Locating the local village within the global village

Assessment possibilities and practical challenges

Presented at the 4th Sub-Regional Conference on a Assessment in Education, hosted by Umalusi from the 26th to the 30th June 2006

Vanessa Scherman, Elizabeth Archer and Sarah Howie, Centre for Evaluation and Assessment, Faculty of Education, University of Pretoria, South Africa

Abstract

The Centre for Evaluation and Assessment (CEA) situated at the Faculty of Education, University of Pretoria in South Africa has been working collaboratively with the Curriculum, Evaluation and Management (CEM) Centre at the University of Durham in the United Kingdom on an assessment project since 2003. The CEM centre has developed a suite of monitoring projects catering for learners from primary school, through to A-levels. The CEA has been researching the feasibility of adapting and implementing two projects, one for the primary school and one for the secondary school, for the South African context. The instruments that were developed by the CEM centre are currently being used as baseline assessments in a number of countries, including Australia, New Zealand, Scotland and Germany. In contrast to these countries, South Africa is a developing country, with vast discrepancies in terms of schooling conditions and resources with the additional challenges of multilingualism in the classroom. These issues complicate the implementation of equitable assessment practices. The tension arises between adequately mapping the instruments in terms of context specific monitoring of achievement within South Africa, while maintaining the integrity of the instrument for the purpose of international comparisons. In this regard issues of validity, reliability, fairness and practicality are highlighted. These issues pertain to the quality of the instruments and the research question addressed is: **To what extent can an international monitoring system be adapted for the South African context and implemented effectively**. This paper addresses these issues as part of an ongoing research project, funded by the National Research Foundation (NRF).

Introduction

This paper aims to explore the possibility of using monitoring systems developed internationally for 'national monitoring' to a developing world context such as South Africa. The guiding research question is: **To what extent can an international monitoring system be adapted for the South African context and implemented effectively**. Here certain issues come to the fore, namely to what extent are the assessments valid and reliable, the issue of equity is raised in addition to that of fairness and practicality.

These issues (validity, reliability and equity) are discussed against the backdrop of quality education. The challenge of any education system is to be able to provide quality education for participants in the system and it is not surprising that internationally there has been a reemphasis on quality education. Two of the recent United Nations conferences namely the Jomtien Declaration in 1990 and the Dakar Framework for Action in 2000 have recognised that the quality of education is imperative if goals and objectives of developing countries are to be met (UNESCO, 2005). However, there is little consensus on what quality education is, as the concept could be understood differently by different stakeholders (Fitz-Gibbon, 1996) and when asked to describe quality many would use the terms useful, good, efficient or measuring up (Botha, 2002).

In 2003, the Centre for Evaluation and Assessment (CEA) at the University of Pretoria in collaboration with the Curriculum, Evaluation, and Management Centre (CEM) at the University of Durham, in the United Kingdom, embarked on a research project to investigate the possibility of adapting existing monitoring systems established in the United Kingdom for the South African context. This project is funded by the National Research Foundation, a national funding body in South Africa, to investigate the possibility of adapting existing monitoring systems established in the United Kingdom for the South African context. The aim of adapting the monitoring systems is to provide information on the quality of education learners receive, quality here specifically referring to whether academic gains are made.

The CEM centre is a research centre in the United Kingdom and has developed a number of monitoring systems at various stages of the United Kingdom schooling system. Most well known are the Primary Indicators at Primary Schools (PIPS), Middle Years Information System (MidYIS), Year 11 Information System (YELLIS) and finally A-level Information System or ALIS (CEM, 2002a). Although there were several projects which could be investigated the CEA decided to focus on PIPS, which would be implemented at the beginning of primary school and MidYIS which would be implemented at the beginning of secondary school which were strategically the two grades most in need of baseline measures as they are the beginning of primary and secondary school (the South African version of PIPS is referred to as PIPSSA which is Primary Indicators at Primary Schools in South Africa, while the South African version of MidYIS is referred to as SASSIS which is South African Secondary School Information System).

The monitoring systems developed by CEM were feasible options as the CEA identified a potential need for monitoring systems for schools as there is no specific policy or programme in place to monitor or evaluate learners at the beginning of primary and

secondary school and the CEA recognised that it was precisely at these levels collecting baseline information would be invaluable in order to track the progress of learners. There are currently a couple of policy initiatives that are relevant to the discussion on quality of education and monitoring. These are the policy on Integrated Quality Management System (IQMS) (including the Whole School Evaluation programme) and Systemic Evaluation. As part of the Whole School Evaluation component of the Integrated Quality Management Systems schools should be evaluating themselves on a yearly basis and many schools felt ill-prepared to undertake the official self-evaluation programme under the IQMS. It was believed that schools would welcome a system of monitoring that would permit them to evaluate their learners and then track them as well as collecting other information (e.g. attitudinal). This would provide them with an insight into their learners, their behaviour and performance and therefore not only enhance the school's ability to intervene where weaknesses were identified but this would also provide them with important information ahead of the IQMS evaluation.

The paper begins with an outline of national monitoring systems developed in other countries (Section 2), followed by the rationale for why the monitoring systems developed by CEM were selected (Section 3). The concept of equitable assessments practices is then elaborated on specifically highlighting issues of validity, reliability and fairness. The paper concludes with recommendations on the way forward.

Monitoring in education

School success has often been thought of in terms of achievement and tools used to monitor progress of learners in order to ensure achievement (Safer & Fleischman, 2005). However, school success is not just achievement and the concept of monitoring needs to be defined whilst presently there is little agreement in literature on the definition of monitoring (Sammons, 1999). Even though there is little agreement of what the concept means, monitoring is constantly mentioned in school effectiveness research and is often linked to the achievement of learners (Scheerens, Glas & Thomas, 2003:14):

> …frequent monitoring and evaluation of learners' progress stands out as a factor that is consistently mentioned in research reviews as a correlate of educational achievement.

Scheerens *et al* (2003) are of the opinion that monitoring can be defined as a systematic gathering of information in order to make judgments about the effectiveness of schooling. Furthermore, monitoring stresses ongoing gathering of information as a basis for making decisions. Raffan and Ruthen (2003) further elaborate on the gathering of information by linking the activity to learning and keeping an eye, if you will, on learning in terms of difficulties experienced and progress made or in other words focusing specifically on the learner and classroom level providing a mechanism of formally regulating the desired level of quality (Scheerens *et al*, 2003) by means of informed planning, teaching and assessment. Monitoring assesses achievement trends over time (Lockheed, 1996) and in the words of Hager and Slocum (2005: 58) "a system for ongoing progress monitoring is critical to ensure the student is continually moving toward mastery." For the purpose of this paper monitoring is seen as gathering relevant information on learner performance at various stages in order to

ascertain whether academic gains have been made in order to identify strategies were necessary (Scherman, 2006).

There are a number of monitoring systems internationally that illustrate the characteristics required of good monitoring programmes, namely that the monitoring system includes a manageable unit of education, has an explicit rationale underpinning the system as well as a primary aim, is negotiated among stakeholders and has a positive affect on behavioural aspects, as well as should not interfere with the system that is being monitored (Fitz-Gibbon, 1992). These monitoring systems include the ZEBO-project developed in the Netherlands. The ZEBO-project that consists of three elements namely a pupil monitoring system, (ZEBO-PM), an assessment of educational content covered (ZEBO-CC) as well as measures of school process indicators (ZEBO-PI) (Hendriks, Doolaard & Bosker, 2001; Hendriks, Doolaard & Bosker, 2002). Also, in Australia the Victorian Certificate of Education data project can be identified which aims to assist schools to monitor effectiveness of teaching and learning in 53 subjects over a period by providing schools with performance data (Rowe, Turner & Lane, 2002). While in the United States of America the ABC+ (Attitudinal/ Behavioural/ Cognitive Indicators plus Context) model can be identified which aims to provide process data to schools and districts at the classroom-, grade-, and school level in order to develop school improvement plans that are driven by best practices in school effectiveness and staff development research (Teddlie, Koshan & Taylor, 2002). A similar project to the ABC+ model is the Assessment Tools for Teaching and Learning or asTTle that was developed in New Zealand which also focuses on school improvement. asTTle aims to provide educators with a resource which will assist in the creation of tests in reading, writing and mathematics, includes an input function for performance as well as national norms and comparisons to cohort groups but perhaps more importantly provides diagnostic information for individual learners and the class as a whole. The diagnostic information can then be used for future teaching based on the strengths and the weaknesses of learners (Ward, Hattie & Brown, 2003). Finally, the Tennessee Value-Added Assessment System (TVAAS), which was developed in the United States, can be identified. This system is called a value-added system where value added refers to a model in which academic gains made by learners are investigated. The primary purpose of TVAAS was to provide information for summative evaluations pertaining to how effective a school or educator has been in leading learners to achieve academic gains over a period of time (Sanders & Horn, 1998), reflecting growth regardless of initial levels of performance (Sanders, Wright, Ross & Wang, 2000).

A monitoring system for South Africa

A monitoring system in the South African context has to serve the same purposes as the examples briefly discussed thus far. The lessons for South Africa that can be taken from these examples are clear in that it is pertinent to consider that although the classroom and the school-level are the primary focus, other areas of the system such as the district and provincial level cannot be ignored. Thus, one has to consider the inclusion of the parents or community in addition to higher levels of the education system such as district, provincial or national level. Furthermore, the rationale has to be clear in that is the goal to develop tools for self-evaluation to monitor effectiveness or is the goal to make use of already developed tools in order to develop self-improvement plans. Finally, the level of participation of the school has to be identified, does the school collect the information themselves, send the information for capturing and transformation and then analyse the data or does the school

liaise with research consultants who collect the data, analyse the data and provide detailed feedback reports.

For South Africa, and in light of policy initiatives, it would be important to include other levels of the system as well so as to ensure that vital elements within the system are included. For example, without inclusion of the district office schools may not be able to obtain the support they need to carry out improvement plans. Furthermore, in light of the uncertainty as to what is expected in terms of self-evaluation as well as the timelines associated with self-evaluation processes, it may be beneficial to make use of instruments which are already developed but can be adapted to the South African context. As this may take the least time in terms of development but could potentially yield effective results. Finally, with the demands placed on schools it is not likely that they will have the time to collect and analyse the information themselves but rather make use of researchers who will be able to collect the necessary data as well as supply the information that is needed tailored, to the school's needs.

The CEM centre has also developed a number of monitoring systems using value-added systems at various stages of the United Kingdom schooling system as mentioned earlier. Not only has the CEM centre developed monitoring systems at ever level of the schooling system, the Centre also enjoys substantial support from the educational community and the schools in the United Kingdom particularly schools pay for the services offered by CEM. Furthermore, the monitoring systems developed by CEM are what one would call a ground-up approach as schools have chosen to participate in the projects. This approach is in contrast to the top-down systems that are imposed on schools by the Education system. Moreover, the development of the monitoring systems was determined by the need to measure outcomes along with covariates so that fair comparisons can be made as well as process variables from which hypotheses could be generated. This approach invariably is appealing especially considering South Africa's apartheid past and now where there is a need to make fair comparisons and equitable assessment practices.

The South African Qualifications Authority (SAQA) specifically aims to establish equitable assessment while remaining cognisant of the history of South Africa and the reconstruction and development goals of the new democratic government of this country as well as the need to align the South African education and training system to emerging international trends of best practice in the provision of quality education and training and lifelong learning (SAQA, 2001). Thus the South African version of adaptation of the PIPS and MidYIS instruments from the CEM centre would have to be aligned with equitable assessment practices through accommodations for the unique South African context. Equitable assessment practices are discussed in the following section.

Equitable Assessment Practices

According to Borg (2001) equitable assessment allows "for learners (i) who learn in different ways, such as we see in multiple intelligence theory, (ii) who have different backgrounds which act as unique learning frameworks, (iii) who may be at different developmental stages and (iv) who develop a different understanding of the instructional process, such as a

learning difficulty or lateral thinking." This clearly illustrates the wide range of concepts incorporated in equitable assessment.

The National Center for Research on Evaluation, Standards and Student Testing (CRESST) (1999:1) in the USA defines equity in assessment as follows:

> Equity is the concern for fairness, i.e., that assessments are free from bias or favoritism. An assessment that is fair enables all children to show what they can do. At minimum, all assessments should be reviewed for (a) stereotypes, (b) situations that may favor one culture over another, (c) excessive language demands that prevent some learners from showing their knowledge, and (d) the assessment's potential to include learners with disabilities or limited English proficiency.

The suite of instruments developed for the CEM centre was developed and specifically designed for the English context. As the South African context differs widely from that in the United Kingdom the unique learning context of South Africa may influence how children perform on these instruments. Certain accommodations and adaptations of the PIPS and MidYIS instruments were thus needed in order to develop instruments, which would provide equitable assessment information. Furthermore, the idea of equitable assessment is encompassed in the SAQA principles of good assessment namely fairness, validity, reliability and practicability (SAQA, 2001: 16). Each of these principles is discussed separately in the following sections.

Fairness

SAQA explains fairness as taking account of and addressing of issues pertaining to the inequality of opportunities, resources and appropriate teaching and learning approaches in terms of acquisition of knowledge understanding and skills. Here issues of bias in respect of ethnicity, gender, age, disability, social class and race in the assessment approaches, instruments and materials are important. In addition, what is being assessed has to be clear (SAQA, 2001).

The idea of fairness in equitable assessment obviously stretches well beyond only cultural fairness. Fairness in assessment is often accomplished through accommodations to an existing assessment where adjustments are made in terms of settings and procedures or controlling of intervening factors such as culture, which complicate the assessment of a specific construct. Effective accommodations and adaptations boost the performance of learners influenced by these intervening factors, but not that of learners unaffected by these factors (Elliott, McKevitt and Kettler, 2002; Bowen & Ferrell, 2003; Thompson & Ouenemoen, 2003; Fuchs, Fuchs, Eaton, Hamlett & Karns 2000). A multitude of accommodation possibilities have been highlighted by authors (Bowen & Ferrell, 2003; Hofstetter, 2003; Polloway, Epstein & Bursuck, 2003; Eliott et al., 2002; Taylor et al., 2002, Ysseldyke et al., 2001), these accommodations to establish fairness encompass changes in scheduling, setting, equipment or technology, presentation and response.

Issues of validity and reliability are intrinsically related to appropriate accommodations and adaptations of assessment. Eliott et al. (2002: 155) sees accommodations as providing access to the instrument and assessing a child without exposure to social practices would thus translate into inequitable assessment practice as the child will be limited in the use of reading

strategies such as reading for meaning and utilising context clues. Any adaptation of the CEM centre instruments to the South African context would thus need to consider of exposure to specific contexts and cultural practices and even types of representation to ensure that these do not act as intervening variables and thus undermine the validity of the instrument in the South African context by confounding the underlying constructs being examined.

Issues of validity

The nature of this paper precludes a comprehensive discussion on this aspect, but a fuller and more detailed discussion may be found in Scherman (2006). What follows here is a brief summary.

The central validity issue in adapting an assessment for the South African context is determining which adaptations and accommodations would preserve the meaningfulness of the scores (Fuchs, et al. 2000: 66). When accommodations produce scores for children in South Africa, which measure the same attributes as the original assessment measures for children in the country for which it has been developed, the instrument can be said to be valid for the South African context. It is thus the removal of the irrelevant construct variance created by the difference in culture, context, language, social practices, etc. which results in validity.

According to Cohen, Manion and Morrison (2003: 105), validity is basically the view that "…a particular instrument in fact measures what it purports to measure…" Validity addresses the question: to what extent is the interpretation of results appropriate as well as meaningful (Gronlund, 1998), and is a unitary concept that is based on various forms of evidence, with construct-related validity being the central concept, and ultimately is concerned with the consequences of using the assessment or questionnaire (Gronlund, 1998; Linn & Gronlund, 2000). Under the unitary concept face validity and content-related validity can be identified.

Face validity or the superficial appearance of what the test measures from the perspective of the participant is subsumed under content-related validity (Urbina, 2004). While content-related validity is generally understood as the extent to which how well the questions in the assessment matches the field within which the assessment can be located (Coolican, 1999). Thus, the sampling of items from the broader domain and items included is important (Gronlund, 1998) in terms of relevance as well as representativeness (Urbina, 2004). Content-related validity, which includes face validity and curriculum validity, where curriculum validity refers to the extent to which the abilities or competency assessed matches the curriculum (Thorndike, 1997), is established by means of drawing up tables of specifications or by consulting content specialists (Suen, 1990).

Reliability

Generally, reliability refers to the consistency of scores, which are obtained by the same individuals when they are requested to complete the assessment on different occasions (Anastasi & Urbina, 1997). Furthermore, reliability is important, as unless results are stable one cannot expect the results to be valid. Reliability not only gives an indication of how

much confidence can be placed in a particular score obtained but also how constant the scores will be which are obtained in different administrations (Owen & Taljaard, 1996).

The consistency gives an indication of the ability of items to measure the same variable or construct where inconsistent items do not measure the same construct. Internal consistency is used in this study and is a pre-requisite for construct validity, where one would expect a high item-total correlation since items measuring the same construct contributes to the total score of a test (Kline, 1993).

Issues of practicability

According to SAQA practicability refers to taking available financial resources, facilities, equipment and time into account. This speaks directly towards issues of sustainability and is very closely related to the specific context where the instrument is being employed. Any accommodations and adaptations have to be closely related to the instructional approach and material utilized in the classroom. If accommodations are unfamiliar to learners such as the use of technology or specific presentation mediums, the accommodations and adaptations in themselves can decrease the performance achievement of the learners on the assessment (Ysseldyke et al., 2001; Elliot et al., 2002; Wasburn- Moses, 2003).

The CEA through its knowledge and experience of assessment and evaluation in the South African context sought to adapt the CEM centre instruments to be as valid as possible for the South African population. The specific experience in the adaptation of the PIPS and MidYIS instrument to the South African context is discussed below.

Research design

A mixed methods approach was followed in this research, namely the integration of both quantitative and qualitative methods. This provided the researchers with additional opportunities for answering the research questions adequately (Teddlie & Tashakkori, 2003). A detailed discussion of the design and methods used may be found in Scherman, 2006.

Sample

Several schools were purposefully selected to participate in this project for maximum variation in their characteristics and background. As the aim of the research is to develop a monitoring system, which would be appropriate for South African schools regardless of the variation in schools, it is was imperative to include schools from various backgrounds. Due to financial constraints a limited number of schools could be accommodated. The sampling for the PIPSSA and SASSIS projects is discussed below.

PIPSSA

The PIPSSA project sampled seven schools of which four were former White schools[1], two of these were English medium schools and two Afrikaans medium, two of the schools were from the former African while one school from the former Indian participated. Schools

[1] For the study, due to financial constraints schools were selected in order to provide maximum variation in order to see if the instruments were suitable. White schools are the former Model C schools while African schools are the former Department of Education and Training. Indians schools are the former House of Delegates while Coloured Schools are Former House of Representatives.

generally requested that most of their Grade 1 classes be included, thus between one to four classes were assessed per school depending on the needs of the specific institution. In total, 426 learners participated in the 2005 PIPSSA study, of which the average age was 7 years (minimum age 6 years and maximum age 8 years) and 54% were male.

SASSIS

For SASSIS three former White schools of which two were English medium and one school dual medium were included as well as three African schools, two Indian schools and finally two Coloured schools. Two classes from every school were randomly selected[2] by means of WinW3S. Thus, all learners had an equal and independent chance of being selected (Gay & Airasian, 2003). In total 794 learners participated, of which the average age was 14 (minimum age 12 and maximum age 19) and 51% were female.

Instruments implemented in this research

PIPSSA

The PIPSSA assessment is computer-based and was loaded onto laptop computers and, with the help of trained fieldworkers, administered to learners at participating schools. The PIPSSA instrument consists of 17 subtests, which are combined into three different scales: early phonics, early reading and early mathematics. The scales are generated as follows:

SASSIS

The assessment instrument for SASSIS is paper-based and consists of seven subsections which were collapsed into four different scales namely the vocabulary scale, the mathematics scale, the skills scale, and the non-verbal scale each of which were designed to measure certain skills and abilities (the scales and the subtests are discussed below). The seven subtests were timed and consist of multiple-choice items with the exception of the mathematics subsection, which included both constructed response items as well as multiple-choice items. The scales are:

1. The Vocabulary scale is derived from the subtest with the same name in the assessment and measure abilities in vocabulary as well as fluency and speed.
2. The Mathematics scale is derived from the subtest with the same name in the assessment and measure abilities in mathematics as well as fluency and speed.
3. The Skills scale comprises two subtests namely the Proof Reading subtest and the Perceptual Speed and Accuracy subtest. The Proof Reading and Perceptual Speed and Accuracy subtests are designed to measure fluency and speed in finding patterns and spotting mistakes and as such rely heavily on the learner's scanning and skimming skills.
4. The Non-Verbal scale comprises three sections namely Cross Sections, Block Counting and Pictures. These tests attempt to measure 2-D and 3-D visualisation, spatial aptitude and pattern recognition. The Non-verbal score is a useful indicator of ability for learners for whom English is a second language, as there is no reliance on language (CEM, 2002b).

[2] WinW3S was used for this and it is a within-in school sampling package developed by the Data Processing Centre of the International Association for the Evaluation of Educational Achievement (IEA). Special permission was obtained to use the program as the program is normally only used in IEA studies.

The assessment is a combination of a speed assessment and power assessment where a speed assessment measures the speed with which participants perform tasks and the difficulty of tasks are manipulated through timing. While, a power assessment on the other hand has no time limit and difficulty is manipulated by increasing or decreasing the level of complexity of items. As the assessment is a combination of a speed assessment and a power assessment, the time limits typically allow the majority of participants to attempt most or all of the items (Urbina, 2004).

Data Collection

PIPSSA

For the PIPSSA component, laptops were used for the data collection. A team of between 4-6 people went out into the schools after being trained on how to operate the software and how to conduct an assessment. Each fieldworker assessed one learner at a time. The fieldworkers were given a venue (whether it is an unused classroom or the school hall) to then assess the learners. The use of the computer-based assessment meant that standardised procedures could easily be followed as the assessment was guided by the program itself. Assessments took place in English, Afrikaans and Sepedi, depending on the language of teaching at each school. The learners were assessed and the data was captured immediately onto the computer in the form of DAT Text files. Once the data was collected, the data was downloaded from the laptops.

SASSIS

Each school was visited on a separate day and fieldworkers administered the instruments. Each classroom had a fieldworker overseeing the standardised administration procedure. The fieldworker read a script explaining the assessment and questionnaire as well as the time limits for each subsection. This ensured that the administration procedures were standardised across the schools and that each learner receives exactly the same information. The assessment as well as the questionnaire took approximately two and a half hours to complete. The English script was translated into Sepedi and Afrikaans (the two additional language of instruction for the sampled schools) in order to ensure that each learner would understand what was expected. Two groups of translators were used for the translation of the administration script. The first group translated the English script into Sepedi and Afrikaans while the second group of translators checked the Sepedi and Afrikaans translations against the English version. Any changes or corrections were made and the scripts finalised. Thus, administration of the assessments took place in English, Sepedi and Afrikaans depending on the school that was visited.

In order to capture the administration process the fieldworkers completed an administration questionnaire detailing the administration process, which includes problems experienced, comments made by learners and general impressions as well as time taken for the majority of learners to complete the subsection.

Data analysis

Document analysis

The document analysis for both the PIPSSA and SASSIS instruments included examination of the curriculum policy documents, specifically the Language Learning Area and the

Mathematics Learning Area curriculum documents. The documents were imported into Atlas *ti*, and analysed over a two-week period, the results were used in conjunction with the evaluation reports by expert evaluators.

Analysis of the validity and reliability of instruments

In order to investigate the different aspects of validity (in this case face and content-related validity) specialists in the field of psychology and education were approached. Two research psychologists as well as an educational psychologist evaluated the assessment instrument for content-related validity. The psychologists were asked to complete an evaluation form. A meeting was scheduled to discuss the results of the evaluation.

Furthermore, specialists in the field of education, specifically mathematics and language, were also approached and the assessment was evaluated from a curriculum perspective. The specialists were asked to complete an evaluation form in addition to drawing up a table of specification. Once the evaluation task was completed, a meeting was scheduled with each specialist to discuss the results of the evaluation.

Internal consistency reliability[3] was used in the analysis, which is a pre-requisite for construct validity, where one would expect a high item-total correlation (above .70) since items measuring the same construct contributes to the total score of a test (Kline, 1993). Over and above indicating the stability of measures over time, this would also strengthen the inferences that could be made by the researchers on the content-related validity of the assessment (Suen, 1990).

Results

Document analysis

Initial indications are that it would appear from the policy documents that there is a reasonable overlap between the assessments for both primary and secondary school and the intended policy documents. The overlap for the primary school components was better than the overlap of the secondary school component. The result is perhaps not surprising as the primary school assessment was developed with a curriculum in mind while the secondary school assessment was developed as an "abilities assessment".

In the primary school assessment the Early Phonetics scale relates to the objectives of Listening Reading and Viewing as well as Language as denoted in the Revised National Curriculum for Grade-R and Grade 1 (National Department of Education, 2002a). The Early Reading scale addresses the outcomes as set out in the Revised National Curriculum for Grade-R and Grade 1 such as understanding the purpose of print, distinguishing letters, awareness of directionality and the ability to identify words. The Early Mathematics scale

[3] Generally, reliability refers to the consistency of scores, which are obtained by the same individuals when they are requested to complete the assessment on different occasions (Anastasi & Urbina, 1997). Consistency gives an indication of the ability of items to measure the same variable or construct where inconsistent items do not measure the same construct.

addresses the outcomes of the Revised National Curriculum for Grade-R and Grade 1 (National Department of Education, 2002b). The following skills are assessed: numbers, operations, relationships, space, quantity, counting, simple calculation, working with money, fractions, simple division and shape as well as measurement.

For the secondary school assessment it was found that the type of skills assessed was present in the language and mathematics curriculum. Of the six outcomes in the language curriculum, three learning outcomes are highlighted namely *Listening* as learner have to listen to the instructions, *Reading and Viewing* as well as *Language Structure and Use*. The learning outcomes mentioned correspond to the Instructions of the assessment as well as Proof Reading and Vocabulary where Proof reading corresponds well with *Reading and Viewing* as well as *Language Structure and Use* and Vocabulary with *Language Structure and Use*. There is a greater overlap with the mathematics curriculum and the secondary school assessment as four of the five outcomes are represented, namely Numbers, Operations and Relationships, Patterns, Functions and Algebra, Space and Shape, Measurement. The four learning outcomes correspond well with Perceptual Speed and Accuracy (Patterns, Functions and Algebra), Mathematics (Numbers, Operations and Relationships, Patterns, Functions and Algebra as well as Measurement), Block Counting (Space and Shape), Pictures (Patterns, Functions and Algebra) and Cross Sections (Space and Shape). However, before any final decisions can be made the results from the expert evaluation has to be considered.

Expert appraisals

The specialists in Education from mathematics and languages were approached to evaluate the assessments for both primary and secondary school. The same brief was given to the specialists namely to evaluate the subtests and items in terms of content validity as well as curriculum validity. The specialists were asked to develop assessment frameworks to match items to learning outcomes. The results are given separately for the two components below:

PIPSSA

The external reviewers found that there was comprehensive overlap between the curriculum and the instrument. The mode of presentation of the items, in terms of using laptops however leads to some difficulty in the PIPSSA project.

1. Financial demands. Most of the schools involved in this project do not have computer laboratories; as such laptop computers need to be rented for the fieldwork. The cost of renting laptop computers for the fieldwork represents a major part of the expenditure in this project.
2. Security. Travelling with valuable equipment such as laptops present a serious security risk in South Africa. This may negatively impact the safety of the fieldworkers.
3. Administrative burden. The process of booking, renting collecting and returning the laptop computers as well as having to repeatedly upload the necessary software for fieldwork represents a large administrative burden to the CEA team. This translates into many person-hours of labour, which may have been used more productively.
4. Administration time. The administration time of twenty minutes projected per child (Tymms & Wylde, 2003) is greatly increased in the PIPSA project as laptops must be set up and fieldworkers are often not as computer literate as the educators in the UK.
5. Sustainability. In order to achieve true sustainability for this project it would be necessary to empower educators to administer this test and relay the data to the CEA.

> In order to achieve this, it would be essential to ensure that the necessary infrastructure is in place. Currently there are vast discrepancies in the availability of computer facilities for schools in South Africa. The South African Department of Educations Draft White Paper on e-education of August 2003 indicates that in 2002 only 26.5% of schools had access to computers for teaching and learning. The availability of computers for educational purposes varied from only 4.5% in the Eastern Cape to 56.8% in the Western Cape. Although some provinces have launched ambitious programmes to equip all schools in the province with computers such as the Gauteng Online project, the targets set have not yet been realised (Gauteng Department of Education, 2005). It may be more prudent to follow the same course of action as the American Dietetics Association which delayed the switch to computer based assessment due to inaccessibility of test centres with computer facilities until such a time that they had established an appropriate infrastructure (Ruiz, Fitz, Lewis & Reidy, 1995).

Some of the items in the assessment were however deemed as being too Euro-centric by external evaluators. The expert evaluators of the primary school assessment made suggestions to make the assessment more appropriate for the South African context. Examples of suggestions for the primary school component included:

1. The expert evaluation reports indicated that the computer-based PIPS was likely to disadvantage learners who have not been exposed to cartoons, animations and three-dimensional overlays. These elements are less pronounced in the paper-based version of the PIPS assessment. Learners may be distracted or misled by the graphic representations and the assessment would thus not truly be assessing what it purports to.
2. The reviewers indicated that some of the graphic representations were very Euro-centric and may have to be replaced with more South African representations. For example replacing the beach balls with soccer balls.
3. The reviewers found some of the phonetic items inappropriate for all language groups as the pronunciation amongst various language groups can differ widely and diagraphs and diphthongs are often found in different placements in African languages than in English.
4. The reviewers indicated that the vocabulary section would have to be revisited. The specifically indicated items such as gnome, toadstool, castle and cherries as possibly being inappropriate for the South African context.

SASSIS

For the secondary school component the task of matching the curriculum and the assessment was easier to accomplish for mathematics then for language. Although it would be possible to construct a similar table for the Vocabulary subtest which forms part of the Language Learning Area it is more complicated for the Proof Reading subtest as the learners are provided with a passage that they have to correct and not singular items which can be neatly characterised as easy, moderate or difficult. It is for this reason that a similar table for the Language Learning Area is not provided.

The analysis of overlap between the Mathematics Learning Area and the SASSIS assessment proved to be very fruitful. The mathematics specialist indicated that skills needed for four out of the five learning outcomes where represented in the assessment namely Learning

Outcome 1: Numbers, operations and relationships, Learning Outcome 2: Patterns, functions and Algebra, Learning Outcome 3: Space and shape and finally Learning Outcome 4: Measurement. The specialist however raised a concern that certain items were excessively easy, that Learning Outcome 1 and 2 were over represented in the mathematics subtest of the assessment and that the time limits needed to be revised. Furthermore, the mathematics specialist indicated that certain items were not present in the mathematics curriculum but that the items would be accessible to an average Grade 8 learner as a result of general knowledge, experience and problem solving strategies.

Upon analysing the content of the assessment, the language specialists indicated that the Instructions, Vocabulary subtest, and Proof Reading subtest were of relevance for the Language Learning Area in that the skills assessed are taught in the curriculum specifically Learning Outcome 1: Listening, Learning Outcome 3: Reading and viewing and Learning Outcome 6: Language structure and use. Furthermore, one of the specialists indicated that the items were not bias in terms of gender or race and that the language used is age appropriate. However, the other specialist indicated that although the basic skills are present in the curriculum that certain items would prove difficult for second language learners and that these items should be evaluated.

In order to make the assessments for both primary and secondary school relevant for the South African context suggestions included in the expert evaluations were effected before data collection. Examples of suggestions for the secondary school component included:

1. The expert evaluation reports indicated that the instructions could be ambiguous and difficult to follow. Thus, the instructions were rewritten so that learners would understand what was expected of them but that the rewritten version would still be comparable to the original.
2. The reviewers indicated that should a learner be unsure of what to do that they would have to page to the beginning of the subtest in order to reread the instructions. This wastes time. Thus, the instructions were included at the top of the page throughout the assessment so that learners if uncertain could reread the instructions without wasting time.
3. The reviewers were not happy with the time limits allocated to the subtests however the majority of the learners were able to complete 90% of items per subtest with the exception of Mathematics and Proof Reading. Therefore, the time allocated for each subtest was increased so that the majority of the learners would be able to or almost be able to complete the subtest.
4. The reviewers indicated that certain words in the vocabulary section were ambiguous and that the way in which the words were presented was not in line with how vocabulary was taught in South Africa. As a result, the vocabulary subtest was revised not only were ambiguous words replaced but also the core word for which a synonym had to found was placed within the context of a sentence. It is suspected that the as a result the items may be easier but more accessible to second language learners.

In addition to specialists in Education, specialists in the field of Psychology were also asked to evaluate the assessment from a psychological assessment point of view. The brief was to review the instruments for content-related validity. An Educational psychologist as well as

Research psychologists formally reviewed the instrument. The outcome of the reviews indicated that the subtest do correspond with the domain of items found in ability assessments.

To conclude, the decision of the education specialists indicated that the assessment was relevant for the South African curriculum although certain changes would need to be effected. Additionally, the specialists in Psychology indicated that the items included in the assessment adequately sampled the domain of abilities assessments.

Reliability analysis

PIPSSA

Reliability analysis was undertaken for 16 of the subtests (excluding the handwriting subtest) as well as the three scales of the assessment (Table 1 and Table 2). As can be seen the reliability coefficients are quite high all above .79 with the reliabilities reported for the Letters (0.97) and Stories subtest (.96) as well as the Early mathematics scale (0.96) and Early reading scale (0.95) The reliability coefficient for Shapes was by far the lowest with .79. Vocabulary and Reading items, which were deemed inappropriate for the Afrikaans and Sepedi learners, were not included in the analysis.

Table 1 Reliability coefficients for the fifteen PIPSSA subtests

Subtest	Cronbach Alpha
Rhyming Words	.85
Repeating words	.83
Vocabulary	.93
Ideas about reading	.85
Letters	.97
Mix up words	.88
Quiz words	.92
Stories	.96
Sentences	.91
Sizes	.87
Counting	.82
Sums A	.85
Numbers	.91
Shapes	.79
Maths	.86
Sums B	.95

Table 2 Reliability coefficients for the three PIPSSA scales

Scale	Cronbach Alpha
Early Reading	.95
Early Phonics	.86
Early Mathematics	.96

Table 3 Reliability coefficients for the three PIPS scales for the UK

Scale	Cronbach Alpha
Early Reading	.97

Early Mathematics	.90
Total	.98

(CEM, 2002c)

As can be seen from the Table 3 the reliability coefficients for South Africa compare well with the reliability coefficients for the United Kingdom. The results are encouraging as although the assessment needs to be developed further in terms of face validity the items themselves are sound for our context.

SASSIS

Reliability analysis was undertaken for the seven subtests as well as the four scales of the assessment (Table 4 and Table 5). As can be seen the reliability coefficients are quite high all above .71 with the reliabilities reported for the Perceptual Speed and Accuracy and Proof Reading subtest (.94) as well as the Skills scale which comprises the two subtests (0.95). The reliability coefficient for Cross Sections was by far the lowest with .71. Upon inspection and exploring the item statistics it was found that two items were problematic and was removed from the analysis. The two items in question were similar in nature and as a result caused some confusion.

Table 4 Reliability coefficients for the seven SASSIS subtests

Subtest	Cronbach Alpha
Vocabulary	.91
Mathematics	.91
Proof Reading	.94
Perceptual Speed and Accuracy	.94
Cross Sections	.71*
Block Counting	.78
Pictures	.82

* After Two items deleted

Table 5 Reliability coefficients for the four SASSIS scales

Scale	Cronbach Alpha
Vocabulary	.91
Mathematics	.91
Non-verbal	.88
Skills	.95

Table 6 Reliability coefficients for the four MidYIS scales for the UK

Scale	Cronbach Alpha
Vocabulary	0.90
Mathematics	0.93
Non-verbal	0.89
Skills	0.84

(Source CEM, 2002c)

A similar trend can be seen for the MidYIS/SASSIS component of the project as with the PIPS/PIPSSA component in that the reliability coefficients for both South Africa and the

United Kingdom are comparable. Thus one could tentatively conclude that the items included seem to be consistent across different contexts.

Discussion

Monitoring systems are important mechanisms that schools can use to gauge their effectiveness. If quality education is to be investigated then some form of monitoring is needed. The type of monitoring system used depends on the aim, purpose, or the rationale of the system. In section 2 of the paper several national monitoring systems from other countries was briefly described. What is clear from the brief description is that the monitoring systems had a clear aim or purpose. At the heart of the systems described was to provide accurate information upon which decisions for teaching and learning could be based. For this research, the aim is not all that different in that the aim is to develop a system, which schools and educators could use to monitor learner performance as well as be used as a tool for internal evaluations, for improvement purposes.

However, monitoring systems implemented by schools and used to assist in self-evaluation processes in the context of South Africa are not available. The schools within South Africa vary greatly and schools within rural areas and townships are still disadvantaged in terms of resources and facilities. However, current assessments do not reveal the complexities within which disadvantaged schools work and in order to evaluate the true performance of a school more appropriate monitoring and measurement systems are necessary. Moreover, with the increasing demand of the provincial and national education departments to ensure that schools become accountable for their learners' performance, the need for a system, which monitors learner performance, has become imperative. Schools will have to develop the capacity to monitor their own effectiveness in order to be accountable for their learners' performance. By means of using systems such as PIPSSA and SASSIS with adaptations for the South African context school processes as well as outputs can be monitored.

Furthermore, equitability is a matter of degree and is achieved through a combination of fairness, validity, reliability and practicability. Explorations into the feasibility of using existing monitoring systems are promising especially when evaluating in terms of validity, reliability, and fairness, which are discussed separately.

Validity per se is inferred from evidence as well as ultimately depends on many different types of evidence from which inferences are drawn and expressed by degree such as high, moderate and low and is specific to a particular use. In terms of the evidence considered the validity of these assessments could be said to be moderate as all though there is considerable overlap between the content domain and the assessment for both PIPSSA and SASSIS there is also room for improvement, highlighted perhaps by the Euro-centric nature of diagrams included in the PIPSSA assessment as well as by the overlap of skills taught in the curriculum and the skills tested in the SASSIS assessment. However, considering that the assessment was developed in another country with similar and well as different objectives in mind the result is heartening especially as is that the use of the value-added approaches, as is used in these systems, contributed to establishing fairness and validity Furthermore, issues pertaining to validity of research are an important aspect more so now that there are a variety of methodological choices available. According to Newman *et al* (2003):

> …researchers strengthen validity …when they can show the consistency among research purposes, the questions and the methods they use. Strong consistency grounds the credibility of research findings and helps to ensure that audiences have confidence in the findings and implications of research studies.

A key issue, and a discussion, which has been taking place in South Africa for a while now is that of fairness particularly in terms of cultural fairness. As previous mentioned the assessments particularly the PIPSSA assessment, may not be as culturally fair in South Africa as it would in other another country perhaps. However, the assessment is accessible to learners and a key challenge and part of the recommendations from the project team is to adapt the assessment so that it is not bias in terms of culture.

Furthermore, the reliability of these assessments, which ascertains the consistency of the results and gives an indication of how stable the results are were comparable to the results in analyses undertaken in the United Kingdom. Reliability coefficients should be high, above .7 for assessments. From the analysis undertaken it appears as if both the PIPSSA assessment and the SASSIS assessment are reliable as the coefficients are above 0.7. The result is perhaps not surprising as these are well-established assessments. However, the result contributes to the assertion that the assessments are valid for the South African context, thus positioning South Africa within the international arena.

The issue of practicality though has not been directly addressed thus far however; these have implications for the way forward. Practicality, in light of the current discussion, is thought of as the financial resources, facilities, equipment and time. Presently, the project is funded by the NRF so there are no financial resources from the school per se are required. However, in the future this may be an important consideration for school participation. Equipment and time are considerations however, especially in terms of the PIPSSA assessment. The assessment is a computer-based assessment and the participating schools do not always have the computers available and so far the CEA has been making use of laptop computers. The use of laptop computers also makes it possible for the project team to load the assessment and download the data at the end of every testing session. However, this is a time consuming and often a laborious task. Furthermore, the PIPSSA assessment takes anywhere between 25 minutes and 45 minutes per child to complete. This means that certain children are taken out of the classroom for extended periods of time, which is not always ideal.

Way Forward

This research is in progress and the data are not yet fully explored. Nonetheless, there are number of plans underway for the near future PIPSSA and SASSIS and these include:

1. **Further development of the monitoring system to include contextual indicators.** Different inputs, process and outputs should be included if the monitoring system is to be comprehensive in nature as well as tap different domains such as affective, cognitive, and behavioural. However, if the monitoring system is to be comprehensive then information from more than one level should be included. Thus, additional contextual information will have to be collected from the learner-, classroom- and school-level. The system in its present form only provides learner-level information. Thus, questionnaires will have to be developed and evaluated to ensure validity.

2. **Extended exploration of construct validity:** Problematic items have to be determined as well as underlying data structure to evaluate construct validity to ensure that the constructs or scales in the assessment are sound. Thus, factor analysis will have to be undertaken in order to ensure that the items in each subtest are testing the same construct. Furthermore, for the PIPSSA assessment a move back to the paper-based version is suggested as the graphical presentations in the paper based version makes use of less three-dimensional overlays that may distract from the construct validity of the instrument. Many of the graphic representations still maintain a cartoon or animation characteristic in the paper-based format all graphic representations will be changed in terms style as changing only the most problematic items would result in inconsistencies in the style of representation throughout the assessment.
3. **Predictive validity has to be established for the South African context**: The assessment is used for prediction purposes in the context of the United Kingdom. If predicative validity is to be established for South Africa, the results from the assessment will have to be correlated with academic results, specifically language and mathematics, obtained from school-based assessments.
4. **Analysis procedures to be undertaken**: Analysis procedures used to provide information given to schools would have to be evaluated and appropriate analysis procedures for the initial validation phase as well as more developed phases will have to be identified. For example, standardised feedback cannot be given initially, as the assessment has not been standardised for the South African context and currently due to financial constraints and as a result small sample sizes the standardisation will not take place in the initial stages of the project. However, the aim is to standardise the assessment for the South African context and to develop national norms.
5. **The feedback reports to schools**: The feedback provided would need to be simplified and narratives added so that the results are presented in a comprehensive manner. Individual school reports are more appropriate in a South context that is presented to the schools during information sessions as well as follow-up telephone calls. The report should include background information on the assessment and how the learner results should be interpreted. Individual learner results should be provided as well aggregated scores. Exceptional learners should be identified as well as learners who may require additional attention. As far as possible visual representations in the form of graphs should be given, possible reason for poor performance is given as well as key areas where learners had difficulty. The report should also include attitudinal data as to the problems learners are experiencing at school, views towards the school and classes.

In conclusion, it is clear that exploring International systems has promise for South Africa. Not only can the instruments be used when adequately adapted but also lessons learnt from the development and implementation of such systems can be used in order to obtain a monitoring system that is fair, valid and reliability.

References

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

Botha, R. J. (2002) Outcomes-based education and educational reform in South Africa. *International Journal of Leadership in Education 5*(4), 361-371.

Creswell, J. W. (2003) *Research design: Qualitative, quantitative and mixed method approaches* (2nd ed). London: Sage Publications.

Coolican, H. (1999) *Research methods and statistics in Psychology* (3rd Ed). London: Hodder & Stoughton.

Curriculum, Evaluation and Management Centre (CEM) (2002a) Introduction to the MidYIS Project. Retrieved 17 May, 2002, from http://midyis.cem.dur.ac.za.uk/default.asp.

Curriculum, Evaluation and Management Centre (CEM). (2002b) The MidYIS Baseline Tests. Retrieved 3 December, 2002, from http://www.midyisproject.org/thetests.asp

Curriculum, Evaluation and Management Centre (CEM). (2002c) Reliabilities. Retrieved 3 December, 2002, from http://www.midyisproject.org/reliabilities.asp

Curriculum, Evaluation and Management Centre (CEM). (2006) PIPS Reliabilities. Retrieved 5 May, 2006, from http://www.cemcentre.org/

Fitz-Gibbon, C. T. (1992) 'Empower and monitor: The EM algorithm for the creation of effective schools'. In J. Bashi & Z. Sass (Eds), School effectiveness and improvement. Proceedings of the Third International Congress for School Effectiveness. Jerusalem: Magnes Press.

Fitz-Gibbon, C. T. (1996) *Monitoring education: Indicators, quality and effectiveness*. London: Continuum.

Gay, L. R., & Airasian, P. (2003) *Educational Research: Competencies for analysis and application* (7th ed). New Jersey: Merrill Prentice Hall.

Gauteng Department of Education (2005) Gauteng online to reach 1000 Gauteng Public Schools. Retrieved 18/01/20006 from http://www.education.gpg.gov.za/publications/gauteng%20online%20to%20reach%20public%20schools.htm

Greene, J. C. (2005) Mixing methods in evaluation workshop notes. Centurion Country Club, Gauteng. 23-24 June 2005.

Gronlund, N. E. (1998) *Assessment of student achievement* (6th ed). Boston: Allyn and Bacon.

Hager, K.D., & Slocum, T.A. (2005) Using alternative assessment to improve educational outcomes. In *Rural Special Education Quarterly 24*(1), 54-59.

Hendriks, M. A., Doolaard, S., & Bosker, R. J. (2001) School self-evaluation in the Netherlands: Development of the ZEBO-instrumentation. *Prospects XXXI* (4), 503 – 518.

Hendriks, M. A., Doolaard, S., & Bosker, R. J. (2002) Using school effectiveness as a knowledge base for self-evaluation in Ditch schools: the ZEBO-project. In A. J. Visscher & R Coe (Eds.), *School improvement through performance feedback*. Lisse: Swets & Zeitlinger Publishers.

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

Linn, R. L., & Gronlund, N. E. (2000) *Measurement and assessment in teaching* (8th ed). New Jersey: Prentice Hall.

Lockheed, M.E. (1996) International context for assessments. In P Murphy, V Greaney, M.E. Lockheed and C Rojas (Eds), *National Assessments: Testing the system*, pp 9-20. Washington DC: The World Bank.

Murphy, K. R., & Davidshofer, C. O. (1994) *Psychological testing: Principles and applications*. New Jersey: Prentice-Hall International.

National Center for Research on Evaluation, Standards and Student Testing. (1999) CRESST Assessment Glossary. Retrieved, June 5, 2004, from http://cresst96.cse.ucla.edu/CRESST/pages/glossary.htm.

National Department of Education. (2002a) National Curriculum Statement - Grades R-9: Languages - English Home Language. Retrieved 17 August, 2005 from: http://www.education.gov.za/mainDocument.asp?src=docu&xsrc=poli

National Department of Education. (2002b) National Curriculum Statement - Grades R-9: Mathematics. Retrieved 17 August, 2005 from: http://www.education.gov.za/mainDocument.asp?src=docu&xsrc=poli

Neuman, W.L. (1997) *Social research methods: Qualitative and quantitative approaches*. Boston: Allyn and Bacon.

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

Plomp, T. P. (2004) Quality assurance in Netherlands education. Workshop held at the University of Pretoria, 5 August.

Raffan, J., & Ruthen, K. (2003) Monitoring and Assessment. In J Beck & M Earl (Eds), *Key issues in secondary education: Introductory readings*, pp 28-40. New York: Continuum.

Rowe, K. J. (1999) 'Assessment, performance indicators, league tables, value-added measures and school effectiveness? Consider the issues and let's get real!' Paper presented at Australian Association for Research in Education. Retrieved July 5, 2005, from http://www.aare.edu.au/99pap/row99656.htm.

Rowe, K. J., Turner, R., & Lane, K. (2002) Performance feedback to schools of students' year 12 assessments. The VCE data project. In A. J. Visscher & R Coe (Eds.), *School improvement through performance feedback*. Lisse: Swets & Zeitlinger Publishers.

Ruiz, B., Fitz, P.A, Lewis, C. & Reidy. C. (1995) Computer-adaptive testing: a new breed of assessment. In *Journal of the American Dietetic Association, 95*(11), 1326-1328.

Safer, N., & Fleischman, S. (2005) 'How student progress monitoring improves instruction'. In *Educational Leadership*/February 2005.

Sammons, P. (1999) *School effectiveness: Coming of age in the twenty-first century*. Lisse: Swets & Zeitlinger Publishers

Sanders, W. L., Wright, S. P., Ross, S. M., & Wang, L. W. (2000) Value-added achievement results for three cohorts of roots and wings schools in Memphis: 1995-1999 outcomes. Retrieved July 20, 2004, from: http://www3.sas.com/govedu/edu/research.html.

Sanders, W. L., & Horn, S. (2003) An overview of the Tennessee Value – Added Assessment System (TVAAS): with answers to frequently asked questions. Retrieved August 22, 2003, from http://www.mdk12.org/instruction/ensure/tva/tva_1.html.

Scheerens, J., & Hendricks, M. (2002) School self-evaluation in the Netherlands. In D. Nevo (Ed), School-based evaluation: An international perspective. Amsterdam: JAI Elsevier Science.

Scheerens, J., Glas, C., & Thomas, S.M. (2003) *Educational evaluation, assessment and monitoring: A systemic approach*. Lisse: Swets & Zeitlinger Publishers.

Suen, H. K. (1990) *Principle of test theories*. New Jersey: Lawrence Erlbaum Associates.

Teddlie, C., Kochan, S., & Taylor, D. (2002) The ABC+ model for school diagnosis, feedback and improvement. In A. J. Visscher & R Coe (Eds.), *School improvement through performance feedback*. Lisse: Swets & Zeitlinger Publishers.

Teddlie, C., & Tashakkori, A. (2003) Major issues and controversies in the use of mixed methods in the social and behavioural science. In A. Tashakkori & C. Teddlie (Eds), *The Handbook of mixed methods in the social and behavioural research.* London: Sage Publications.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed). New Jersey: Prentice Hall.

Tymms, P. & Wylde, M. (2003) Baseline assessment and monitoring in primary schools. Paper presented at the Symposium Connectable Processes in Elementary and Primary Section – Bamberg April 2003

Urbina, S. (2004). *Essentials of psychological testing*. New Jersey: John Wiley and Sons, Inc.

United Nations Educational Scientific and Cultural Organisation (UNESCO). (2005). Education for all: Global monitoring report. Retrieved 3 June 2005 from http://www.unesco.org/education/efa.

Ward, L., Hattie, J.A., & Brown, G.T. (2003, June). The evaluation of asTTle in schools: The power of professional development. asTTle Technical Report, #35, University of Auckland/Ministry of Education.

Williams, K. (1999). Mixed quantitative and qualitative evaluation tools: A pragmatic approach. Retrieved March 17, 2003 from: www.cemcentre.org/ebeuk/papers2/Williams%20Kevin.doc.