



NEW ZEALAND QUALIFICATIONS AUTHORITY  
MANA TOHU MĀTAURANGA O AOTEAROA

# **Standards-based assessment in the senior secondary school**

**A research synthesis**

**Final report: February 2005**

ISBN: 978-1-877444-03-6

# **Massey University College of Education**

**Te Kupenga o Te Matauranga**

**Nick Zepke, Linda Leach, Jill Brandon, Jan Chapman, Guyon Neutze, Peter Rawlins, Adèle Scott**

## **Contents**

List of figures	4
Abbreviations used	4
Executive summary	5
Introduction	8
Synthesis of literature	17
Theme one: selected policy issues in standards-based assessment	17
Theme two: technical matters in standards-based assessment	21
Theme three: the impact of standards-based assessment on teaching	27
Theme four: the impact of standards-based assessment on learning	32
Theme five: impacts of standards-based assessment on diversity	37
Summary templates	41
Index of templates	
Bibliography	223

## **List of figures**

Figure 1	Selection strategy	<b>9</b>
Figure 2	Matrix of located studies	<b>11</b>
Figure 3	Matrix of empirical studies	<b>11</b>
Figure 4	Relationship of findings to focus areas	<b>13</b>

## **Abbreviations used**

CBA	Competency-based assessment
ERO	Education Review Office
HSC	Higher School Certificate (Australian senior secondary school qualification)
MOE	Ministry of Education (New Zealand)
NCEA	National Certificate of Educational Achievement
NQF	National Qualifications Framework
NRT	Norm-referenced testing
NZCER	New Zealand Council for Educational Research
NZPPTA	New Zealand Post Primary Teachers' Association
NZQA	New Zealand Qualifications Authority
SBA	Standards-based assessment
UK	United Kingdom
USA	United States of America

Note: Where abbreviations are specific to one study, they are only given in that template.

## Executive summary

1. In May 2004, the New Zealand Qualifications Authority (NZQA) contracted the authors from the College of Education, Massey University to review and synthesise research literature about standards-based assessment (SBA) in the senior secondary school.
2. In its request for proposal, NZQA indicated that the focus of the literature review would be on SBA in the senior secondary school, including issues such as:
  1. high-stakes assessment and SBA
  2. impacts of SBA on teachers' classroom teaching, assessment and workload
  3. impacts of SBA on students' learning and perceptions of achievement
  4. what works particularly well with SBA
  5. gender and ethnicity issues
  6. low/medium/high achiever issues
  7. motivation and self-efficacy issues
  8. meeting student learning and assessment needs
  9. validity, reliability and manageability issues
  10. synergy between teaching and assessment
  11. impact on lifelong learning.

After discussion with NZQA officials, a twelfth focus area was added to address the policy context within which SBA operated in New Zealand.

12. policy issues including an historical perspective on SBA in New Zealand.

Assessment for learning, in relation to the senior secondary school, impacts on motivation, learning and achievement were also to be considered.

3. Discussion with NZQA staff resulted in the identification of five key themes to address the focus areas:

*Theme one: policy issues*

Evidence to contextualise the use of SBA in New Zealand secondary schools.

*Theme two: technical matters*

Evidence that informs the technical debates surrounding SBA.

*Theme three: teaching*

Evidence and perspectives to inform understanding of the impact of SBA on teaching.

*Theme four: learning*

Evidence and perspectives to inform understanding of the impact of SBA on learning.

*Theme five: diversity*

Evidence on how SBA impacts on diverse learners.

4. A search strategy using international databases was developed. The initial search located 90 pages of bibliographic information about studies of interest. From this list, those studies that seemed most closely related to the previously identified themes were selected and accessed. Bibliographic references in accessed items alerted the research team to further material. This search iteration continued for over four months.
5. Items to be reviewed and synthesised were selected using a set of inclusion/exclusion criteria. Of about 130 items initially deemed suitable, 88 were selected for analysis and 80 were used for the synthesis of the literature.
6. Items selected for systematic review were summarised on templates and included in the report. The template format captured the following information:
  - Standard bibliographic information.
  - Country of origin.
  - Key words.
  - Abstract.
  - Type of research (quantitative, qualitative, descriptive, theoretical, analytical, critical and other, eg action research/opinion piece/literature review/prescriptive).
  - Key themes (details of any findings related to any of the five identified themes, eg policy, technical, teaching, learning, diversity).
  - Methodology, including information on the scale of the project.
  - Evaluative comments (the reviewers' comments on the value, place, worth and relevance of the article for the New Zealand context).
  - Other (any other information about the article that had particular interest but was not summarised under other headings).
7. The completed templates were used to inform the synthesis. The analysis of templates generated 23 statements as potential findings.

*Theme one: policy issues*

1. SBA is a dominant assessment paradigm in English speaking countries.
2. A single SBA framework for all learning remains contentious.
3. Achievement standards are better at assessing some learning than unit standards.
4. Teacher training and development, good practice exemplars and moderation must be resourced adequately.

*Theme two: technical matters*

5. It must be clear which purposes of assessment are being served by SBA.
6. Standards emerge from a consensus-seeking process affected by ambiguities and subjectivities.
7. It is difficult to achieve both validity and reliability together in SBA.
8. Alignment between curriculum, teaching and assessment achieves better results.
9. SBA can atomise learning and make integration of learning difficult to manage.

*Theme three: teaching*

10. SBA can place students at the centre of teaching.
11. SBA can focus the enacted curriculum.
12. SBA can influence pedagogy positively.
13. Standards-based pedagogy can have a positive impact on student outcomes.
14. SBA initially increases teacher workload.

*Theme four: learning*

15. SBA can affect student learning positively.
16. The formative nature of SBA can be motivating.
17. High-stakes SBA is not uniformly motivating.
18. SBA raises student workload issues.
19. SBA can help meet student needs through programme flexibility.

*Theme five: diversity*

20. The impact of SBA on diverse students is under-researched.
  21. The impact of SBA systems on diverse students' academic performance is variable.
  22. The effects of high-stakes assessment and accountability on diverse students are marked in SBA systems.
  23. Alternative assessments need to be considered for diverse students.
8. Given the large amount of research available on this topic, only a proportion of relevant literature could be synthesised in the time and with the funding available. Hence, this synthesis does not claim to capture all influences and effects.

# Introduction

## Background

According to Scriven (2003, p 15), educational evaluation is ‘the process of determining the merit, worth or significance of things’. Such a clear-cut definition obscures the contentious and complex nature of assessment. There is no uniform way of naming the process so neatly defined by Scriven, as terms such as measurement, testing, assessment and evaluation are used interchangeably to describe it. Different names suggest that different philosophical frameworks and processes may be at play. Gipps (1994) uses the notion of paradigm change to describe differences in the way assessment purposes and processes have been theorised and practised over the last half century. However, paradigms rarely change lineally and a metaphor of ‘paradigm wars’ may be more appropriate. One paradigm is positivist and outcomes-based. It attempts to evaluate scientifically and objectively the goals and outcomes of programmes. Student achievement is tested and described using approaches from scientific psychology, notably psychometric methods and models. A second ‘paradigm’ constructs learners as consumers, and assessment attempts to judge whether learners’ needs have been met. A third focuses on uniqueness of context. It questions objectivity and generalisability of data and places a subjective evaluator within a unique context at the centre of the process. A fourth paradigm emerges from the increased demand for accountability placed on educational programmes by government organisations and the public at large. It seeks, using scientific methods, to certify that learners meet the expectations of society and the economy.

Cutting across these different philosophical paradigms are two quite different approaches to determining the merit, worth or significance of things. Both compare things, but what they compare differs. Biggs (2003) refers to these two ways of comparing as norm-referenced assessment and criterion-referenced assessment. In norm-referenced assessment, each learner is compared with others in a particular group. At its simplest, it ranks learners according to their marks, grades or percentages. Sometimes the actual score a learner achieves is adjusted so the group results fit a normal distribution curve, ensuring that a set proportion achieves the top and bottom grades, while most fit around the middle. Criterion-referenced assessment compares an individual’s performance with pre-set standards, stated as learning outcomes and defined by performance criteria that clearly describe what a learner has to know or be able to do in order to succeed. All learners can succeed or not, depending on whether they meet the criteria. Two forms of criterion-referenced assessment have been identified (Leach, Neutze & Zepke 2003). Competency-based assessment produces a competent or not-yet-competent result. Achievement-based assessment has arisen mainly in New Zealand out of the criticism that competency-based assessment denies recognition of excellence (Philips 2003). To address this, grade-related criteria have been developed.

The two approaches are not mutually exclusive. They can work together such as when norm-referenced tests are used to assess learners on predetermined national standards and criteria (Wolf 1995). But as Gipps (1994) points out, there is an ever-present tension between educational and political paradigms. One seeks to provide continuous learning and assessment for enhancing learning. The other dictates national standards for accountability to national and local stakeholders. SBA is used for both. It can consequently carry conflicting expectations.

## The project

NZQA is researching SBA practices in the senior secondary school. To ensure that findings from this research build on and link to international and national literature, it commissioned this literature review. Yet SBA is not a simple concept to review. Croft (1993) points out that SBA is difficult to define. The term has been used in numerous ways to mean different things, even within the NZQA literature. Then, as the reference to Biggs highlights, some authors do not use the term at all. Others use different words to describe similar meanings. For example, criterion-referenced assessment has been used synonymously with criteria-based assessment, SBA, standards-referenced assessment, competency-based assessment and achievement-based assessment. On other occasions these terms have different meanings. Competency-based assessment and achievement-based assessment are frequently interpreted as different versions of SBA. Moreover, the set-subset relationship among terms is often conceptualised differently (Tognolini et al 2001). For example, at times *standards* are conceived as a subset of *criteria*, at other times this hierarchy is reversed.

In this review, all paradigms have been considered in the synthesis of the literature. Hence, research on formative, summative and evaluative SBA has been sought. Reports of empirical research, theoretical treatises as well as policy documents and articles arguing uncritically for just one point of view have been reviewed. Diverse terms in the literature to fit the New Zealand context have been interpreted. Thus, the approach to assessment is *standards-based* (SBA); the subsets of SBA used in the NCEA are *achievement-based* (ABA) and *competency-based* assessment (CBA).

In commissioning the review, NZQA indicated that the focus of the literature review should be on SBA in the senior secondary school, including issues such as:

1. high-stakes assessment and SBA
2. impacts of SBA on teachers' classroom teaching, assessment and workload
3. impacts of SBA on students' learning and perceptions of achievement
4. what works particularly well with SBA
5. gender and ethnicity issues
6. low/medium/high achiever issues
7. motivation and self-efficacy issues
8. meeting student learning and assessment needs
9. validity, reliability and manageability issues
10. synergy between teaching and assessment
11. impact on lifelong learning.

After discussions with NZQA officials, a twelfth focus area was added to address the policy context within which SBA operated in New Zealand.

12. policy issues including an historical perspective on SBA in New Zealand.

Assessment for learning, in relation to the senior secondary school, impacts on motivation, learning and achievement were also to be considered.

Discussion with NZQA staff resulted in the identification of five key themes to address the focus areas:

*Theme one: policy issues*

Evidence to contextualise the use of SBA in New Zealand secondary schools.

*Theme two: technical matters*

Evidence that informs the technical debates surrounding SBA.

*Theme three: teaching*

Evidence and perspectives to inform understanding of the impact of SBA on teaching.

*Theme four: learning*

Evidence and perspectives to inform understanding of the impact of SBA on learning.

*Theme five: diversity*

Evidence on how SBA impacts on diverse learners.

These themes focused the systematic literature review, which was conducted between May and October 2004.

## **The methodology of a systematic review**

Being a systematic review of the literature, this project required a methodology that would enable the research team to locate relevant literature, select studies that were likely to be of most interest to NZQA, summarise the findings of those studies in a way that would be accessible to others, and write a critical review of the studies.

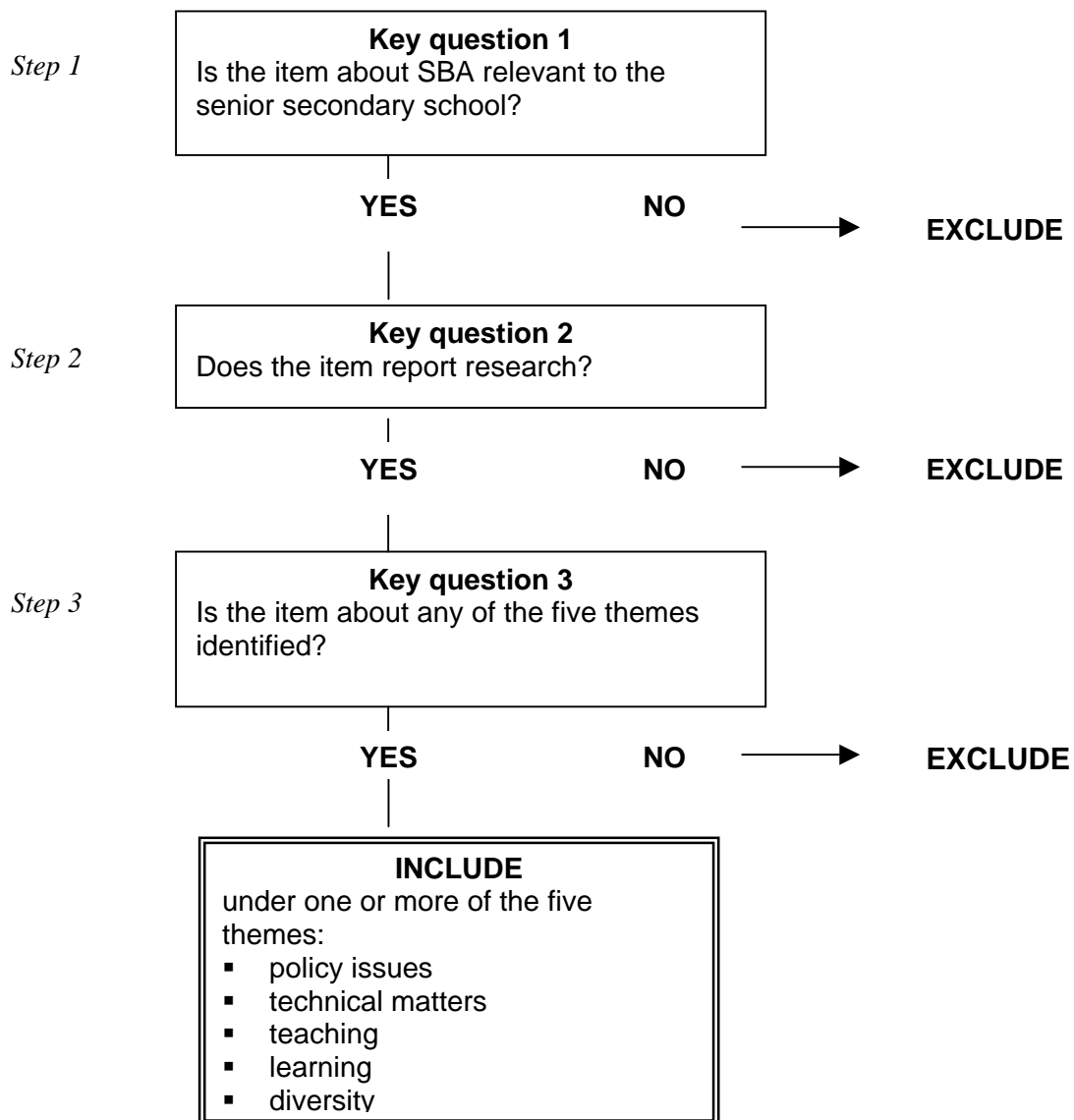
### **Search strategy**

A qualified librarian was employed to conduct the searches. He focused on literature written in English from Australia, Canada, New Zealand, the UK and the USA, concentrating on publications between 1990 and 2004. He searched journal articles, monographs, Internet databases, published research and evaluation reports. He was provided with a list of search terms that might capture the studies. These terms were: SBA, standards-referenced assessment, competency-based assessment, criterion-referenced assessment and achievement-based assessment. He extended and developed this list as he searched and familiarised himself with the literature and specific terminology of individual databases. He also searched specifically for literature on sub-populations such as Māori and Pasifika students. The initial search located 90 pages of bibliographic information about studies of interest. From this list those studies that seemed most closely related to the themes were selected and accessed and inclusion/exclusion criteria were applied to them. The search iteration continued for over four months until the writing of the critical review of the studies began. From that point on, only articles that were considered too important to leave out were included.

### Selection strategy: inclusion/exclusion criteria

A selection strategy was developed to determine which studies would be analysed for inclusion in the templates. Three steps were used in the inclusion/exclusion process. Each step was based around a key question. Figure 1 summarises the strategy and criteria. These were discussed with, and approved by, NZQA representatives.

Figure 1: selection strategy



The *research* criterion in the second key question posed some difficulties. It was decided to include not only empirical studies but also literature reviews and theoretical work in this cut.

Once written, each completed template was read critically by another member of the research team to ensure that all selection criteria were met. If there was doubt about a study it was *included* in the first instance.

This process was planned as an iterative one. The search for articles was ongoing, as was the application of the inclusion/exclusion criteria, sometimes relegating included

articles to the excluded pile. Date of publication emerged as another important criterion. While some studies of SBA were conducted in the 1980s, this review focused on more recent work, particularly that done since the mid 1990s.

### **Summarising strategy: the templates**

The review team was contracted to complete an annotated, evaluative bibliography and proposed to achieve this by developing a template format that would summarise key information from each article reviewed. The template format was designed to enable the reader to glean at a glance the information on the themes covered by the research, the nature of that research and some evaluative comments. The templates were not intended to be a substitute for reading the original. Indeed, the way material was treated varied with the type of literature reviewed. All material was reviewed with NZQA's focus areas in mind. This meant that, at times, valuable information in the book or article was not included on the template. A decision whether to complete one template for an edited book or a separate one for a number of chapters proved difficult. As a general rule, it was decided that where chapters were diverse, suitable chapters were summarised, but where the chapters were similar in content, the book as a whole was templated.

The template format was developed and agreed to by NZQA representatives. The headings for the template were:

- Standard bibliographic information.
- Country of origin.
- Key words.
- Abstract.
- Type of research (quantitative, qualitative, descriptive, theoretical, analytical, critical and other, eg action research/opinion piece/literature review/prescriptive).
- Key themes (details of any findings related to any of the five identified themes, eg policy, technical, teaching, learning, diversity).
- Methodology, including information on the scale of the project.
- Evaluative comments (the reviewers' comments on the value, place, worth and relevance of the article for the New Zealand context).
- Other (any other information about the article that had particular interest but was not summarised under other headings).

Eighty-eight studies were templated for the review.

### **Matrix**

In order to gain a visual impression of the information collected, and to help identify any gaps, two matrices were developed. The first (figure 2) shows the range of countries from which information was collected in relation to the five themes. The second (figure 3) shows the number of empirical studies informing each theme.

Figure 2: matrix of located studies

	<b>Policy</b>	<b>Technical</b>	<b>Teaching</b>	<b>Learning</b>	<b>Diversity</b>
<b>USA</b>	10	14	13	15	10
<b>UK</b>	3	10	9	11	2
<b>Australia</b>	3	6	2	3	0
<b>New Zealand</b>	13	12	15	14	3
<b>Other</b>	0	0	3	2	2

Most studies are listed under more than one theme. Indeed some studies contained evidence for all themes. While the New Zealand capture was relatively rich, the studies from Australia were few. Despite numerous emails and a full second bibliographic search, the yield from Australia remained small. This might be explained by Australia's relatively late entry into SBA in secondary schools. The number of studies dealing with diversity was also small. This was not surprising as the debates about SBA and its alternatives are often still at a philosophical level. Many of the works reviewed were concerned with defending and attacking the principles of SBA rather than investigating its effects.

The inclusion/exclusion criteria used a very broad definition of research. Included under this broad heading were quantitative and qualitative empirical works as well as systematic literature reviews. Consequently, it was thought important to maintain a good ratio of studies reporting empirical research against theoretical work. Figure 3 shows the number of empirical studies informing each theme. The total number of studies in that theme is given inside brackets.

Figure 3: matrix of empirical studies

	<b>Policy</b>	<b>Technical</b>	<b>Teaching</b>	<b>Learning</b>	<b>Diversity</b>
<b>USA</b>	7 (10)	7 (14)	10 (13)	10 (15)	6 (10)
<b>UK</b>	1 (3)	3 (10)	4 (9)	6 (11)	1 (2)
<b>Australia</b>	2 (3)	1 (6)	0 (2)	0 (3)	0 (0)
<b>New Zealand</b>	6 (13)	6 (12)	5 (15)	6 (14)	0 (3)
<b>Other</b>	0 (0)	0 (0)	2 (3)	1 (2)	0 (2)

The research team decided that at least half of the literature reviewed should consist of empirical studies. This goal was not achieved for the themes dealing with technical matters and diversity. This was not surprising as technical matters are usually dealt

with by analysing concepts rather than through empirical research, and the field in general is not advanced enough to systematically investigate diversity issues. For the remaining three themes the 50 percent empirical quota was achieved. Figure 3 also shows that almost 65 percent of studies originating in the USA were empirical while only about 40 percent of New Zealand studies were. This could indicate that SBA in the USA is more acceptable to the academic community. The literature in New Zealand tended to be more conceptual, even political, including many opinion pieces. This suggests that in New Zealand the debate about the acceptability of SBA still has to be resolved. Certainly there is room for more empirical research in New Zealand on this controversial subject.

### **Critical review strategy**

Eighty of the templated studies were used in the critical synthesis. Each member of the team reviewed a theme. All of the templates were read, with each member separately making notes on each one in an effort to identify *findings* for the five themes. Each *finding* consisted of a statement that addressed one or more of the focus areas specified by NZQA. Initial reports to support the *findings* for each theme were drafted. After a group discussion of these draft reports, each member redrafted her/his theme, rewording, amalgamating *findings* and identifying new *findings*. The number and titles of *findings* were not finalised until the draft was sent for editing. The second draft and the templates were quality assured and edited by another team member. Finally, a draft of the completed review was read and commented on by an independent peer reviewer.

### **The findings**

The findings from this research synthesis are:

#### *Theme one: policy issues*

1. SBA is a dominant assessment paradigm in English speaking countries.
2. A single SBA framework for all learning remains contentious.
3. Achievement standards are better at assessing some learning than unit standards.
4. Teacher training and development, good practice exemplars and moderation must be resourced adequately.

#### *Theme two: technical matters*

5. It must be clear which purposes of assessment are being served by SBA.
6. Standards emerge from a consensus-seeking process affected by ambiguities and subjectivities.
7. It is difficult to achieve both validity and reliability together in SBA.
8. Alignment between curriculum, teaching and assessment achieves better results.
9. SBA can atomise learning and make integration of learning difficult to manage.

*Theme three: teaching*

10. SBA can place students at the centre of teaching.
11. SBA can focus the enacted curriculum.
12. SBA can influence pedagogy positively.
13. Standards-based pedagogy can have a positive impact on student outcomes.
14. SBA initially increases teacher workload.

*Theme four: learning*

15. SBA can affect student learning positively.
16. The formative nature of SBA can be motivating.
17. High-stakes SBA is not uniformly motivating.
18. SBA raises student workload issues.
19. SBA can help meet student needs through programme flexibility.

*Theme five: diversity*

20. The impact of SBA on diverse students is under-researched.
21. The impact of SBA systems on diverse students' academic performance is variable.
22. The effects of high-stakes assessment and accountability on diverse students are marked in SBA systems.
23. Alternative assessments need to be considered for diverse students.

Figure 4 maps how the *findings* address the original *focus areas* set out in the NZQA *Request for proposal* brief as well as the additional focus area agreed to in subsequent discussions. The 12 focus areas are displayed on the vertical axis. The 23 findings are shown on the horizontal axis. Inspection shows that each focus area is addressed by at least one finding.

Figure 4: Relationship of findings to focus area

Findings	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Focus area																							
1																	*						*
2										*	*	*	*	*				*					
3															*	*	*	*	*				
4								*		*						*			*				
5															*						*	*	*
6															*						*	*	*
7															*	*	*	*	*				
8								*					*	*	*				*				*
9				*		*	*	*	*					*		*	*						
10										*	*	*						*					
11	*			*							*												
12	*	*	*	*																			*

### Limitations of this literature review

The literature reviewed here is, of course, a selection. There will be studies that were not able to be located which may provide quite different perspectives on some findings; indeed, different findings altogether. Also, projects published in other languages may offer alternative results to those discussed here as only research published in English was considered for this review.

### Recommendations for further research

It is clear from the matrices that there is a dearth of New Zealand empirical research on SBA in the upper secondary school. Further research on its effects on learning generally and on motivation in particular is needed. The effects of high-stakes assessment in the New Zealand context should also be investigated. However, perhaps the most important area for empirical research is the impact of SBA on diverse learners.

### Structure of this report

The critical review of the literature is now presented followed by the individual, annotated, evaluative templates. A bibliography of the templated studies and other references mentioned concludes this review.

## References

- Biggs, J (2003). *Teaching for quality learning at university*. (2nd ed). Maidenhead: SRHE and Open University Press.
- Croft, C (1993). *The conflicting world of standards-based assessment*. Paper presented at the fifteenth national conference of the New Zealand Association for Research in Education, Hamilton, New Zealand.
- Gipps, C V (1994). *Beyond testing: towards a theory of educational assessment*. London: The Falmer Press.
- Leach, L, Neutze, G & Zepke, N (2003). Course design and assessment for transformation, in N. Zepke, D. Nugent and L. Leach. (eds). *From reflection to transformation*. Palmerston North: Dunmore Press.
- Philips, D (2003). Lessons from New Zealand's National Qualifications Framework. *Journal of Education and Work*, 16(3), 289–303.
- Scriven, M (2003). Evaluation theory and metatheory. In T. Kellaghan, D. L., Stufflebeam & L. A. Wingate (eds). *International Handbook of Educational Evaluation*. Dordrecht: Kluwer Academic Publishers.
- Tognolini, J, Andrich, D & Ball, S (2001). *International best practice in outcomes-based assessment related to post-compulsory education*. Osborne Park, WA: Curriculum Council of Western Australia.
- Wolf, A (1995). *Competence-based assessment*. Buckingham: Open University Press.

## Synthesis of literature

### Theme one: selected policy issues in standards-based assessment

#### 1. Standards-based assessment is a dominant assessment paradigm in English speaking countries

Gipps (1994) identifies psychometrics, educational measurement and educational assessment as three assessment paradigms that have dominated assessment policy and practice since the 1950s. Within the educational assessment paradigm she identifies three forms: criterion-referenced, performance and teacher assessment. These three forms of assessment have become widely accepted in English speaking educational systems.

Criterion-referenced assessment emerged in the USA in 1963 (Gipps 1994). By the 1980s it had reached a dominant position, which Birmingham (2001) suggests was due to concerns about teacher accountability in raising educational standards. Apparently, all states in the USA have now developed a set of standards and assessment regimes. 'Even so with 50 assessment instruments each playing its own tune, there is at least as much noise as there is harmony' (Tognolini et al 2001, p 39). They also report that Canada, in introducing SBA, has experienced similar dissonance problems. In the UK, assessment traditionally focused on academic curricula. Wolf (1995) notes an increasing interest in vocational curricula in the 1980s with the development of National Vocational Qualifications (NVQs) which spelled out what a student should be able to do and to what standard (Philips 1998). Tognolini et al (2001), suggest that in the UK, centralised standards setting is believed to ensure comparable marking. In Australia, national curriculum reform followed the Hobart Declaration on Schooling in 1989. This included frameworks for assessment 'based on a defined progression of learning outcomes' (Philips 1998, p 74). At state level, the New South Wales Government's White Paper called for a standards-based approach in 1997 (Tognolini et al 2001). In the Australian Capital Territory, colleges teaching Year 11 and 12, trialled SBA in 1996 and 1997 (Francisco 1999). In summary, the most recent senior secondary initiatives in Australia are firmly fixed on SBA (Strachan 2001). He summarises the trend in English speaking countries as follows.

In education as in other fields, *standards* have been a catch-cry of the 1980s and 1990s. Educational reform initiatives for improved achievement, including demands for accountability and reporting of what students know and can do, increased interest in assessment against standards during the 1990s (p 250).

Indeed, SBA has been theorised as just one aspect of a wider educational reform movement (WestEd, 2000) Cowan et al 2002; Bennett & Merrick 2004). Clune (2001) takes up this view. He proposes a holistic model of standards-based reform, which through purposeful activities leads to standards-based policy, which then leads to a rigorously implemented standards-based curriculum for all students. This, in turn, leads to high student achievement through SBA. Clune argues that some kind of deepening of the curriculum is needed to inform SBA.

New Zealand was strongly influenced by overseas ideas during the 1980s. In 1990, after a decade of debate, the New Zealand government moved towards SBA and with that set up the New Zealand Qualifications Authority (NZQA), an agency responsible

for development of a National Qualifications Framework (NQF). One of its functions was to redefine education and training, subsuming both in a universal framework of competencies (Barker in Peddie & Tuck 1995). Many of the features of the NQF were migrated from the Scottish Vocational Education Council (SCOTVEC) and the National Council for Vocational Qualifications (NCVQ) in England (Eagle & McDonald 2000).

To explain this policy migration in New Zealand, Philips (1998) uses a seven-stage model. In the 'accumulation' stage there was an awareness of the need for change and overseas ideas were gathered. During 'incubation', these ideas were discussed and modified by policy makers but were not yet incorporated. 'Assimilation' saw a favoured structure for a New Zealand qualifications framework emerge. The details of the policy were filled out during 'translation' and these were integrated at the 'contextualisation' stage. 'Refraction' saw the consequences of policy in the local context, including modifications. The final stage, 'resolution' sees the integration of the imported policy or its abandonment. At the time of writing, some issues about SBA have not been resolved. These are now discussed.

## **2. A single standards-based assessment framework for all learning remains contentious**

A distinctive feature of the New Zealand migration of assessment policy was the decision to create a single, National Qualifications Framework (NQF) to recognise and register both academic and vocational learning. Philips (2003) considers the NQF to be possibly the most comprehensive in the world, embracing all nationally agreed achievements for qualification credit. He refers to this as the 'Big Idea' (Philips 1998). Eventually, the NQF was expected to consist of an unknown number of vocational and academic unit standards sorted into content domains and ten levels of difficulty.

However, the concept of a unitary framework combining vocational and academic learning has not found universal favour. For example, Hall and Codd et al in Peddie and Tuck (1995), argue that unit standards are more suited to assessment of technical or practical skills than higher order mental operations, are unsuited to university learning and are not flexible enough to cater for different contexts. Others claim their use threatens to fragment or atomise integrated knowledge and distort the curriculum (Elley & Hall in Peddie & Tuck 1995; Hearn 1997; Lee & Lee 2000). A further concern was the separation of doing from understanding (Elley in Peddie & Tuck 1995). Codd et al (in Peddie and Tuck, 1995) argued that more easily measured learning would dominate assessment while critical, creative and integrative thinking would be neglected. Kearnes and the New Zealand Vice Chancellors' Committee (cited in Peddie & Tuck 1995) argue that the use of unit standards rewards mediocrity and has no place for excellence. Despite this criticism, NZQA decided in 1993 that its unit standards based qualifications framework should remain the one 'Big Idea' (Locke 1999).

### **3. Achievement standards are better at assessing some learning than unit standards**

This ongoing critique led the New Zealand government in 1996 to approve the introduction of achievement standards as a more meaningful way of assessing some learning. This resulted in the broadening of the NQF (Lee & Lee 2000). A first step saw the inclusion of non-unit standards-based qualifications in the framework. A second step saw the announcement of the National Certificate in Educational Achievement (NCEA) in the Achievement 2001 policy. This introduced graded achievement – not achieved, achieved, achieved with merit and achieved with excellence – for both the internally and externally assessed achievement standards (Lee & Lee 2000; Philips 2003). However, it should be noted that a form of achievement-based assessment had been trialled in the 1980s.

Achievement-based assessment initiatives were not new in 1999. For example, Gilmore (1991) reviewed an initiative in assessing English using achievement standards, reporting that both students and teachers were positive about achievement-based assessment. Despite such strong endorsements, achievement-based assessment did not gain traction. However, groups of teachers in certain subjects continued to trial achievement-based assessment (Locke 1999). The English Study Design Project is an example of such activity. It continued to conduct small-scale research projects among English teachers. Results supported the achievement-based version of SBA (Locke 1999) and so the project team continued to develop and modify the English curriculum towards one based on achievement standards (Hall 1999).

### **4. Teacher training and development, 'good practice' exemplars and moderation must be resourced adequately**

SBA, it has been claimed, is costly to implement (Gilmore 1991; Wolf 1995; Linn & Herman 1997). Wolf (1995, p 133) suggests that SBA 'is a system of assessment which is extremely expensive. It is time-consuming in absolute terms, for both candidates and assessor and it imposes major costs in terms of equipment, provision of wide-ranging assessment situations and repeated evidence of mastery. Becker and Rosen (1992) address the issue of cost effectiveness from a different perspective. While arguing in favour of national testing of standards, they use economic reasoning to suggest that SBA is too costly to sustain in the classroom because it cannot motivate learners as well as norm-referenced assessment. In summary, three areas are identified as being particularly costly to implement: initial teacher training and ongoing professional development; exemplars to implement the standards; and moderation.

A number of authors argue that classroom assessment should be given greater prominence in initial teacher training (Assessment Reform Group 1999) and continuing professional development (Linn & Herman 1997; Lee & Lee 2000; NZPPTA 2002). Sizmur and Sainsbury (1997) argue that teachers must have a good grasp of the underlying construct in all its complexity so that their professional knowledge is central to the process of assessment. Shymansky et al (1997) affirm the complexity of the work needed to design, refine and score performance tasks. In their study, Florian et al (2000) identified five American state policies that contributed to district level capacity for standards-based reform. Two of these were providing professional development activities and allocating state and federal flexible funding.

The need for government-led funding is echoed by the Assessment Reform Group (1999) and NZPPTA (2002).

Another area of resourcing discussed in the literature relates to the development of teaching resources and exemplars. NZPPTA (2002) identifies problems with availability of assessment tasks, problems with their consistency and concerns that many standards provide surface learning only. Its report recommends that exemplars demonstrating good practice be provided. This could address the concerns about the variability in marking of the same piece of work. Sadler (1987) looked at four ways of describing standards and advocated that exemplars and verbal descriptions be used together. When used together, teachers may make sound qualitative judgements about the achievements of their students, both for improving learning and for summative reporting. From their study, Cowan et al (2002) identify the benefits of a library of exemplars for staff and students. Such a library improved the reliability of teachers' scoring and learnability for students.

Gipps (1994) links exemplars and moderation in her paradigm of educational assessment. Consistency of scoring, identified as a problem by NZPPTA,

... can be achieved through a process of moderation and provision of exemplars. These exemplars and moderation procedures need to be made available to all the teachers involved in any particular assessment scheme (p 160).

However, moderation threatens to be prodigiously complex and expensive and may threaten the usability of SBA (Peddie & Tuck 1995). Gipps (1994) lists statistical moderation, moderation by inspection, panel review, consensus moderation and group moderation as common moderation methods. In New Zealand, a number of these methods operate, although both Gilmore (1991) and Hall (2000) advocate a test-based moderation process. Hipkins et al (2004) found in their study of six schools that designing and moderating new assessment tasks was one factor contributing to increased teachers' workloads. In contrast to these views, some authors highlight the professional development benefits that accrue from moderation processes and offset the costs incurred (Wilson & Floden 2001).

## **Theme two: technical matters in standards-based assessment**

### **5. It must be clear which purposes of assessment are being served by standards-based assessment**

William (2001) cites four purposes of assessment identified from the work of the United Kingdom Task Group on Assessment and Testing. These are formative, diagnostic, summative and evaluative. Formative assessment enables the positive achievement of students to be recognised and appropriate next steps planned. Diagnostic assessment identifies learning difficulties so that appropriate remedial help and guidance can be provided. Summative assessment records students' overall achievement in a systematic way. Evaluative assessment reports on how well a school meets accountability requirements. The Assessment Reform Group (1999) makes a clear distinction between assessment *of* learning for the purposes of grading and reporting, and assessment *for* learning 'which calls for different priorities, new procedures and a new commitment' (p 2). While there are tensions between

evaluative and formative assessment (Black & Wiliam 1998; Winch & Gingell 1996), Brookhart (2001) found formative and summative assessment to be quite compatible.

SBA is but one aspect of what has become a standards-based movement. Clune (2001), for example, theorises an integrated process of standards-based education. A first stage in this holistic process is the desire for standards-based reform which, through purposeful activity, leads to standards-based policy. This in turn leads to a rigorous standards-based curriculum for all students and to SBA. Hall (1999) echoes the integrated standards-based education view of assessment when he suggests that there is more to standards than merely specifying what a student needs to know, do or value. He suggests the need for 'systems standards' that define what needs to be done to support the assessment process. This integrated model stands in contrast to the New Zealand experience. The holistic standards-based education model theorised by Clune did not migrate to New Zealand in its full form. Rather, as Philips (1998) suggests, New Zealand imported only the SBA aspect. Government, through NZQA, created policy and assessment standards while curriculum matters were addressed elsewhere by the Ministry of Education (Hall in Peddie & Tuck 1995).

SBA is used both in assessment *for* and assessment *of* learning, so serves formative, summative and evaluative purposes (Assessment Reform Group 1999; Wiliam 2004). There is considerable debate in the literature about which of these purposes should take priority. Winch and Gingell (1996) argue that the evaluative aspect is the most important because teachers must demonstrate that they are meeting the standards set by society. Hall (1999) too sees public confidence in the system of assessment as important. National standards, as implemented in the USA (Younghee 1998; O'Neill & Stansbury 2000) and in the UK (Winch & Gingell 1996; James 2000) primarily serve evaluative purposes. Others critique the prevalence of assessment for evaluation and accountability as they lead to teaching to and learning for the test (Boss et al 2001; Wiliam 2004). Davis (1995) argues that evaluative assessment, in the form of national standards and their criteria, cannot capture the complexity and variety in 'rich' knowledge and skills. SBA also has formative purposes that can lead to significant improvement in student achievement (Black & Wiliam 1998; Assessment Reform Group 1999; James 2000; Supovitz 2001; Wiliam 2004). Formative assessment focuses on the improvement of learning, on immediate feedback and on learners' understanding of what they have to do to meet the standards (Brookhart 2001). However, it is not certain that the different purposes of SBA are mutually exclusive. Brookhart (2001) argues that they complement one another. Preece and Skinner (1999, p 24) conclude, '... if the two forms of assessment [national tests and teacher assessments] are to complement each other satisfactorily, it is important that they have equal status and that their complementary roles are understood'.

## **6. Standards emerge from a consensus-seeking process affected by ambiguities and subjectivities**

Setting standards is an uncertain and ambiguous process, relying on the subjectivities of individuals and groups. Sadler (1987) argues that standards-referenced assessment makes direct and extensive use of teachers' qualitative judgements. Sadler (2005) draws on an extensive array of literature from 65 universities in seven countries to show that the words criterion-referenced describe four different models of grading work. Linn and Herman (1997) and O'Neill & Stansbury (2000) note that a set of standards should represent consensus among stakeholders, on what students must

know and be able to do. They suggest that standards must be acceptable to key constituencies. Linn and Herman (1997) add that the consensus includes the opinions of teachers, business people, and parents who meet to set standards but must also be involved in reviewing student work to enable it to be placed on appropriate proficiency levels. This in itself is problematic. According to Cowan et al (2002), two groups of experts tasked with developing standards for the same subject would come up with different sets. Indeed, Croft (1993) points out that the term standards itself has been used in numerous ways to mean different things, even within NZQA literature.

Different types of standards have been identified. O'Neill & Stansbury (2000) specify two in their handbook on developing implementation systems in the USA. The first type; content standards, identify the areas of knowledge, understanding and skills students are expected to learn in key subjects and career areas. The second type, performance standards, indicates how well individuals perform, by defining and illustrating levels of expected accomplishment of content standards. In a New Zealand context, Locke (1999) identified two different types of standards again; unit standards and achievement standards. Unit standards assess individual competence. Achievement standards specify levels of achievement. Further, Croft (1993) argues that achievement standards may be potentially shaped by 'reasonable performance levels' rooted in norm-referenced considerations.

While the technical aspects of SBA are complex, the debate about their value is even more so. There are avid proponents of this type of assessment and equally passionate critics. Advocates stress: improved transparency and understanding of the assessment process (Barker in Peddie & Tuck 1995; Francisco 1999; Tomlinson 2002); higher levels of student achievement (Supovitz 2001); improved links between knowledge and performance (Barker in Peddie & Tuck 1995); improved generic skills (Gfroerer 2000); more stability and robustness of teacher judgements from diverse assessment methods (Pitman 1985); enhanced international comparability (Peddie & Tuck 1995); and the potential democratisation of learning and the erosion of traditional barriers and quotas (Barker in Peddie & Tuck 1995). Indeed, Gipps (1994) argues that SBA ameliorates competition, reduces anxiety, increases intrinsic motivation, promotes achievement and cooperation, self-efficacy, metacognition and deep learning.

Critics, on the other hand, have many complaints. Croft (1993) attacks one of the main arguments for SBA – that it departs from norming – when he points out that it can be a hidden form of norm-referencing. Lee and Lee (2000) and Sismur and Sainsbury (1997) identify issues of proliferation, atomisation and specificity in criterion-referenced assessment as resulting in manageability and workload problems for teachers. Moreover, they quote Dearing: 'many teachers feel that the mechanics of recording teacher assessment information have interfered with teaching and learning' (p 137). Singh-Morris (1997) also contends that unit standards are reductionist. Students do not learn in discretely defined bits so assessing pre-specified skills may lead to a narrowing of the curriculum, over-assessment and the growth of an assessment curriculum rather than a learning curriculum. Wilson and Floden (2001) agree, reporting fears that standards will trivialise education and deskill teaching by being too prescriptive, creating an inflexible delivery system that is incapable of coping with diversity. There are also claims (Hall 1999; NZPPTA 2002) that SBA is neither reliable nor objective (Elley in Peddie & Tuck 1995; Davis 1995).

The complexity reported in the literature is confounding. However, Gipps (1994), Hager et al (1994), Hearn (1997) and NZPPTA (2002) agree that all the complexities and criticisms mentioned can be addressed.

### **7. It is difficult to achieve validity and reliability together in standards-based assessment**

Validity and reliability presume to guarantee, on the one hand, that assessments are fit for the purposes they are set, and on the other, that these purposes are consistently met. Among a number of technical worries, Hager et al (1994) list questions about validity and reliability raised by commentators who suggest that SBA is inherently invalid and unreliable. Davis (1996) partially agrees, arguing that validity is achievable only at a cost, particularly in the learning of 'rich' knowledge. However, this seems too strong a position. Hager et al (1994) refute each of the eight worries identified, including those about validity and reliability, arguing convincingly that a well-designed SBA system can overcome each. Moreover, it seems generally accepted that validity, understood as fitness for purpose (Gipps 1994; Hall 1999), is achievable. The notion of fitness for purpose implies that learning, teaching and assessment take place in specific contexts. Gipps (1994) develops a context-related model that focuses on consequences of assessment for students. Using the notion of consequential validity, she argues that inferences and uses of assessment will be valid if interpreted for the local context. Khattri et al (1998) support this line of argument from data obtained in their research. Black and Wiliam (1998) buttress the argument for validity in SBA with the notion of ecological validity which concerns the degree to which the behaviours observed and recorded in assessment reflect the behaviours that actually occur in natural settings. In short, the balance of opinion in the literature seems to be that SBA can be valid as long as it assesses course learning outcomes and is fit for the context for which it is intended.

The case of reliability is not so straightforward, for context related validity challenges the whole basis of reliability. Davis (1995, 1996) goes so far as to argue that, conceptually, SBA can never be both valid and reliable. He, Gipps (1994) and Hall (1999) suggest that validity is more important than reliability. Crombie, in Peddie and Tuck 1995, points out that concern with reliability can lead to an over-restrictive view of what is being tested, and may result in invalid assessment (Linn & Herman 1997). Gipps (1994) argues that the traditional correlational idea of reliability does not apply to SBA. SBA does not require traditional reliability criteria based on measures such as reliability coefficients, which she argues can be misleading and inappropriate. Rather, she wants assessments to be comparable. The level of comparability depends on the purpose of the assessment. In high-stakes tests, this needs to be greater than in formative assessments. While these writers seem comfortable with jettisoning or altering traditional notions of reliability, Hall (1999) points out that the public may not be happy to do so. He suggests that public confidence in SBA will be undermined if reliability cannot be demonstrated. Hager et al (1994), arguing for an integrated form of SBA, suggest that reliability suiting the needs of SBA is achievable. They point out that informed professional judgement has been found to have a high level of consistency. Gipps (1994) puts her faith in moderation to achieve comparability. While she excludes statistical moderation, she advocates moderation by consensus groups and external consultations.

While in a technical sense, fitness for purpose side-steps traditional notions of validity and reliability, Hall's concern about public perception cannot be dismissed. Gipps' notion of comparability is flexible, and addresses both fitness for purpose and the concerns about reliability in high-stakes assessments. Where assessments are high-stakes, comparability needs to be more rigorous than for formative assessments. In a high-stakes situation, Hall's idea of integrated moderation tests, would seem to answer the need for greater rigour. For example, in his evaluation of English study design moderation tests, he found an average consistency of 0.92 with teacher assessments (Hall 2000).

## **8. Alignment between curriculum, teaching and assessment achieves better results**

Alignment is a key technical matter addressed in the literature. The concept has different constructions. It may specify alignment between standards and assessments, between standards and teaching, between standards and curriculum and between teaching and assessments. Any or all of these may be present in specific cases. The evidence is that the closer the alignment between these factors, the better students achieve.

Linn and Herman (1997) argue that alignment depends on teachers' ability to understand expectations set out in the standards and to obtain resources and expertise to help students meet those standards. Birmingham (2001) argues that the best way to make this link is to develop assessment tools that derive directly from the curriculum and cites a number of research studies that show that a stronger link between assessment and teaching results in higher success rates, and the link between them allows for more student-focused teaching. Porter and Smithson (2001) argue that there should be close alignment between curriculum and assessment. They warn that assessment tasks should not be taken as indications of the 'intended curriculum' as they represent only a sample from the content domain the assessment is intended to represent. Clune (2001), arguing for an integrated standards-based education, highlights the importance of alignment between standards-based curriculum, standards-based learning and SBA. In seven of nine case studies reported, there were moderate to strong gains in student achievement. Clune (2001, p 14) concludes that better alignment produced 'wide-spread and substantial gains in the quality of teaching and learning for all students ...'. In their research, Boss et al (2001) found that continuous juggling between misaligned assessments and curriculum were a threat to teaching and learning. Supovitz (2001) found evidence that poor alignment between what is taught and the way it is tested, or between content and what is tested, may affect gains in student achievement.

## **9. Standards-based assessment can atomise learning and make integration of learning difficult to manage**

There are many questions about the manageability of SBA. As they often address quite different manageability issues they are addressed in this report under their respective themes. For example, manageability issues, such as increased workloads for both teachers and students, are addressed under the findings relating to teaching and learning. Manageability questions about the cost-effectiveness of context-based SBA and the cost of establishing reliability, for example through expanded

moderation processes especially for high-stakes assessment, have been addressed under the policy issues theme. In this section, manageability is addressed from a technical perspective, identifying issues that arise from the complex structure of standards and the atomisation of learning.

A number of authors, particularly in the New Zealand context, critique the atomisation of learning in SBA (Peddie & Tuck 1995; Davis 1996; Hearn 1997; Hall 1999; Boss et al 2001; Strathdee & Hughes 2001; NZPPTA 2002; Barrington 2004). They agree that it is educationally unsound and creates a management burden for both teachers and students. They argue that standards are often represented as being statements of discrete knowledge and skills to be achieved. The assumption here is that discrete skills are generic and can be transferred to multiple contexts. Further, discrete skills do not represent the holistic nature of knowledge. Davis (1995) argues that rich knowledge and skills are complex, relational to many other things. To know something or to be able to do it involves complex and holistic systems of belief connected in diverse ways. Learning does not occur in discrete bits but is an integrated process. Hall (1999) argues that, in any case, assessment against separate standards is unlikely to satisfy public credibility and at the same time runs the risk of fostering a 'bricks without mortar' approach to course design, delivery and assessment. Atomisation results in an explosion of potentially infinite numbers of standards, which immediately impacts on manageability. Boss et al (2001), suggest that a 'relentless tyranny' of perpetual assessment results.

Other writers show that SBA does not necessarily have to create such manageability issues or be atomistic. It can be managed in a number of ways. Davis (1995) implies that broad general standards, allowing for variety and richness in contexts, are preferable to atomistic standards. Hearn (1997), reporting the views of the Qualification Framework Inquiry, also argues that atomisation is not an inevitable consequence of SBA, that it can elicit sophisticated skills and knowledge and does not inhibit them. Hager et al (1994), argue that it is counter-productive to atomise knowledge, skills and tasks because it destroys the distinctive character of a body of knowledge. Bodies of knowledge and skills are '...much richer than sequences of these isolated tasks and the overall approach fails to provide any synthesis of the tasks' (Hager et al 1994, p 4). Consequently, they recommend an integrated approach in which discrete standards are combined to better represent the field of knowledge. Hall (2000) echoes Hager et al and refers to the integration of unit standards in Year 12 English achieved by English Study Design (ESD) teachers. Locke (1999) reports that integration of units in ESD and the addition of achievement-based assessment (ABA) was supported by 88% of teachers in the project.

### **Theme three: the impact of standards-based assessment on teaching**

#### **10. Standards-based assessment can place students at the centre of teaching**

Assessment is one of the most powerful educational tools for promoting effective learning. But it must be used in the right way. There is no evidence that increasing the amount of testing will enhance learning. Instead, the focus needs to be on helping teachers use assessment, as part of teaching and learning, in ways that will raise pupils' achievement. (Assessment Reform Group, 1999, p 2)

In recent years, the traditional view of assessment as being separate from learning has been challenged by emerging views that learning and assessment are inextricably linked. The New Zealand Ministry of Education has published a number of policy documents that recognise and emphasise the importance of assessment in informing learning, for example: *Assessment for better learning* (New Zealand Ministerial Working Party on Assessment for Better Learning 1989), *The New Zealand Curriculum Framework* (MOE 1993) and *Assessment: policy to practice*, (MOE 1994). This shift in the focus of attention towards greater interest in the interactions between assessment and classroom learning is coupled with the hope that improvement in classroom assessment will make a strong contribution to the improvement of learning (Black & Wiliam 1998).

Standards-based reform is consistent with this emerging view of assessment *for* learning rather than assessment *of* learning (Black & Wiliam 1998; Assessment Reform Group 1999; Crooks 2002). Assessment *for* learning places formative assessment at the centre of the interaction between student and teacher and calls for deep changes both in teachers' perceptions of their own role in relation to their students and in their classroom practice. In particular, it suggests a move to a more student-centred teaching approach, placing the student in a more active role in the learning, teaching and assessment cycle, thus creating a partnership between student and teacher. The clarity and transparency of assessment standards help teachers provide students with information about what they know and can do and, more importantly, a clear picture of what they need to do to improve so they can take charge of their own learning (Black & William 1998; Crooks 1988).

Black and Wiliam (1998) note a number of factors that limit the effective implementation of assessment *for* learning. Firstly, formative assessment is not well understood by teachers and is weak in practice and secondly, national or local requirements for certification and accountability will exert a powerful influence in classroom practice. This last point has also been observed by James (2000) noting that assessment is increasingly being used for measuring and judging teachers and that the publication of assessment results puts great pressure on teachers to draw high test performances from their students. Winch and Gingell (1996) support this trend towards greater accountability of teachers. In their view, ensuring accountability is so important in education that assessment for accountability overrides other purposes of assessment.

## **11. Standards-based assessment can focus the enacted curriculum**

There has been much criticism directed at standards-based reform and its perceived effect on pedagogy and the enacted curriculum. One of the main criticisms is that it atomises the curriculum and fragments intricately integrated knowledge (eg Peddie & Tuck 1995). Holistic knowledge and understanding gives way to knowledge that is more easily measured at the expense of critical, creative and integrated thinking.

Two studies from the UK found that assessment has a constricting effect on curriculum and pedagogy, with teachers tending to teach for assessment rather than learning, resulting in individual topics being covered in less depth and some topics in the curriculum not being covered at all (Preece & Skinner 1999; Harlen & Crick 2003). Kannapel et al (2001) found in Kentucky that the 'Core Content for Assessment' was used in preference to the full 'Kentucky Curriculum Framework'.

Although the ‘Core Content’ was meant to be only a part of a comprehensive curriculum, many participants in the research reported that testing core subjects took precedence over teaching from the full curriculum. This was partly due to the full Curriculum Framework’s lack of specific alignment to the assessment instrument – the Kentucky Instructional Results Information System (KIRIS) – and partly to its sheer size, over 500 pages. It should be noted, however, that while the emphasis of teaching to the test narrowed the curriculum, in some instances it had the effect on some schools of expanding curriculum into areas that had previously received little attention, for example arts and humanities. As already discussed under technical matters, alignment between curriculum and assessment is a key factor in raising student achievement. If there is close alignment between curriculum and assessment then teaching to the test will result in teaching the curriculum.

The above findings are mitigated by the results of an extensive piece of research conducted by Wilson and Floden (2001). This research found little evidence to support critics’ fears that certain aspects of the curriculum were being under or over-emphasised and that teaching was becoming more narrowly focused. Although every teacher reported that tests affected instruction, classroom observations showed that assessments were neither predominant nor entirely absent. Many teachers admitted that assessment was having a significant influence on the enacted curriculum in their classroom but although teachers were aware of the debate about ‘teaching to the test’, some rejected the claim while others felt it was a non-issue because the test helped teachers to focus and raised expectations for all students.

## **12. Standards-based assessment can influence pedagogy positively**

The literature provides mixed messages about the impact of SBA on how teachers teach. In a study by Boss et al (2001), teachers reported that assessment was impinging on teaching time and reported the need for continual adjustment to goals and teaching plans in order to inject assessments into the daily routine. There was a growing perception of assessment as a ‘relentless tyranny’, with curriculum material being covered summarily to meet short-term goals, threatening high-level mastery. Teachers felt that there was an insidious effect of assessment becoming more important than the ‘joy of learning’. In a study by Preece and Skinner (1999), teachers felt that their teaching was more didactic with less emphasis on practical work and other student-centred activities. Teaching seemed to be more convergent than divergent.

In contrast to these findings, Wilson and Floden (2001) found in their American study that, although standards-based reform was having an effect on pedagogy, the effects were not extreme and in essence teaching was not being ‘re-invented’ in the image of assessment. The authors found that the clarity of assessment standards created a catalyst for teachers using their professional judgement to create a more coherent teaching practice embracing the old and the new.

Teachers had rationales for what they taught and when, and a clear sense of direction and obligation. But most teaching remained more familiar than new, more ordinary than challenging (p 214).

Consistent with these results are the findings of a recent Education Review Office (ERO) report. ERO (2004) found that 93% of schools reported that the introduction of the NCEA had not led to substantial changes to teaching practices and concluded that significant changes were not required in schools where teaching practice was previously effective.

The balance of the evidence is that the effects of SBA on pedagogy are moderately positive. According to Kannapel et al (2001), teachers in Kentucky reported that standards-based reform has had a positive impact on their pedagogy. Many teachers attempted new pedagogical approaches aimed at introducing more variety, subject matter integration, thematic instruction, hands-on experiences and group activities. These changes to pedagogy are consistent with the intent of the standards-based reform initiative in Kentucky. Preece and Skinner (1999) found that 60% of the teachers in their study reported that that the curriculum and assessment structure gave a much better focus to their teaching resulting in it being more organised, systematic and standardised.

These views are supported by the results of a New Zealand study by Bushnell (1992). This study looked at the use of grade-related criteria to assess drama and found that it led to: more effective assessment; teachers having an increased awareness of what they were doing; more effective reporting to students and parents about their progress and achievement; and a positive impact on their own teaching at other levels of the school. A similar New Zealand study by Eng (1992) found a weak indication that using grade-related criteria may have led to a shift to a more student-centred approach on the part of the trial teachers. In addition to this finding, 64% of the teachers reported that their teaching style had changed in a positive way and that the criteria focused goals and course requirements. Barrington (2004) also reports that under SBA, teachers are more focussed on higher order thinking skills and are more concerned with teaching for learning rather than teaching for high-stakes.

### **13. Standards-based pedagogy can have a positive impact on student outcomes**

The balance of evidence from the literature suggests that standards-based reform, and its implied pedagogical changes, has a positive impact on student outcomes (Black & Wiliam 1998; Clune 2001; Kannapel et al 2001; Hipkins 2004; Hipkins et al 2004). There are, however, some studies that report that the effects may only be modest (Supovitz 2001) or not unequivocal (Khattri et al 1998). There are also some indications that it may create a performance orientation in students which is contra to the philosophical underpinnings of SBA (eg James 2000; Stefanou & Parkes 2003).

Studies by Supovitz (2001) and Khattri et al (1998) in the United States address the question why the positive impact of SBA on student outcomes may be modest. Supovitz (2001) warns that changing teachers' practices may not translate into gains in student achievement in the short term and has identified a number of possible arguments as to why this might be so. The first two of these are: poor alignment between what is taught and the form by which it is tested; and poor alignment between the content of what is taught and what is tested. Supovitz (2001) also argues that our preoccupation with instant results leads us to seek effects before they can reasonably be expected to show up. Education is a cumulative process and small effects compounded over multiple years may translate into much larger impacts for both students and teachers. While it may be reasonable to expect professional

development to influence teacher practice almost immediately, it may be asking too much to expect students to absorb and act on a new pedagogical style with which they are not familiar. In addition to this, it will take time for the principles and ideas underpinning assessment reform to be clearly defined and understood at all levels of educational organisations and the wider community (Khattri et al 1998). Schools in the USA are now extending standards-based reform into the junior school and conducting public information campaigns so that students and their parents are more familiar with the implied pedagogical style and the associated methods of assessment inherent in standards-based reform.

#### **14. Standards-based assessment initially increases teacher workload**

One key area of criticism relates to the impact of SBA on teacher workload (eg Singh-Morris 1997; Shymansky et al 1997; Sizmur and Sainsbury 1997). One of the key reports from New Zealand's point of view is the Te Tiro Hou report on the Qualifications Framework Inquiry commissioned by the NZPPTA in 1997. This report considered the likely impact on secondary school qualifications of the government's Green Paper on the National Qualifications Framework. In summarising the main findings of the report, Hearn (1997) noted the following possible impacts on teacher workload. Firstly, atomisation of the curriculum resulted in complex assessments that impact on teachers' workload. Secondly, although the educational value of re-assessment was recognised, teachers identified re-assessment as a key workload issue, arguing that it will be necessary to get a reasonable balance in the way that re-assessment is managed.

In a submission by the NZPPTA in 2002 to the Education and Science Select Committee inquiry into the implementation of the NCEA, it was noted that a higher than anticipated increase in teacher workload was not expected to reduce due to the complexity of assessment tasks and an increase in administrative paperwork. The recent ERO report (2004) on the implementation progress of NCEA acknowledged the substantial work required by schools to prepare for NCEA, but also noted that a surprisingly low number of schools, only 16%, reported ongoing workload issues in terms of additional work required to develop resources.

Research by Shymansky et al (1997) into the design of science assessment tasks in Iowa affirms the complexity of the work in designing, refining and scoring assessment tasks that capture important evidence about what students think and can do. The authors argue, however, that enhancement of pedagogical practices and the improved assessment of learning justify the expense of effort and resources. Similar results have been found in studies by Bushnell (1992) and Francisco (1999) confirming that, although the process of implementing SBA was initially time consuming, the benefits to improved assessment and the re-evaluation of teaching strategies, outweighed the cost. In a recent report, Hipkins et al (2004) noted that although teacher workloads remain high, they were more manageable when schools set aside time for professional discussion and course development during the school week. Indeed, ERO (2004) have found that 73% of teachers reported that time allocated for professional development was a factor assisting in the implementation of the NCEA.

The improvement to professional dialogue between teachers has been highlighted as a significant positive consequence of standards-based reform (eg Harlen & Crick 2003;

James 2000; Clune 2001; Wilson & Floden 2001). For example, Wilson and Floden (2001) reported that assessment standards have led to increased professional dialogue between colleagues. Many teachers reported that conversations with peers were the most productive and meaningful professional development in recent times. Two major areas of discussion were highlighted from this research. One involved discussion about the match between the enacted curriculum and state standards and the other, discussions about using test results to pinpoint topics where students are doing worse than expected.

## **Theme four: the impact of standards-based assessment on learning**

### **15. Standards-based assessment can affect student learning positively**

From a teaching perspective, SBA can affect positively both how teachers teach and the achievement of their students. Principals in New Zealand secondary schools have also observed that students are becoming more empowered by SBA (Barrington 2004). In Canada and the USA, researchers have observed that students appear to be gaining more knowledge of the learning process itself (Brush 1997; Boss et al 2001). They are applying higher thinking skills to learning and the assessment process is more transparent than the previous system (O'Donovan et al 2000; Barrington 2004). NZCER's *Learning Curves* research project (Hipkins & Vaughan 2002), also reported that teachers have observed an increase in student confidence. This is particularly evident in English and mathematics where achievement is seen to be rising (Hipkins 2004; Hipkins & Vaughan 2002; Hipkins et al 2004). More students are returning to school to attempt to meet the learning outcomes in a belief that they can gain the necessary qualifications to equip them for their future aspirations.

A number of other projects have found that SBA and standards-based instruction enhance learning. Ronis (1999) used brain-imaging techniques to see whether any instructional techniques and assessment processes were superior in supporting learning in mathematics. He identified performance-based instruction and assessment as more effective than more traditional techniques. Bushnell (1992), Gipps (1994) and Kannapel et al (2001) also indicate that clear standards enhance learning, with quality tasks promoting self-monitoring. The Education Commission of the States (2002) and Barrington (2004) place emphasis on clear goals for learning and teaching and argue that this is beneficial in the shaping of the performance of teachers, learners and state school systems. Accountability is now about teachers supporting students in their learning. This involves teachers making clear links with the curriculum and gathering evidence that learning meets the achievement levels of the standard.

Many schools in New Zealand are beginning to acknowledge the value of SBA to learning (Barrington 2004; Gibson 2004; Mallard 2004). Assessment is no longer about students competing against one another, but about achieving to a set standard. Given the range of diversity of the New Zealand student population in schools, Gibson (2004) draws attention to issues that have previously had negative effects on student learning. He criticises the effects on students who, under the norm-referenced system, ended up with a 'failed subject mark' (p 3), yet may have responded well to individual sections of an examination. Another example he gives is the marginalising of students who participated in Māori immersion programmes and who were not provided with assessment in Te Reo, unless by special arrangement.

Eng (1992), Anderson (1999) and ERO (2004) reported that schools with experience of SBA through previous implementation of unit standards exhibited the most confidence in implementing the NCEA. In some instances new types of learning programmes have emerged. These were of particular benefit to underachieving students. ERO (2004) noted that schools encouraged students to enter learning programmes that matched their abilities. Brush (1997) reported that underachieving learners improved with SBA after becoming familiar with it. Roderick and Engel (2001) found that for the at-risk students in their study, SBA on its own improved neither student performance nor learning. They suggested that to improve learning, schools needed to offer ongoing support, particularly to at-risk students.

Some criticisms still remain such as the atomisation of content, misalignment between SBA and curriculum and over-assessment. According to the critics (Peddie & Tuck 1995; Singh-Morris 1997; Simon & Forgette-Giroux 2000; ERO 2004), the teaching and assessing of multiple discrete content areas in the curriculum, rather than taking a more holistic approach, interferes with learning. For example, ERO (2004) reported some instances whereby too many standards were being attempted at levels 1 and 2, leading to issues of assessment 'crowding out' learning. NZPPTA (2002) also raised concerns over student choice in selection of standards to learn. Too much choice will result in students becoming 'autonomous choosers'; people who function well in the market place but who do not have 'an education' (Singh-Morris 1997). Advice to students should ensure that standards selected are based on a balanced learning programme that meets their needs.

## **16. The formative nature of standards-based assessment can be motivating**

If the literature on high-stakes assessment is mixed, on formative assessment it is very clear. Formative, or classroom assessment as it is sometimes referred to, is a process that can help motivate students to achieve higher standards (Anderson 1999; Barrington 2004). Wiliam (2004) argues that student achievement will improve primarily through what happens in classrooms. The teacher's role in the classroom is not to teach per se, but rather to create situations in which students learn. Black (2000) is of the view that assessment should primarily aim to serve the purpose of promoting student learning.

To motivate students to participate in learning, Bushnell (1992), Eng (1992) and Francisco (1999) note that detailed feedback assists students to understand why they got a particular mark. The feedback provided a transparent account, not only of how to achieve the standard, but also contributed to an increased understanding of the assessment process. ERO (2004) reported that teachers and students were now receiving more information about assessment criteria and expectations than previously. Students appreciated increased opportunities to discuss learning with their teachers. They identified this as an improvement from the previous norm-referenced system.

Formative assessment through a feedback process to students involves a shift from quality control in learning to quality assurance (Black & Wiliam 1998). 'Rather than teaching students, and then, at the end of the teaching, finding out what has been learnt, it seems obvious that what we should do is to assess the progress of learning whilst it is happening' (Wiliam 2004, p 16). Teaching can then be adjusted if students do not appear to be learning. By providing clear objectives, students can recognise

gaps in their learning between the current level and desired level, and can take effective action to close the gap (Ronis 1999; Brookhart 2001). Crooks' (1988) research has shown that classroom evaluation practices have substantial impact on students and their learning. Students have been shown to benefit through feedback, feed-forward and self-evaluation when there is an increased use of formative assessment (ERO 2004). The NCEA provides opportunities for such meaningful and focused formative assessment through the internally administered assessment events that occur throughout the year.

Black and Wiliam (1998), the Assessment Reform Group (1999) and Brookhart (2001) are of the view that if students are aware of their own processes of learning and understanding, then self-assessment becomes part of their learning. They also expressed the view that these skills need to be learnt by students. ERO (2004) noted in its evaluative report of NCEA, that there appeared to be a strengthening between teaching, assessment and learning outcomes, which assisted students to evaluate their own learning. The report identified two approaches to self-assessment. One was the use of tracking sheets by students to monitor progress towards achieving components of an achievement standard or unit standard. The other involved increasing students' self-assessment skills by encouraging them to reflect on their performance, set goals and discuss these with peers.

In New Zealand, the change to SBA has involved an increase of in-class assessments (unit standards and the internally assessed achievement standards). As more teachers are recognising the value of formative assessment on student learning, they are also exploring more meaningful ways in the presentation of assignments for achieving standards (Education Commission of the States 2002). Through engagement with performance tasks and portfolio assignments (Anderson 1999), students are able to observe and define boundaries of their learning as well as observe and determine the quality of their outcomes. This impacts on their motivation to learn as well as developing their writing and thinking skills (Khattari et al 1998).

### **17. High-stakes standards-based assessment is not uniformly motivating**

The literature regarding student motivational states indicates that the responses to assessment of individual students are complex (Crooks 1988). Motivation is affected by the ability, personality, experiences, current attitudes and self-perception of the learner (Crooks 1988). Harlen and Crick (2003) identified three overarching attributes which influence the motivational state of the learner. The attributes are largely dependent on the view of the learner's self-concept, their energy for the assessment task and their self-view of the capacity to complete the task. Learners with high levels of the three attributes have demonstrated that they are successful in achieving standards. Students, who acquire standards cumulatively and are not subject to high-stakes testing, not only enhance and increase their self-esteem but are also more motivated and become competent learners (Gipps 1994; Harlen & Crick 2003).

Becker and Rosen (1992), using econometric modelling, argue that competitive assessment (norm-referencing) has considerable motivating powers and should be used more commonly. Mixing students of different abilities working towards single standards impedes motivation. The more capable students know that they do not have to work very hard, whereas the less able may not even try and therefore give up. What

are needed are variable standards for students to select from, as this is more likely to lead to learning success (Becker and Rosen 1992).

While high-stakes assessment does not necessarily have singular effects (Roderick & Engel 2001), the majority of the literature on high-stakes testing suggests that, on balance, it has more demotivating effects than motivating outcomes. It motivates some students to greater efforts, but also demotivates others, particularly those most at risk of not achieving (Roderick & Engel 2001). Reay and Wiliam (1999) report very negative effects of high-stakes testing using SBA in a class of students. They observe that the panoptic surveillance embedded in high-stakes assessment resulted in students judging themselves as if some external eye was constantly monitoring them. James (2000) similarly highlighted some negative impacts of high-stakes SBA. When she reviewed the national curriculum in England, she considered that the high-stakes assessment used for this curriculum had considerable destabilising and stressful effects on students. This occurred when students became performance orientated rather than learning orientated.

Harlen and Crick (2003) report that, in general, high achieving students are less affected by high-stakes assessment than those achieving more moderately. However, it has marked effects on all students, with particularly devastating effects on those who receive low grades. For this group in particular, Harlen and Crick (2003) note that: high-stakes assessment increased assessment anxiety and placed pressure on students through the expectations of teachers, parents, and communities; assessment focused teaching impacted on preferred learning styles; and the use of repeated practice assessments impacted negatively on the confidence of the learner. Crooks (1988) also identifies test anxiety as impacting on task performance in high-stakes assessment.

While recognising that SBA has positive impacts on motivation, the literature examined on high-stakes SBA was critical of its effects on student motivation. Students with high self-efficacy were thought to be the most likely group to achieve when participating in high-stakes SBA. Students possessing low self-efficacy were considered to be the least motivated and the most likely to be at risk of not achieving (Harlen and Crick 2003).

## **18. Standards-based assessment raises student workload issues**

Teachers have expressed concern about students' workload under the NCEA (Hipkins et al 2004). Some students are choosing not to take part in some assessments as a way to manage their own workloads. This may prevent students from being able to take subjects at higher levels in the future. NZPPTA (2002) is not only concerned about students being stressed from constant assessment, but also about teachers being overworked by the amount of administration involved with the increase in internal assessment practices with the NCEA.

ERO (2004) found some concerns that impact on the workloads of students and teachers. The timetabling of assessment activities is the most commonly discussed issue that is related to the planning of units of work. Across the senior levels, schools are developing strategies regarding the timetabling of assessment activities to avoid congestion and to ensure a balanced workload for students. This mainly occurs through the use of wall charts or whiteboard calendars in the staff-room. A balance of assessment activities across the year still requires improvement in most schools, so

students are not continually stressed by intensive ongoing assessment and conflicting demands from different subject requirements.

### **19. Standards-based assessment can help meet student needs through programme flexibility**

SBA has the potential for flexibility of course design. Although various commentators have regretted the atomisation of standards (Davis 1996; Hearn 1997; Hall 1999; Boss et al 2001), others argue that integrating such discrete standards can be useful when planning programmes for diverse learners (Hager et al 1994; Hipkins et al 2004). ERO (2004) recommends learning programmes that provide a mix of standards and levels to meet the diverse range of student needs. This can occur when schools examine their programme structures and delivery.

Hipkins (2004) suggests that there are rich possibilities for reshaping school subjects because the NCEA has supplied the necessary assessment flexibility. However, she and Hipkins and Vaughan (2002) acknowledge that there must be a limit to flexibility. Hipkins found three types of course arrangements in the six schools investigated in the *Learning Curves* project to make her point about the flexibility of the NCEA. The first type of course was the 'traditional approach' to teaching and learning, whereby schools taught within specific disciplines. This was found to be the common approach. Such courses mainly assessed achievement standards, not unit standards. The second type of course identified was the least common found in schools and was 'locally redesigned'. Schools used a mixture of achievement standards and unit standards to assess student learning. The curriculum was usually organised around the assessment instruments used, but most courses allowed for some variation to include broader contexts for learning and reduce traditional curriculum content. The third type, the 'contextually focused' courses were drawn from vocational education. They are characterised by curriculum components being linked to students' everyday lives and offered a reduced number of credits, creating more flexibility and freedom for different types of learning experiences.

Writing in the UK, James and Gipps (1998) argue that assessment for the twenty-first century should emphasise higher-order skills and 'deep learning,' while not neglecting basic learning. On the basis of this, the kind of assessments used should include not only assessments of facts, but also the assessment of deeper understanding of concepts and principles and their application when applied into unfamiliar concepts. The balance of the evidence suggests that the NCEA can and does provide rich possibilities for the reshaping of school subjects into flexible learning programmes.

## **Theme five: impacts of standards-based assessment on diversity**

### **20. The impact of standard-based assessment on diverse students is under-researched**

In examining the impact of SBA on diverse learners, two issues become very apparent. First, there is a dearth of research in this area. Secondly, when there are references to diverse learners, the references are often of a global nature, thus making it difficult to determine how particular groups of learners (eg 'language minority' or special needs students) are affected by the use of SBA approaches. This issue is further complicated by the fact that students can simultaneously belong to more than

one category of diversity (eg language minority and low socio-economic) (National Council on Disability 1996, cited in Ortiz 2000).

In the USA Ortiz (2000) has criticised policy makers and educators for failing to recognise increasing diversity when they designed and implemented standards-based reforms. Black (2000) expressed similar concerns in his review of the proposed NCEA. He noted that no attention was given to problems of gender, cultural or socio-economic bias and therefore it was not possible to judge whether bias problems would be alleviated or exacerbated with the NCEA. However, while the combination of a lack of research and a lack of specificity have limited the conclusions which can be drawn, the literature does indicate areas of concern and areas for further investigation.

## **21. The impact of SBA systems on diverse students' academic performance is variable**

The academic achievements of diverse learners within SBA systems have been mixed. Ortiz (2000) has argued that, in spite of SBA reforms in the USA, there is still a significant gap between the achievement of students with special needs and their middle class majority peers. Culturally and linguistically diverse students and students with disabilities also experience widespread underachievement, high rates of being held back a grade and high drop out rates (Robertson et al 1994; National Centre for Educational Statistics 1995 both cited in Ortiz 2000). A study undertaken by Kannapel et al (2001) on the impact of standards-based reforms in Kentucky led the authors to conclude that, while the reforms were considered generally successful, they did not result in high achievement for all students, particularly minority students and those from low income households. Madaus and Clarke (2001) also found that achievement rates of Hispanic, Black and White students have remained the same over a thirty-year period, despite the introduction of high-stakes assessment (including SBA). They conclude that high-stakes, high standards tests do not have a markedly positive effect on learning and teaching.

However, a review of 24 competency testing programmes from 19 American states has led Marion and Scheinker (1999) to conclude that low performing students benefited if the entire system focused on bringing all students to a high standard, rather than just focusing on a few students demonstrating minimum competency. Several states now have longitudinal disaggregated data for students with disabilities, which provide consistent evidence of improved performance over time (Bielinski & Ysseldyke 2000; Trimble 1998; both cited in Thurlow 2000), but these students still perform below other groups of students (Thurlow, Nelson, Teelucksingh & Ysseldyke 2000, cited in Thurlow 2000).

To date, the findings suggest that diverse students perform better under SBA than under a norm-referenced system, but that as a group they are still performing below their peers. There are also indications that SBA can be used to assess diverse students' achievement more effectively when compared to traditional norm-referenced assessment. For example, Davidson (1992) argues a sound case for employing SBA in assessing English competence in language minority students, as it provides greater opportunity to measure a broader range of language skills when compared to norm-referenced assessment. The variability of findings indicates a need for further research and closer analysis of the data to provide a more accurate and comprehensive picture of how particular groups of diverse learners perform within different SBA systems.

It is also quite possible that simply setting up an SBA system is insufficient to improve the achievement levels of diverse learners (Madaus & Clarke 2001). Other factors, when combined with an SBA system, appear to positively influence student outcomes. Ortiz (2000) has identified high quality instruction, employment of specialised teachers and alignment of curriculum and standards as being important factors influencing academic outcomes for diverse students.

Another factor influential in improving academic outcomes is the use of formative assessment. In an extensive review of research on formative assessment, Black and Wiliam (1998) found that students with low scores on initial basic skills tests, who were involved in formative assessment, showed significant learning gains over their control group peers when retested. The gains were larger than those with medium and high scores on the initial test. Low achievers in another study made significant improvement when provided with a model for developing their self-assessment capabilities (Black & Wiliam 1998).

In a district-wide assessment programme in Colorado, where a minimum competency standards-based system was put into place to address the achievement gap between disadvantaged and affluent students, there was a narrowing of the achievement gap between ethnic and socio-economic groups and both genders (Waters et al 1995). Moreover, an increasing number of students moved to the next level of attainment. These improvements were attributable to both the implementation of SBA and extensive in-service professional development, which focused on pedagogy and classroom management (Waters et al 1995), rather than solely to SBA reforms. Further research is needed into the factors that are influential in improving outcomes for diverse learners and how these can be effectively employed within an SBA system.

## **22. The effects of high-stakes assessment and accountability on diverse students are marked in standards-based assessment systems**

It is evident that when used as high-stakes assessment, SBA can have significant impacts on groups of diverse students, both in the short and the long term (Harlen & Crick 2003). The effects are particularly devastating on those receiving low grades (Harlen & Crick 2003). Thurlow (2000) refers to a number of studies that indicate diverse students in the USA are often held back or drop out as a consequence of high-stakes assessment. This can result in heavy penalties for both the individual student and society. However, it must be added that this can occur whether the high-stakes assessments are standards-based or norm-referenced. School Certificate undoubtedly has had a similar impact in New Zealand. The question that now needs to be answered is whether the NCEA, as a high-stakes national assessment system, can provide more positive outcomes for diverse students than the previous system. Mallard (2004) has indicated that there are initial signs of improved achievement by Māori and Pasifika students. He also noted that three percent fewer students left New Zealand schools in 2003 with no qualifications, with improvements being particularly noticeable for Māori and Pasifika school leavers.

Another adverse reaction to the use of high-stakes assessment for accountability purposes has been the exclusion of groups of diverse students from formal assessment programmes by schools in the USA. Schools engaged in such practices to avoid being financially and publicly sanctioned if their students did poorly (Ortiz 2000; Thurlow

2000). This is in spite of being legally required to report the results. Without results, schools, districts and states are unable to satisfactorily track the progress of diverse students, and as a consequence these students could be missing out on additional support and funding (Thurlow 2000). While schools in New Zealand do not face quite the same sanctions, publication of 'league tables' can have an impact on public perceptions. It would be interesting to ascertain to what degree New Zealand secondary schools discourage students from participating in the NCEA, the background of these students and the factors that have influenced the school's decision.

Not surprisingly, fear of failure has also been found to be negatively linked to high-stakes assessment (Reay & Wiliam, 1999). In his extensive research synthesis, Crooks (1988) concluded that, 'Studies have repeatedly shown substantial negative correlations between measures of test anxiety collected before tests are administered and performance on those tests' (p 461). An interesting picture of the impact of high-stakes assessment on individual students has emerged from another study. Roderick and Engel (2001) found that the threat of being held back a grade did result in 62% of students at risk in a Chicago study working harder. It must be noted that the majority of these students did have support within the school to assist them to make progress and the students were willing to put in effort. A third of the students, while worrying, did not work harder. They also tended to lack support at school. While the fear of failure may act as a motivating force for numbers of students in the short term, it is possible that if experienced for extended periods of time (eg, two or three years), students may give up and drop out rather than rising to the challenge.

### **23. Alternative assessments need to be considered for diverse students**

SBA potentially provides schools with the opportunity to adapt assessment tasks to meet the needs of diverse learners, while still assessing the set standard(s). This is in contrast to norm-referenced assessment, which requires strict adherence to standardised assessment conditions if students' results are to be compared. In order that SBA is fully inclusive, there is a significant number of issues that need to be addressed, particularly surrounding accommodations (eg, more time, special aids) and alternative assessments. For example, Thurlow (2000) argues that it is important that the process is made clear for how decisions are made regarding: participation in regular assessment; accommodations that should be allowed; and acceptable alternative assessment options. She also argues for greater exploration and transparency with regard to the nature of allowable accommodations (eg, the way the test is administered, how the student responds, materials used) and investigation into the availability of alternate assessments and their alignment with general curriculum standards set for all students.

Professional development of teachers is also critical (Gipps 1994; Thurlow 2000). This needs to occur in relation to assessing students' needs and making appropriate adaptations to assessment tasks and assessment conditions. This is necessary, as concerns have been raised about teachers' lack of knowledge and the problem of over-accommodation (Elliott & Thurlow 2000, cited in Thurlow 2000). Greater accountability is needed from educators to ensure that diverse learners are being adequately catered for, both in regard to the provision of equitable assessment opportunities, and the provision of appropriate learning opportunities which allow

them to perform to the best of their ability on assessment tasks (Education Commission of the States 2002).

Professional development is also vital in changing the attitudes of teachers, who believe accommodations and alternative assessments provide some students with an unfair advantage (Thurlow 2000). It is important that the alternative assessments are well designed and moderated so they are viewed as comparable and of equal status to those commonly used, rather than watered-down versions (Gipps 1994). Furthermore, students with disabilities need to be educated to recognise and request the accommodations they need (Thurlow 2000). One could argue that parents should also be included in this process.

In addition to accommodations and alternative assessments, professional judgements are required as to when it is inappropriate to assess students with high-stakes SBA. For example, Bennett and Merrick (2004) have argued that English language learners should not be subjected to high-stakes assessment until they are sufficiently proficient in English. Protocols and policies are needed to ensure the decisions to exclude a student from a formal assessment are being made in the student's best interests and with the student's and parents' approval.

In order to satisfactorily address the issues raised above, research needs to be undertaken to determine what constitutes effective practice and how accommodations may affect the validity of the assessment outcomes. Once research findings are available, clear policies and guidelines will need to be developed to support schools, students and their parents in the process of meeting individual students' learning and assessment needs. Professional development and sound examples of accommodations and alternative assessments will also need to be provided for teachers (Linn & Herman 1997; Thurlow 2000).