Teachers' Use of Diagnostic Testing to Enhance Students' Literacy and Numeracy Learning

Lesley Ljungdahl, University of Technology Sydney, NSW, Australia
Anne Prescott, University of Technology, Sydney, NSW, Australia

Abstract: The importance of literacy and numeracy skills is paramount in most societies, their acquisition essential for communication and employment. This study set out to determine whether teachers using multiple choice assessment tasks could enhance student learning in literacy and numeracy. A software program that gave the teachers access to the results in terms of preset strands was provided to one group of teachers and the other group used the traditional techniques of looking over the students' test papers. It focuses on the testing of students using standardised PAT (Progressive Achievement Test) comprehension and mathematics tests with the intervention of a software tool (AutoMarque) which is intended to expedite analysis of the results. While much research has been carried out on literacy and numeracy testing, relatively little attention has been paid to the significance of speedy feedback and analysis of results which can lead to improved pedagogy. Constructive teacher feedback following assessment tasks assists students' learning and provides them with the skills they need to improve performance in subsequent assessments. This study highlighted the difficulties that time-poor teachers have in implementing new technologies despite their commitment to assessment for learning.

Keywords: Literacy, Numeracy, Diagnostic Testing, Assessment Elementary School

Introduction

THE IMPORTANCE OF literacy and numeracy skills is paramount in most societies, their acquisition essential to reach social and economic goals:

> *... literacy education and assessment have become high-stakes administrative issues. But precisely what counts as adequate literacy levels and how they might best be gauged remains a matter of considerable debate. (Freebody & Wyatt-Smith, 2004, p. 31) .*

Literacy practices are highly diverse: "Literacy is the flexible and sustainable mastery of a repertoire of practices with the texts of traditional and new communications technologies via spoken language, print and multimedia" (Luke, Freebody, & Land, 2000). The traditional view of literacy as 'reading, writing, listening and speaking' is now extended to viewing and representing. Students may engage in visiting an exhibition, transferring a CD of music to their iPod, reading a magazine, texting their friends, and posting information and photographs on FaceBook as well as literacy tasks related to their school work.

Numeracy means more than manipulation of numbers (Doig, 2001). Many countries, including members of the Organisation for Economic Co-operation and Development (OECD), use the term mathematical literacy which is 'an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and

The International Journal of Learning
Volume 16, Number 2, 2009, http://www.Learning-Journal.com, ISSN 1447-9494

COMMON GROUND

to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen' (OECD, 2009).

Strong literacy/numeracy skills are the foundation for successful participation in many spheres of life. Governments and parents demand accountability of their schools by looking at test results as a measure of money well spent and improved learning outcomes (Matters, 2006). As a consequence, governmental effort has focused on summative assessment type initiatives but more information is needed by teachers if they are to assist student learning (Black & Wiliam, 1998). This study explores the use and analysis of multiple-choice test results in order that assessment for learning can take place. Analysis of the responses of each student on each item in summative tests can provide teachers with valuable diagnostic information, especially when the distractors in the multiple-choice items are carefully chosen to reveal misconceptions (Holmes-Smith, 2005). When student performance data are transformed into information to assist teachers to modify their teaching practices, student learning improves (Axworthy, 2005). Thus, good assessment *for* learning practices can assist teachers in developing their students' literacy and numeracy skills (Wiliam, 2006).

Assessment and Learning

Assessment serves a range of purposes – including making judgements about a student's achievement and decisions about that student's progress and placement. National and international research closely link assessment and learning (de Lemos, 2002; Meiers, 2008; National Council of Teachers of Mathematics, 2000). If learning is paramount then the purpose of assessment changes from assessment *of* learning to assessment *for* learning (Stoll, Fink, & Earl, 2003), and assessment *as* learning which focuses on constructive feedback from the teacher and on developing the student's capacity to self-assess and reflect on his/her own learning (Holmes-Smith, 2005).

The growing understanding of the distinction between assessment *for* learning and assessment *of* learning determines the crucial role that both teachers and students play in the learning process. Assessment of learning has often been used to inform changes to the curriculum for the next year and has not necessarily been used to improve the learning of the students who provided the information (Glasson, 2008). Assessment for learning focuses on the role of the teacher and is an integral part of the cycle of teaching/learning (Hattie, 2003). Therefore one of the most powerful ways of improving learning and raising standards is to change the emphasis and use of the assessment practice (Black & Wiliam, 1998; Wiliam, 2006). Consequently, assessment should:

- contribute to effective planning of teaching and learning. A teacher's planning should provide opportunities for both learner and teacher to obtain and use information about progress towards learning goals.
- focus on how students learn. The process of learning has to be in the minds of both learner and teacher when assessment is planned and when the evidence is interpreted.
- be recognised as central to classroom practice. It involves both learners and teachers in reflection, dialogue and decision making.
- be regarded as a key professional skill for teachers. Teachers require the professional knowledge and skills to plan for assessment; observe learning; analyse and interpret

> evidence of learning; give feedback to learners; and support learners in self-assessment (Assessment Reform Group, 2002).

Such a shift in assessment practices and goals entails significant adjustments for many schools and teachers. Honest reflection and shared learning by teachers can be the catalyst for positive change in teaching (Commonwealth Department of Education Science and Training, 2001). It has been observed that:

> *The best teachers constantly monitor what is happening to students as they set about learning and investigate when things do not proceed as planned or expected. They also enquire [into] their own practice so they might get better at ensuring that their students learn successfully.* (Demos, 2004)

Effective teachers are able to analyse the data [from tests], consider the results and possible factors impinging on the students' performance levels, and discuss the implications for their curriculum planning, their teaching practice and learning support systems (Nisbet, 1998, p. 36).

National testing programs provide extensive information on the results of tests. For example, in Australia, the National Assessment Program – Literacy and Numeracy (NAPLAN) encourages teachers to analyse their class results to improve numeracy and literacy standards. Teachers can, however, feel overwhelmed by the quantity of and format in which the assessment data are presented and often believe they lack the skills to transform the data into usable information to improve their classroom practice (Allen, 2005).

Standardised assessment programs can be used to inform teaching and learning strategies in schools and therefore lead to improved student outcomes. Certain conditions should be met (Cooney, 2006):

- The assessment frameworks are based on a well-defined learning and achievement continuum and the tests are related to what is taught in schools, what students learn, how they learn and the standards they are expected to demonstrate.
- The test items allow the achievement levels of all students to be accurately determined on a common scale against standards.
- Parents and schools receive timely and appropriate feedback that identify strengths and weaknesses and which support the improvement of teaching and learning.

Standardised, multiple-choice tests such as the reading comprehension PAT (Progressive Achievement Test) and numeracy PAT (PATMaths) allow teachers to gain an insight into their teaching and their students' learning (Cooney, 2006). It is problematic that there is a considerable time between the students taking the test and the teachers receiving the results as the tests are marked in a central location. Expediting the analysis of results has the potential for constructive feedback to teachers and students.

Multiple-choice Questions

Good assessment practices provide information on what students know and can do as well as recommendations for future learning. Multiple-choice tests are sometimes thought to be simplistic, and unable to test complex cognitive tasks (Athanasou & Lamprianou, 2002).

Multiple-choice questions cannot test students' ability to develop and organise ideas and present these in a coherent argument as in an essay type answer. Nevertheless, with care in their construction, multiple-choice tests can provide a better sample of the topic examined than many other formats (Zimmaro, 2004). The objectivity in scoring is a major factor in reducing inter-teacher and intra-teacher variability. Athanasou and Lamprianou (2002) believe multiple choice questions may have been overused and are difficult to write. Nonetheless, if distractors are used judiciously, they can effectively reveal student misconceptions (Haladyna, 1999). For example: if the stem of the question is "Add 12 and 4", then one of the choices has to be the correct answer (16). The other distractors could be 48 (12×4), 8 ($12 - 4$) and 3 ($12 \div 4$). So a teacher who sees one of those answers will recognise the error and can thus assist the student with an understanding of "add".

Standardised multiple choice tests are commercially available with the advantage of providing

- validity in terms of the accuracy of the content;
- predictability and focus of test results; and
- reliability in terms of the consistency and reproducibility of the test's results (Athanasou, 1997).

Standardised tests allow comparisons to be made between an individual student and the class, but there can be language issues which disadvantage students whose first language is not English (Freeman, 2009).

This study set out to determine whether teachers using and analysing multiple-choice assessment tasks could enhance student learning in literacy and numeracy. More specifically, it sought to understand the ways in which analytical software might influence assessment practice as well as influence student literacy and numeracy learning. Hence, a software program that gave the teachers access to the results in terms of preset strands was provided to one group of teachers and the other group used the traditional techniques of looking over the students' test papers.

Methodology

The aim of this study was to explore the achievement of elementary students on a standardised multiple-choice test and investigate ways teachers could effectively determine the needs of the students in literacy and numeracy. The research is a mixed method study with scores on a standardised test providing information about potential changes in outcomes in a pretest and post test taken eight months apart. Qualitative data from focus groups and interviews provided insight into factors influencing teachers engagement with the diagnostic testing and software as well as their assessment processes and practices in the experimental and control classes.

Participants

Grade 3 and 5 teachers from seven schools across the Sydney metropolitan area participated. Six schools had one class and one school had two classes in the study. The participants consisted of 7 teachers and 148 students. The participating teachers were divided into two groups – the experimental group received the diagnostic software program (AutoMarque)

to mark multiple-choice tests and to determine the areas where students needed extra assistance; the control group manually marked and analysed the students' responses. The teachers participated in a professional learning program about assessment (see below).

To minimise bias, the allocation into control and experimental groups was random, so that each class was given the same chance of selection. Each group was further divided for testing literacy and numeracy in Grades 3 and 5. Hence, CL5 indicates control group in literacy in a Grade 5 class, and EN3 indicates experimental group in numeracy in a Grade 3 class, and so on.

According to the ethics clearance, participants were free to withdraw from the study without giving any reason. One experimental class completed the pretest but later withdrew from the study (EL3), another tested their students at the beginning and end of the study but took no part in the second professional learning day or survey (EN5).

Professional Learning Program

At the beginning of the project all teachers (both of control and experimental groups) took part in a professional learning program on assessment. Summative and formative assessments were described as well as the knowledge about student understanding that can be gained from carefully devised multiple-choice tests. These assessments allow teachers to determine what each student could do, their areas of misconceptions, and so support the teachers in their teaching. The teachers were also helped to write multiple-choice tests using web-based sources of questions. In an acknowledgement that teachers are time-poor, the participating teachers were offered multiple-choice tests on topics of their choice to be used in their normal testing program. The teachers in the experimental program were also shown how to use a software package (see below).

At the end of the study, most of the teachers took part in a second professional learning program on assessment for learning, including practice in evaluating tasks and rubrics.

Software Package

The teachers in the experimental group were given access to a software program to interrogate each student's results from multiple-choice tests (Young, 2009). After scanning the answer sheets, the software could be used to analyse the results in terms of pre-determined syllabus strands, class results and individual student results so that in subsequent instruction the teachers could specifically address areas of weakness and consolidate understanding. The results for each question were also included so that the quality of the pedagogy prior to the assessment could be determined.

Testing Instrument

PAT tests (Progressive Achievement Tests) in reading comprehension and mathematics are standardised tests produced by the Australian Council of Educational Research (ACER). In a four-year longitudinal study, Fogarty (2007) found that the PAT tests were a good indicator of school grades in later years. Students who did well on PAT tests also did well in their school grades and vice versa.

The Mathematics Test consists of three achievement tests each with 35 multiple-choice questions which assess mathematical skills within the areas of numbers, space, measurement, chance and data, and algebra. The tests each take 40 minutes. Grade 3 students took the PATMaths 1 test and Grade 5 students took the PATMaths 3 test.

The Reading Comprehension Test covers vocabulary in context and comprehension questions designed to measure factual and inferential understandings using 38 multiple-choice questions. The test consists of a series of passages which the student is required to read and then answer related questions. The tests each take 40 minutes. Grade 3 students took PAT-R test 1 and Grade 5 students took PAT-R test 2.

Results

The study began with a control group (CL3, CL5, CN3, CN5) and an experimental group. As the study progressed, however, it became obvious that there were two experimental groups. One experimental group participated in the professional learning program but *not* in using the software program (EN5, EN3) and the other experimental group fully participated in the study (EL5). Because only one of the seven classes fully participated in the experimental part of the study, a statistical analysis is less compelling. On the other hand, this pattern of participation provides a rich opportunity to consider issues and affordances involved in initiating new assessment practices and adopting a technology designed to enrich diagnostic analysis.

In the first instance the means and standard deviations were calculated for pre and post tests (Tables 1 and 2). Except for EN3, the mean mark for each class increased from pretest to post test – as would be expected with eight months of teaching between the two tests. To ascertain whether the increases in scores are significant, paired t-tests were used. The data indicate that in mathematics the students in CN3, CN5, and EN5 changed their performance between the post test and pretest at the 0.05% level (Table 2). For the literacy results, the only statistically significant change was in CL5 (Table 3).

Table 1: Analysis of the Numeracy test Data

Class/Teacher	Numeracy Pretest	Numeracy Post test	t-test (Deg of Freedom)
CN5	$\overline{x}$ =71.1 σ=10.5	$\overline{x}$ =75.8 σ=9.8	0.04* (25)
EN5	$\overline{x}$ =51.8 σ=12.4	$\overline{x}$ =58.4 σ=11.0	0* (25)
CN3	$\overline{x}$ =48.4 σ=15.0	$\overline{x}$ =56.8 σ=11.3	0.004* (13)
EN3	$\overline{x}$ =38.1 σ=14.6	$\overline{x}$ =37.7 σ=15.6	0.54 (20)
* Significant at the 0.05% leve			

Table 2: Analysis of the reading Comprehension Data

Class/Teacher	Literacy Pretest	Literacy Post test	t-test (deg of Freedom)
CL3	$\bar{x}$ =46.0 σ =7.8	$\bar{x}$ =48.2 σ =9.4	0.37 (8)
CL5	$\bar{x}$ =68.3 σ =7.8	$\bar{x}$ =80.0 σ =10.7	0* (29)
EL5	$\bar{x}$ =49.5 σ =9.9	$\bar{x}$ =51.0 σ =12.0	0.45 (21)
* Significant at the 0.05% level			

There was no significant difference in the experimental classes and most of the control groups show improved performance. As noted earlier, this study used a quasi-experimental design. The control and experimental groups were not constructed to match ability because the experimental and control groups were made up of existing classes. The results of the pretests in Tables 1 and 2, however, indicate that in each pair, the students in the experimental classes were the low achieving students. This can also be seen in the box-and-whisker plots (Figures 2-5).

Box-and-whisker plot

In descriptive statistics, a box-and-whisker plot graphically depicts the minimum score, lower quartile, median, upper quartile, and maximum score of a set of data. This is often called the five-figure summary. The pretest and post test data for each class were analysed. Using the five-figure summary, box-and-whisker plots allow comparisons between data sets with 25% of data in each whisker and 25% between lower quartile and median, and between median and upper quartile (Figure 1).

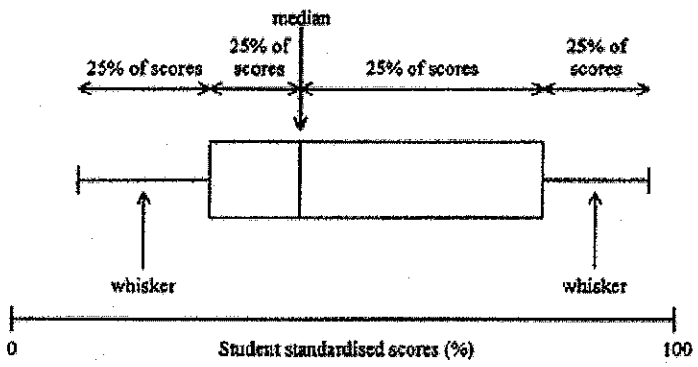

Figure 1: Box-and-whisker diagrams

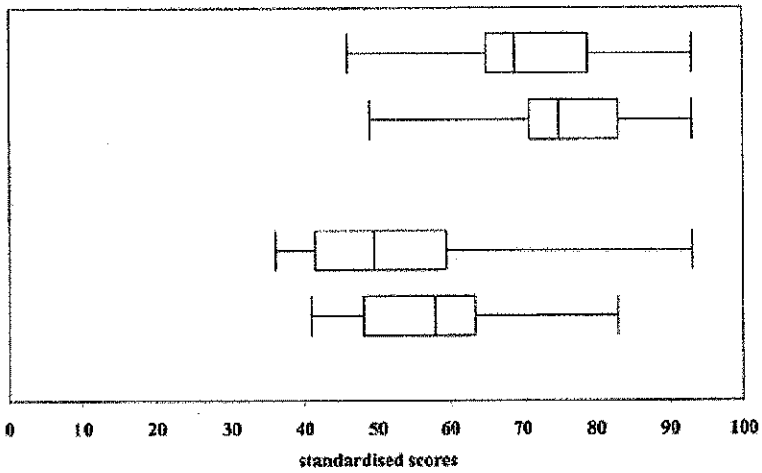

Figure 2: CN5 (top) and EN5 (bottom) pretest and post test

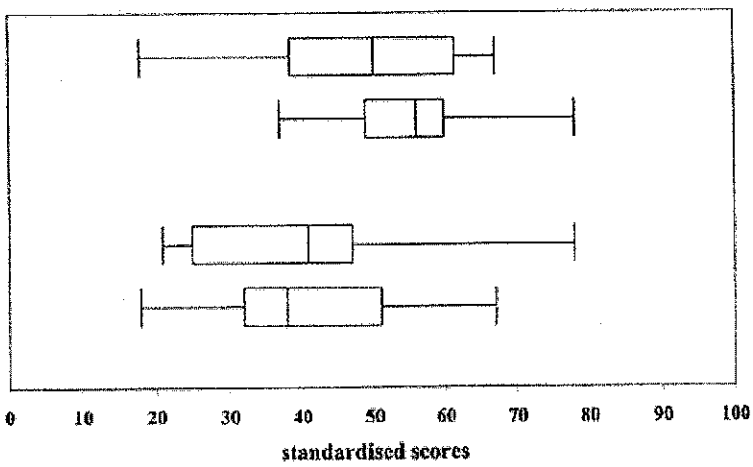

Figure 3: CN3 (top) and EN3 (bottom) pretest and post test.

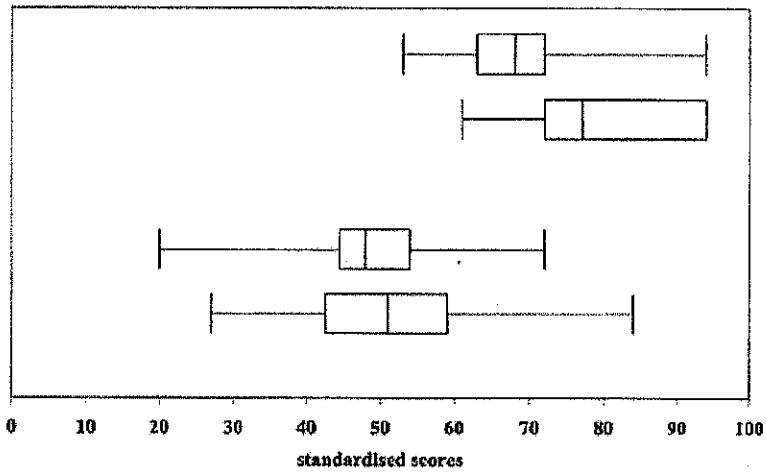

Figure 4: CL5 (top) and EL5 (bottom) pretest and post test

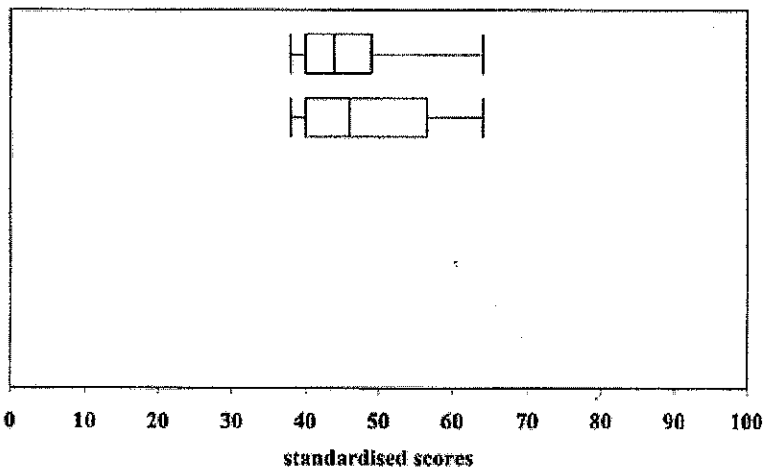

Figure 5: CL3 pretest and post test

The box-and-whisker plots (Figures 2-5) indicate that for all classes except CL3 the range of scores decreased from pretest to post test. The plots show how the experimental group classes were the low achieving classes with scores well below those of the control group classes in both pretest and post test results.

An indication of improvement from one test to the next can be seen by considering the post test class scores in relation to the pretest median - by definition, 50% of the scores in the pretest are above the median. In the post test approximately 75% of the scores are above the pretest median score for classes CN3, CN5, CL5 and EN5.

Multiple Choice Tests

In the interviews the teachers reported that the multiple choice format was seen as a positive feature by some students (rather than an open-ended format) since they believed there was a good chance of getting the correct response to each question if a number of answers was provided [teachers' responses in italics]:

> *If I give [my students] the normal written test they say 'Oh no I don't know it'. But with [multiple choice tests] they can look at it and I see that they're a lot more comfortable.*

The teachers pointed out that some lower-ability students believed the element of luck could help them and filled out the answers randomly to make sure they finished the test. On the other hand, some students did not finish the tests because they were conscientiously trying to get the correct answer and did not allocate their time wisely.

Some teachers were uncertain about the testing process and how the distractors were determined. They perceived that a negative aspect of the multiple-choice testing was not being able to observe the cognitive processes of the students:

> *You don't know the thinking skills that have gone into answering it. If you don't have the working out you can't see where they're coming from or what they've used to answer the question. You just see their results. I want kids to show me the working out.*

In other testing situations, when the teachers had the opportunity to observe the students' thinking processes by closely looking at their written work, they believed they learned more about the students' understanding of a concept:

> *Well you can see that they made a silly error but their process, their thinking strategy was right but just the execution was inaccurate.*

Focus group – the Control Group

The teachers in the control group were given each student's score on each question in the test. The results of the test were no surprise to some teachers – they highlighted individual students' learning and *were the same as I already thought.* Various factors impinged on their ability to capitalise on the information provided in the pretest results - primarily a lack of time for detailed analysis of the test items even though they recognised that such analysis *should be really good for diagnostic purposes.* The marking of multiple-choice questions was seen as less problematic than that for other tests like open-ended questions where more subjectivity was evident, even if criteria for scoring were clearly listed:

> *Different teachers look at different aspects as more important than others and they are going to mark accordingly.*

The teachers believed that multiple-choice tests were useful for testing skills such as recall of facts, basic skills and understanding a concept. They were less confident that problem solving and higher order thinking skills could be tested.

Since they perceived that there is *a high priority to provide feedback to parents on their child's academic performance and progress*, they wished *they had more time to analyse areas of the test i.e. literal/inferential* and *help identify where students need help – further consolidation and also where they need extension work.*

Speed of results was significant for all teachers:

> *I've got an A class and they really like to do well. They're extremely competitive. ... They aim to get as close to 100% as possible and getting the feedback immediately is very important.*

Focus group – The Experimental Groups

The teachers in the two experimental groups received the students' scores on each question in each test as well as an analysis of the results in terms of topics. It is difficult to determine the impact of the teachers having these test results as only one teacher (EL5) attended the second professional learning day where the teachers were interviewed and focus groups were conducted.

One teacher (EL5) was able to use the software in analysing his class results. He could identify *the strands where the children were having the greatest difficulty and ... look back at the questions, seeing which questions were mostly incorrect to help students identify their strengths and weaknesses.* This meant that he was able to isolate errors made by the majority of students and then devise lessons to address these problems. The facility of being able to access and analyse results almost immediately focussed the teaching and learning on the needs of individual children and/or groups of students. He believed that his knowledge of computers enabled him to effectively install and utilise the software to analyse multiple-choice test results. The teachers in the other experimental group did not use the software and it appears that a lack of technological support was a contributing factor.

The teacher who used the software was also able to compare the results of classes in the same age group in a school. Such comparison of class results was welcomed since it could lead to identification of successful teaching practice and/or relative abilities of students. He believed the quick turnaround on getting the results contributed to assessment for learning and thus improved teaching practice. Getting the results back straight away with quick feedback to students was seen as educationally sound.

While the explicit connection between the testing and the use of the software program and improved student performance cannot be fully evaluated in this study, all teachers emphasised the need for timely feedback in their assessment practices. It was possible to identify additional factors whereby teachers contributed to improve learning by providing opportunities for student learning; setting up environments where students could explore; and enhancing learning by explicit and systematic teaching.

Discussion

No matter how good a resource might be, if teachers do not use or poorly use a resource, it is unlikely to enhance learning in the classroom. The complex nature of numeracy and literacy and the comprehensive repertoire of practices were beyond the scope of this study. Analysing the numeracy and literacy learning needs of students is usually carried out through a range

of assessment practices, including diagnostic tests and the use of standardised tests as well as evaluation of the multiplicity of classroom-based tasks. With the information gained from these multiple forms of assessments, teachers can implement differentiated learning plans within the classroom.

Overall the test results from pretest to post test increased in the eight months between the tests (Tables 2 and 3). The one exception was EN3. The improvement of results is not unexpected considering the time lag between the testing, during which period the students received instruction in literacy and numeracy. The details of what individual teachers did in the classroom to improve literacy and numeracy is not known. Other possible factors which contributed to the change in results are the range in student ability (Freeman, 2009), familiarity with multiple-choice tests (Haladyna, 1999), and class instruction in the weeks preceding the test (Zimmaro, 2004).

The relative similarity of the results between the experimental and the control groups was unexpected. A partial explanation of the parity of results between the pre- and post-tests lies in information gleaned from the interviews with teachers. They discussed the combination of different abilities in the classes and difficulties in analysing results. Significantly, they recognised the importance of giving feedback to their students in a timely manner which could help them improve proficiency in literacy and/or numeracy. Their views are supported by Holmes-Smith (2005) and Axworthy, (2005) who suggest that expeditious analysis of results encourages teachers to modify their teaching practices (on the basis of evidence-based testing) allowing students to engage in their own learning. According to the results of the standardised tests, the students in the control group were more able and the students in the experimental groups were the low achieving students. The teachers of the low achieving classes had many more misconceptions and misunderstanding to deal with – did the teachers just run out of time, whereas the control group teachers had fewer issues to cover?

Some of the teachers in the study noted that they were able to predict their students' results on the PAT tests, based on their knowledge of their ability. They suggested that this was based on a general knowledge of their students rather than because they had used similar multiple-choice questions or the students were familiar with the external tests. The software tool could clearly help in expediting results and leave the teacher with more time to analyse their students' individual learning needs.

The nature of the PAT tests had obvious limitations if used as a measure of success in literacy. Multiple choice test items do not contain criteria, checklists or rubrics – the answers are only right or wrong. In this project, the focus is on a narrow range of literacy: reading comprehension of written texts. When making judgements about a student's strengths and weaknesses in literacy, a range of contexts over a period of time are needed, rather than the focus on a particular test. While the PAT Comprehension test is standardised, it will only provide partial evidence of attainment, ignoring the wider repertoire of literacy practices and the range of different forms of communication. In this case, reading comprehension was the focus, rather than listening, writing, and speaking skills. The PAT test could not assess, for example, whether the student could use a range of vocabulary and could write paragraphs with cohesive threads or use punctuation correctly to help the reader, but could still give an indication of each student's literacy.

While the explicit connection between the testing and the use of the software program and improved student performance cannot be fully evaluated in this study, the teachers emphasised the need for timely feedback in their assessment practices.

Conclusions

From this study it can be deduced that teachers appreciate the usefulness of tools which allow them to analyse results of tests and to quickly convey instruction to students based on this information. This inference is problematic, however, since some teachers believed they knew the ability of their students so well that they could predict the outcome of their tests. The results suggest that literacy and numeracy skills can be enhanced with strategic feedback from teachers but they are not conclusive.

Assessment is multi-faceted and teachers are challenged by issues surrounding what should be assessed, and how assessment can improve learning. The use of the software tool has the potential to expedite analysis of students' results and thus to highlight students' learning needs which the teacher can address with appropriate literacy and numeracy interventions. Collecting the students' results from tests is in itself not sufficient. Teachers need practice and professional learning in the interpretation of results in order to make the best use of assessment data if they are to move from an overall measure of achievement to more fine-grained indicators of students' specific strengths and weaknesses. While computer software has the potential to help teachers to analyse data quickly, teachers need technological support and professional learning programs to effectively use such resources.

Some of the issues involved in diagnostic assessment for learning in literacy and numeracy have been described in this research project. The use and analysis of multiple-choice responses has potential for teachers interested in targeting the specific needs of their students. Further information is needed to examine ways teachers use multiple-choice test results to enhance their students' learning and why technology designed to assist with this is not fully exploited.

This study raises key questions about both assessment applications and the professional learning associated with the adoption of technology. What professional learning is necessary for teachers to begin to use new technology? How do time-poor teachers gain a detailed understanding of their students' literacy and numeracy needs? How do teachers interpret and make use of data from tests to inform their teaching practices? The answers to such questions could provide the basis for a move towards evidence driven teaching practices that meet students 'where they are at' and provide a pathway to take them forward.

References

Allen, R. (2005). Using the evidence for student achievement for improvements at individual, class and school level. Paper presented at the Using Data to Support Learning research conference, Melbourne.

Assessment Reform Group. (2002). Assessment for learning: 10 principles - Research-based principles to guide classroom practice. [Electronic Version]. Retrieved 19.10.2008 from *www.assessment-reform-group.org.uk.*

Athanasou, J. (1997). Introduction to educational testing. Sydney: Social Science Press.

Athanasou, J., & Lamprianou. (2002). A teacher's guide to assessement. Melbourne: Thomson.

Axworthy, D. (2005). Turning data into information that improves learning: The WA experience. Paper presented at the Using data to suport learning research conference, Melbourne.

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

Commonwealth Department of Education Science and Training. (2001). The use of data to inform effective intervention in literacy and numeracy programs in the early years of schooling.

[Electronic Version]. Retrieved 16.10.2008 from *www.dest.gov.au/sectors/school_education/publications_resources/profiles/documents/litnum_early/datausepaper_rtf.htm.*

Cooney, G. (2006). Review of statewide assessments in the context of national developments. Interim Report. Sydney.

de Lemos, M. (2002). Closing the gap between research and practice: Foundations for the acquisition of literacy. Melbourne: Australian Council for Educational Research.

Demos. (2004). About learning, Report of the Learning Working Group [Electronic Version]. Retrieved 12.3.2009 from *www.demos.co.uk.*

Doig, B. (2001). Summing up: Australian numeracy performances, practices, programs and possibilities. Melbourne: Australian Council for Educational Research.

Fogarty, G. (2007). Research on the Progressive Achievement Tests and academic achievement in secondary schools. Melbourne: Australian Council for Educational Research.

Freebody, P., & Wyatt-Smith, C. (2004). The assessment of literacy: working the zone between 'system' and 'site' validity. Journal of Educational Enquiry, 5(2), 30-49.

Freeman, C. (2009). First national literacy and numeracy tests introduced [Electronic Version]. Research Developments, 20 Article 12. Retrieved 16.3.2009 from *http://research.acer.edu.au/resdev/vol20/iss20/12.*

Glasson, T. (2008). Improving student achievement through Assessment for Learning [Electronic Version]. Retrieved 10.10.2008 from *http://cmslive.curriculum.edu.au/leader/default.asp?id=25374&issueID=11603.*

Haladyna, T. M. (1999). Developing and validating multiple-choice test items (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Assoc.

Hattie, J. (2003). Teachers make a difference [Electronic Version]. Retrieved 12.12.2008 from www.acer.edu.au/documents/RC2003_Hattie_TeachersMakeADifference.pdf.

Holmes-Smith, P. (2005). Assessment for learning: Using Statewide Literacy and Numeracy tests as diagnostic tools. Paper presented at the Using data to support learning research conference, Melbourne.

Luke, A., Freebody, P., & Land, R. (2000). Literate futures: Review of literacy education. Brisbane: Education Queensland.

Matters, G. (2006). Using data to support learning in schools. Students, teachers, systems. Melbourne: Australian Council for Educational Research.

Meiers, M. (2008). A longitudinal study of growth in literacy and numeracy in the primary school years [Electronic Version]. Monitoring Learning. Retrieved 18.3.2009 from *http://research.acer.edu.au/monitoring_learning/12.*

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, Virginia: National Council of Teachers of Mathematics.

Nisbet, S. (1998). Using assessment data to inform planning and teaching: The case of state-wide benchmark numeracy tests [Electronic Version]. Retrieved 16.10.2008 from *www98.griffith.edu.au/dspace/bitstream/10072/2477/1/31483.pdf.*

OECD. (2009). OECD Programme for International Student Assessment (PISA) [Electronic Version]. Retrieved 12.3.2009 from http://www.oecd.org/pages/0,3417,en_32252351_32235968_1_1_1_1_1,00.html.

Stoll, L., Fink, D., & Earl, L. (2003). It's about learning (and it's about time). London: Routledge: Falmer.

Wiliam, D. (2006). Assessment for learning: why, what and how. Paper presented at the Assessment for Learning.

Young, A. K. (2009). AutoMarque: FlickNTick Pty Ltd.

Zimmaro, D. M. (2004). Writing good multiple-choice exams [Electronic Version]. Retrieved 16.3.2009 from www.utexas.edu./academic/mec .