

OFFICE OF LEARNING AND TEACHING

An Environmental



Scan of Tools and Strategies

that Measure Progress in School Reform

Department of Education and Training  
*An Environmental Scan of Tools and Strategies that Measure Progress in School Reform*  
Published by the Department of Education and Training  
© State of Victoria, 2005

All rights reserved. Except under the conditions described in the Copyright Act 1968 of Australia and subsequent amendments, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical and photocopying, recording or otherwise, without the prior permission of the copyright holder.

**Address inquiries to:**

Research and Development Branch  
Office of Learning and Teaching  
Department of Education and Training  
Level 2, 33 St Andrews Place  
GPO Box 4367  
Melbourne 3001

**Project Management**

Sandra Mahar, Research Manager, Office of Learning & Teaching

This report was commissioned by the Research & Development Branch of the Office of Learning and Teaching.

Professor Patrick Griffin of the Assessment Research Centre, University of Melbourne was appointed to produce the report and was supported by Kerry Woods and Nguyen Thi Kim Cuc.

# Scan of Tools and Strategies that Measure Progress in School Reform

## Background Information

The *Blueprint for Government Schools* sets out the reform agenda for the Victorian government school system. The *Blueprint* is a comprehensive reform approach which contains a number of individual strategies which address three priority areas for reform:

- recognising and responding to diverse student needs;
- building the skills of the education workforce to enhance the teaching-learning relationship; and
- continuously improving schools.

As in all systems, the Department of Education and Training is concerned with the assessing the impact of the reform both in an impact sense but also to inform the multi year implementation of the range of strategies contained within the *Blueprint*.

*An Environmental Scan of Tools and Strategies that Measure Progress in School Reform* was commissioned by the Office of Learning and Teaching to inform the development of an evaluation approach to the *Blueprint*, in particular to the Flagship Strategy 1 Student Learning.

The work was undertaken by Professor Patrick Griffin, Doctor Nguyen Thi Kim Cuc and Kerry Woods of the Assessment Research Centre, University of Melbourne. The key research questions for this project were:

1. What methodologies are being used to measure progress in school reform, particularly in the areas of:
  - student learning outcomes in physical, personal and social learning, discipline-based learning and interdisciplinary learning;
  - student learning which focuses on the attainment of deep understanding;
  - teacher practice in relation to pedagogy, assessment (both formative and summative) and reporting; and
  - school organisation and culture.
2. How effective are particular methodologies in evaluating progression school reform and what are the issues?

The report was managed by the Research and Development Branch within the Office of Learning and Teaching.

**Dahle Suggett**

Deputy Secretary

Office of Learning and Teaching

# **An Environmental Scan of Tools and Strategies that Measure Progress in School Reform**

## **Executive Summary**

### **An Overview of Current Tools and Methods**

This report describes and evaluates current methods for measuring the progress and impact of school reform initiatives. The first section of the report reviews strategies at local and systemic levels. It details an assessment of methods of measuring progress in the United Kingdom, Canada, the United States, Japan, Hong Kong and New Zealand.

Two themes emerged from the review. The first is the challenge of aligning the intended focus of school reform strategies and measures of reform outcomes. The second relates to the pervasive problems of accountability and methods for tracking progress.

### ***Measuring outcomes***

Most systems attempt to quantify progress by measuring student outcomes. This often means measurement of student performance on tests of literacy and numeracy, perhaps with science and social studies included in the mix. This approach is widespread across nations and systems. Canada currently has plans to broaden the focus of its testing program, and New Zealand provides an example of a national assessment strategy that relies on multiple measures covering all facets of the national curriculum.

There are many goals of school reform, such as improvements in teacher practice and school administration, community engagement, broadening the curriculum and providing flexible pathways for students, that are not easily measured in terms of student progress on a limited set of tests, or to short-term measures of progress. Some of these areas are currently examined, in the United Kingdom, Hong Kong and New Zealand for example, through inspections or reviews of schools.

### ***Tracking progress***

Controversy surrounds current attempts to monitor progress in school reform by tracking change in student academic achievement. The problems surrounding the use of “value added analyses” in the United Kingdom and United States are discussed in the report. In brief, these problems include:

- doubts about test validity;
- methodological concerns over the handling of missing data;
- the problem of whether or not to include demographic data;
- the fairness of various models, especially where high stakes are attached to outcomes of analyses;
- lack of transparency of analyses for the majority of stakeholders;
- inability to distinguish between true learning and ‘teaching to the test’; and
- inability to indicate the specific practices or reforms to which improvements are attributable, or to support causal interpretations.

## **Guidelines for Evaluation**

The next two sections of the report provide a review of the current trend for “evidence-based” reform in education and use this as a platform to discuss the importance of rigor in the design of evaluation studies, sampling and data collection procedures, analysis and interpretation of evidence.

Section Four of the report then takes the conditions and cautions discussed in previous sections, and applies them to setting out a positive case for a comprehensive program of evaluation. While the previous sections focus on what *not* to do, this section provides detailed guidelines for evaluation that can be applied to a range of reform programs. The advice provided in this section is generic rather than specific, and applicable to the wide range of reform strategies encompassed within the Victorian Department of Education and Training’s *Blueprint for Government Schools*.

Section Five specifically discusses the evaluation of the implementation, processes and outcomes of the *Victorian Essential Learning Standards (VELS)*. The section argues the case for the development of an assessment framework based on the approaches of Rasch, Glaser and Vygotsky. Section Six of the report then provides three case studies where measurement and monitoring of student outcomes have been incorporated into program evaluation.

The report concludes with a statement of the requirements for comprehensive evaluation of the progress of reform strategies. Key issues raised are the imperatives of:

- using professional judgment to identify critical success factors for the reform initiatives, and expected levels of performance for those factors;
- testing professional judgment against empirical data;
- gathering empirical data in a manner that is clear and defensible (i.e., the study design, sampling methods, instruments, data collection, handling and analysis procedures are sound, and claims based on the data analysis can be clearly supported);
- monitoring sustainable or long term trends over a sufficient time period;
- evaluating reforms in terms of outcomes, processes, and impacts;
- documenting impeding and supporting factors, and the unintended outcomes of reform; and
- combining evidence from multiple sources and using these to link reforms to indicators of progress.

In summary, the report argues that evaluation of progress in school reform requires the identification of appropriate indicators of successful implementation, including but not restricted to student academic outcomes. The indicators can be used to comment and reflect upon relationships between styles of implementation and outcomes for students and for sub-groups of students. They provide perspectives from multiple vantage points of students, parents and teaching professionals.

## Table of Contents

<b>Introduction</b>	<b>7</b>
<b>Section One: An Overview of Tools and Methods</b>	<b>9</b>
Major Reform Movements	10
United Kingdom	10
Background	10
Student testing: The Qualifications and Curriculum Authority (QCA)	11
Value added analyses	13
Setting targets	15
School inspection (OFSTED)	16
United States	16
Background	16
The No Child Left Behind Act: Standards-based reform	17
National Assessment of Educational Progress (NAEP)	19
The annual state report card	19
State diversity in accountability measures	20
Maryland: A case study and comparison with Victoria	21
Test alignment	24
Measures of progress: Absolute targets and value added analyses	24
The Tennessee Value Added Assessment System	26
Whole school reform models	27
Summary	28
The Wider Influence of the United States and United Kingdom Reforms	28
New Zealand	28
Education Review Office	29
National Education Monitoring Project (NEMP)	30
Assessment Tools for Teaching and Learning (asTTle)	30
Canada	31
National monitoring of student outcomes (SAIP, PCAP, PISA)	32
Evaluation of the Manitoba School Improvement Program: A case study	33
European Reforms	34
School Reform in Asia	35
Japan	35
School self-evaluation	36

Monitoring student achievement	36
Hong Kong	36
Surveys of teaching professionals	37
Quality Assurance Framework	37
OECD Interests in Education Reform	38
Summary of the Review of Current Methodologies and Procedures for Measuring Progress in School Reform	39
Measuring outcomes	39
Tracking progress	40
Conclusion	40
<b>Section Two: Tools and Methods for Evaluating Reform Initiatives</b>	<b>41</b>
Overview	41
The Challenge of Monitoring Progress in Reform	42
<b>Section Three: Evidence-based Reform</b>	<b>44</b>
Evidence-based Education	44
<b>Section Four: Guidelines for Evaluation</b>	<b>47</b>
Five components of an evaluation study	48
1. Review program and assess needs	48
2. Plan the project and identify resource needs	48
3. Monitor the implementation of the reform strategy	49
4. Evaluate the impact of the reform	49
5. Report the cycle of evaluation and development	49
Component 1. Review program and assess needs	50
1.1: What is reviewed?	50
1.2: Who should be involved in the review?	50
1.3: How can the review be carried out?	51
1.4: When should the review be carried out?	51
1.5: What should be reported about the review?	51
1.6: What is the outcome of the review of the reform strategy?	52
Component 2: Planning implementation of the reform strategy	52
2.1: What should be the focus of planning?	52
2.2: Who is involved in planning?	53
2.3: How should the planning be undertaken?	53
2.4: When should planning occur?	53
2.5: What should be reported about planning?	53

2.6: What should be the outcome of the project planning?	53
Component 3: Monitor the implementation of the reform strategy	54
Monitoring	54
Component Evaluation	54
Problem Study	54
Benefit Survey	54
Status Survey	54
Levels of Implementation	55
3.1: What should be monitored in the implementation of the reform strategy?	55
3.2: Who should be involved in monitoring implementation?	55
3.3: How should the implementation be monitored?	56
3.4: When is the implementation monitored?	56
3.5: What should be reported as a result of monitoring the implementation?	56
3.6: What should be the outcome of monitoring implementation?	57
Key questions for evaluation of the implementation stage of a reform strategy	57
Component 4: Impact evaluation	58
4.1: What is the focus of the impact evaluation?	58
4.2: Who should be involved in the impact evaluation?	59
4.3: How can impact be evaluated?	59
4.4: When should the impact be evaluated?	59
4.5: What should be reported about the impact evaluation?	60
4.6: What should be the outcome of the impact evaluation?	60
Component 5: Reporting the cycle of evaluation	60
5.1: What is the purpose of reporting?	60
5.2: Who should be involved in reporting the cycle?	61
5.3/4: How and when should the cycle be reported?	61
5.5: What should be reported about the cycle?	61
5.6: What is the outcome of reporting?	61
Summary	62
<b>Section Five: The Victorian Essential Learning Standards (VELS)</b>	<b>62</b>
Student learning outcomes as an indicator of progress	63
Theoretical background: The OECD DeSeCo project	64
Glaser, Rasch and Vygotsky: A synthesis of approaches to develop a framework for assessment of student outcomes (Griffin, 2004)	65
Assessing student outcomes against VELS	68
Monitoring change in student outcomes and other indicators of success	68

<b>Section Six: Case Studies</b>	<b>69</b>
The National Education Monitoring Project (New Zealand)	69
The Maryland School Performance Assessment Program (MSPAP)	70
Evaluation of the Hong Kong Primary Native English-speaking Teacher (PNET) scheme	72
Profiles in English as a Second Language	73
Using student achievement data to inform a program of evaluation.	74
<b>Conclusion</b>	<b>75</b>
<b>References</b>	<b>77</b>
<b>Glossary of Terms</b>	<b>82</b>

# **A Review of Current Methodologies and Procedures to Measure Progress in School Reform**

## **Introduction**

The purpose of this report is to describe and evaluate current methodologies and procedures for measuring the progress and impact of school reform initiatives. Section One of the report provides an overview of strategies at local and systemic levels with an emphasis on standards based and multi method approaches. It details an assessment of methods of measuring progress in school reform in the United Kingdom, Canada, the United States, Japan, Hong Kong and New Zealand.

Section Two of the report presents a discussion of evaluation studies and the problems of monitoring progress, with particular emphasis on the type of data collected and the use of value-added analyses. This picks up from the previous section of the report, and summarises the challenges of these types of evaluation studies.

Section Three discusses the current trend for "evidence-based" reform in education, and uses this as a platform to discuss the need for rigor in the design of evaluation studies, sampling and data collection procedures, analysis and interpretation of evidence.

Section Four takes the conditions and cautions discussed in the previous sections, and applies them to setting out a positive case for a comprehensive program of monitoring and evaluation. While the previous sections focus on what *not* to do, this section provides detailed guidelines for evaluation that can be applied to a range of reform programs. The advice provided in this section is generic rather than specific, and applicable to the wide range of reform strategies encompassed within the Victorian Department of Education and Training's *Blueprint for Government Schools*.

Section Five builds on information set out in the previous sections to consider the evaluation of the implementation, processes and outcomes of the *Blueprint for Government Schools, Flagship Strategy One: Victorian Essential Learning Standards (VELS)*. The section presents a discussion of the challenges of using student learning outcomes as indicators of progress, argues the case for the development of a framework for assessment based on the approaches of Rasch, Glaser and Vygotsky, and then links back to the difficulties of monitoring change in student outcomes and attempting to associate change with specific reform initiatives.

Section Six provides three case studies where measurement and monitoring of student outcomes have been successfully incorporated into program evaluation. The first and second of these examples describe large-scale assessment programs (currently in New Zealand, and prior to 2003 in Maryland) that have not been confined to standardized testing across a limited range of basic skills. The third example describes a longitudinal study of a system-

wide school improvement program in Hong Kong that is monitoring change over time, and linking different stages of reform implementation to a range of indicators of success that extend beyond simple description of change in terms of student outcomes.

The report concludes with a statement of the requirements for comprehensive evaluation of the progress of reform strategies. Key issues raised are the imperatives of:

- using professional judgment to identify critical success factors (key indicators) for the reform initiatives, and expected levels of performance for those factors;
- testing professional judgment against empirical data; and
- gathering empirical data in a manner that is clear and defensible (i.e., the study design, sampling methods, instruments, data collection, handling and analysis procedures are sound, and claims based on the data analysis can be clearly supported).

A case is made for:

- the importance of monitoring sustainable or long term trends over a sufficient time period;
- evaluating reforms in terms of outcomes, processes, and impacts;
- documenting impeding and supporting factors
- recording the unintended outcomes of reform;
- combining evidence from multiple sources; and
- using these to link reforms to indicators of progress.

In summary, evaluation of progress in school reform requires the identification of appropriate indicators of successful implementation, including but not restricted to student academic outcomes. The indicators can be used to comment and reflect upon relationships between styles of implementation and outcomes for students and for sub-groups of students. They provide perspectives from multiple vantage points of students, parents and teaching professionals.

## **Section One: An Overview of Tools and Methods for Measuring Progress in School Reform**

Current international trends in thinking about education have centred upon the imperative of improving the quality of education in schools. During the 1970s and 80s, this debate placed most emphasis on curriculum development and small-scale reform initiatives aimed at helping individual schools to change (Elmore, 1996). The rhetoric of the time suggested that schools were the most important units of change, and identified the key characteristics of 'effective schools' as professional leadership, shared goals, professional development for teachers, high expectations of student learning, monitoring of student progress, and encouragement of parent involvement (Earl, Torrance, Sutherland, Fullan & Ali, 2003; Edmonds, 1979; Reynolds & Packer, 1992). These goals and aspirations for schools have not altered. However, as Earl et al. pointed out, overall results of a range of activities directed at helping schools in the process of change proved to be disappointing. In particular, it seemed difficult to find evidence of robust and sustainable improvement based on curriculum-defined reform strategies in individual schools (Elmore, 1996; Fuhrman, 2001; Teddlie & Reynolds, 2000). Small-scale reform initiatives were criticised as sporadic, piecemeal and vulnerable to changes of key school personnel (Earl et al.).

As a result, school improvement strategies in many countries shifted from an emphasis on local reform at the level of the school or classroom to larger scale reform efforts (Earl et al., 2003). Over the past ten years, these have typically included reform initiatives taken at the national or state level, as well as commercial models of whole-school reform. In the United States, for example, the government has funded implementation of externally developed whole-school reform programs and, unsurprisingly, availability of funding has coincided with a proliferation of such programs. These have taken the form of external agencies developing and validating school improvement strategies and providing continuing professional development and support to school staff.<sup>1</sup> At national or state levels, reform efforts have typically included provision of funding for designated purposes (such as support for the education of under-privileged minorities by the U.S. government), and legislated changes to and centralization of curriculum, as well as introduction of national standards for student learning outcomes and centralized assessment and accountability programs (Fuhrman, 2001; Whitty, Powers & Halpin, 1998).

The trend towards establishment of national standards and national testing of students as a primary method of evaluating progress in reform is commonly described as a *standards-based reform* strategy. However, while setting national, state or even district standards, and monitoring and reporting proportions of students who achieve those standards, is a widely-used methodology for measurement of progress, few systems or governments rely on any one form of measurement procedure or data upon which to base their assessments. This reflects conservatism about the reliability and validity of current methods for assessing student outcomes and, in particular, for tracking progress<sup>2</sup>. Rather, most gather performance data from a range of sources and via a range of methods. Also, while increased centralisation of control of standards for student achievement has been adopted by some systems, others have

adopted reforms designed to devolve more, rather than less, responsibility for school performance onto individual schools.

School reform strategies, and the methodologies used to evaluate them, cannot be simplistically isolated from their particular social and political contexts. For example, in England under the Blair government public debate about educational reform has included a strongly-voiced concern about equity (Hopkins, 2005) and, in terms of assessment methodologies, this has been translated as one of the primary indicators against which reform outcomes are evaluated (Barber, 2002). In New Zealand, dissatisfaction of Maori and Pacific Islander communities with the ability of schools to meet the aspirations of their students has shifted the focus of school reform to include strategies built upon increased community involvement and recognition of diverse needs (Fancy, 2004). The following scan of current methods for assessing school reform strategies covers a range of procedures used in countries chosen for inclusion because of their demographic similarity to the Australian educational context. Other reasons for inclusion in the analysis were consistently high student outcomes demonstrated in international evaluations of student achievement such as the OECD's Program of International Student Assessment (PISA) studies, and the constraints of language on accessibility of sufficient information upon which to base the report.

### *Major Reform Movements*

#### *United Kingdom*

##### *Background*

A national inquiry into school management in the United Kingdom produced the Taylor Report (*A New Partnership for Our Schools*), in 1977, which led to the 1980 Education Act. This marked the beginning of a decade of wide-ranging reform during which new certification procedures and new assessment methods were introduced and a national curriculum with key learning areas was established. Of particular importance, national testing of school performance was proclaimed, with an emphasis on mathematics and science. In general, the government proposed reforms designed to centralize control over education policy. It introduced a new category of 'grant-maintained schools,' allowing schools to draw their operating funds directly from central government, and moved to distance post-secondary education from local authorities, to reform teacher education and to privatize school inspections. A 1984 Green Paper setting out the policy options for a consumer-led education system was particularly controversial, and it was followed by an equally controversial White Paper entitled *Better Schools* (March 1985), which foreshadowed the 1986 Education Act. A key phase in school reform followed, with passage of the Education Reform Act of 1988. This imposed a centrally mandated national curriculum on all local school districts.

The attention to a common curriculum was significant, as it marked a rebuttal of progressive education, of curricula tailored to the needs of individual students, and of approaches emphasizing process rather than standards and content. From *The National Curriculum 5-16: A Consultation Document (1987)*, ‘foundation subjects’ were developed for the compulsory years of schooling, with ‘attainment targets’ set for each ‘Key Stage,’ and assessment at ages 7, 11, and 14, to coincide with the Key Stages, and 16 where the General Certificate of Secondary Education (GCSE) operated. The government developed a Parents' Charter, which promised parents would have information about their children’s progress. By the start of this century, changes had been adopted by the Labor Party which attempted to bridge the gap between the teacher unions and the government by giving more independence and authority to schools but demanding more accountability. There was an acknowledged lack of evidence of the impact of school reform.

In 2002, Michael Barber argued that, prior to the establishment in England of a national curriculum and standards, and the associated national testing of all students at ages 7, 11, 14 and 16, combined with independent inspection of schools, government attempts to reform the education system were hampered by a dearth of reliable evidence. In his role as Head of the Prime Minister’s Delivery Unit since 2001 and his former position as Head of the Standards and Effectiveness Unit at the Department of Education and Employment, Barber has been a consistent advocate of standards-based reform and the use of indicators of success in reform strategies such as:

- increase in the proportion of students who meet standards set out in the National Curriculum;
- decrease in the proportion of students who leave school with no qualification; and
- improvement in results from the OECD PISA studies.

The National Curriculum sets standards of achievement for all students in state schools in England, Wales and Northern Ireland. It prescribes the subjects that must be taught and also the stages of learning (Key Stages) through which students are expected to progress. Standards of student achievement in each of the curriculum-mandated subjects are directly related to student age so that, for example, level 2 of Key Stage 1 is described as the average standard that should challenge a 7 year old student. Similarly, level 4 of Key Stage 2 is described as the average standard that should challenge an 11 year old student. The outcomes of national testing of students at each of the relevant ages is reported in terms of the proportion of students who achieved the expected levels in the core curriculum subjects (i.e., reading, writing, mathematics). The development of national tests of student achievement is the responsibility of the Qualifications and Curriculum Authority.

*Student testing: The Qualifications and Curriculum Authority (QCA)*

The Qualifications and Curriculum Authority specifies continuity of national tests of student achievement from year to year in terms of test length, coverage of a programme of study,

characteristics of test questions, marking scheme, balance of type of questions and the characteristics of students to be tested (i.e., the extent and form of inclusion of students whose home language is not English and students with special needs). Test questions are piloted with two samples of students to check the wording of test items, time taken to complete the tests, and teachers' and markers' views of the tests. Test items are also analysed in terms of item difficulty and student response patterns. Questions are then selected for a second round of pre-testing based on the tests' ability to

- cover a range of skills;
- provide a range of difficulty levels to discriminate appropriately between students;
- provide a range of response types including extended responses and multiple choice questions; and
- produce a distribution of student performance with a mean score approximately equal to the observed mean score from previous years.

The purpose of the second round of pre-testing is to monitor the comparability of tests from year to year. A sample of students takes both the current year's test and the test planned for the following year, to permit equating of scores. For some subjects, students also take an anchor test, which remains constant over years, so that tests can be linked. The pre-test is also used to monitor performance of students whose home language is not English.

The tests are marked by classroom teachers, and the accuracy and consistency of test marking is monitored by local education authorities. Each school must subject their marking procedures to a full audit every four years.

Teachers are also involved in the setting of level thresholds for standards of achievement using an Angoff (1971) procedure. Explicitly, according to the QCA, this means that classroom teachers are asked to consider a child working at a specified level and to work through every question on the test, indicating the probability that the child would be successful in answering the question. The set of teacher judgments is used to indicate the score a child would need to achieve on the test to be considered as working just within the target level. The procedure is repeated for all levels covered by the tests. QCA uses statistical data from the pre-test and the teacher judgment data to establish draft thresholds for standards.

Level thresholds for achievement standards are confirmed following administration of the tests. As part of this process, external markers compare a sample of scripts from the most recent round of testing with scripts sampled from each level for the previous three years of testing. A meeting of stakeholders, including the test developers, chief markers, and representatives of teacher associations and academic institutions, re-examines and confirms each of the level thresholds. As a final stage, tests are subjected to independent reviews via desk analyses of test questions, validity and design, and detailed item response analyses.

These are circulated to teachers, with advice about the changing performance of students over time and implications for teaching practice.

### *Value added analyses*

Since September 2002, the publication of school-by-school test results has included a 'value added' analysis. These measures are intended to permit comparisons, between schools with different student intakes, that are fairer than those provided by comparisons based on percentages of students who attain a particular standard of achievement. They are also expected to provide information to track student progress.

Traditionally, the label of 'value added analysis' has been attached to a range of statistical procedures that attempt to compare students (and, when student results are aggregated, schools, teachers or systems) against their own performance measured longitudinally or against the average or predicted progress made by a specified reference group. However, these procedures are subject to methodological constraints which can be a trap for the unwary. They are also the source of considerable controversy when attached to high-stakes outcomes such as funding or staffing decisions or the public dissemination of comparisons of schools<sup>3</sup>. Indeed, it could be argued that some of the current enthusiasm for value added analysis results from an overly optimistic interpretation of the name given to the group of statistical procedures that support the analysis. Value added analysis is, in its simplest form, a comparison of residual scores, or aggregated residual scores, in a regression analysis. They are not, as the name would suggest, a clear and unequivocal indication of the 'value' added to baseline student ability by schools, teachers, or education reform strategies. Rather, in the United Kingdom the analysis is an indication of how much a student's performance on a specified battery of tests differed from predicted performance when a group of predictive factors were entered into the regression equation and, more recently, a 'shrinkage' factor was applied to compensate for the impact of variable sample sizes on the analysis (Evans, 2005; Fischer Family Trust, 2004).

In the United Kingdom, value added analyses are used as evaluations of school performance as much as indications of the progress made by individual students. For example, schools in the lowest quartile in the value added analyses were recently categorised by David Hopkins, Chief Adviser on School Standards to the Department for Education and Skills (DfES), as 'under-performing' schools (Hopkins, 2005). The publication of school standing in terms of value added outcomes through the Achievement and Attainment Tables (formerly referred to as Performance Tables) has a derogatory impact on schools. The extent to which this is warranted should be cautiously weighed against the validity of measures of student outcomes and the sophistication, or lack thereof, of the statistical model used to derive judgments.

The UK Department for Education and Skills (DfES, 2004) described the value added measures they use in terms, for example, of comparisons of each student's Key Stage 2 (KS2, aged 11 years) performance in the national assessments with the median performance of other students with the same or similar results as that student at Key Stage 1 (KS1, aged 7

years). A similar comparison is also made when students are aged 14 years. Individual student scores are then averaged for a school to give a score that is standardized with a mean of 100. A current example of value added percentiles is shown in Table 1.

Table 1

*Value Added Percentile Scores (DfES, 2004)*

Value Added Score	Percentile Ranking
102.1 and over	Top 5% of schools nationally
100.9-102.0	Next 20% of schools nationally
100.3-100.8	Next 15% of schools nationally
99.8-100.2	Middle 20% of schools nationally
99.3- 99.7	Next 15% of schools nationally
98.0- 99.2	Next 20% of schools nationally
97.9 and below	Lowest 5% of schools nationally

Thus, value added scores above 100 represent schools where students on average are considered to have made more progress than similar students nationally. The DfES asserted that, for the KS1 to KS2 value added analysis, a score of 101 indicated that on average each of a school's students made one term's more progress between KS1 and KS2 than the median value for students with similar KS1 attainment (DfES, 2004). This measure is, however, particularly vulnerable to distortion based on the number of students included in the calculation. For example, DfES cautioned that, for schools with only 30 students in the value added analysis, measures of 99.1 to 100.9 represent broadly average performance. By comparison, for schools with more than 50 students included in the analysis, measures of 99.3 to 100.7 are described as broadly average. Perhaps to address some of these problems, the most recent round of reports on school performance were based on tests of statistical significance of results (using 95% confidence limits and correcting for number of students in the sample) in addition to percentile ranking (Fischer Family Trust, 2004).

It should also be noted that, under this system, the performance of individual students is not being tracked longitudinally. Rather, students are being compared against the average performance of a group of students who passed through the system earlier. Changes in curriculum emphases in the intervening periods are not factored into the equation, unless they can be accounted for in the design of test questions. Similarly, performance against standards of achievement is being monitored only in relation to tests of core skills in reading, writing, mathematics and science.

The fairness of the value added measure has come under considerable scrutiny, leading to the piloting by DfES of a value added measure that also takes a complex range of contextual variables into consideration when attempting to calculate the difference between a student's actual and predicted test score. The variables to be added into the proposed *contextual value*

*added analysis* include student characteristics such as mobility, gender, ethnicity and age, and school characteristics such as level of deprivation of the school community. Student scores are averaged to provide a score for the school, and a 'shrinkage factor' is then used to correct for the impact of sample size.

The factors included in the new regression model are:

- student's prior attainment taken from the student's average point score in English, mathematics and science on earlier tests;
- school prior attainment taken from the average point score of all students on earlier tests;
- student mobility (dichotomously categorised as whether or not the student had joined the school prior to the start of the school year);
- school community deprivation (currently defined as the proportion of students who qualify for free school meals); and
- student characteristics such as gender, ethnicity and age (Evans, 2005).

This new measure was piloted by DfES with a sample of schools in 2004, but outcomes of the pilot have not yet been published. The statistical analysts, however, have indicated continued concern over the usefulness of measures such as access to free school meals to indicate 'social deprivation' (Fischer Family Trust, 2004). In addition, other variables such as geographic location of schools seem to influence the results of analyses, but are not currently factored into the regression equation, nor are the relationships between ethnicity and home language background taken into consideration. As a conservative statement, the contextual value added measure of schools in the United Kingdom is currently a 'work in progress'.

### *Setting targets*

The Department for Education and Skills sets performance targets against which to measure the success of its strategies for school improvement. For example, for the period 2002-2007, the national performance targets included a statement of the percentages of students expected to achieve the required levels of performance in English, mathematics, ICT and science, and commitment to reducing the number of schools where fewer than 60% of 14 year old students achieved the expected levels in English and mathematics, and to increasing the percentage of students who spend at least two hours each week on high quality physical education and school sport (DfES, 2002).

Other important indicators of successful reform, as defined by Barber (2002), included reductions in truancy and exclusion rates, a narrowing of the achievement gap between pupils at age 16, and improved outcomes of school inspections conducted through the Office for Standards in Education (OFSTED). Indeed, reduction of the number of schools assessed by

OFSTED as ‘failing’ and reduction in the amount of time taken to turn ‘failing’ schools around were cited as key indicators of reform success (Barber; Hopkins, 2005).

### *School inspection (OFSTED)*

OFSTED is a non-ministerial government department responsible for the inspection of all schools (both state and independent) in England, and its authority extends to inspection of local education authorities, teacher training institutions and colleges, and regulation of childcare facilities. School inspections are a process of gathering evidence on the performance of a school, and rely upon observation of lessons, interviews with teachers, analysis of pupil work, questionnaire responses by parents, and meetings with parents, pupils and school governors. The broad research questions posed by the team of inspectors are:

- What does the school do well?
- How high are the standards?
- How well are pupils taught?
- How well is the school led and managed?
- How much has the school improved since the last inspection?
- What should the school do to improve further?

The process of school inspection also includes provision of both oral and written reports to the school principal, teaching staff and the school’s governing body, and a summary report to parents. The school is then required to provide a response detailing how it intends to implement the inspectors’ recommendations.

## *United States*

### *Background*

In the 1980s, the reform movement in the United States was closely associated with the effective schools movement (Clark & Astuto, 1986). In its early phases, emphasis was placed upon issues of equity, but this shifted to a debate about the provision of ‘excellence in schooling’. Much of the research supporting reform initiatives was based on studies of urban rather than rural schools, large rather than small schools, elementary rather than secondary schools, and those enrolling children from low socioeconomic status rather than from middle- or upper-class backgrounds. The indicators of performance were basic skills tests, not higher-order thinking or creativity scales. The reforms of the early 1980s reasserted the primacy of schools, of traditional forms of schooling, of the conventional curriculum, and of basic literacy and numeracy skills.

In 1983, release of the report *A Nation at Risk* (United States National Commission on Excellence in Education, 1983) cautioned that United States schools were failing, that academic standards were falling, and that school graduates compared unfavorably with their peers in trading-partner nations. This led to political intervention aimed at raising the quality of education and of the nation's workforce. Many states raised graduation requirements, revised student testing, provided academic enrichment programs, upgraded teacher certification, and introduced professional development programs for teachers and school administrators. Many also adopted reforms to set more homework, to write demanding textbooks, and to lengthen the school day and year. By the early 1990s, a new wave of reform strategies had emerged, based on the metaphor of 'markets' and private business analogies suggested by Chubb and Moe's (1990) *Politics, Markets, and America's Schools*. Features of the proposed reforms were deregulated and self-governed schools, operating autonomously around parental and student choice, driven by competition, and guided in curriculum by customer preferences.

### *The No Child Left Behind Act: Standards-based reform*

The concept of standards-based reform, or reform anchored on policy level statements of expected student achievement, emerged in the United States in the 1980s, fuelled by the publication of *A Nation at Risk* and the low scores of American students on international achievement tests (DeCoker, 2002; Fuhrman, 2001). It is currently the dominant ideology of educational improvement across American states. The establishment of 'challenging academic content standards and challenging student academic achievement standards' is federally mandated through the No Child Left Behind (NCLB) Act of 2001 (1111.b.1.A). The Act requires that each state submit a plan to demonstrate that challenging academic standards will:

- apply to all children and all schools (1111.b.1.B);
- encompass mathematics, reading or language arts and, from 2005-2006, science (1111.b.1.C); and
- encourage the teaching of advanced skills (1111.b.1.D.i.III).

Further, assessment of student academic achievement must be aligned with the state's academic standards (1111.b.1.D.ii). States are also required to implement an accountability system to monitor 'adequate yearly progress' in all schools. Accountability systems must:

- be based on the state's academic standards and assessments (1111.b.2.A.i);
- include sanctions and rewards for schools (1111.b.2.A.iii); and
- demonstrate adequate yearly student progress, while also closing the achievement gap between high and low achieving groups (1111.b.2.B).

Adequate yearly progress is defined in the Act as a measure based primarily on state academic assessments, but also including graduation rates for secondary students and at least one other academic indicator (1111.b.2.C). These may include achievement on additional state or locally administered assessments, grade-to-grade retention rates, attendance rates, and percentages of students completing gifted and talented placements and college preparatory courses.

As part of the monitoring of progress, each state is required to establish a starting point for measurement of the baseline percentage of students meeting or exceeding the state's minimum standard of proficiency, which must be the higher of two standards. These are the standard achieved by either the state's lowest achieving group of students from economically disadvantaged backgrounds, major racial and ethnic groups, students with disabilities, or students with limited English proficiency, or the school at the 20<sup>th</sup> percentile in the state (1111.b.2.E). States must also establish a timeline for adequate yearly progress to demonstrate that, within a twelve-year period, all students will meet or exceed state proficiency standards (1111.b.2.F).

Explicitly, the requirements for assessment of student academic achievement are that they:

- align with the state's standards for academic achievement;
- from 2005-2006, measure student achievement in mathematics and reading in each of grades 3 – 8, and at least once in grades 10 – 12;
- from 2007-2008, measure student achievement in science at least once during grades 3-5, 6-9 and 10-12;
- involve multiple up-to-date measures of student achievement, including the measurement of higher order thinking skills and understanding;
- produce individual student descriptive and diagnostic reports; and
- enable results to be disaggregated within states and schools by gender, major ethnic groups, English proficiency, migrant status, level of economic disadvantage and disability status (1111.b.3.C).

From 2003, states were also required to annually assess the English proficiency (in oral language, reading and writing) of all students from other language backgrounds, and to participate in the biennial assessments of grades 4 and 8 reading and mathematics conducted by the National Assessment of Educational Progress (NAEP) (1111.c.2).

Currently, student achievement is monitored through NAEP at both national and state level. Thus, individual state jurisdictions evaluate the *outcomes* (in terms of student test scores) rather than the *process* of education reform in schools (Goertz, 2001). Standards establish expectations and, ideally, accountability procedures based on measures of student performance are expected to provide feedback to schools and to systems to support the achievement of those standards (Goertz).

### *National Assessment of Educational Progress (NAEP)*

National and periodic testing of American students in reading, mathematics, science, writing, civics, geography, the arts and U.S. history has been conducted since the 1960s by the National Assessment of Educational Progress (NAEP), which is also referred to as the 'Nation's Report Card'.<sup>4</sup> The testing framework and specifications are set by the National Assessment Governing Board (NAGB), a group of governors, state legislators, local and state education officials, representatives of the business community and members of the general public.

NAEP samples students at grades 4, 8 and 12, from public and non-public schools, and across geographic regions. It does not provide scores for individual students or schools, but rather reports on subject-matter achievement for sampled populations and relevant sub-groups of those populations. The assessments include constructed-response questions and some performance tasks.

However, the nationally mandated NAEP assessments specifically designed to monitor change over time in student achievement are restricted to the domains of mathematics and reading. NAEP support this restriction by asserting that precise measurement of trends in student achievement is dependent on replication of past procedures, which would obviate changes in test design based on the evolution of curriculum or educational practice. These tests are administered nationally to students at the ages of 9, 13 and 17 years. NAEP provides state-level results to participating states.

### *The annual state report card*

Section 1111 of the No Child Left Behind Act requires that states submit an annual report card that documents:

- aggregated information on students at each proficiency level on state academic assessments;
- disaggregated information on proficiency levels for minority groups where statistically viable;
- comparisons of actual student proficiency with state targets for progress;
- indication of the percentage of students excluded from the analysis, disaggregated against minority groups;
- two-year trends in student achievement for each subject area and grade level tested;
- graduation rates for secondary students; and
- information about teachers including professional qualifications, percentage of classes not taught by highly qualified teachers, and a comparison of classes not taught

by highly qualified teachers disaggregated across high-poverty and low-poverty schools (1111.h.1.C).

States may also include other information in the annual report card to demonstrate progress, such as school attendance rates, average class size, gains in English proficiency of students from other language backgrounds, percentage of students passing advanced placement courses, and incidence of school violence, substance abuse, student suspensions, student expulsions, and parental involvement in schools (1111.h.1.D).

### *State diversity in accountability measures*

While adoption of some form of standards-based accountability is pervasive across states, there is considerable diversity between states in terms of who is held accountable for meeting standards of student achievement, how performance against expectations is gauged, and the consequences that are attached to not meeting expected standards (Goertz, 2001). In the 1990s, standards-based reform models held schools and/or teachers accountable for student outcomes, but a perceived lack of incentive for students to take the tests seriously prompted several states and school districts to enact ‘promotion gates’: students could not progress to the next level of schooling if they did not meet district or state performance standards (Goertz). The introduction of high-stakes assessments for students, however, has been vigorously blamed for a narrowing of curriculum and instructional practices or, in other words, a tendency to ‘teach to the test’ (Pedulla, Abrams, Madaus, Russell, Ramos & Miao, 2003). High stakes assessment for students has also generated considerable political controversy (Goertz et al., 2001).

There is a marked diversity, across states, in terms of the types of tests (criterion- or norm-referenced), number and mix of subject areas, and grade levels included in assessments (Goertz, 2001). Approximately half of the states use criterion-referenced tests; others use only norm-referenced tests or a mixture of both. All of the states assess student performance in reading and mathematics (e.g., NAEP), but not all include tests of writing, social science and science.

States also have a variety of interpretations of the legislated requirement that they use multiple up-to-date measures of student achievement, including the measurement of higher order thinking skills and understanding. There is no clear, shared interpretation of what this means (Goertz et al., 2001). Some states use only multiple-choice format assessments, while others include performance measures and open-ended questions. Some meet the requirement through the inclusion of non-cognitive measures, while others report on formative assessments such as early literacy tests (Goertz et al.).

An example of a comprehensive state accountability system is provided below for the state of Maryland, on the basis of data gathered in the late 1990s. This is an optimal, rather than typical, example.<sup>5</sup>

*Maryland: A case study and comparison with Victoria*

Victoria, in Australia, and Maryland, in the United States, have approximately the same student cohort size, similar geography and similar populations. Both systems assess at years 3 and 5. Maryland also assesses at year 8 whereas Victoria assesses at year 7. A comparison of the approaches is shown in Table 2.

Table 2

*Comparisons of Accountability Systems: Maryland and Victoria*

Index	Victoria	Maryland
Cohort Size	~ 60,000	~60,000
Year levels assessed	3,5,7	3,5,8
Number assessed	~120000	~180000
Subjects assessed (1995)	English Mathematics	English (state and national) Mathematics (state and national) Science (State and national) Social Education (state)
Test type	Paper and pencil MCQ, small performance assessments in writing Discrete subject tests	Paper and pencil performance Integrated task assessment
Sample type	Full cohort Single administration Secure design and scoring	Multiple matrix, sampled by rotated task, teachers involved in design
Scoring	Central scoring predominant	Local scoring predominant
Standards achieved	80% of students at or better than expected	45% of student at or better than expected
Reports	State results public Disk-based data sets provided to schools to generate predetermined reports	District results public
Data publicly available	Performance Gender Language Racial origin	Enrolments by year level Student mobility Assistance schemes ESL proportions Economic support

Index	Victoria	Maryland
		District wealth per capita
		Expenditure per pupil
		Teachers per 1000
		Assistants per 1000
		Support staff per 1000
		Students with kindergarten
		National test reading score
		National test language score
		National test mathematics score
		Days at school per year
		Hours of school per day
		Percent meeting satisfactory standard
		Percent meeting excellent standard
		Participation rates at each year
		Completion rates at year 12
		Percent dropout rates
		Percent promotion rates
		Percent meet university entrance standard
		Percent attend university
		Percent attend post secondary two years
		Percent attend post secondary four years
		Percent enter employment

Two impressions are gained immediately from an inspection of the Maryland assessment program. The first is that the extent of the monitoring program is far broader than that focused on student outcomes in the local system. The second is the breadth of accountability and open disclosure of information. The community appears to have a better chance of making informed decisions about the schools and the education system.

A third difference is not immediately obvious. Prior to 2003, the Maryland state-wide assessment program was entirely performance-based and the items integrated curriculum areas. It was unusual, for example, for mathematics or science or English to be assessed in isolation. This is a similar system to that espoused in the Victorian Essential Learning

models. The tasks in Maryland were set as integrated performance exercises and students were advised regarding how each component was to be marked. Levels of performance were set for each domain of learning using a five-point scale. For example, Level 3 of a five-point scale, as shown in Table 3, was established as a satisfactory level of performance for Grade 3 Reading. The similarity to benchmarks is noted, but there were five levels for each task rather than a single benchmark level to be attained.<sup>6</sup>

Table 3

*Maryland Standards for Grade 3 Reading Performance Tasks*

Level	Literary Experience	To be Informed	To Perform a Task
3	Adequate understanding of text, some connections, supports responses with text based references, some understanding of literary elements	Adequate understanding of text, limited connections between ideas and text, supports responses with text information	Adequate understanding of the text, evidence of constructing meaning, applies graphic information, some extension between ideas and text

In Victoria, the standard is set at a designated profile level for the grade level on the CSF, supplemented by the National Benchmarks. For grade 3 this is level 3. Discrimination within levels requires some external form of assessment. In the case of Victoria, centrally set tests are used for the purpose. Victorian schools are assessed against the proportion of students achieving satisfactory levels of performance against the profile level for a specified grade. For grade 3 reading it is level 3 of the English profile and its sub-strands.

In Maryland, classroom teachers were recruited and trained in marking and moderating performance-based assessments using samples of student work as exemplars. They marked all the tests, reported scores to state level authorities (who scaled tests and items to match a common score scale) and retained the individual student scores in the school. The importance of scaling to the system is that it provides the possibility of comparing performance over time and monitoring student performance.

The teachers recorded school level performances against benchmark standards set at a state level. For a school to be classified as performing at a satisfactory level, 95% of the students must be marked as achieving at the mid range of the assessment scale for the performance tasks. The standards were difficult to achieve. For example, an average of 35% of schools was identified as having reached the ‘satisfactory level’.

In Victoria, the teachers mark a small performance task and administer multiple-choice tests. Marks for the performance assessment are entered onto the student test booklets and returned for centralised marking. Results are fed back to the school some two months later. No moderation of assessments is included and the writing performance task is replicated for

central marking. An average of 80% of students is designated as being at or above the required standard. Detailed analyses of items, student diagnostic information through performance maps, and individual, class and school level analyses are made available to schools through a disk-based database. Schools are then encouraged to use the data to plan a curriculum improvement program based on the Assessment Improvement Monitoring (AIM) program.

### *Test alignment*

A challenge to the effective monitoring of standards-based reform strategies in the United States, whether through a national assessment such as NAEP or tests developed by individual states, is the issue of alignment, or lack thereof, between curriculum, standards and assessments (Porter & Smithson, 2001). This problem is pervasive across nations and systems. If assessments have not been strongly linked to standards, they clearly cannot be claimed to assess student performance against those standards. Certainly, the reliance of some states on norm-referenced tests of mathematics and reading, as the base of state assessment procedures, has been criticised for lack of commensurability between tests and the academic standards they are purported to measure (Goertz et al., 2001).

### *Measures of progress: Absolute targets and value added analyses*

Different states and districts have tended to use one of three approaches to monitoring school reform in terms of progress towards expected standards: measurement against absolute targets, relative growth or value added comparisons, and/or narrowing of the achievement gap (Goertz et al., 2001). Wide definitions of acceptable performance have been rife. For example, in 2001 schools in Texas were rated against absolute targets that included at least 50% of students in each of the major ethnic sub-groups passing the state assessment in reading, writing and mathematics, student attendance rates exceeding 94%, and the dropout rate not exceeding 6%. By comparison, states defining school progress in terms of relative growth based the assessment on improvement against each school's past performance and, in some cases, the distance from state goals. In California, schools were assigned individual annual growth targets based upon the difference between their student achievement outcomes and the state performance target. In other states, schools were required to demonstrate that they had narrowed the achievement gap between low- and high-performing students, in addition to meeting absolute or relative performance standards.

A pervasive problem with the federally mandated method of measuring school progress is the weight placed on single-year changes in test scores (Kane, Staiger & Geppert, 2002). Under the No Child Left Behind legislation, grade-level test scores of students are compared to those of students in the same grade level from previous years. These are referred to as *current status indicators* or described as 'snapshots' of student performance taken at a particular point in time (Drury & Doran, 2003). However, at individual school level the average cohort of students in any grade level in any year is less than 70 (Kane et al.). With a sample this small, test scores can show large annual fluctuations, many of which can be

attributed to yearly change in the demographic composition of the grade and have little to do with school performance.

Some states are currently challenging federal legislation on these grounds, and claiming the right to prioritise state-level measurements of progress over those required by the federal legislation (Foy, 2005). For example, dissatisfaction with methods of quantifying progress under the No Child Left Behind legislation has recently led the state of Utah to demand the right to report improvement at the level of individual students as they progress from grade to grade. However, states that currently rely on such longitudinal value-added analyses to track progress and meet accountability requirements also face difficulties because of the technical limitations of these analyses. Indeed, the consensus of contributors to a special issue of the *Journal of Educational and Behavioural Statistics* (Spring, 2004) was that it might be impossible to find a value added model that does not have inherent technical problems and can thus be defended against criticism. The overall conclusion was that use of value added models is ill-advised at the moment. Specific challenges to the accuracy and usefulness of value added analyses include that:

- they do not ameliorate doubts about the relevance of measures based on tests of a limited range of subject domains (Herman & Golan, 1993). Value added measures are only as valid as the tests on which they are based;
- they are expensive to set up and maintain (WEAC, 2004);
- there are many different models of value added analysis that vary in the ways they handle missing data (which is an important issue, because data from struggling students are more likely to be missing than data from high-achieving students) (Rubin, Stuart & Zanutto, 2004), whether or not they include demographic data (Tekwe et al., 2004), and the analytic units (i.e., schools, grade levels or classrooms) on which they base reports (Drury & Doran, 2003);
- they may inadvertently create lower performance expectations for disadvantaged students, if they are based on estimates that adjust for student background characteristics (Tekwe et al.). Value added analyses that exclude socioeconomic and demographic characteristics, whether at school- or student-level, tend to be biased against schools with an over-representation of students from disadvantaged backgrounds, while value added analyses which include these factors tend to be biased against schools with an under-representation of students from disadvantaged backgrounds (Tekwe et al.). The decision of which factors to include in the model, then, is based on political intent rather than objectivity in measurement;
- they rely upon complex statistical procedures that are not transparent or easily explained to the majority of people who are affected by them;
- within-school or within-classroom variability adds significantly to the contextual factors that may need to be entered into an analysis. Current models only take student, and possibly between-school, variability into account. Some researchers have concluded that separating out contextual effects from teacher effects poses major technical challenges, and would necessitate an extremely complex value added model (e.g., McCaffrey et al., 2004);

- they may inflate the degree of random error involved in measuring student progress (Drury & Doran);
- they cannot distinguish between true learning and ‘teaching to the test’ (Crane, 2002);
- they do not account for shifts across grade levels in the nature of underlying skills (Reckase, 2004). For example, tests of grade 3 mathematics may centre on skills in arithmetic while by grade 8 the tests encompass problem solving and algebra;
- they typically can identify schools that are well above or below average, but do not reliably identify schools that are slightly above or below average (Crane); and
- at best, they provide information on which schools are producing the best outcomes in terms of yearly improvement on state assessments. However, as Raudenbush (2004) cautioned, they cannot indicate the specific practices or reforms to which those improvements are attributable, nor do they support causal interpretations (Rubin et al., 2004).

Some of these problems are elaborated upon in the following review of the Tennessee Value Added Assessment System (TVAAS).

#### *The Tennessee Value Added Assessment System*

Since 1991 in Tennessee, students in grades 3 to 8 have taken state assessments in five core subjects of reading, language arts, mathematics, science and social studies. Following analysis, schools receive a report card describing how each grade fared in every subject in comparison to the previous year, based on national norms. This information is then used, in conjunction with other indicators, to determine state rewards and sanctions for schools and districts. Teachers also receive a report card (that is not made public) describing the test outcomes for his or her students.

As developer of the Tennessee Value Added Assessment System (TVAAS), William Sanders (1998) claimed the model minimizes measurement error by relying on up to five years of results for each student, provides accurate results even when data are missing by estimating test outcomes on the basis of a current score, and eliminates the need to control for contextual variables. Sanders argued that each student acts as his or her own control for the analysis (Ballou, Sanders & Wright, 2004), thus making the assumptions that socioeconomic and demographic factors are constant over a student’s school years and that deprivation and disability do not have cumulative effects on learning. Sanders and his colleagues have conceded, however, that TVAAS does not control for demographic and socioeconomic variables at the school level (Ballou et al.).

As the requirements of accountability set out in the NCLB Act have become increasingly topical in the United States, a number of researchers have begun to question the assumptions and claims of TVAAS (e.g., Bracey, 2002, 2004; Kupermintz, 2002). In particular, both Bracey and Kupermintz noted that:

- there has been no independent review of the underlying methodology of the TVAAS, as it is proprietary and held in secret. This hampers any attempt to critique the model;
- there is no evidence to support claims that analysis of previous test scores alone can control for all other outside influences;
- demonstration that students of certain teachers show greater or lower gains on average than students of other teachers does not explain why this may be the case;
- TVAAS relies on the acceptance of multiple-choice tests as adequate measures of educational outcomes;
- there is no attempt to use independent, corroborating evidence to support claims that teachers judged as effective by TVAAS are also judged as effective by students, parents, peers or administrators (see also, Bock & Wolfe, 1996); and
- the model applies circular logic. The strong claim made by Sanders and Rivers (1996) that teacher effectiveness is the dominant factor affecting student academic gain is misleading because teacher effectiveness is defined by student academic gain.

Concerns about the reliability of TVAAS were heightened by reports of unstable assessments of teacher quality (Bracey, 2002). Teachers who had been identified as very effective in one round of testing were not necessarily identified as effective in the next year (Bracey). In 2004, the TVAAS came under more intense political scrutiny in Tennessee. Opponents of the system argued that it was an expensive and flawed measure of teacher and school effectiveness. In particular, concerns were expressed over large differences between students' raw score test results and the value added measures of their performance. This may have been caused in part by confusion over weighting procedures used to equate test items from one round of testing to another (WEAC, 2005), but this is impossible to verify because the model and data are not available for scholarly review.

#### *Whole school reform models*

The provision of federal funds for school improvement programs has led to a proliferation of 'research-based, school-wide' reform programs (Houston et al., 1999). Most of these programs monitor impact in schools through outcomes on one or more of a range of measures including standardized tests, state assessments, NEAP, curriculum-embedded assessments, and teacher-designed assessments. In addition, some report student achievement in terms of students' choice of more or less challenging courses, attendance rates, graduation rates, retention rates, time taken to complete units of instruction, scores on self-esteem and attitude inventories, and teacher ratings of performance. As noted by Houston et al., most of the program developers report data supporting claims of positive effects on student test scores. In a system where the success or failure of students, teachers, schools and state education authorities is gauged primarily in terms of student scores on a sometimes quite restricted battery of tests, this could be (mis)interpreted as useful evidence.

In general, most weight seems to be placed upon one or more of a battery of commercial, norm-referenced tests, including the Iowa Test of Basic Skills, Metropolitan Achievement

Test, Stanford Achievement Test, Comprehensive Test of Basic Skills, or upon state- or district-wide standardized tests. The domains tested are predominantly reading, writing and mathematics, although some state assessments include social studies and science.

### *Summary*

In the United States, strategies of school reform and methods of measuring progress in reform vary from state to state and, as a further complication, there is an implicit pressure for observable improvements within short periods of time. This situation places considerable demands on educators at every level of the system, without necessarily improving outcomes for students.

### *The Wider Influence of the United States and United Kingdom Reforms*

The United Kingdom and United States school reform movements have affected other countries including both Australia and New Zealand. In Australia, reform through the aegis of the Australian Education Council (AEC) over a period of approximately fifteen to twenty years has pushed for more conformity in primary, secondary, technical, and further education provision. Specifically, the AEC developed national goals for schooling, a 'curriculum mapping' exercise to develop consistency among core subjects, a proposal for national assessment of student performance, national award restructuring of teachers' salaries, and consideration of national teacher accreditation. Every state and territory revised its Year 12 certification and assessment procedures to cope with rising retention rates. The period also saw a shift of enrolments towards private schools.

A similar radical reform of educational management went on in New Zealand. In the late 1980s, a national task force to review educational administration recommended that 95% of education funding be provided directly to schools, that each school negotiate a charter, that a board of trustees at each school govern its functioning, that each school be regularly audited on performance outcomes, and that parents be able to exercise choice over which school their children attended.

### *New Zealand*

In 2004 the Secretary for Education, Howard Fancy, reviewed the process of school reform in New Zealand from the start of the 1990s. He noted that, although students in New Zealand have performed consistently well on international achievement measures, the wide gaps between the strongest and weakest students, and the variation within schools, motivated the agenda for reform. School reform in New Zealand was originally aimed at decentralisation of administrative control of schools, and has continued to evolve over a fifteen-year span in

response to changing community perceptions and international trends. Reforms to the curriculum and qualifications framework have been addressed towards inclusion of cultural groups, with an emphasis on the establishment of clear statements of expected learning outcomes and standards and of different pathways through the education system to accommodate the diverse needs of students (Fancy).

In the early years of reform, most evidence of progress was gathered through school inspections conducted and publicly reported by the Education Review Office (ERO). More recently, these have been supplemented by a wide range of measures including student achievement and other education indicators such as incidence of absenteeism, exclusions and bullying.

### *Education Review Office*

An ERO Education Review is an external evaluation of the quality of education provided for students in all New Zealand state schools. Education Reviews focus on improvements in student achievement. Evidence is gathered via the school's self-evaluation, following a process of setting priorities for the review in negotiation with the school and local community. Schools are expected to engage parents and students in the process of self-evaluation, and a self-audit checklist is also provided by ERO to schools. Key questions posed by ERO are based upon:

- the extent and quality of information about individual student achievement in relation to essential learning areas, skills, attitudes and values, what the school's community thinks is important in student achievement, and how well this information is used to support learning;
- individual student ability to reach satisfactory standards of achievement, including provision for the needs of under-performing students;
- the usefulness of the school's reporting to parents on student achievement;
- cooperation between teachers;
- tailoring of teaching to the needs of individual students;
- teacher responsiveness to curriculum change, pedagogy and classroom practice;
- effectiveness of school leadership in terms of setting expectations, monitoring teaching quality, providing feedback and professional development for teachers, engaging community participation; and
- school effectiveness in self-monitoring and taking action to improve.

In addition, schools self-evaluate against a series of questions probing their ability to provide a safe and supportive environment, and to foster positive relationships between students, staff and the community. Review officers carry out investigations in schools by asking questions and seeking evidence to support school statements. This information is assembled into a report for schools and made publicly available via the ERO website.

### *National Education Monitoring Project (NEMP)*

Measures of student achievement include testing of a national sample of students at grades 4 and 8 in all curriculum areas and on a four-year cycle through the National Education Monitoring Project (NEMP) (Flockton, 1999). A random sample comprising approximately 2.5% of students in the relevant grade levels is selected for testing. Additional samples of children attending Maori-language immersion schools are also selected at year 8 level, to permit comparison of their achievement with Maori students in the main sample.

Each year, the assessments cover approximately 25% of the national curriculum. They are not restricted to 'priority' areas such as literacy, mathematics and science, but also encompass art, information skills, technology, music, social studies, health and physical education. One third of the assessment tasks are held constant from one cycle of testing to the next, to permit trends in achievement to be described.

To minimise load on students, different groups of students attempt different tasks. Experienced teachers, working within their own regions, are trained to conduct assessment tasks. Most assessment activities are presented orally by the teacher, on videotape or via computer, and most responses are provided orally or by performance rather than in writing, often recorded on videotape for later analysis and marking.

NEMP reports achievement at a national level and not the performance of individual schools, teachers or students. Detailed descriptions of student performance on a task-by-task basis are provided, rather than summaries for subject areas. Teachers are encouraged to use some of the tasks with their own students, to compare their results with the national sample.

### *Assessment Tools for Teaching and Learning (asTTle)*

In conjunction with the University of Auckland, the Ministry of Education has developed an assessment tool that allows teachers to track the progress of their students in reading, writing and mathematics (in both English and the indigenous Maori language) in years 4 to 12 (Brown & Hattie, 2003; Hattie & Brown, 2004). Teachers use an online program to construct paper and pencil tests targeted to the needs of their students. Student scores are entered into the computer program to generate reports in graph format that allow teachers to analyse results for their students against curriculum levels and objectives, and to compare results for their students against national standards. Within this framework, responsibility for monitoring student achievement is given to the classroom teacher, who is also provided with feedback on the strengths and weaknesses of students, and with suggested strategies for the development of learning plans.

The concept of a teacher-managed assessment system is built upon the principles that, ideally, a national assessment strategy should:

- follow from, and not dictate, the curricula;
- assist in communicating what the curricula intends;
- generate debate about standards of performance embedded within curricula domains;
- help teachers to understand the levels of performance outlined in the curricula;
- help advance the debate about improvements to the curricula;
- reduce the stakes of assessment so that the assessment becomes a tool for teaching, learning and feedback to stakeholders, rather than the driver of instruction;
- be part of a broader range of evidence used to demonstrate educational outcomes;
- ensure comparative information is available to teachers, without resorting to high-stakes comparisons such as league tables which over-emphasise school-level comparisons and are open to criticism on the basis of methodological flaws; and
- provide useful information to students, parents, teachers, principals and systems that is valued as part of the teaching and learning process (Brown & Hattie, 2003).

Thus, asTTle is promoted as one component of a broad strategy for monitoring student achievement, which fulfils requirements for accountability information while also helping to inform teaching practice (Brown & Hattie, 2003).

In contrast to other systems, then, New Zealand has not taken up national testing of core skills (predominantly literacy and numeracy in most instances) as the primary outcome measure for assessing progress in school reform. Instead, emphasis has largely been placed on identifying and supporting effective reform strategies through a mixture of clear articulation of expectations, development of national benchmarks, and development of sophisticated assessment resources for the use of teachers and the system (Fancy, 2004). Evidence is gathered through a range of methods including the National Education Monitoring Project (NEMP), the Trends in International Mathematics and Science Study (TIMSS), the school entry assessment (SEA) surveys of young children's early literacy, early numeracy and oral language, the National Certificate of Educational Achievement (NCEA) and teacher-managed assessment tools.

### *Canada*

In Canada, responsibility for education largely devolves upon provincial governments. However, the federal government wields influence over educational reform both through funding incentives and its role as constitutional protector of minorities (Anderson, 2000). Further, a national system of performance indicators, standardized national testing as part of an outcomes approach to appraising school success, a shift towards a Pan Canadian curriculum and adoption of 'essential learning outcomes' for both course approval and

graduation requirements, form the basis of national monitoring of school reform initiatives in Canada (Anderson).

*National monitoring of student outcomes (SAIP, PCAP, PISA)*

In 1989, the Canadian Council of Ministers of Education introduced the *School Achievement Indicators Program* (SAIP) to provide a national assessment of achievement in reading, writing, science, mathematics content and mathematics problem solving for 13 year old and 16 year old students.<sup>7</sup> SAIP presents student achievement results aggregated at both national and province level, and results for English and French school systems within jurisdictions. Since 1999, contextual information has also been collected to support the interpretation of results. Students respond to survey questions about their opportunities to learn the subjects being tested, their attitudes towards the subject, and other information about their interests. Teachers and school principals provide additional contextual information through questionnaires.

SAIP assessments run on a cyclical program (similar to that adopted by the OECD's PISA), and assessment criteria and curriculum frameworks are jointly established by teams representative of all participating provinces and territories. Student performance is reported in terms of standards referenced frameworks, designed to reflect the skills and knowledge Canadian students are expected to have achieved by the end of secondary schooling. In each assessment, both age groups respond to components of the same test to facilitate cross-age comparisons. However, changes in assessment design, sample selection and scoring procedures have compromised the ability to conduct longitudinal comparisons. This methodological flaw has been acknowledged and steps taken to avoid it in the conduct of future assessments.

In 2007, the SAIP will be replaced by the *Pan-Canadian Assessment Program* (PCAP) which will initially assess student achievement in reading, mathematics and science, with intentions to add the subject areas of information and communications technology, second languages and the arts over the longer term. PCAP will assess achievement outcomes for 13 year old students across all Canadian jurisdictions. Two years later, the same cohort of students will be reassessed using OECD PISA. It is intended that a Canadian component of PISA will use items reflecting Canadian content to link with the pan-Canadian and jurisdictional assessments. This is expected to provide a cross-validation of information about student achievement.

At the provincial level, various strategies are implemented for monitoring the outcomes of reform initiatives. For example, Ontario publicly ranks school performance on the standardized tests contained in the national assessment program and imposes an external test of literacy which Year 10 students must pass to meet requirements for graduation (Anderson, 2000). The evaluation of the Manitoba School Improvement Program, by contrast, relied upon a broad range of indicators and data collection methodologies and conducted a

longitudinal study of the impact of reform initiatives (Earl et al., 2003). These are detailed below:

*Evaluation of the Manitoba School Improvement Program: A case study*

The evaluation study conducted by Lorna Earl and colleagues (2003) into the effectiveness of the Manitoba School Improvement Program (MSIP) has taken a longitudinal approach to monitoring both the implementation and the longer term sustainability of an educational reform initiative. MSIP is an example of a reform strategy for secondary schools, instigated by a provincial government and taking a whole-school reform approach, with a particular emphasis on building schools' capacity to enhance student learning experiences and outcomes. It is also a (rare) example of a large scale study which explicitly attempted to link reform strategies to outcomes, rather than assessing student outcomes and then retrospectively holding schools accountable without any rationale linking outcomes to reforms. Procedures adopted in the study for assessing progress in school reform were wide-ranging, with data collected via multiple methods, including:

- Annual interviews with school principals, and/or school improvement coordinators, focused on current school improvement priorities, perceived impact on students, processes of change and plans for continued improvement.
- Focus group discussions with teachers that covered understandings of current school improvement policies and the process of change, impact on teaching practice and students, staff involvement in school improvement activities, and measures of success.
- Focus group interviews with students discussing attitudes to school, perceptions of change within the school, effective learning practices, student involvement in decision-making and community support for the school.
- Surveys of principals to ask about the impact of school improvement strategies and perceptions of progress.
- Surveys of teachers focused on the process of school improvement in the school, the impact on teaching practice and on students, and community involvement.
- Surveys of students focused on student engagement in the school and with their own learning, perceived changes in the school, and an assessment of how well the school prepared them for future work or study.
- School achievement data in the form of graduation rates, summaries of student grades in English, mathematics, history and science, and summaries of student achievement on provincial exams in English and mathematics.

These data were used to define a set of constructs and establish scales. The key constructs for longitudinal comparisons across a four year period included school improvement processes, student learning and student engagement. Schools were grouped according to the length of time they had been engaged in the school improvement program. For each school, Earl et al.

(2003) detailed a school profile that included the school's goals and improvement initiatives, as well as scale scores for each of the key constructs. Qualitative data were used to interpret and support findings from the quantitative data analyses. The outcomes of the school improvement program were assessed in terms of overall change in scale scores.

### ***European Reforms***

As the United States and the United Kingdom are both members of the OECD group of countries, of which the majority are European, it is clear why some of the reforms to be observed in European countries share features with those described above. However, it is useful to look at some of the smaller countries, such as Denmark, Spain, Greece and Norway, rather than Germany, France, and Italy which were less threatened by political developments in Europe.

For example, during the 1980s, the Danish government favored policies of privatization, deregulation and quality controls that also characterized the United Kingdom and the United States. By 1990, the Danish central administration was most concerned with frameworks and quality assurance. School administration had been decentralized to the individual school where a school board of seven parents, two teachers, and two students answered to a municipal council. The board's duties included issuing guidelines to the school principal, approving the budget, establishing educational plans, timetabling, teaching materials and school rules, and making nominations for teacher appointments (Bjerg, 1991).

The imperative for national economic competitiveness also sparked major reforms in Spanish education after the transition from dictatorship to democracy in the late 1970s. Responsibility for education was delegated from central to regional government, and reform strategies centred upon better teacher preparation programs, more flexible curricula, parental involvement, inclusion of vocational training in senior secondary education, and a uniform certificate after the *Bachillerato* (upper secondary) years to incorporate technical/ vocational as well as academic courses (Carabana, 1988).

Development of a single European economic community influenced Greece to remodel both its curricula and its school administration. School reforms were, at least in part, motivated by a perceived need to be compatible with the patterns elsewhere in the European Community. It is of note that decentralization of accountability was a major theme of these reform strategies. For example, Norway, traditionally among the most centralized administrations in Europe, transferred power to make determinations on services such as education to regional and municipal bodies, but also devised a national curriculum plan to oversee and guarantee educational quality (*Mønsterplan for Grunnskolen*) (Rust & Blackmore, 1990).

## *School Reform in Asia*

The imperative to compete in an international economy forced the education community of the 1980s and 1990s to live and work in a 'borderless world.' Even Japan, with one of the most successful economies, moved in the 1980s to internationalize its schools. Prime Minister Yasuhiro Nakasone established a National Council for Educational Reform which delivered four reports recommending a more liberal approach to the curriculum, less formality but tougher discipline, and a concentration on individual differences. These suggestions were rejected in the mid-1980s, but have since been re-worked into a new raft of reform strategies.

The debate over school reform in Japan can be viewed as part of a movement towards reconstruction in countries based upon a Confucian (and to a lesser extent Muslim) rather than European ethos. Confucian economies have traditionally placed emphasis upon responsibility to the nation, skills in mathematics and science, development of a technologically literate work force, and on rewards for student academic achievement rather than equity of student opportunity. Examples include Deng Xiaoping's policy in China of providing 'keypoint schools' as centres of excellence for the best students from kindergarten to university (Lewin & Xu, 1989; Pepper, 1990), and the policy in Singapore to favor the best schools and to encourage private schools (Kwong & Kool, 1990).

### *Japan*

Despite the consistently high scores of its students on international achievement tests such as OECD PISA, there has been a growing dissatisfaction with the education system in Japan (DeCoker, 2002). Ironically, the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) has flagged intentions to soften many of the policies that have been cited by other governments as evidence of the superiority of the Japanese school system (Kawamura, 2004). For example, the Minister has recently issued policy statements directed towards broadening the curriculum, reducing emphasis on entrance examinations, creating a five-day school week, establishing a more flexible education system to meet the needs of a wide range of students, de-centralising control over schools and encouraging a process of self-evaluation in schools (Kawamura; MEXT, 2005).

In Japan, the national government's influence on educational standards is mediated through its textbook-approval process such that the textbook is the definitive statement of national standards (DeCoker, 2002). The Ministry mandates the national curriculum, and disseminates the curriculum through the approved textbooks, but it does not have a mechanism for checking that the curriculum is followed in schools (DeCoker). Rather, pressure to conform to the national curriculum is exerted mainly through the high school and university entrance examinations, which are based upon the curriculum (DeCoker).

### *School self-evaluation*

In the most recent statement of reforms, the Minister has mooted intentions to establish evaluation systems for schools and teachers, and claimed national responsibility for assurance of educational standards (Kawamura, 2004). However, the claim is mitigated by a simultaneous, and seemingly contradictory, commitment to encouraging increased flexibility and autonomy for local authorities and schools (Kawamura). Explicitly, all Japanese schools are required to conduct self-evaluations of educational activities and school management, and to report these publicly (Kawamura). Teacher certificates are accredited for a fixed term and teachers are evaluated prior to renewal; teachers are also rewarded for demonstration of excellent performance (Kawamura). However, the evaluation of teachers needs to be considered in the context of the high salaries and high status of the teaching profession in Japan, and the accompanying competition for teaching positions (Ellington, 2001).

### *Monitoring student achievement*

Japanese students take high school and university admission examinations, participate in international assessments of achievement (e.g., PISA) and since 2002 the Ministry has administered achievement tests to a sample of students in grades 5 and 9 (DeCoker, 2002). Students may also take tests administered through the private education system, the *juku* (i.e., after-school coaching or extension colleges that help prepare students for entrance admission tests). These tests are primarily practice tests for students who aspire to enter the next level of education in an hierarchical and intensively competitive system. They are aimed at extending students to achieve at high academic levels, and can be taken by middle and high school students every few months. As such, they provide students with regular feedback on their progress. The *juku* publish the test results in a range of formats, such as comparative percentiles and estimates of students' levels of potential university enrolment (DeCoker). In addition, schools often display their students' results on the *juku* tests as indication of their success in preparing students for university enrolment. Indeed, the quality of education in Japanese schools is judged primarily by the proportion of graduates who achieve enrolment in prestige universities (DeCoker).

### *Hong Kong*

At the end of 2004, the Education Commission of the Hong Kong Special Administrative Region of The People's Republic of China issued its third report on the progress of territory-wide educational reform. The broad scope of the reform covered the curriculum, assessment mechanisms and the admission systems for different stages of education. The seven areas targeted for specific reform were curriculum, language education, support for schools, teacher professional development, student admission systems, assessment, and the delivery of increased access to senior secondary and post-secondary education.

### *Surveys of teaching professionals*

The progress of reform has been monitored through large-scale surveys of school heads, heads of Key Learning Areas and teachers (the most recent of which involved 148 primary schools and 101 secondary schools) (Education Commission, Hong Kong SAR, 2004). The surveys examined the impact of school reform on students, the school as a whole, and the professional development and confidence of teachers and school heads. Survey respondents were asked about the amount of diversification of student learning activities within the new curriculum (with emphasis upon moral and civic education, reading to learn, project learning and use of information technology), student improvement in areas such as communication, independent thinking, motivation and commitment, and student participation in extra-curricular activities at the school. Information was also gathered on respondents' attitudes to curriculum reform, and the perceived effects of reform on relationships between the school and parents, relationships between teachers and students, deployment of school resources, teacher workload, collaboration between teaching staff, and teacher confidence.

In terms of the impact of reform strategies on student outcomes, school heads were surveyed on their perceptions of improvements in student learning since the implementation of reform. The areas of possible improvement posed to the school heads were communication skills, critical thinking skills, creativity, learning motivation, learning interest, national identity, responsibility, perseverance, respecting others, commitment and overall learning performance. Individual subject teachers were also asked about whether or not they thought that students had improved in specific subject areas.

### *Quality Assurance Framework*

Although not explicitly discussed by the Education and Manpower Bureau, Hong Kong, as part of its monitoring strategy for school reform initiatives, there is nevertheless a well-established protocol of school inspections in Hong Kong. The Quality Assurance Framework is based upon a comprehensive statement of performance indicators for Hong Kong schools (Quality Assurance Division, EMB, 2002). These performance indicators are presented as a checklist to facilitate both school self-evaluation and external reviews. They cover such diverse topics as student attitudes and behaviour, student participation and achievement, support for student development, links with parents, school culture, curriculum, teaching practices, performance assessment, school management and organisation, planning and administration, leadership, staff management, planning and management of resources, and self-evaluation procedures.

School inspections engage the whole school community in the process of gathering information. For example, questionnaires are sent to school staff, students and parents, school documents are scrutinised to provide evidence of performance, lessons and other school activities are observed, and samples of student work and examination papers are inspected. The outcome of an inspection is a public report and school action plan to support the development and improvement of the school.

### *OECD Interests in Education Reform*

Since the introduction of the PISA project in the 1990s and increasing interest in education as an economic driving force, the OECD has implemented several studies in cooperation with the UNESCO Institute of Statistics (UIS). Of particular interest is a global study planned to monitor 'World Education Indicators'. This project examines more than learning outcomes, and launches a global study of education systems addressing inputs, procedures and outputs in an effort to examine on a global basis the indicators of effectiveness of school reform. Its emphasis initially will be on primary education and essentially aimed at the fourth year of education.

In 2001, WEI National Coordinators noted the importance of assessing educational quality and its equitable distribution to students and of obtaining school-level information related to these issues. OECD decided that a school survey should focus on primary schools, and ask school principals and teachers of grade 4 students a series of policy and curriculum relevant questions.

The questions to school principals cover the following issues:

- the basic characteristics of the school;
- school resources;
- school activities;
- evaluation in the school;
- decision-making in the school;
- possible behavioural problems in the school;
- parental involvement in the school; and
- instruction time in the school.

The questions to teachers cover the following issues:

- teacher training;
- teaching time;
- characteristics of students;
- characteristics of classrooms;
- teaching practices;
- professional satisfaction;
- perception of teacher status in their country;
- types of texts and reading activities typically covered in their reading lessons; and

- types of topics and mathematics activities typically covered in their mathematics lessons.

Questions to systems cover the following issues:

- the intended national curriculum;
- the education system;
- the characteristics of the teaching force; and
- current reforms affecting education for grade 4 students.

### ***Summary of the Review of Current Methodologies and Procedures for Measuring Progress in School Reform***

This section of the report has reviewed current methodologies and procedures in measuring progress of school reform initiatives, with an emphasis on standards based and multi method approaches across a range of systems. Two themes have emerged from the review. The first is the lack of alignment between the intended focus of school reform strategies and measures of outcomes. The second relates to the pervasive problems of accountability and methods for tracking progress. A third issue, which is currently given relatively scant consideration in the literature, is the largely unresolved problem of how to link specific reforms to outcomes.

#### ***Measuring outcomes***

Most systems attempt to quantify progress in reforms by measuring student outcomes. However, this often simply means measurement of student performance on tests of literacy and numeracy skills, perhaps with science and social studies included in the mix. This sends a message to educators that certain aspects of teaching and learning are to be prioritised over others. The problem is pervasive across nations and systems. Canada has plans to redress the situation in its Pan Canadian Assessment Program, and New Zealand provides an example of a comprehensive national assessment strategy that relies on multiple measures covering all facets of the national curriculum. The New Zealand National Education Monitoring Project, for example, is not restricted to ‘core’ areas such as literacy, mathematics and science, but also encompasses art, information skills, technology, music, social studies, health and physical education.

There are many goals of school reform, such as improvements in teacher practice and school administration, engagement of the local community, broadening the curriculum and providing flexible pathways for students, that are not amenable to measurement in terms of student progress on a limited set of tests, or to short-term measures of progress. Some of these areas are currently examined, in the United Kingdom, Hong Kong and New Zealand for

example, through inspections or reviews of schools. Others are simply not quantified or tracked. Few strategies for monitoring progress acknowledge the importance of taking a longer term view of the impacts, on students, teachers, schools and the community, of reform strategies.

### ***Tracking progress***

Controversy surrounds current attempts to monitor progress in school reform by tracking change in student academic achievement. The problems surrounding the use of value added analyses in the United Kingdom and United States have been raised in this report. In brief, these include:

- doubts about the validity of measures based on tests of a limited range of subject domains;
- methodological concerns over the handling of missing data;
- the problem of whether or not to include demographic data and what demographic data should be included (i.e., which student, classroom or school background variables are the most important, and how sufficient are these factors);
- the fairness of various models, especially where high stakes are attached to outcomes of analyses;
- lack of transparency for the majority of people;
- inability to distinguish between true learning and ‘teaching to the test’; and
- inability to indicate the specific practices or reforms to which improvements are attributable, or to support causal interpretations (Rubin et al., 2004).

### ***Conclusion***

This section of the report has identified tools and strategies currently used across a range of international systems to measure progress in school reform, particularly in the areas of student learning outcomes and, where possible, in teacher practice and school organisation. The effectiveness of a range of methods has been reviewed, and constraints and issues have been discussed.

The next section of the report will examine considerations for evaluating progress in school reform being driven by the *Blueprint for Government Schools* in Victoria, with particular emphasis on *Flagship Strategy One: The Victorian Essential Learning Standards*.

## Section Two: Tools and Methods for Evaluating Reform Initiatives

### *Overview*

Evaluation of programs of school reform has, both traditionally and currently across many systems, relied upon summative assessment, often in the guise of the measurement and comparison of standards of student achievement in a limited number of learning domains. This is particularly noticeable in the United States, where there is an insistent political demand that teachers, schools, districts and states be held ‘accountable’ for demonstrable annual improvement in student academic outcomes. In other systems, such as New Zealand, the emphasis is less on summative assessment (i.e., whether or not the reform is presumed to be effective in terms of student achievement or other outcomes), and is weighted towards a more formative style of evaluation with the goal of improving program effectiveness.

As Patton (2001) noted, however, whether evaluation is targeted primarily towards accountability or program improvement, it is dependent upon development of comprehensive program information systems that ideally should identify, describe and monitor critical success factors, include both qualitative and quantitative information, and rely on both case and aggregate data. This combination of factors is necessary to ensure that information is both sufficiently rich to capture the complexity of the relationships between reform initiatives and their outcomes for individuals and groups, and sufficiently targeted to permit meaningful analysis. The temporal relationship between school reform initiatives and observable outcomes also needs to be clearly established. If changes in response to reform initiatives, whether in student outcomes, teaching practices or school management, are expected to occur over the longer term, then they will need to be monitored over the longer term. If the interest is in sustainable, rather than short-term, improvements, then longitudinal monitoring is essential (Earl et al., 2003).

In short, monitoring progress in a program of school reform may encompass:

- evaluation of the *outcomes* of the reform;
- tracking the *process* or *implementation* of the reform;
- consideration of the *impact* of the reform on a range of stakeholders;
- exploration of external factors that impede or support the success of the reform;
- documentation of unintended outcomes of the reform; and/or
- measurement of the presumed net impact of the reform by comparing program outcomes with estimates of outcomes in the absence of the program (U.S. General Accounting Office, 1998), or through comparisons of empirically-derived groups that differ in styles of implementation of the reform.

Each of these approaches may contribute, in a complementary fashion, to the development of a substantive understanding of progress in school reform. None, however, is sufficient when taken in isolation. Rather, as cautioned by Benson, Hinn and Lloyd (2001), each approach is explicitly connected to a different goal or objective of measurement and monitoring. Further, as Fouts (2003) commented on the evaluation of the implementation of the Essential Academic Learning Requirements in Washington State, it is the compilation of findings from a range of studies that provides the most useful and valid understanding of progress in education reforms.

This section of the report will look at approaches to evaluation in relation to the Victorian Department of Education and Training's *Blueprint for Government Schools, Flagship Strategy One*, with particular emphasis on monitoring progress in relation to student learning and the Victorian Essential Learning Standards (VELS). VELS is taken as a case study, to permit the analysis to extend beyond the superficial level and to provide specific advice on tools and methods that can reliably and validly monitor progress in school reform. It is acknowledged, however, that this is a partial and incomplete approach to the evaluation of the reform agenda proposed in the *Blueprint for Government Schools*. A comprehensive and extended analysis, as described in Section Four of this report, would include a detailed discussion of methods for monitoring impact on stakeholder groups and sub-groups and analysis of supporting and impeding influences within schools and the broader social context. It would look at the reform initiative holistically, rather than in its component parts, and as a set of inter-connected strategies that extend beyond the establishment of a shared statement of standards of student achievement to include strategies for planning curriculum, teaching practices and assessment to support the standards, funding schools and programs adequately so that they can achieve the standards and providing excellence in leadership and professional development opportunities for the education workforce.

### ***The Challenge of Monitoring Progress in Reform***

Earl (2004) commented on the tension created by the conflicting ends to which assessment is directed, and her insights are cogent for analysis of the evaluation of progress in school reform. Assessment, Earl noted, has traditionally been used as a sorting mechanism although it should ideally be regarded as a way to enhance learning. Similarly, measurement of student outcomes against a set of standards is currently used in many systems as a way of sorting schools into those to be criticised as 'under-performing' (Hopkins, 2005) or, under the United States' No Child Left Behind Act for example, to attract sanctions or rewards. There are several reasons, detailed below, why this is counter-productive to efforts to monitor progress in school reform.

First, the best-designed assessments of student academic achievement measure just that - student academic achievement. Taken in isolation, they do not measure the quality of schools, of teachers, or of programs of reforms. Observation that students in one school have performed at a higher standard (on a particular set of competencies assessed by a particular

set of tests) than students in another school does not explain why this should be so. Rather, it begs the questions: ‘Why have students in this school achieved better results than students in other schools? Can we find differences between the groups to satisfactorily explain this observation, or are we asking the wrong questions?’ As a further complication, tests of student achievement are particularly vulnerable to flaws of misalignment between the intended and assessed curriculum. So it must be acknowledged that tests of academic achievement do not, of themselves, allow inferences of the quality of reform initiatives.

Second, attempts to monitor progress in school reform by comparing schools on changes in student outcomes over time are beset by methodological problems and constraints. The multiple, extraneous contextual factors that contribute to the academic and social development of students cannot be sufficiently controlled to permit causal or deterministic conclusions to be drawn. This is obvious in the current search in Great Britain to identify an appropriate set of contextual indicators to support school monitoring schemes (Fischer Family Trust, 2004). Further, to attempt to evaluate progress in school reform on the basis of student outcomes, and then to claim that improved student outcomes are a reflection of the success of school reforms, is to fall into the trap of circular logic for which the Tennessee Value Added Assessment System has been roundly criticised (e.g., Bracey, 2004; Kupermintz, 2002). In summary, attempts to infer the success or failure of school reform initiatives by monitoring change over time in student outcomes are hampered by:

- the potential for misapplication of statistical analyses, that are appropriate to monitor broad trends over large samples, to the problem of defining progress among a small sample of students within a school or classroom, and related controversies over the ‘fairness’ of such measures;
- the misuse of a procedure, designed for tracking developmental trends across the longer term, for the monitoring of progress within an illogically short period;
- the misinterpretation of analyses that are fundamentally *exploratory and probabilistic* as though they were *explanatory and deterministic*. As Raudenbush (2004) warned, the most sophisticated multi-level models of progress in student achievement do not indicate the specific reforms or factors to which improvements are actually attributable; and
- the fundamental assumption of normal distribution of both outcome and predictor variables required for value-added analyses (Raudenbush & Bryk, 1986). This prompts the use of standardized tests of student achievement as outcome measures, which in turn contributes to misalignment between outcome measures and the skills and capacities they purport to monitor<sup>8</sup>.

The measurement of student outcomes can appropriately be used as *one* indicator among a range of indicators of progress in an evaluation study. Monitoring student progress over time can be used for evaluation of broad trends over large samples, over extended time frames, and as part of an exploratory study that highlights greater or less than expected improvement as the basis of a more in-depth evaluation. These analyses are the starting-point, and not the end-point, of evaluation studies.<sup>9</sup>

Further, measures of student outcomes best serve as indicators of progress when they are clearly linked to item response modelling (i.e., Rasch, 1980), criterion referenced continua (Glaser, 1963, 1981) and Vygotskian (1996) notions of development including scaffolded learning and the importance of zones of proximal development. This decisively shifts the underlying logic of the measurement to a formative, rather than summative, outcome.

### **Section Three: Evidence-based Reform**

Evidence is more than assessment data. Gathering evidence is a package that requires a clear design, purpose, process of collection and interpretation and finally a way of informing stakeholders about the results. It is not often that all these requirements are adequately met.

Assessment evidence is often derived from a mixture of system-designed test data and local observations combined with a cocktail of intuition and folk wisdom. These are applied to a confrontational set of data analyses which in many cases are extremely difficult to interpret and almost impossible to report to key players. So a specialist is sometimes employed to visit schools to offer teachers a brief and surface interpretation of results.

This is neither an elegant nor suitable use of data, nor is it an appropriate accumulation of evidence of learning and teaching to be used for continuous improvement. Education policy based on poor quality evidence is severely compromised. However, evidence-based education policy and practice is currently being enthusiastically promoted around the world, and simultaneously coming under fire not because it lacks rigor but because it is too hard, too rigorous and too scientific for the ‘art’ of education. Its critics demand less rigor and less design in data collection in order to advise and inform policy development. In light of this dilemma, this section is addressed to the challenge of gathering evidence that is properly collected, adequately analysed and appropriately interpreted.

#### ***Evidence-based Education***

The term *evidence-based* seems set to become a fashion in education although, if taken to extremes as in the United States, the approach may collapse under the weight of its own expectations. But *evidence-based* refers to an approach which argues that policy and practice should be justified in terms of sound evidence about their likely effects. Education is not an exact science, but it is too important to be determined by opinion based on flawed or inadequate evidence, surface examination of data, ‘trust-me’ interpretations of results and anecdotes from classrooms being forwarded as solid and verifiable bases on which school, and even system, policy is to be developed and implemented. The imposition of policies without adequate evidence about their likely effects and costs devalues teaching as a profession and leads to the waste of public resources.

The term *evidence-based* (as coined by Hargreaves, 1996) refers to the idea that the methods that practitioners use should be evaluated to show whether the interventions actually work, and that the results should be fed back to influence policy and practice. This was initially forwarded as an attempt to move from individualistic approaches to education policy to an approach in which ideas were tested and, where necessary, changed. The term *evidence-based* also reflects attempts to adopt successful interventions based on what is known to work, both in health and in education. Davies (1999) argued that an evidence-based approach to policy is not a panacea but a set of principles and practices for enhancing educational practice. However, critics point to the complexities and subtleties of teaching (and learning), in which teachers' own experiences, beliefs and values are often more influential than research findings.

The idea of an evidence-based approach (Davies, 1999; Hargreaves, 1996) was taken further by Whitehurst (2001) who defined evidence-based education as '(t)he integration of *professional wisdom* with the best available *empirical evidence* in making decisions about how to deliver instruction'.<sup>10</sup> Professional wisdom was described as the judgment that individuals acquire through experience, leading to the reliance on development of consensus views. Whitehurst argued that the pool of professional wisdom could be increased in a range of ways, including the identification of local successes and their incorporation into broader fields of instruction. Empirical evidence, on the other hand, depends upon the systematic collection and careful analysis of data. Evidence-based education thus relies on a combination of professional wisdom drawing on individual experience and consensus and empirical evidence based on research data. Both are needed, because professional wisdom supports adaptation to local circumstances and the idiosyncratic nature of the classroom, the school, the community and the region, while empirical evidence supports comparison of competing methods and generates cumulative knowledge. More importantly, it avoids the fad, missionary zeal and the 'good idea' approach to education in which today's bright idea is lampooned as tomorrow's catastrophe.

A key point, then, is that the evidence underpinning policy decisions should be based on professional judgment tested against empirical data. For evidence to be sufficient, however, it must also be able to be defended. This means that the design, methods of sampling, instruments, data collection, handling and analysis procedures must all be open and transparent, and that claims and conclusions based on the data must be clearly supported. All evidence is *not* created equal. Indeed, Ross and Rust (1997) described four levels of study design that can provide different types of evidence:

- randomized allocation and control group designs are called *experiments*;
- *surveys* have known probability samples and might lead to such things as correlation studies;
- *investigations* in which the subjects are neither randomised nor controlled in pre-post comparisons; and
- *observations*, in which no systematic approach is taken to data collection or cleaning, include approaches such as case studies or action research.

There is a need for rigor in the collection of evidence, including design of data collection procedures and instruments, if claims about the effectiveness of an educational intervention are to be given credence. Currently, comparisons of two or more conditions that differ in levels of exposure to an educational intervention are commonplace. Investigations are also commonly used to support assertions, but this cannot be extended to system-wide policy without some form of generalisable design. For example, rigor in evidence collection can ensure that the participants being compared have at least the same characteristics across conditions, that the rules of chance mean that the smart, motivated, or experienced have the same probability of being in condition one as in condition two, and thus that differences between two interventions do not result from pre-existing difference in the participants and/or subtle selection biases. Policy makers and educators have to be able to show that claims of success for programs of school reform involve a similar intervention and implementation for all groups, and that the participants in the project were representative of a specified target population. In addition, it is important to question whether there were other ways that observed results may have been achieved, and to consider the consequences of the intervention as well as the influence of local conditions.

Well-designed assessment instruments are pivotal to the collection of evidence upon which sound decisions can be based, but the procedures used in collecting assessment data often undermine the careful and meticulous development of assessment instruments. This is compounded when the inadequate design of data collection procedures is accompanied by biased judgments and inadequate recording procedures. Adherence to principles of good study design assure that instrumentation is appropriately developed and used, but the demands placed on evidence collection should also require that the interpretation of evidence is relevant, useful and accurate. In short, assessment evidence should be:

- based on learning and assessment theories;
- transparent and externally verifiable in both process and interpretation;
- resource sensitive in development and use;
- familiar to teachers so that their existing valid assessment protocols may be used with enhanced rigor;
- accessible to teachers as experts in that the process needs to be of low cost and effort for development and directly interpretable in terms of skills and capacities;
- fair, equitable and unbiased for groups;
- informative and evidential such that needs of stakeholders are met;
- able to accommodate system, student and teacher requirements and be capable of both statistical and consensus moderation;
- combined with a rigorous and defensible approach to data collection and control of sources of error to enhance reliability;
- free of alternative explanations for results; and
- controlled to yield demonstrable face and construct validity.

The next section of the report takes the conditions and cautions discussed in the previous sections, and applies them to the task of monitoring and evaluating progress in school reform. Rather than focusing on the complexities and challenges of evaluation from a negative perspective, the section attempts to set out a positive case for a comprehensive program of evaluation. The intent is to provide guidelines for evaluation that can be applied to a range of reform programs. The focus of the reform may shift from the establishment of standards of achievement for students to the development of professional standards to describe and manage teacher performance, or from the implementation of new technology to the trialing of mentoring programs for teaching professionals; the fundamental guidelines for evaluation are designed to apply across a broad range of programs.

### **Section Four: Guidelines for Evaluation**

The procedural guidelines discussed in this section focus on the identification of the goals of a reform strategy and the needs to which the reform is targeted, and the development of plans to address those goals and needs. It sets out how implementation, processes and outcomes may be evaluated. The term *system* is used throughout, although the model is not restricted to any specific type of delivery unit such as a state-level system, school, or other educational institution. The guidelines are designed to help in the planning, monitoring and reporting of the impact of reform strategies, and to be adapted to the specific needs and purposes of a range of reform initiatives.

Listed within each section of the evaluation model is a set of possible methods, persons, goals and so on. These lists are advisory, rather than prescriptive. They provide a frame of reference in which to address a series of basic questions that can be used to review and monitor a reform strategy. These are:

- What goals and needs were set for the reform?
- What plans were developed?
- What processes were used?
- What impact did the reform have?
- What effect did the reform have?

The approach is based on a belief that a corporate approach to program delivery and evaluation needs to be taken. That is, each reform strategy should be seen to be an integral part of a system's overall strategic plan and approach to program delivery. The model identifies five components of a cyclical process of review and development as shown below. Reporting is both an independent component and an integral aspect of each other component.

### The Model

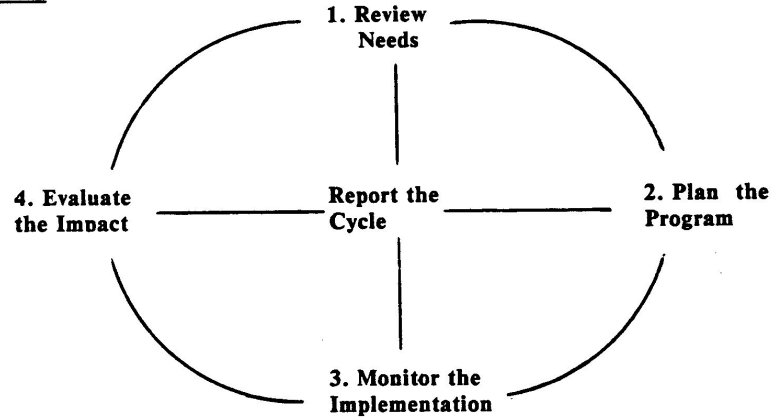


Figure 1. A cyclical model of review and development.

### *Five components of an evaluation study*

#### *1. Review program and assess needs*

In this component the goals and needs to which the reform initiative is targeted are assessed and problems are addressed. This includes an examination of policies and goals, and the clear statement of change objectives. In particular, explicit performance indicators, measures, benchmarks and targets are set for the reform strategy. These are designed to answer the questions:

- What are the key indicators of success for the reform strategy?
- What are the expected levels of performance against those indicators of success?

#### *2. Plan the project and identify resource needs*

In this component, a plan is developed to implement the reform strategy. This must clearly incorporate resource needs, budgets and timelines. It should also provide direction and set priorities for pursuing the targeted outcomes and outputs. At the school level, for instance, the plan might enable the formulation of a school improvement project, an outline of the budgeting details for that project, and a statement of the timeline for implementation. At the system level, the plan might state the goals and indicators of success for the implementation of a reform strategy, the resources needed to implement the reform and to monitor implementation, and a timeline for monitoring progress.

### *3. Monitor the implementation of the reform strategy*

In this step the reform strategy is implemented, and monitored in order to determine how well it has been implemented. The monitoring procedures help to improve the reform strategy and the way it works through a formative evaluation process.

### *4. Evaluate the impact of the reform*

At appropriate stages in the implementation of the reform, an evaluation is undertaken focusing on the outcomes, processes and impacts including the costs/benefits and effectiveness. These are assessed at multiple levels and timeframes but directly linked to the goals of the reform strategy, using the indicators and benchmarks established in the initial review phase of the cycle.

### *5. Report the cycle of evaluation and development*

In this component, evaluators might assess the effectiveness of reporting on the reform strategy, the plan, its delivery and impact. They would also examine the decision-making patterns, and the reporting and communication to both internal and external audiences for both short term and long term consequences as well as for different sub-groups of stakeholders in the project.

Review of programs of reform should be continuous during their life cycle. This means that there is never a final report on the progress of the reform. Rather, evaluators should have built in continuous internal review procedures and a means of recording and acting on observations and decisions. A review team needs to meet regularly, depending on the amount of work to be done and the specific tasks to be addressed at different stages of the implementation of a reform strategy. The size of the review team also varies depending on the size and complexity of the reform initiative, but is nevertheless active in all phases of the project from planning to evaluation and reporting. Each of the five components of the model has six procedural questions that need to be addressed.

- What is the task or focus of attention?
- Who is involved?
- How is it addressed?
- When is the appropriate time to address the matter?
- What should the report include?
- What are the outcomes and outputs of the component?

Linking the five component questions with the six procedural questions gives a thirty-cell matrix and each of the thirty cells has a range of activities and responsibilities. Each cell is numbered for reference purposes in subsequent sections of this report.

Table 4

*Matrix of Evaluation Components and Procedural Questions*

	What	Who	How	When	Report	Outcome
Review	1.1	1.2	1.3	1.4	1.5	1.6
Plan	2.1	2.2	2.3	2.4	2.5	2.6
Implement	3.1	3.2	3.3	3.4	3.5	3.6
Evaluate	4.1	4.2	4.3	4.4	4.5	4.6
Report	5.1	5.2	5.3	5.4	5.5	5.6

***Component 1. Review program and assess needs***

*1.1: What is reviewed?*

This is the beginning of the planning process and involves an assessment of the needs to which the reform strategy is addressed, linked back to the strategy’s mission, goals, purpose and plans. The review places the reform strategy within its socio-cultural, organizational and educational context. The following list provides suggestions for planning:

- Is there congruence between the goals of the reform strategy and those of other relevant levels of the system?
- Is the status of current achievement against the goals examined and reported?
- Are the problems and constraints faced by the system identified?
- Have the needs been identified, analysed and prioritised? Are they *declared, comparative, gap analysed or felt needs?* (See glossary).
- Have the needs been translated into change objectives?
- Have priorities been established among the change objectives?
- Have performance indicators been defined for each objective?
- Are there target levels of performance for each indicator using explicit measures or evidence?
- Are the targets observable and measurable?

*1.2: Who should be involved in the review?*

Administrators should ensure that all stakeholder interest groups are involved in the review and planning for the implementation of the reform strategy.

### *1.3: How can the review be carried out?*

There are many ways to carry out an initial review of a reform strategy. Choice depends upon the approach that best permits identification of change direction to achieve the overall goals of the reform. They include:

- environmental analyses;
- goal analyses;
- consultant input;
- professional development for substantive and process skill acquisition sessions run by consultants and/or core groups;
- problem identification and analyses;
- needs assessments and analyses – gap analyses;
- identification of areas of change and areas of maintenance;
- identification of change objectives;
- ranking and prioritising of change objectives;
- definition of indicators and measures of success;
- setting targets to be incorporated in statements of change objectives; and
- status surveys of goal achievement levels.

### *1.4: When should the review be carried out?*

The timing involves more than just when the implementation of the reform commences. It involves the duration of the whole process of reform. A comprehensive reform strategy cannot be reviewed by short-term methods. So planning for the review needs to take this into account.

### *1.5: What should be reported about the review?*

The review might lead to a proposal for the implementation of the reform strategy with background information about its development and, in particular:

- a mission statement and a list of related goals for the reform;
- a list of needs in prioritised order;
- a set of change objectives developed from needs;
- performance indicators for each goal and objective;
- information about the current levels of achievement with respect to the goals of the reform strategy; and
- achievement targets for the next time cycle (e.g., school year) for the goals.

### *1.6: What is the outcome of the review of the reform strategy?*

The outcome of the review should be a detailed contextualised statement of what the outcomes of the proposed reform should be. The context statement should be able to identify the following:

- a set of goals agreed upon and owned by all key groups;
- a set of prioritised change objectives;
- known and agreed indicators of success for reform objectives;
- targets agreed upon by all parties involved in pursuing the goals;
- a review committee which agrees to meet regularly and to monitor the project;
- a context statement for the reform strategy outlining the mission, goals, needs and prioritised change objectives; and
- an understanding of the constraints operating on the reform strategy.

### ***Component 2: Planning implementation of the reform strategy***

In this stage, the detailed proposal for implementation of the reform strategy should be prepared.

#### *2.1: What should be the focus of planning?*

The planning stage is an opportunity to determine the evaluation and dissemination potential of a reform. It is built upon the identification of alternative solutions to meet goals and objectives, and specification of criteria to judge the progress of the reform. Criteria are applied to a detailed model or to a series of alternative models for implementation of the reform. The model is usually an existing entity, which may be identified in a review of literature dealing with research pertinent to the problem identified in the context evaluation. Alternatively, the reform strategy may be completely innovative in which case a new model needs to be developed.

The set of reform objectives prioritised in the review stage now have to be translated into implementation plans aimed at achieving those objectives. These include:

- identifying a range of possible strategies for each goal;
- searching for or developing appropriate models for each strategy;
- evaluating each of the strategies and models;
- developing an action plan;
- resourcing the action plan;
- setting timelines;

- evaluating the action plan for its appropriateness, comprehensiveness and suitability; and
- establishing responsibilities for component development and implementation.

### *2.2: Who is involved in planning?*

The development of an action plan for implementation of a reform strategy involves a range of people. As each component of the plan is identified or developed, the personnel responsible for implementation or development are also identified and involved in finalising each step of the plan. A representative core group is key to the development of action plans, but their consultation and co-operation with their peers and those they represent is essential.

### *2.3: How should the planning be undertaken?*

Planning procedures can take a variety of forms, depending largely on the nature of the project and the planning timeframe.

### *2.4: When should planning occur?*

Planning cannot proceed until after the goals, problems, needs, change objectives and strategies of a reform initiative have been identified and, of course, the priorities and indicators of success have been established. The detailed planning needs to get the proposed reform strategy to an operational stage with resources identified and made available. Planning should also identify timelines for key stages of the reform implementation.

### *2.5: What should be reported about planning?*

The report of this component should detail the planning procedures for implementation of the reform strategy and contain all the characteristics so far discussed, including:

- a detailed proposal outlining the goals, strategies, resources and timelines as well as the expected outcomes and outputs and their appropriate indicators, measures and targets;
- details of resources required for successful implementation; and
- an action plan for implementation of the reform strategy.

### *2.6: What should be the outcome of the project planning?*

A detailed action plan that provides information about goals, indicators of success, appropriate measures and targets for the implementation of the reform strategy should be developed by this stage.

### ***Component 3: Monitor the implementation of the reform strategy***

Formative evaluation needs to be conducted while the reform strategies are being implemented. This provides information for program decisions, maintains a record of procedures, and helps to detect possible flaws in the implementation strategy and/or the reform initiatives themselves. In so doing, the evaluation can make use of a variety of techniques, which are listed below.

Formative evaluation helps to study both outcome and process criteria in order to fully evaluate and monitor the progress of the reform strategy. Formative evaluation provides the possibility that the program of reform can be changed. Six approaches to the monitoring and formative evaluation are described below.

#### *Monitoring*

This is continuous in the evaluation process and focuses on specific tasks and components of the reform strategy. Its purpose is to check whether the tasks are being completed according to the action plan, which in practice needs to be developed during the planning phase.

#### *Component Evaluation*

This focuses on critical sections of the action plan and sets up evaluation and feedback cycles. Component evaluation is demanding but it can be simplified if there are only one or two critical components identified in the plan. If organisational change is involved, there may be a need for staff development materials. All projects consist of components. Each component requires a different type of monitoring and evaluation.

#### *Problem Study*

This uncovers problems involved in implementing the reform strategy, and is usually coupled with a benefit study to avoid engendering a negative mind set.

#### *Benefit Survey*

A benefit survey concentrates on identifying both unanticipated and intended benefits of the reform. The objectives should indicate the initial benefits. If the benefit survey is conducted too soon, respondents are likely to cite only the objectives as the benefits or outcomes and will not detail additional, unintended benefits. For example, if the project is large and public relations oriented then the key people may have had career advancement as a side benefit. Evaluators need to be wary of the objectives being restated as benefits or outcomes without clear and measurable evidence in corroboration.

#### *Status Survey*

This is used to determine the rate of progress in implementing the reform strategy using the planned rate as a benchmark. It is also necessary to regularly monitor the progress towards

the intended outcomes. If student achievement is one of the goals, for instance, then regular but infrequent assessments would be necessary.

### *Levels of Implementation*

Implementation studies are not merely reports of what is done in a program of reform. Where teaching and classroom practices are part of a reform strategy, for instance, the actual implementation and levels of use of procedures and materials by staff cannot be assumed. Rather, this needs to be empirically verified and monitored. By monitoring the staff use and involvement in the program, the actual implementation of the project can be assessed. If surveys are used, then the instruments would also gather information on the staff perception of implementation. It is important to make this distinction and to validate staff perceptions. The innovation configuration helps to define the nature of the overall innovation as implemented.

#### *3.1: What should be monitored in the implementation of the reform strategy?*

The project action plan should have detailed the reform strategy and all its components. The strategy and its detailed plan now need to be implemented and monitored throughout the life of the reform. The following aspects need to be documented and people assigned to monitor each of them:

- the progress of each component of the reform strategy;
- professional development programs for all staff involved in the implementation;
- timelines for the implementation of the reform;
- resource use and supply;
- problems and benefits for teaching staff, students and parents;
- concerns of the teaching staff, students and parents;
- levels of implementation of the reform in various schools or by various teachers;
- appropriateness of goals, indicators of success and targets; and
- status with respect to goals, indicators of success and targets.

#### *3.2: Who should be involved in monitoring implementation?*

- The review committee
- Area or level coordinators in schools
- Evaluation staff both internal and external
- Clients (in schools, students and parents)
- Those responsible for key components in the action plan

### *3.3: How should the implementation be monitored?*

Monitoring can best be conducted using a range of techniques such as:

- action research;
- maintaining log books and diaries;
- component evaluations;
- problem and benefit surveys from teaching staff, students and parents;
- status surveys to determine overall levels of progress;
- teaching staff surveys of concerns;
- individual levels of use of the reform strategy; and
- examination of administrative records.

### *3.4: When is the implementation monitored?*

It is important to use procedures to monitor throughout the life of the reform initiative as it is implemented so that time lines, resources and materials are obtained as soon as they are needed.

- Action research is continuous throughout the implementation. Various problems and approaches can be monitored through the stages of action research used as a reflective improvement procedure.
- Maintaining log-books should be continuous throughout the implementation.
- Status surveys are conducted after several months to allow effects to become evident. They should be timed to allow some gains to have been made.
- Problem surveys are conducted early but care needs to be taken to prevent negative attitudes. After the initial problem survey, these are always accompanied by benefit surveys.
- Benefit surveys are conducted after considerable time has elapsed to permit the benefits to be apparent.
- Component evaluations are used to develop the components during or before their implementation.

### *3.5: What should be reported as a result of monitoring the implementation?*

Formative evaluation reports are essentially internal to the project team. This is not to say that internal members of the project team produce all formative evaluation reports. The reports should focus on:

- changes in the original plan for the implementation of the reform;
- refinements to the success indicators and targets;

- adjustments of goals and change objectives ;
- resource usage for each component of the policy implementation;
- indicators of involvement by teaching staff and students in the implementation;
- specifications of the reform strategy and the context in which it operates;
- problems and unanticipated benefits;
- strengths and weaknesses of the implementation and the reform strategy;
- essential features of implementation for the success of the reform strategy; and
- achievements in relation to goals and indicators of success at intervals of implementation.

### *3.6: What should be the outcome of monitoring implementation?*

Formative evaluation procedures used during the implementation stage of the reform strategy can also be used to ensure the success of the reform. These include:

- determination of how well the reform is being implemented;
- development of a refined project plan;
- evaluation of the indicators, targets, measures and benchmarks;
- strategies for engaging teaching staff, students and parents;
- identification of weaknesses and strengths of the reform strategy; and
- documentation of reform development, components, achievable outcomes, adjustments and resource requirements.

### ***Key questions for evaluation of the implementation stage of a reform strategy***

Questions in this section are based upon the monitoring of ongoing progress towards the achievement of the goals of a reform strategy. They may identify the need for, or interim effects of, any changes in approach.

1. What is the nature of the program of reform that was actually implemented?
  - Does it differ from the desired program?
  - Does it differ between schools and/or teachers? In what ways? How do different styles of implementation relate to other indicators of progress for the reform?
  - Does the program continue to be relevant to the original needs? To changing needs?
2. What human resources are needed to implement the reform strategy?
  - What special skills and qualifications are needed?
  - What are the roles of ancillary staff and outside experts?

- What professional development is necessary for teaching staff?
  - How and when is professional development organised?
  - Are there any problems with staffing (i.e., changing attitudes, level of engagement, level of implementation)?
3. What physical and material resources are being used to implement the strategy?
- How are the physical facilities being used? Any problems?
  - Is there sufficient stability or access?
4. What are the frequencies and types of communication concerning the reform strategy?
- Between the system and key stakeholders?
  - Are these random or organised?
  - Do they encourage interest and participation?
  - Do they produce criticism, positive or negative?
  - How is the feedback used constructively?
5. Have there been any major changes to the overall reform strategy?
6. Are there any noticeable changes evident in participants (i.e., students, teaching staff)?
- How were these changes monitored?
  - Can these changes be considered as improvements? On whose values? With respect to what criteria of success?
  - Are there any unexpected changes? Are they beneficial?
  - Are there negative reactions by the participants (i.e., students, parents, teaching staff)? Why?
  - Are the changes evident to people outside the project team? To the participants themselves?
7. What records are being kept on the project?
- Are they sufficiently objective? If subjective, are the value criteria clearly stated?

#### ***Component 4: Impact evaluation***

##### *4.1: What is the focus of the impact evaluation?*

The motivation for reform is to improve and change. The desired change might be in people, procedures, materials, or in combinations of these. However, change itself might not be sufficient if it is transient rather than sustainable. Similarly, a program of reform needs to identify the changes within its own boundaries and the extent to which these can be scaled up

and disseminated beyond the limits of the immediate program. In order to determine the impact of a program of reform, evaluators need to focus on:

- achievement of targets using measures defined in the planning stage and/or refined in the implementation stage of the project;
- cost benefit and efficiency;
- appropriate processes and implementation strategies;
- comparative effectiveness against other programs of reform;
- constraints on implementation and processes;
- nature of the client group (i.e., teaching staff, students and parents);
- likely dissemination strategies; and
- relevance to other settings and contexts.

#### *4.2: Who should be involved in the impact evaluation?*

All key stakeholders, and evaluators who are both internal and external to the project, should be involved in the evaluation of reform impact.

#### *4.3: How can impact be evaluated?*

Usually an impact evaluation is summative in nature and involves formal and rigorous procedures, usually conducted by evaluators external to the team of people responsible for the implementation of the reform. The purpose of this type of evaluation is to establish that the program of reform can be clearly linked to outcomes. The extent to which an impact study is needed is dependent to some extent on the level of investment in the reform strategies. Without question, it is reliant on the explicit identification, early in the establishment of the reform program, of what the outcomes (or critical indicators of success) are expected to be. These studies can take the form of:

- comparisons of achievements against targets and benchmark standards;
- comparative studies against similar systems;
- cost benefit studies;
- surveys of community and stakeholder reactions;
- survey data from client groups (i.e., teaching professionals, students, parents); and
- control group or quasi/control group studies where possible.

#### *4.4: When should the impact be evaluated?*

The appropriate timing for the evaluation of the impact of a reform strategy is determined by several factors, including that:

- the implementation and monitoring data should indicate that the reform strategy has been fully and appropriately implemented; and
- the client group should have had the opportunity to have been affected by the reforms.

#### *4.5: What should be reported about the impact evaluation?*

Clearly the most important things to report about the impact of the reform strategy are changes in terms of key indicators of success that have been set for the reform and identified in the planning and proposal stages. Changes need to be understood in the context of the overall program of reform and hence other things need also to be reported, including:

- resource consumption;
- targets achieved;
- targets not achieved;
- constraints on implementation and processes;
- changes in the client group (i.e., teaching staff, students, parents);
- changes to policy; and
- decisions about the ongoing development of the program of reforms.

#### *4.6: What should be the outcome of the impact evaluation?*

Decisions need to be made regarding the extent to which programs of reform should be revised, recycled or abandoned. The report should provide information that can be used for a review of the program of reform and an evaluation of the goals of reform.

### ***Component 5: Reporting the cycle of evaluation***

The report component is a public one, with the purpose of dissemination. It details the communication to stakeholders of the reform strategy, its development, resourcing and evaluation. This component assesses the effectiveness of the reporting strategy in terms of its purpose, its delivery and effectiveness.

#### *5.1: What is the purpose of reporting?*

The purposes of this component include:

- informing stakeholders;
- disseminating successful practices to other systems;
- documenting development of the reform strategy;
- increasing public confidence in the system;

- defence against inaccurate reporting by third parties;
- setting a proper context for interpreting the achievements of the system; and
- facilitating meta-evaluation at system levels.

*5.2: Who should be involved in reporting the cycle?*

There is a need to distinguish between the people who are involved in developing the variety of reports and those who receive the reports. External evaluation and reporting is valuable because of perceptions of objectivity.

*5.3/4: How and when should the cycle be reported?*

The timing and format of reports depends on the nature of the information and when it becomes available. Reporting should be continuous as it pertains to the stakeholder and audience needs and the dissemination strategy planned in the earlier stages of the project.

*5.5: What should be reported about the cycle?*

A comprehensive report on a program of reform might address:

- purposes and audiences for reporting;
- expected and achieved impact of the reform;
- changes to the original goals of the reform strategy;
- changes in the implementation plans needed to increase impact;
- methods of assessing needs;
- indicators of meeting needs;
- processes of the reform;
- resources allocated and adjustments;
- consideration of the effects of reporting on a range of stakeholders;
- feedback from audiences; and
- the nature of the information reported to various audiences.

*5.6: What is the outcome of reporting?*

This is the 'bottom line' of the evaluation. In essence the impact of the report should be the dissemination and professional acceptance of the reform strategy. There are some procedural outcomes but the most important involve the impact of the report on the community, the profession and the nature of education itself.

## *Summary*

This section of the report has presented procedural guidelines for a comprehensive program of evaluation with a focus on the identification of the goals of a reform strategy and the needs to which the reform is targeted, and the development of plans to address those goals and needs. It has explained how implementation, processes and outcomes may be evaluated. The next section will build on this information to consider the evaluation of the implementation, processes and outcomes of the Victorian Essential Learning Standards (VELS).

### **Section Five: The Victorian Essential Learning Standards (VELS)**

There are several important questions that must be addressed in an evaluation study of the progress and impact of VELS in schools. These are:

- What are the most appropriate indicators of success for the reform initiative?
- How can progress in these indicators be monitored?
- Can the reform initiatives be appropriately linked to indicators of progress?
- What professional development is available to teachers?
- What are teachers expected to do in relation to:
  - teaching against VELS?
  - assessment against VELS?
  - embedding the notion of developmental assessment in VELS?

These questions refer back to Patton's (2001) caution that the first requirement of an evaluation study must be the identification of critical success factors for the program of reform. In other words, the progress of a reform initiative cannot be effectively monitored without a clear explication of what success would be expected to entail.

The first of the questions can be addressed, in part, through the measurement of student learning outcomes, with the proviso that these outcomes must be clearly linked to the objectives of the reform strategy. The problem with this, however, is that there are as yet no clear procedures articulated for the assessment of VELS. The second question can be addressed by charting trends over time in student learning outcomes, and by combining these analyses with longitudinal evaluation studies drawing upon a range of other indicators of progress. This suffers from a flow on of the problem outlined above.

### *Student learning outcomes as an indicator of progress*

Just as any coherent assessment system must start from a clear understanding and statement of the intended curriculum (Matters, 2004), it is axiomatic that the tools and methodologies used to evaluate a program of reform cannot be meaningfully separated from the intention of those reforms. Accordingly, progress in student learning outcomes in terms of the implementation of VELS must be judged in relation to the purposes of the reform. White (2005) advocated two fundamental questions intrinsic to both the development of the Essential Learning Standards and, by extension, to attempts to measure progress in student outcomes related to the standards. These were:

- What are the purposes of schooling?
- What are the things that students at different stages should be achieving?

He argued that student learning outcomes should reflect an *interconnection* between:

- processes of physical, personal and social development, including the ability to develop positive social relationships, work in teams, manage personal learning and engage appropriately with the broader community;
- discipline-based learning; and
- interdisciplinary capacities including analytical, evaluative, reflective, productive, creative, communicative and other generic skills.

Two clear points explicated in VELS are:

- assertion of the interwoven nature of the components of the standards, which establishes that these should not be assessed independently but rather in relationship to, and combination with, each other; and
- the recognition that different priorities for learning should be set for different ages and stages of development.

In terms of tools and methods of measurement, this signals that progress in terms of student outcomes related to VELS might not be best monitored through standardized tests of core skills. Bentley (2004) argued, with reference to the current situation in England, that even if broad qualities and generic skills are specified in a curriculum document, assessment practices which rely on conventional, subject-based testing will marginalise those qualities and skills so that learning becomes defined by the areas that are explicitly mandated and formally tested. Similarly, in the United States a perceived lack of alignment between professed standards and the tests used to assess progress against those standards has been the source of considerable criticism (Goertz, Duffy & LeFloch, 2001; Porter & Smithson, 2001).

In Victoria, AIM testing and benchmarking are currently used to establish standards of essential elements of literacy and numeracy that are agreed upon as the ‘minimum acceptable’ (Rowley, 2003). However, it is possible to assess some higher order capacities and skills with AIM results if they are appropriately analysed (Griffin, 2004; Glaser, 2005 (personal correspondence)), and to use these within a broad framework of assessment.

Further, as Geoff Masters (2004) pointed out, the Essential Learning Standards recognise qualitative, in addition to quantitative, differences in the expected standards of achievement for students across age levels. This enhances the power of measurement, as it enables the measured skills to be translated into levels of development on an underlying construct. Different qualities of performance or capacity can also be described in some of the domains, but additional empirical work is needed to verify the expected trajectories of developmental progress in, for example, domains such as team work, conflict resolution, self-management and community involvement. Nevertheless, current approaches at least suggest part of the way forward (Griffin, 2004). Clearly, more than a single measure is needed but the framework exists for such developmental assessments<sup>11</sup>.

### ***Theoretical background: The OECD DeSeCo project***

From a theoretical viewpoint, the development of appropriate measures of progress in terms of the Essential Learning Standards might draw upon the insights of the OECD’s DeSeCo project (OECD, 2002). This project was an attempt to define the sorts of capacities, beyond the basic or core skills of reading, writing and numeracy, which are deemed necessary for a successful life and a well-functioning society. The project established that a well-designed, large-scale assessment program of complex skills would need to:

- derive data from multiple sources, including but not limited to the collection of data through large-scale assessments;
- recognize the existence of continuous scales, so that levels of proficiency can be identified for analytic purposes and to aid in the interpretation of scores;
- ideally be based upon longitudinal studies, although broad trends can also be monitored through cross-sectional surveys; and
- develop and validate measures beyond the current raft of assessment tools and strategies that focus upon, for example, reading literacy, mathematical literacy and numeracy, and scientific literacy (e.g., PISA, ALL).

Each of these is possible given current technology and theoretical approaches to educational measurement. Each can be achieved using current methodologies, as evidenced in the Western Australian basic assessment program. It is not a simple process, however, as cohort assessment is involved, but it can be managed with sample studies. Certainly, it would require some basic research and development to establish the breadth of skills that can be meaningfully measured (OECD, 2002). As part of this research agenda, the DeSeCo project suggested that appropriate test development should:

- build upon successful experiences with similar activities;
- encourage the use of methodologies that differ from traditional assessment techniques, including the use of portfolios or performance assessment; and
- explore computerized modes of delivery to establish fully adaptive and interactive assessment.

However, what the OECD (2002) does not establish is how these complex data collection exercises can yield reliable and valid data. Caution should be exercised in following these admonitions without such constructive advice. The current technology and methodologies used in Victoria can be extended to develop methods of monitoring VELs in ways that will have credibility for the community and government (Griffin, 2004). The following section supports this argument, and provides an overview of an approach to the development of a framework for the assessment of student outcomes.

***Glaser, Rasch and Vygotsky: A synthesis of approaches to develop a framework for assessment of student outcomes (Griffin, 2004)***

It is commonplace to attempt to describe student outcomes in terms of a list of capacities and skills, and to expect that these capacities or skills can be assessed by simple observation of students performing specific tasks. This approach to assessment is grounded on the mistaken assumption that certainty is attainable in assessment of human skills or development. In contrast, this section of the report argues for the recognition of an *inferential* and *probabilistic* model of assessment that draws together the work of Rasch (1980), Glaser (1963, 1981) and Vygotsky (1996) to provide an appropriate framework for the reporting of student outcomes. This is particularly useful when the goals of education encompass such things as creativity, teamwork, social communication, problem solving and other quite difficult attributes to observe in action.

The prerequisites for development of a framework of assessment are:

- clear specification of the desired outcomes of education programs;
- clear specification of a framework describing what is meant by growth in competence or achievement. This needs to be used as a frame of reference for interpreting assessment and identifying how to improve learning; and
- systematic assessment to permit reporting in a meaningful way to students, parents and other stakeholders.

In 1960, the Danish mathematician Rasch reasoned that the underlying nature of development on latent traits could be defined probabilistically in relation to the observable tasks that students performed. In particular, if tasks could be arranged in order of the increasing amount of capacity required for success, then the nature of the trait could be

defined by the nature and order of the tasks. Rasch's argument was grounded on the algorithm that, on balance, students who perform well on a test or set of tasks have a lower probability than other students of responding to easy items or tasks incorrectly, while students who perform poorly on a test or set of tasks have a lower probability than other students of successfully responding to difficult items or tasks. If this logic holds, then student development in a particular domain could be traced along a continuum described in terms of ordered clusters of tasks. In other words, the statistical characteristics of a group's responses to a set of items could be used to build a measurement scale.

At approximately the same time, Glaser (1963) described the concept of criterion-referenced testing which, like Rasch's approach, sought to describe development in terms of an ordered set of tasks. He originally used the term 'criterion' to describe a defined domain of content or behaviours to which test items were referenced, but this tended to direct assessment towards a checklist of unrelated skills that provided scant information for educators. As a result, Glaser (1981) expanded his original definition of criterion referencing to advise that tasks should be ordered in a progression that led to an overall interpretation of increasing proficiency. This meant that development could be described in terms of a continuum of tasks or stages of increasing competence. There was no need to define tasks as having only one outcome, approach or solution, and no need to restrict tasks to paper-and-pencil exercises scored in a pre-determined way. Rather, judgment could be used to interpret performance on complex tasks, not only in terms of the tasks undertaken but also of the manner in which they were completed.<sup>12</sup> A *criterion* could then be thought of as a threshold on the developmental continuum, rather than the content of a domain of skills (Glaser, 2005).

The theoretical underpinning of this framework for assessment is based on the empirical and statistically-derived relationship between the demands of a task and the capacity of a person to perform it. Unlike traditional forms of testing, the tasks used in this form of assessment do not necessarily have single correct answers and, as they become more complex, criteria defining thresholds or levels of performance can be used. There is thus no restriction on the nature of the tasks used for assessment and, in the most general of the Rasch models (Linacre, 1990), there are few restrictions on scoring procedures.

The framework for assessment is congruent with Vygotskian (1996) theories of learning as a socially mediated activity, in which a principal task for educators is the scaffolding of student learning and the recognition that students may differ in terms of a 'zone of proximal development' (ZPD). In particular, it supports educational objectives of targeting opportunities for learning and the development of individual learning plans for students. A synthesis of these ideas about education and educational measurement directs the purpose of assessment towards locating a student on an empirically-derived continuum of development. Student outcomes are interpreted in terms of the things the student has just been learning, is currently engaged in learning and is about to learn (with support and assistance), rather than focusing on a single score or comparisons with a group. This provides a substantive interpretation of the measurement for reporting purposes.

From a practical perspective, item response modelling, with Rasch model computer software, is used to build a variable map of the underlying latent trait (skill or competency). These maps simultaneously display a distribution of student performances on a group of tasks or items, and a distribution of tasks ordered in terms of their relative difficulty for a group of students. Each item can then be reviewed in terms of the skills involved in responding to the item, and this is a matter of substantive interpretation. It relies on the professional judgment of experienced teachers and subject specialists to interpret the levels on the developmental continuum. The mapping of students and items onto a common scale identifies a 'transition point' where an increase in item difficulty can be associated with a change in the kind of skill or capacity required to achieve the desired outcome for that item. When student ability and item difficulty are mapped onto equal points along the continuum, the probability of success for the student on that item is 50%. Thus, if the student were to improve a little in terms of the capacity or skill, he or she would have a better than even chance of success on the task or item. For the teacher, this provides information upon which to base an individual learning plan that aims to increase the student's probability of success.

At a system level, a combination of item response modelling and criterion-referencing, interpreted in terms of Vygotskian (1996) theories of scaffolded learning and targeted teaching within a zone of proximal development for each student, can be used to describe and validate an appropriate assessment framework for VELs. At a school level, teachers can be shown how to develop an assessment framework that encompasses these theories of measurement and consists of the following steps:

- describe the teaching and learning sequence that would facilitate learning through a series of levels;
- define the levels that are expected to develop;
- devise a set of tasks or activities that would track these levels;
- develop a set of specifications for the assessment tasks;
- identify the steps of the components of each task;
- anticipate, for each step, different qualities of performance from students at different levels;
- define codes for the different quality of performance;
- place the codes on a continuum of increasing levels of performance;
- identify clusters of codes and define overall levels of performance;
- interpret the clusters of codes using performance rubrics;
- define the characteristic knowledge and skills for each level and interpret these in terms of readiness to learn;
- identify intervention strategies for each level;
- set target levels for mastery and/or competence; and
- design a report that indicates achievement, readiness to learn and intervention strategies.

The final step in the process of building a framework of assessment is the development of an appropriate strategy for reporting to students, parents, schools and systems. It is assumed that a criterion referenced analysis and interpretation requires a criterion referenced reporting format. This can proceed in one of two ways. The first is an item level content analysis, pioneered by the Australian Council of Educational Research (ACER). This is an item analysis and provides a report on each item with success rates. The second, developed by the Assessment Research Centre (ARC) at the University of Melbourne, is a profile approach, whereby stages of increasing competence are defined and used for reporting procedures. This method is not reliant on the use of any specific set of tasks or tests.<sup>13</sup>

### ***Assessing student outcomes against VELs***

The assessment of student outcomes in relation to VELs ought not proceed until the standards have been validated, evaluated and reviewed in terms of the hypothesised developmental progressions within learning domains, consistency of expectations across learning domains, quality of performance associated with each standard level and benchmarking of standards against other national and international curricula (VCAA, 2005a). The validation of the quality of performance that characterises each of the standard levels across learning domains is of crucial importance. This involves a process of negotiation and moderation between expert teachers within and across learning domains, to establish both the theoretical underpinnings and empirical validation of expected standards of student achievement.

As the VCAA (2005b) pointed out, validation of standards of achievement is the first stage, which must then be followed by the question: ‘When has a student achieved the standard?’ The answer to this question, according to the VCAA, can best be established by *teachers* making an *on-balance professional judgment* of student performance *over time* and *across a range of tasks and situations*. Teacher judgment will be guided by exemplars of student work to illustrate the expected standard, and moderated through collaboration with colleagues to help teachers make judgments that are both valid and reliable for students within a class, and within a school, and within the state. One needs to be cognisant, however, of the previous fates of teacher judgment as a means of assessment. The Maryland approach offers one successful way of developing and moderating assessments, and if linked to the current calibration methods used in Victoria and then to the developmental assessment process of interpretation (Griffin, 2004), there is a relatively clear way forward.<sup>14</sup>

### ***Monitoring change in student outcomes and other indicators of success***

When the Essential Learning Standards have been subjected to review and evaluation, and empirically and theoretically validated, a possible method of monitoring student outcomes is to compare over time the proportions of students who are assessed as achieving each of the standards. Of course, this begs the question of how that judgment is to be made in a valid and

reliable way. Further, if change in student performance is expected to occur in the short term, such as one school year, then it is reasonable to monitor shifts in the proportions of students at each level on an annual basis. If some outcomes of the reform are envisaged to reflect cumulative years of schooling, then this should also be reflected in the time-scale for monitoring progress.

Comparison of the proportions of students across the entire system who achieve each of the standard levels does not require individual tracking of students, although it does require sound sampling and study design procedures. Provided the intention is to evaluate the program of reform, rather than to attempt to monitor individual schools, teachers or students, the process is not so vulnerable to the methodological constraints of value-added analyses. If used and interpreted with discretion, value added forms of analysis can provide useful information as part of an overall evaluation program.<sup>15</sup> However, such an approach does not obviate the necessity to link progress in student outcomes in terms of VELs to the reform initiative itself. This requires the inclusion of a broad range of indicators of successful implementation of the reform, which can then be used in conjunction with the student data to comment and reflect upon the relationships between styles of implementation and outcomes for students, for sub-groups of students, and from the perspective of students, parents, and teaching professionals.

## **Section Six: Case Studies**

The following discussion describes three examples in which measurement and monitoring of student outcomes have been successfully incorporated into program evaluation. The first and second examples describe large-scale assessment programs that have not been confined to standardized testing across a limited range of basic skills. The third example describes a longitudinal study of a system-wide school improvement program that is monitoring change over time, and linking different stages of reform implementation to a range of indicators of success that extend beyond simple description of change in terms of student outcomes. This study uses the ARC's method of profiling of student outcomes, discussed above, as one of its key indicators of success.

### ***The National Education Monitoring Project (New Zealand)***

In terms of practice, the National Education Monitoring Project (NEMP) in New Zealand provides an example of a large-scale program of testing that monitors trends in student achievement across a broad range of skills and capacities. Its utility, for the evaluation of a reform initiative such as VELs, is to illustrate a method of monitoring complex and inter-connected skills and capacities, without reliance on standardized, paper-based forms of testing.

The project monitors progress in student outcomes at a national level, rather than the performance of individual students, teachers or schools. It is designed to evaluate and report on progress across the entire system rather than for each student, and thus can incorporate a very wide range of skills and tasks. Annual testing covers approximately 25% of the national curriculum, with one third of assessment tasks held constant from one cycle of testing to the next to permit equating and comparison across years. Performance assessment is monitored through videotaping or via computerized adaptive testing. It would be nigh impossible to collect data of such richness and complexity if the performance of every student was monitored in every subject domain. Rather, NEMP randomly samples students for testing, and then produces detailed information about the performance of the national sample on each task. Teachers are then encouraged to use appropriate test items with their own students and monitor their results against the national sample.<sup>16</sup>

### ***The Maryland School Performance Assessment Program (MSPAP)***

Prior to 2003, the state of Maryland provided an example of large-scale evaluation based on assessments of student learning outcomes that were wide-ranging and performance-based, and inclusive of other indicators of success, thus providing a comprehensive body of evidence upon which to base decisions.

The Maryland School Performance Assessment Program (MSPAP) consisted of criterion-referenced performance tests in mathematics, reading, writing, language use, science and social studies for students in grades 3, 5 and 8. The tests were solidly grounded in unambiguous statements of the goals and expected outcomes for students at each grade level, which contained a general statement of the expected standard for each topic supported by a set of indicators and specific performance objectives. Of particular interest for VELs, the MSPAP was designed to assess how well students relate and use knowledge from different subject areas, and how well they apply knowledge to solving problems and demonstrate that ability in performance tasks. Tests used to assess basic skills and knowledge, such as reading for general comprehension, writing to communicate clearly, making accurate arithmetic calculations and identifying and understanding scientific, historical and geographic information, also emphasised higher order abilities such as supporting an answer with evidence, predicting an outcome and comparing and contrasting information.

The first stage in the development of the assessment program was the establishment of the key indicators of successful teaching and learning for students in grades 3, 5 and 8: the Maryland Learning Outcomes. These were based upon national and international studies of student achievement and on Maryland's curriculum framework, and then reviewed by local curriculum supervisors, advisory groups, school system superintendents, and empirically validated through field testing.

MSPAP assessment tasks were inter-related activities and items organised around a common theme. Students might be required to respond to questions or directions that led to a solution, a recommendation or an explanation. Many tasks were designed to assess multiple content

areas, and activities could involve groups or individuals, practical, observation or reading-based activities and/or activities that required written responses, production of lists, charts, diagrams and/or drawings.

Annual test development consisted of five phases: planning, design, development, review and revision, and then a field test. In the planning and design phases, tasks were selected from previous test administrations for re-use, and compared against the learning outcomes to identify gaps in the tests. New items and activities could then be planned. Approximately 50% of a test consisted of tasks that were carried forward from previous rounds of testing. The test specifications were then designed in terms of the task outlines, topic areas covered, time allotted for each task, and specific outcomes to be assessed.

In the development phases, teachers of students in grades 3, 5 and 8 were recruited to help write specific tasks and activities, develop scoring tools and write instructions for test administration. These teachers were trained on the principles of performance assessment, the problems of test bias, and the Maryland Learning Outcomes. Then the tasks and activities were subjected to a comprehensive review process, in which they were screened for technical soundness, controversial or sensitive topics, developmental appropriateness, ability to be appropriately scored, and clarity. In addition, assessment specialists conducted feasibility reviews that considered, for example, whether sufficient time had been allotted for each task, directions were clear and concise, materials were appropriate and all classrooms could accommodate the administration of each task or activity. In the final phase, field tests were conducted in schools outside Maryland but with comparable student populations, to verify the feasibility of task administration.

In each content area, MSPAP results were reported through five proficiency levels, with level 1 being the most proficient. Performance standards set for schools and local systems were described as ‘satisfactory’ if 70% of students scored at proficiency level 3 or better, or ‘excellent’ if 70% of students scored at level 3 or better and at least 25% of students scored at level 2 or better.<sup>17</sup>

MSPAP is an example of a large-scale assessment program, developed with the primary purpose of improving the standard of teaching in schools. Its relevance to VELs is apparent in its focus upon how well students solve problems both cooperatively and individually, how well they apply knowledge to ‘real world’ problems, and how well they integrate and use knowledge from different subject areas.

Maryland has recently redesigned its state-wide assessment program in order to meet the requirements of standards-based reform mandated by the No Child Left Behind Act. This has unfortunately led to the abandonment of their performance-based program of testing and the adoption of a criterion-referenced and norm-based assessment program that allows them to conform to the strictures of national accountability under the Act.<sup>18</sup>

## *Evaluation of the Hong Kong Primary Native English-speaking Teacher (PNET) scheme*

In 2003, the Government of the Hong Kong Special Administrative Region, Education and Manpower Bureau (EMB) requested the Assessment Research Centre at the University of Melbourne, in collaboration with the Institute of Education in Hong Kong, to undertake a detailed evaluation of the deployment of native-speaking teachers of English in primary schools. This reform initiative was part of the Hong Kong government's strategy to address a perceived decline in standards of English language proficiency.

The study design was based upon both cross-sectional and longitudinal monitoring of a range of indicators of success. These included, but were not restricted to, measures of student outcomes in English gathered via teacher observation against profiles of student achievement and by interview test. Other indicators included the styles of understanding and opinions that the native English-speaking teachers (NETs), local teachers of English, and school principals held about the role and responsibilities of the NET in the school, the integration of the NET into the school community, and the overall implementation of the reform initiative.

The study is currently in its second of three years of data collection. Baseline data were collected and analysed in 2004. Student achievement and attitude data were scaled using item response (Rasch, 1980) modelling, and other indicators were grouped using cluster analysis to uncover interpretable patterns of difference in response. The fundamental questions for the analysis include:

- Are there observable differences between schools, and within schools between groups of students or teachers, in the way that the reform is implemented? How do these change over time?
- Are any characteristics of schools, school principals, teachers (both local and native English-speaking), and/or student demographic background predictive of differences in reform implementation? For example, how do attitudes to the PNET scheme, or the educational background of the local English teachers or school principals, relate to the style of implementation in the school?
- Is there a relationship between differences in reform implementation and student attitudes or achievement outcomes? Are these relationships mediated by levels of student achievement or attitude? Are they mediated by different characteristics of the teachers, such as training, access to support and professional development, and school and classroom resources? For example, are there different relationships between reform implementation and progress for high- and low-achieving students?

A central feature of the study has been the choice of materials and methodologies for the identification of student proficiency levels. Constraints on data collection in schools required that the assessments did not disrupt teaching and that they were welcomed by teachers as contributing to, rather than distracting from, their classroom practice and professional development, while also providing a rich sense of students' language proficiency. The

sample of students included those from the early years of primary schooling (P1 and P2) so standardized tests of reading and writing were clearly inadequate for the collection of information about student performance. Two well-established instruments were used, and empirically validated to verify their suitability to the Hong Kong context. These were the *Profiles in English as a Second Language* (Griffin, Smith & Martin, 2003) and the *Interview Test of English Language (ITEL-ed)* (Griffin, Tomlinson, Martin, Adams & Storey, 2004). In particular, the *Profiles* instrument provides a useful example of assessment against validated standards of achievement, using the on-balance judgments of teachers taken over time and across a range of tasks and situations.

### *Profiles in English as a Second Language*

The development of literacy profiles (Griffin, Smith & Burrill, 1995) commenced in Victoria in 1986, when the research section of the Ministry of Education began a search for a system of monitoring achievement in schools. After almost five years of research and development, the first *Victorian Literacy Profiles Handbook* appeared, and this was followed by several years of successful monitoring of student achievement. It was also apparent that profiling as a system of monitoring student outcomes had a positive effect on teaching and learning.

Of particular note, profiles are descriptive rather than prescriptive. They illustrate the expected trajectory of development for students, and can be used by teachers to monitor a broad range of achievement levels. They are holistic, incorporating many kinds of learning and communicating a wide range of learning outcomes including the cognitive, affective, aesthetic and practical, and higher order outcomes of knowledge and skills. They permit a wide range of formal methods of student assessment (i.e., tests and related assessment tasks) and informal methods (i.e., observation and descriptive judgments) to be calibrated and mapped onto a common developmental scale. Using item response modelling, they enable students' performances on different tasks to be compared both from student to student and from year to year. Profiles serve both summative and formative assessment requirements, and include quantitative components that allow data to be aggregated across subjects and/or students. Moderation, or teacher comparison of evidence and justification of judgments, is central to the use of profiles, which in turn supports the professional development of teachers (Griffin et al., 1995).

When teachers in Australia began to use the literacy profiles designed for students whose first language is English, it became apparent that students from other language backgrounds followed a different developmental pathway in their acquisition of English language skills. The *Profiles in English as a Second Language* (Griffin et al., 2003) were developed in workshops with specialist teachers, and empirically scaled and validated to provide robust indicators of performance for these students.

To suit the specific requirements of the evaluation study, the *Profiles in English as a Second Language* were re-validated with a sample of more than 2,500 primary school students in Hong Kong, to confirm their suitability as measures of student performance in this context.

Student achievement was also independently monitored and cross-checked through an oral test of English language (the *Interview Test of English Language*).

*Using student achievement data to inform a program of evaluation.*

The quantitative data of student achievement forms the basis of a much broader program of research in the evaluation of the PNET scheme in Hong Kong primary schools. Indeed, this serves as an illustration of an evaluation of a school reform initiative where student outcomes are used as one indicator of progress, within a more comprehensive framework of interconnecting indicators. The evaluation starts from an assumption that development in student achievement, improvement in student attitudes, changes in teachers' attitudes and teaching conditions, and other contextual variables including characteristics of the school, classroom and home background, are inter-related so that over-emphasising one or other of these areas would compromise the value of the research for the provision of policy advice. Measuring change over time in students' achievement and attitudes would not, in isolation, provide evidence of the success of the reform strategy. Monitoring change and contrast in teacher attitudes, practices and access to resources enables these to be related to changes in student achievement and attitudes. This in turn allows policy and professional development strategies to be identified and recommended to the system and to schools.

In particular, analysis of the quantitative data is expected to indicate schools where relationships between styles of implementation and student outcomes suggest the need to gather more in-depth information through fieldwork and case studies. Schools are chosen to participate in case study analyses where:

- larger than expected proportions of students achieve high outcomes 'against the odds' (i.e., despite the absence of common indicators of student achievement such as well-educated parents who take a keen interest in their children's studies, access to many books at home and school, enjoyment of reading, and opportunities to speak English outside of school);
- aggregated data and value-added analyses indicate that students make significantly more or less progress than their similar counterparts;
- aggregated school data indicate that student performance is consistently above or below the average for their similar counterparts.

Case study data are based upon interviews with teachers and school principals, observation of classroom practice, and the interpretation of teacher diaries and lesson plans. The teachers have been actively engaged as co-researchers in this evaluation study. They have been trained in data collection procedures and moderation of student achievement scores with their colleagues, encouraged to use their students' data to plan teaching strategies and intervention plans for individual students, and to compare their school results against those for the system to reflect upon the progress of reform implementation.

## Conclusion

This report has discussed the requirements of a comprehensive evaluation of the progress of reform strategies, with particular emphasis on the challenge of monitoring implementation of the Victorian Essential Learning Standards. Key issues raised were the imperatives of:

- using professional judgment to identify critical success factors (key indicators) for the reform, and expected levels of performance for those factors;
- testing professional judgment against empirical data; and
- gathering empirical data in a manner that is clear and defensible (i.e., the study design, sampling methods, instruments, data collection, handling and analysis procedures are sound, and claims based on the data analysis can be clearly supported).

Other important considerations included the usefulness of:

- evaluating reforms in terms of outcomes, processes, and impacts on sub-groups;
- monitoring sustainable or long term trends over a sufficient time period;
- gathering both qualitative and quantitative data, and relying upon both case and aggregate data (Patton, 2001);
- documenting impeding and supporting factors, and the unintended outcomes of reform (U.S. General Accounting Office, 1998);
- encouraging the use of methodologies that differ from traditional assessment techniques, including performance assessment, portfolios and computerized modes of delivery (OECD, 2002);
- combining evidence from multiple sources (Fouts, 2003; OECD, 2002); and
- using these to link reforms to indicators of progress

In summary, evaluation of progress in school reforms requires the identification of appropriate indicators of successful implementation, including but not restricted to student academic outcomes, which can be used to comment and reflect upon relationships between styles of implementation and outcomes for students, for sub-groups of students, and from the multiple vantage points of students, parents and teaching professionals.



## References

- Anderson, K.D. (2000). School improvement and school reform in Canada: Whose perspective is it? *Way towards quality in education: International Conference Proceedings*. Brdo pri Kranju, Slovenia, 8-10 April.
- Angoff, W. (1971). *Educational measurement*. Washington: Educational Council on Measurement.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value added assessment of teachers. *Journal of Educational and Behavioural Statistics*, 29(1), NEA Research.
- Barber, M. (2002). *The next stage for large scale reform in England: From good to great*. Paper presented at the Federal Reserve Bank of Boston 47<sup>th</sup> Economic Conference: Education in the 21<sup>st</sup> Century: Meeting the Challenges of a Changing World. 19-21 June.
- Benson, A.P., Hinn, D.M., & Lloyd, C. (2001). *Advances in program evaluation. Visions of quality: How evaluators define, understand and represent program quality*. Oxford: Elsevier Science.
- Bentley, T. (2004). *Coherence beyond control: Essential learning for the 21<sup>st</sup> century*. Seminar report of the VCAA Invitational Seminar on Curriculum Reform: Towards a Framework of Essential Learning, Melbourne.
- Bjerg, J. (1991) Reflections on Danish comprehensive education, 1903-1990. *European Journal of Education*, 26(2): 133-41
- Bock, R., & Wolfe, R. (1996). *A review and analysis of the Tennessee Value-Added System*. Nashville, TN: Comptroller of the Treasury.
- Bracey, G.W. (2004). Serious questions about the Tennessee Value-Added Assessment System. *Phi Delta Kappan*, 85(5): 716-717.
- Bracey, G.W. (2002). Value added, value lost? *Rethinking Schools*, 15(1), Milwaukee, WI.
- Brown, G.T., & Hattie, J.A. (2003). *A national teacher-managed, curriculum-based assessment system: Assessment tools for teaching and learning (asTTle)*. Project asTTle Technical Report 41. University of Auckland/Ministry of Education.
- Carabana, J. (1988). Comprehensive educational reforms in Spain: Past and present. *European Journal of Education* 23(3): 213-28.
- Chubb, J.E., & Moe, T.M. (1990). *Politics, markets, and America's schools*. Washington, DC: The Brookings Institution.
- Clark D L, & Astuto T A (1986). The significance and permanence of changes in federal education policy. *Education Researcher* 15(8): 4-13
- Crane, J. (2002). *The promise of value added testing*. Washington, DC: Progressive Policy Institute.
- Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, 47(2): 108-121.
- DeCoker, G. (2002). *National standards and school reform in Japan and the United States*. New York and London: Teachers College Press, Columbia University.
- Department for Education and Skills (2002). *Education and skills: Delivering results. A strategy to 2006*. Retrieved May 2 from <http://www.defs.gov.uk>.
- Department for Education and Skills (2004). *The Key Stage 1 (KS1) to Key Stage 2 (KS2) value added measure*. Retrieved April 29, 2005, from <http://www.defs.gov.uk>.

- Department for Education and Skills (2005). *School and College Achievement and Attainment Tables*. Retrieved on 10 May, 2005 from <http://www.DfES.gov.uk/performance/tables>.
- Drury, D., & Doran, H. (2003). The value of value added analysis. *Policy Research Brief*, 3(1), Alexandria, VA: National School Boards Association.
- Earl, L. (2004). *Classroom assessment to maximise learning*. Seminar report of the VCAA Invitational Seminar on Curriculum Reform: Towards a Framework of Essential Learning, Melbourne.
- Earl, L., Torrance, N., Sutherland, S., Fullan, M., & Ali, A.S. (2003). *The Manitoba school improvement program: Final evaluation report*. University of Toronto: Ontario Institute for Studies in Education.
- Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership*, 37(10), 15-24.
- Education Commission, Hong Kong SAR of The People's Republic of China (December, 2004). *Progress report on the Education reform (3): Learning for life, learning through life*. Retrieved on 3 May 2005 on [http://www.e-c.edu.hk/eng/reform/index\\_e.html](http://www.e-c.edu.hk/eng/reform/index_e.html).
- Education Review Office, New Zealand (2005). *Framework for reviews in schools*. Retrieved on 6 May 2005 on <http://www.ero.govt.nz>.
- Ellington, L. (2001). *Japanese education in grades K-12*. Retrieved on 5 May 2005 on <http://www.ericdigests.org/2002-2/japanese.htm>.
- Elmore, R. (1995). Structural reform in educational practice. *Educational Researcher* 24(9), 23-26.
- Evans, H. (2005). *Contextual value added model and factors*. Paper presented at the Contextual Value Added Conference for Secondary Schools, Department for Education and Skills. Retrieved on 10 May 2005 from <http://www.defs.gov.uk>.
- Fancy, H. (June, 2004). *Education reform: Reflections on New Zealand experience*. Presented at the Education for Change Symposium, Melbourne.
- Fischer Family Trust (2004). Contextual value added guidance. Retrieved on 8 May 2005 from [http://www.slamnet.org.uk/assessment/Fischerweb/Oct2004\\_Contextual\\_VA\\_Guidance.doc](http://www.slamnet.org.uk/assessment/Fischerweb/Oct2004_Contextual_VA_Guidance.doc).
- Flockton, L. (1999). *School-wide assessment: National education monitoring project*. Wellington: New Zealand Council for Educational Research.
- Fouts, J.T. (2003). *A decade of reform: A summary of research findings on classroom, school and district effectiveness in Washington State*. Lynnwood, WA: Washington School Research Center.
- Foy, P. (May 2, 2005). Utah governor defies No Child Left Behind Act. *Washington Post*. Retrieved on 4 May, 2005 on <http://www.washingtonpost.com/wp-dyn/content/article/2005/05/02/AR2005050201413.html>.
- Fuhrman, S. (Ed.). (2001). *From the capitol to the classroom: Standards-based reform in the States*. Chicago: University of Chicago Press.
- Glaser (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519-521.
- Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, 36, 923-936.

- Glaser, R. (2005). Personal correspondence.
- Goertz, M.E. (2001). Standards-based accountability: Horse trade or horse whip? In S. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the States*, pp. 39-59. Chicago: University of Chicago Press.
- Goertz, M.E., Duffy, M.C., & Le Floch, K.C. (2001). *Assessment and accountability systems in the 50 States: 1999-2000*. CPRE Research Report Series RR-046. Consortium for Policy Research in Education, University of Pennsylvania.
- Griffin, P. (March, 2004) *The comfort of competence and the uncertainty of assessment*. Keynote lecture at the Hong Kong Principals' Conference, Institute of Education, Hong Kong.
- Griffin, P., Smith, P.G., & Burrill, L. (1995). *The literacy profile scales: Towards effective assessment and reporting*. Melbourne: Robert Andersen and Associates.
- Griffin, P., Smith, P.G., & Martin, L. (2003). *Profiles in English as a second language*. Melbourne: Profiles Press International.
- Griffin, P., Tomlinson, B., Martin, L., Adams, R.J., & Storey, P. (2004). *An interview test of English language (ITEL-ed)*. Melbourne: Profiles Press International.
- Hargreaves, D. (1996). *Teaching as a research-based profession: Possibilities and prospects*. London, Teacher Training Agency.
- Hattie, J.A., & Brown, G.T. (2004). *Cognitive processes in asTTle: The SOLO taxonomy*. Technical Report 43, University of Auckland/Ministry of Education.
- Herman, J.L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*. Winter edition.
- Hopkins, D. (March, 2005). *Large scale reform in England: Towards a high excellence, high equity education system*. RM Education in the Future Conference, Manchester. Retrieved on 30 April 2005 on <http://www.edu.yorku.ca>.
- Houston, P.D., Feldman, S., Sava, S.G., Koerner, T.F., & Chase, R. (1999). *An educators' guide to school reform*. Arlington, VA: Educational Research Service.
- Kane, T.J., Staiger, D.O., & Geppert, J. (2002). *Randomly accountable*. Retrieved on 4 May, 2005 on <http://www.educationnext.org>.
- Kawamura, T. (2004). *Reforming compulsory education*. Ministry of Education, Culture, Sports, Science and Technology, Japan. Retrieved on 5 May, 2005 on <http://www.mext.go.jp/english/topics/04091701.htm>.
- Kupermintz, H. (2002). Value added assessment of teachers: The empirical evidence. In A. Molnar (Ed.), *School reform proposals: The research evidence*. Information Age Publishing.
- Kwong, J., & Kool, S. (Eds.) (1990). *Evolution of educational excellence: 25 years of education in the Republic of Singapore*. Singapore: Longman.
- Lewin, K., & Xu, H. (1989) Rethinking revolution: Reflections on China's 1985 educational reforms. *Comparative Education*, 25(1): 7-17.
- Linacre, J. M. (1990). *Many faceted Rasch model*. Chicago: MESA Press.
- Masters, G. (2004). *Standards and assessment that serve essential learning*. Seminar report of the VCAA Invitational Seminar on Curriculum Reform: Towards a Framework of Essential Learning, Melbourne.
- Matters, G. (2004). *From a thousand flowers blooming through clotted cream to an assessment backbone*. Seminar report of the VCAA Invitational Seminar on

- Curriculum Reform: Towards a Framework of Essential Learning, Melbourne.
- McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., & Hamilton, L. (2004). Models for value added modeling of teacher effects. *Journal of Educational and Behavioural Statistics*, 29(1), NEA Research.
- Ministry of Education, Culture, Sports, Science and Technology, Japan (2004). *Reforming compulsory education*. Retrieved on 5 May 2005 on <http://www.mext.go.jp/english/topics/04091701.htm>.
- Organisation for Economic Cooperation and Development (2002). *Definition and selection of competences: Theoretical and conceptual foundations*. Retrieved on 18 May 2005 on <http://www.portal-stat.admin.ch/desecco/news.htm>.
- Patton, M.Q. (2001). Use as a criterion of quality in evaluation. In A. Benson, D.M. Hinn, C. Lloyd (Eds.), *Visions of quality: How evaluators define, understand and represent program quality*. Oxford, Elsevier Science.
- Pedulla, J.J., Abrams, L.M., Madus, G.F., Russell, M.K., Ramos, M.A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. National Board on Educational Testing and Public Policy.
- Pepper, S. (1990). *China's education reform in the 1980s: Policies, issues and historical perspectives*. Berkeley, CA: Institute of East Asian Studies, University of California Press.
- Porter, A.C., & Smithson, J.L. (2001). Assessing reform implementation and effects. In S. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the States*, pp. 60-80. Chicago: University of Chicago Press.
- Qualifications and Curriculum Authority. (2005). *Test development, level setting and maintaining standards*. Retrieved on 11 May 2005 on <http://www.qca.org.uk/12333.html>.
- Quality Assurance Division, Education and Manpower Bureau, Hong Kong SAR (2002). *Performance indicators for Hong Kong schools*. Retrieved on 3 May 2005 on <http://www.emb.gov.hk>.
- Rasch, G. (1980). *Some probabilistic models for the measurement of attainment and intelligence*. Chicago: MESA Press.
- Raudenbush, S. (2004). What are value added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioural Statistics*, 29(1), NEA Research.
- Raudenbush, S., & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59(1): 1-17.
- Reckase, M. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioural Statistics*, 29(1), NEA Research.
- Reynolds, D., & Packer, A. (1992). School effectiveness and school improvement in the 1990s. In D. Reynolds and P. Cuttance (Eds.), *School effectiveness: Research, policy and practice*. London: Cassell.
- Rowley, G. (2003). *National benchmarks: Reporting to parents*. Transcript of presentation delivered on 15 September, 2003. Retrieved on 18 May 2005 on <http://www.vcaa.vic.edu.au/prep10/aim/parents/nationalbenchmarks/nationalbenchmarks.html>.

- Rubin, D., Stuart, E., & Zanutto, E. (2004). A potential outcomes view of value added assessment in education. *Journal of Educational and Behavioural Statistics*, 29(1), NEA Research.
- Rust, V., & Blackmore, K. (1990). Educational reform in Norway and in England and Wales: A corporatist interpretation. *Comparative Education Review*, 34(4): 500-22.
- Rust, K. & Ross, K. (1997). Sampling in survey research. In J.P. Keeves, J.P. (Ed.), *Educational research methodology and measurement. An international handbook*, (2nd edition) (pp. 663-670). Oxford: Pergamon Press.
- Sanders, W. (1998). Value added assessment: A method for measuring the effects of the system, school and teacher on the rate of student academic progress. *School Administrator*. Retrieved on 9 May 2005 on <http://www.aasa.org>.
- Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: UTVARC.
- Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.
- Tekwe, C., Carter, R., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M. (2004). An empirical comparison of statistical models for value added assessment of school performance. *Journal of Educational and Behavioural Statistics*, 29(1), NEA Research.
- United States National Commission on Excellence in Education, Gardner, D.P. (Ed.). (1983) *A Nation at risk: The imperative for educational reform*. Washington DC: US Government Printing Office.
- United States National Governors' Association 1986 *Time for Results: The Governors' 1991 Report on Education*. United States National Governors' Association, Washington, DC
- U.S. General Accounting Office (1998). *Performance measurement and evaluation*. Washington, DC.
- VCAA (2005a). *Validation of the Victorian Essential Learning Standards*. Retrieved on 25 May 2005 from <http://vels.vcaa.vic.edu.au/standardsvalidation.html>.
- VCAA (2005b). *Victorian Essential Learning Standards: Overview*. Retrieved on 26 May 2005 from <http://vels.vcaa.vic.edu.au/VELSoverview.html>.
- Vygotsky, L. (1996). *Thought and language*. Trans. A. Kozulin. MIT.
- White, M. (2005). *Introduction to the Victorian Essential Learning Standards*. Victorian Curriculum Assessment Authority. Retrieved on 18 May 2005 from <http://vels.vcaa.vic.edu.au/MWinterview.html>.
- White, M. (1987) *The Japanese educational challenge: A commitment to children*. Free Press, New York
- Whitehurst, G. J. (2001). *Educational research and improvement*. Washington: United States Department of Education.
- Whitty, G., Powers, S., & Halpin, D. (1998). *Devolution and choice in education: The school, the state, the market*. Buckingham: Open University Press.
- Wisconsin Education Association Council (WEAC) (2004). *Impacts of value added assessments on teachers and schools*. Retrieved on 9 May 2005 on <http://www.weac.org/Resource/2004-05/valueassess/sanders.htm>.

## **Glossary of Terms**

**ACTION RESEARCH:** A reflective and cyclic approach to review and development of projects

**BENEFIT SURVEY:** A study of the intended and unintended benefits of a project

**CHANGE OBJECTIVE:** An objective related to the identification of a need.

**COST BENEFIT:** The gains of the project set against the losses or costs required to achieve the target.

**COST EFFECTIVENESS:** The comparative costs benefits of the chosen strategy and model compared to potential alternatives.

**COST FEASIBILITY:** The cost benefit of the proposed project set against the availability of resources

**ENVIRONMENTAL ANALYSIS:** An analysis of the goals, the resources, the relationships to other agencies, levels of the system and to organisations outside the agency or system.

**EVALUATION – COMPONENT:** An evaluation of the development of key sections of the project.

**EVALUATION – INPUT:** An evaluation of strategies, models and plans established to achieve goals and objectives.

**EVALUATION – PROCESS:** An evaluation of the project as it is implemented.

**FORCE FIELD ANALYSIS:** An examination of facilitating and blocking forces associated with the resolution of problems and action plans developed to overcome them.

**GANTT CHART:** A time line with all major activities plotted along it.

**IMPROVEMENT PLAN:** An action plan which included the strategies, and detailed activities to be implemented in order to implement the improvement policy.

**LOG BOOKS:** Record keeping methods associated with monitoring the development of project elements

**NEEDS ASSESSMENT:** This is a description of the needs of the agency. There are four types of needs which can be described. Each is valid but they have different origins.

- **Comparative needs** are defined from the resources seen to be available from other similar agencies.
- **Defined needs** are those which designated as essential by a central authority and are lacking in the agency
- **Felt needs** may be the same as wants. The clients or staff feel that they need some resource or project.
- **Measured needs** are defined as the gap between ideal states and observed states.

**PERFORMANCE INDICATOR:** Evidence that an objective has been achieved.

**PERT CHART:** A critical path diagram which shows the relationships between activities and events in the project, and the amounts of time needed to complete the project successfully.

**STATUS SURVEY:** An assessment of the current levels of achievement of the agency with respect to the goals set for its success.

**SWOT ANALYSIS:** This is simply a review of the strengths, weaknesses, opportunities and threats of the agency.

---

<sup>1</sup> For a comparison of whole-school programs of reform in the United States, refer to Houston et al., 1999.

<sup>22</sup> Reliability refers to the likelihood that a test would produce the same results if repeated with the same students. It acknowledges that tests are imperfect measures of complex constructs, and is used to flag the estimated amount of measurement error in a test. Validity is a judgment of whether or not the test actually measures the construct it purports to measure. Thus, do tests of reading, writing and mathematics used by many systems really measure proficiency in reading, writing and mathematics, or are they also measuring some form of rote-learning or general knowledge? To extend this idea, can it be argued that the tests of reading, writing and mathematics validly measure student proficiency across complex learning domains?

<sup>3</sup> Refer to the section in this report on value added analyses in the United States (pp 24-26) for an expanded discussion.

<sup>4</sup> Detailed information about NAEP can be accessed at <http://nces.ed.gov>.

<sup>5</sup> Since 2003, Maryland has abandoned its performance-based assessment program to allow it to conform to the accountability requirements of the No Child Left Behind Act.

<sup>6</sup> Refer to page 70 of this report for an expanded discussion of the Maryland performance-based assessment program.

<sup>7</sup> A detailed description of SAIP can be accessed at <http://www.cmec.ca>.

<sup>8</sup> For a comprehensive discussion of the constraints upon value-added analyses and other methods of tracking student academic achievement over time, refer to pages 13 and 24-26, and to Raudenbush (2004) and Tekwe et al. (2004).

<sup>9</sup> An illustration of the use of student outcome data as part of an evaluation study is given in this report on pages 72-74 (the evaluation of the Hong Kong Primary Native-speaking English Teacher Scheme).

<sup>10</sup> Italics added.

<sup>11</sup> This discussion is expanded on pages 65-68 of this report.

<sup>12</sup> This point is particularly relevant for the evaluation of student outcomes in terms of VELs.

<sup>13</sup> Refer to the discussion of the evaluation of the Hong Kong PNET scheme (pp. 72-74) for an example of the use of profiles to describe student outcomes.

<sup>14</sup> Refer to the discussion of the Maryland School Performance Assessment Program (pp. 21-23, 70-72).

<sup>15</sup> Refer to the discussion of the use of value added analyses in the evaluation of the Primary Native English-speaking scheme in Hong Kong on page 74 of this report.

<sup>16</sup> The National Education Monitoring Project is described in more detail on page 30 of this report.

<sup>17</sup> A full description of the MSPAP can be retrieved from <http://mdk12.org/mspp/mspap/index.html>.

<sup>18</sup> Refer to the website for the Maryland School Improvement Program, <http://mdk12.org/> for more information.