85.

Gronlund NE 1998. *Assessment of student achievement,* 6th edn. Boston: Allyn & Bacon.

Howie SJ 1997. *Mathematics and Science Performance in the Middle School Years in South Africa*. Pretoria: Human Sciences Research Council.

Howie SJ 2001. *Mathematics and Science Performance in Grade 8 in South Africa 1998/1999. TIMSS-R in South Africa*. Pretoria: Human Sciences Research Council.

Howie SJ 2002. Levelling the playing field: an investigation into value-added assessment in South African schools. Unpublished paper. NRF Grant Proposal.

Howie SJ, Venter E, Van Staden S, Zimmerman L, Long C, Scherman V & Archer E 2008. *Progress in International Reading Literacy Study (PIRLS) 2006 summary report: South African children's reading literacy achievement*. Pretoria: University of Pretoria.

Kane MT 2006. Validation. In: RL Brennan (ed.). *Educational Measurement*, 4th edn. Westport, CN: American Council on Education and Praeger.

Killen R 2003. Validity in outcomes-based education. *Perspectives in Education*, 21:1-14.

Kline P 1993. *The handbook of psychological testing*. Routledge: London.

Lekgoko O & Winskel H 2008. Learning to read Setswana and English: Cross-language transference of letter knowledge, phonological awareness and word reading skills. *Perspectives in Education*, 26:57-73.

Linn RL & Gronlund NE 2000. *Measurement and assessment in teaching,* 8th edn. New Jersey: Prentice Hall.

Martin MO, Mullis VX, Gonzalez EJ & Chrostowski SJ 2004. *TIMSS 2003 International Science Report*. Boston: TIMSS & PIRLS International Study Centre.

National Department of Education 2006a. *Grade 3 Foundation Phase National systemic evaluation report*. Pretoria: Government Printer.

National Department of Education 2006b. *Grade 6 Intermediate Phase Systemic Evaluation Report*. Pretoria: Government Printer.

National Treasury, Republic of South Africa 2005. *Provincial Budgets and Expenditure Review: 2001/02-2007/2008 September 2005*. Retrieved from www.treasury.gov.za on 21 April 2006.

Owen K & Taljaard JJ 1996. *Handbook for the use of psychological and scholastic tests of the HSRC*. Pretoria: Human Science Research Council.

Scherman V 2007. The validity of value-added measures in secondary schools. Unpublished PhD thesis. University of Pretoria.

Smith RM 2004. Detecting item bias in the Rasch model. In: EV Smith Jr & RM Smith (eds). *Introduction to Rasch measurement: Theory, models and applications*. Maple Grove: JAM Press.

Suen HK 1990. *Principle of test theories*. New Jersey: Lawrence Erlbaum.

Tymms P & Coe R 2003. Celebration of the success of distributed research with schools: The CEM Centre, Durham. *British Educational Research Journal*, 29:639-653.