Assessing Student Progress at the National Level

Jeffrey K. Smith
Lisa F. Smith
University of Otago
9 August, 2008

Assessing Student Progress at the National Level

Jeffrey K. Smith
Lisa F. Smith
University of Otago

Executive Summary

National and state/provincial assessment programmes are commonplace throughout the world, but they are realised in a wide variety of fashions. Based on an analysis of eight programmes at the state and national levels, this paper examines some of the ideas underlying major assessment programmes, and makes recommendations for how the current configuration of assessment programmes in New Zealand might be made more effective.

National assessment programmes vary on a number of important dimensions. Among these are the stakes involved, whether programmes are mandated, the format of the assessments used, what years and subjects are assessed, and how comparisons are made across year levels of students and over time. Each of these variations has a consequence for how the programme is received and how it affects the educational process.

The Assessment is for Learning programme in Scotland and the New Jersey Statewide Assessment Program are examined in detail. The Scottish programme is very similar to the one in use in New Zealand today, and it is therefore interesting to review the variations in that programme that might be informative to New Zealand's efforts. The New Jersey programme is instructive as well, but more as a cautionary tale. New Jersey's programme is high stakes, costly, and does not appear to have had any perceivable beneficial effect in terms of student achievement.

We move from this analysis to a consideration of what might be done going forward in New Zealand. We are particularly interested in the potential of combining work done by the Ministry of Education in Learning Progressions with the existing assessment programmes of NEMP, asTTle, and AtoL. Other programmes may be able to join such an effort, as well. The goal of the collaboration would be to empirically verify the Learning Progressions and develop sets of assessments that would accompany them. This would result in a better understanding of the progressions, development of useful information on how they might be used in classrooms, and would provide a mechanism for monitoring progress in the progressions at a national level.

The model is presented as one possibility for how the assessment programmes might work in a collaborative fashion to enhance national assessment. Other approaches are possible and can, and should, be explored.

Introduction

Assessment and testing have historically been used for a wide variety of purposes (Haertel & Herman, 2005; Linn, 2001). These purposes include putting students into different educational tracks; monitoring progress of students at the individual, class, school, district, regional, national, and international levels; awarding degrees and other levels of qualification; diagnosing learning weaknesses; determining whether students should be assigned to academic or vocational tracks; making admissions decisions to university; promoting students from one grade to the next; and, as a source of political debate. Linn (2006) added to this list with the inclusion of clarifying expectations for teachers and learners, motivating greater effort by all involved, and providing a basis for rewarding schools that do well and punishing those that do not.

In the past twenty years or so, many nations have started or increased their national assessment programmes (Phelps, 2000). There are commonalities and singularities among the various approaches that different nations have adopted. The purpose of this paper is to look at some of the things that are being done internationally in terms of national assessment programmes, and then to look at what options might be available in the context of New Zealand education. In particular, we examine the following issues:

- What are the variables/dimensions of national assessment programmes?
- What are some of the characteristics of programmes in nations that have some similarity to New Zealand?
- What are the purposes and consequences of different characteristics or approaches to assessment at a national level?
- Might it be possible to combine data from assessments at the national level with theoretical and substantive knowledge in various subject areas to better understand how children progress in general, and in particular how New Zealand children are doing?
- Can a national assessment programme also be a programme that has a research base that informs learning at the classroom level?

The paper begins with a discussion of the characteristics or dimensions that one might consider in looking at national assessment programmes. Then, a number of national and US state programmes (under No Child Left Behind) are considered. Two of these programmes, from Scotland and New Jersey, are considered in depth. Scotland was chosen because of the similarity of its programme to New Zealand, and because they are engaged in a number of activities that might be useful to New Zealand. New Jersey was chosen because it represents a fairly typical approach to mandated testing, and because of the authors' familiarity with the programme. Then, we turn to the final two issues under consideration: first, how New Zealand might combine theoretical and/or substantive ideas along with statistical results to form a stronger and more pragmatic notion of assessment; and second, how a national assessment programme might incorporate an active research and development component to address current issues of importance, as well as to provide a strong basis for investigating innovative ideas that would potentially have significant impact on education in New Zealand.

The Dimensions of National Assessment Programmes

National assessment programmes have been in existence in many countries, or states/provinces within countries for over thirty years in the western world. In some countries (e.g., United States, Australia, Sweden, France), programmes are well-established if continually evolving (Wolf, 2002). In other countries, national assessment is either lacking or more recently initiated (e.g., Ireland, Denmark), or comprises of a mix of high stakes assessments at some levels and very little assessment at all at other levels (Eglend, 2005; Looney, 2006). In addition to differences in history and current level, there is also wide variation in the realisation and execution of assessment programmes. A review of national assessment programmes in a variety of countries has led us to develop a set of twelve dimensions that might be used to effectively characterise the important distinctions among assessment programmes.

1. Stakeholders/Development

The first dimension has to do with who is considered to be an important stakeholder in the development process of assessment programmes. In many countries and states, the agency in charge of education is also in charge of assessment (the ministry of education, the department of education, etc.). In some countries, such as the US and Canada, there is a certain level of direction of assessment activities that occurs at the national level, and a certain level that has been devolved to the state or provincial level. As the assessment programme is developed, the agency charged with its development may or may not include various groups in the discussions and deliberations that lead to the final product. These groups might include politicians/office-holders, members of the public, parents, educators, special interest groups, the business community, for example. The level of influence over the development of the programme by these groups, alone or in combination, can have substantial impact on what the final programme looks like.

In Nebraska, for example, the State Department of Education developed a highly innovative approach to responding to the mandate of NCLB, which involved devolving responsibility for designing the assessment system down to what might be considered the most basic educational level, that of schools and teachers. This approach was known as the Nebraska STARS programme (Standards-Based Teacher-Led Assessment System) (Bandelos, 2004; Plake, Impara, & Buckendahl, 2004). In STARS, the responsibility for the development and nature of the approach for assessment was placed entirely in the hands of schools and teachers and called for using locally chosen or developed assessments instead of a statewide mandated test, which was the approach taken by other states. How did it work out? Nebraska abandoned the system following pressure from a variety of sources, including the federal government, to have a more traditional approach (Impara, personal communication, 2008).

2. Purpose

One of the most interesting and important aspects that the review of various assessment programmes revealed was the wide variety of purposes for such programmes. Some of the notion of purpose has to do with subjects and levels

covered and will be discussed in more detail below. But, at the heart of the issue lies the question: What is this programme for? In many, perhaps most cases, the purpose is accountability. Interestingly, accountability has a pejorative connotation to it that is probably appropriate in most instances. That is, the purpose of many national assessment programmes is not so much to see how the nation is doing, but to hold some group or groups responsible if the results are not satisfactory. Herman (2007) and Darling-Hammond (2006) have argued for a broader conceptualisation of accountability that expands the notion to include capacity building and a concern for the welfare of others. But this more noble definition is not currently at the heart of most approaches. An accountability rationale as it is typically encountered might be contrasted to a programme whose primary purpose is to see how well the nation is doing as a whole, where the strengths and weaknesses lie, and what to work on next. We see such this type of purpose in the National Education Monitoring Project (NEMP) in New Zealand and in the Scottish Survey of Assessment, discussed in more detail below.

Some national assessment programmes have distinct purposes other than accountability. The NCEA programme in New Zealand has as its primary purpose the awarding and accumulation of certification of achievement in various school and life competencies. The SATs in the US and the national testing programme in China are used in the transition of students into institutions of higher education. Thus, these programmes have selection as their primary purpose. Still other programmes are diagnostic in purpose, such as readiness programmes for school entry.

3. Basis

Much of what is seen in terms of assessment programmes is called standards-based, but it is not always clear what that means. Many of the programmes that claim to be standards-based utilize psychometric models appropriate for norm-referenced testing in their programmes, or are adaptations of such programmes. A number of programmes use multiple approaches to assessment, some for credit or qualifications, others for monitoring within the same programme.

4. Level of mandate

Most national assessment programmes have at least some level of requirement or mandatory application. Many programmes are a mix of mandatory components and voluntary components. The mandatory components tend to be those that are controlled by the state or national level, whereas the voluntary components are controlled at the school or classroom level. Mandated programmes are not necessarily high stakes programmes, although it is the case that most mandated programmes are high stakes.

5. Stakes

One of the most important and controversial aspects of national assessment programmes has to do with the stakes that are involved in the programme. Often, the stakes are a consequence of how the results are reported, as opposed to anything that is expressly mandated. It is also important to consider that the issue of stakes exists at a variety of different levels. For example, in most assessment programmes in the United States that are required by the No Child Left Behind (NCLB) legislation, there

is little or no consequence of performance for the individual student. Students are almost never retained in a grade due to poor performance on the statewide assessment, certainly never solely because of performance on that assessment. Teachers may exhort students to do their best, but that is primarily due to the consequence of the assessment for the teacher and the school. Thus, testing under NCLB is frequently low to medium stakes for the student, but high stakes for the teacher, the school, and the school district. The nature of the legal ramifications of the NCLB legislation with regard to stakes is exemplified in the discussion of the New Jersey programme below.

Nations with multiple components to their national assessments may see different levels of stakes with the different components. Furthermore, the stakes associated with the assessments may be an artefact of the publicity and reporting that accompanies the programme, rather than the legal requirements of the programme. NCEA is a good example of this. The stakes of the programme are high for the students by design. That is, a student's receipt of credits toward qualifications is dependent upon individual performance on the assessment. However, the stakes at the school level are more attributable to the publication of league tables in the media and the consequence that such publication may have on the reputation of the school and its ability to attract students.

6. Levels

An important decision concerning national assessment programmes concerns which years or grades will be assessed in the programme. This is often closely tied to the purposes of the assessment programme. Nations that are more interested in a monitoring purpose are more likely to sample in alternate years, or even less frequently. Programmes that have accountability and increase in student growth as the rationales for the programme are more likely to assess more frequently, sometimes every year.

7. Subjects

Determining which subjects or content areas will be assessed in the programme is one of the most critical decisions that is made, and holds the potential for the most far-reaching effects. Subjects that are not included in the assessment programme are often discounted within the realised curriculum of the school. Not being included in the assessment programme can be tantamount to being eliminated from the curriculum. One of the authors of this report served on a committee charged with examining the feasibility of an assessment in fine arts as part of the New Jersey assessment programme. The members of the committee from the arts education community demanded (ultimately unsuccessfully) that there be a multiple choice test in the fine arts in the statewide testing programme, rather than a performance assessment. Their argument was that if there were no fine arts assessment that looked like the assessments in reading and mathematics, then fine arts would ultimately cease to be part of the curriculum.

8. Sampling

There are several approaches to sampling students for inclusion in an assessment programme. The first, and simplest, is to mandate that assessment occur

for all children. A second approach is often referred to as a light sampling approach, where a representative sample of students is assessed. This approach is used in the National Education Monitoring Project in New Zealand. Another approach is called a matrix sampling approach. This is a bit complicated but basically involves sampling at both the student and the item level. Not all students take all items; instead they take a sample of items. The items are rotated among students such that each item is taken by a representative sample of students. The National Assessment of Educational Progress in the United States uses an approach of this type. For many programmes, sampling is done in a "norming" study that tries to develop a sample of examinees that is similar to a national sample of students. When the assessment is taken, performance is related back to the results from the norming study in order to report norm-referenced scales (such as percentiles, stanines, etc.).

9. Format

The format of the items on an assessment varies considerably among countries, and is related to issues of the validity of the assessments, as well as the cost of the assessments to administer and score. In the United States, the multiple choice format dominates statewide assessment programmes mandated by NCLB. Multiple choice items have many admirable qualities, including objectivity of scoring, the ability to administer a large number of items in a relatively short period of time, coverage of a wide variety of objectives or standards, and low cost in terms of administration and scoring. On the other hand, there are many weaknesses to multiple choice items. They are often thought to be artificial in terms of what is required of students, and may suffer from a focus primarily on lower level objectives (knowledge and recall). Because the correct answer has to be provided to examinees on multiple choice items, the nature of the task for multiple choice items is frequently considered to be one of recognising correct answers instead of generating them (although this criticism may be overstated to a degree). Furthermore, guessing on items reduces the reliability of information that can be extracted on any given item.

A popular alternative to the multiple choice format is the constructed response format. Constructed response is a very broad term that basically means that the student has to generate a response. The term is usually reserved for short answers, such as those that would be produced from solving a mathematics problem that has a correct answer. Essay assessments are another approach that is widely used in national assessment programmes.

Some assessment programmes, particularly those that employ a sampling design as opposed to assessing all students, use performance assessments, with student responses being recorded in some fashion (e.g., paper and pencil, products, audio or video tape, interviews), and then marked at a later date. This is a very rich source of information, but can be expensive to administer and score.

10. Equating/comparability/technical quality/stability

Most assessment programmes involve some level of comparability over time. This can be accomplished through statistical equating or through the retention of some items/tasks from an initial administration to a later administration for purposes of comparison. Most high stakes assessment programmes used for accountability purposes examined for this paper use statistical equating of assessments both

horizontally (equating this year's assessment to last year's assessment at the same grade or level) and, to a lesser degree, vertically (equating performance on the Year 4 assessment to performance on the Year 5 assessment). At some level of analysis, this equating is not difficult to accomplish, especially if the assessment is not high stakes. But when new assessments have to be developed on a yearly basis and when the stakes for performance are high, requiring security of the assessments, a number of complications arise. Issues of test equating and test security have become two of the three most popular topics (along with standard setting) in the annual meeting of the US National Council for Measurement in Education.

11. Quality

Efforts to ensure the quality of the assessments used in the national assessment programmes reviewed for this paper was one of the most interesting findings in looking at the programmes considered. Although there is an extensive literature on how to conduct validity analyses for assessments, and although the assessment programmes examined all used well-trained assessment specialists, there was little to no information provided concerning the validity of the programmes. Additionally, in many of the programmes, there was little to no information on reliability. Programmes that make decisions about the competence or lack of same in individual students do not regularly report the standard error of measurement of their programmes – nor can this information be garnered typically by looking at technical reports. In most cases, information concerning the quality of the measures being used simply does not exist.

12. Reporting

The final topic for consideration has to do with reporting. In some programmes examined for this paper, results were only presented at the national and regional levels (including demographic characteristics of the students such as gender and ethnicity) in a similar fashion to the NEMP programme. Other assessments returned the original examination or script to the students to allow for an item-by-item analysis, in a fashion similar to the NCEA in New Zealand, or the SATs in the United States. The level of reporting is just one issue of importance here; the timing of the reporting is another. Many national assessment programmes that report at the student level take so long to do so that the students are no longer with the teachers who had them when the assessment was administered, making it difficult for the assessments to be of much use.

Looking at the Nature and Consequences of Two Assessment Programmes

For purposes of this paper, we examined national assessment programmes in Scotland, Canada (at the provincial level), Sweden, Ireland, and at the statewide level under NCLB in Maryland, Nebraska, Michigan, and New Jersey. It seemed to us that rather than reporting briefly on each of these entities, it would be more beneficial to look at two in more depth. We chose Scotland because of the promise that we think the Scottish approach holds for New Zealand, and New Jersey, because it is fairly typical of programmes in the United States and because we know it best of the state programmes. Information from the remaining programmes will be referred to as appropriate.

Scotland

Scotland may be the country most similar to New Zealand in many respects (especially for those of us who live in Dunedin), including their approach to national assessment. Their overall approach to assessment is reflected in the name given to their programme: Assessment is for Learning. Scotland has basically three major components to their national assessment system. First, there is a qualifications system at the end point of schooling, with students needing to demonstrate competence in various areas in order to earn their qualifications. Thus, the assessment programme is high stakes for students at the secondary school level. Then, there is the sampling of achievement at Primary 3, Primary 5, Primary 7, and Secondary 2 levels. These are low stakes assessments that are used for monitoring and accountability at the national level. The third element is the National Assessments, which are used in classrooms and combined with local measures and teacher judgements to monitor achievement.

National Qualifications: Starting with students in high school, there is a National Qualifications system that is quite similar to New Zealand's NCEA. Assessment includes both internal and external assessments; students can gain credits in a wide variety of areas to suit their individual needs and goals; and, there are various levels of performance possible. The programme is currently undergoing an intensive re-examination and upgrading. The statement for public consultation on the new programme states that all children should:

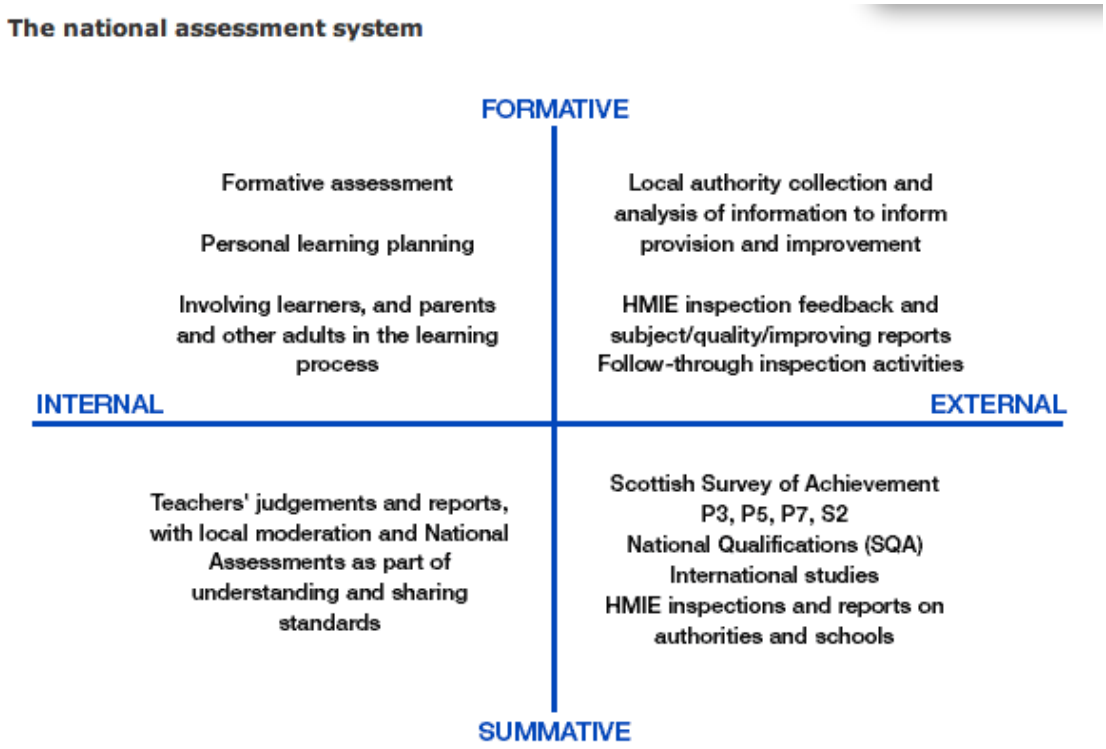
"…*benefit from an assessment system that supports the curriculum rather than leads it and ensures that their transition into qualifications is smooth"* (The Scottish Government, 2008, p. 5).

Scottish Survey of Achievement: This is an assessment programme given at year levels P3, P5, P7, and S2, in the areas of reading, writing, science, and mathematics. It is somewhat analogous to the NEMP programme in New Zealand. The survey is based on a sample of schools and is administered yearly. Some of the assessment material is re-used in subsequent assessments in order to provide for comparisons over time. The programme involves paper and pencil assessment as well as some practical assessments. Questionnaires for teachers and students are also involved. Additionally, information on the educational attainments of students are requested from the schools along with examples of classwork that have been marked by the schools. The programme is designed to provide information about the strengths and weaknesses found in the schools as well as information about instructional practices and the students' broader social milieu.

National Assessments 5-14: These are assessments in literacy and mathematics developed by the government that teachers can download and use in their classrooms for assessment purposes. They are for classroom use and are intended to supplement and support the judgements of teachers. Students who correctly respond to roughly two-thirds of the questions at a given level are thought to have mastered that level and are ready to move on to the next. Although there are differences, the National Assessments, which are intended for use by teachers, are similar in purpose to the asTTle programme in New Zealand.

Analysis: According to government websites, the national assessment programme can be conceptualised as existing along two fundamental axes, one concerning whether assessment is internal or external, and a second concerning whether the assessment is formative and summative. These axes form four quadrants, into which the government places the assessment activities that occur at the local and national levels. A schematic taken from the website, http://www.scotland.gov.uk/Publications/2005/09/20105646/56474, is presented in Figure 1 below.

Figure 1
National Assessment System of Scotland



What is shown here is a programme that tries to integrate what occurs in the classroom to facilitate learning (upper left quadrant) with summative teacher judgements and the use of National Assessments (lower left quadrant), and information collected by external authorities, both local and governmental, for developmental purposes at the local level (upper right quadrant). Finally, there is the summative and external assessment, such as the Scottish Survey information and the National Qualifications that are used as external and summative assessments in the system (lower right quadrant).

Hutchinson and Hayward (2005) provided an excellent discussion of how Scotland got to where it is today (or at least, up to 2005) in terms of assessment. Of particular importance in their discussion is the influence that teachers and other educators have had on the development of the programme, as well as the commitment in a variety of sectors to open debate on the issue of assessment, and the willingness to make adjustments and modifications in the system as evidence indicated might be necessary. Finally, it was interesting to see how the Scottish approach included a range of stakeholders engaged in serious discussions about the development of the

programme, and how classroom teachers actively participated its development. This appears to be a model that is high on trust, which would seem to be an essential element in a successful programme. (It should be noted the Carol Hutchinson has been participating in the NZ assessment review and can provide a more extensive and up to date analysis of the progress being made in Scotland.)

Of particular interest for consideration in New Zealand might be an adaptation of the combination of the National Assessments that are used in classrooms, when teachers think that children are ready for them, and the Scottish Survey of Achievement. The National Assessments allow for an ongoing picture of progress that starts at the student level and can be aggregated to whatever higher level is desired. The Survey serves as an extension of assessment to all areas of the curriculum, and provides a low stakes, external validation of the National Assessments.

New Jersey

New Jersey in many respects represents a typical US state-level response to the mandates of the federal No Child Left Behind act. Under NCLB, each state has to design its own system for assessing all students in grades 3-8 (years 4-9), and once at the high school level. In New Jersey, the 3-8 assessment component is called the New Jersey Assessment of Skills and Knowledge (NJASK) and at the high school level, it is called the High School Proficiency Assessment (HSPA). New Jersey was one of the first of the US states to develop standardized testing; their programme began in the mid 1970's, well before any federal mandates (State of New Jersey, Department of Education, 2008a). The programme has had many political ups and downs over the decades, as education and educational testing have become active political footballs, particularly in gubernatorial races. Part of this political history has led to a fairly low level of trust between the public in New Jersey and the schoolteachers of New Jersey, who are viewed by large segments of the population as unionists who do not have the best interests of children at heart. Additionally, New Jersey Supreme Court rulings over the past twenty years have led the New Jersey State Legislature to spend a rather large amount of money on education. As a consequence of these and other issues, it would be fair to say that the there is a sense of distrust and misgiving between New Jersey's schools and the public.

The current assessment programme in New Jersey is a direct result of the No Child Left Behind legislation. New Jersey uses a standards-based approach to assessment, which means that there is a set of standards for each area assessed, and the assessments are built from those standards. These are extensive documents. The standards for K-12 mathematics alone are 47 pages, single-spaced. The standards are so extensive that New Jersey has developed a "Standards Clarification Project," which consists of over twenty documents and several training videos. The standards for New Jersey were developed by committees of teachers, university faculty, State Department of Education officials, and members of interested groups (including business and community leaders). But the standards that were produced were primarily organized by smaller segments of those committees, and in large part echo the standards developed by national organizations such as the National Council of Teachers of Mathematics. Once the standards had been developed, the State Department of Education solicited bids for developing the actual tests and for running

the testing programme. New Jersey's programme is run by Measurement Inc. in combination with Harcourt Assessment.

NJASK assessments are given in mathematics, reading, and writing at grades 3 through 8, science at grades 4 and 8. The HSPA assesses mathematics, reading, and writing at grade 11. Passing the grade 11 test is required for high school graduation. Three scores are possible for each of the tests: partially proficient, proficient, and advanced proficient. A score of "partially proficient" means that the student did not pass the test. Scores are reported at the level of the student, the school, and the district. They are also separated out by gender, socio-economic status, ethnicity, home language, and special education classification. The NCLB legislation mandates that yearly growth toward 100% passing rates be made by each school and in each of the classifications (referred to in the legislation as Adequate Yearly Progress, or AYP). Failure to reach prescribed progress (determined by each state subject to federal approval) leads to a series of actions that start with support for a given school and move toward sanctions. These sanctions are from the federal legislation and apply to all states. The list of sanctions is presented in Figure 2 below.

Figure 2
Progressive Sanctions under NCLB

NCLB/Title I
School Improvement Continuum Chart

Year	Status	Interventions for Title I Schools
Year 1	**Early Warning** – Did not make AYP for one year	None
Year 2	First year of **school in need of improvement** status. Did not make AYP for two consecutive years in the same content area.	Parent notification, public school choice (or supplemental educational services), school improvement plan, technical assistance from district.
Year 3	Second year of **school in need of improvement** status. Did not make AYP for three consecutive years in the same content area.	Parent notification, public school choice, supplemental educational services, school improvement plan, technical assistance from district.
Year 4	Third year of school in need of improvement status – **corrective action**. Did not make AYP for four consecutive years in the same content area.	Parent notification, public school choice, supplemental educational services, school improvement plan, technical assistance from district and state, corrective action, participation in CAPA.
Year 5	Fourth year of school in need of improvement status – **school restructuring** plan. Did not make AYP for five consecutive years in the same content area.	Parent notification, public school choice, supplemental educational services, school improvement plan, technical assistance from district and state, development of restructuring plan (governance).
Year 6	Fifth year of school in need of improvement status – **implementation of restructuring plan**. Did not make AYP for six consecutive years in the same content area.	Parent notification, public school choice, supplemental educational services, school improvement plan, technical assistance from district and state, implementation of restructuring plan.
Year 7	Sixth year of school in need of improvement status – **implementation of restructuring plan**. Did not make AYP for seven consecutive years in the same content area.	Parent notification, public school choice, supplemental educational services, school improvement plan, technical assistance from district and state, implementation of restructuring plan.

In addition to the pressures coming from potential NCLB sanctions, results for the testing programme each year are published in league tables in the State's news

media. The State also makes the results available to the public via its website on a school level basis. In Figure 3, the results for a school from an urban school district are presented. This school is currently at level 3 in the sanctions from the federal government. The results contrast school performance to district and national results.

Figure 3
Results for an Urban School in New Jersey

Adequate Yearly Progress (AYP) Status	
School made AYP:	NO
School classified as "in need of improvement":	YES
"In need of improvement" status:	3.00
District classified as "in need of improvement":	YES

Attendance Rates			
	School	District	State
2006-07	94.7%	94.7%	94.4%
2005-06	95.0%	93.9%	94.5%
Target for AYP 90%			

SCHOOL DATA COMPONENTS

Elementary Grade Span

LANGUAGE ARTS LITERACY		Year	Percent Not Tested	Proficiency Percentages Partial	Proficient	Advanced
State Standard for AYP = 75% Proficient						
All Students	School	2006-07	0.5%	52.7%	46.7%	0.5%
		2005-06	0%	47.2%	52.8%	0%
» details for subgroups for LAL	District	2006-07	0.3%	36.9%	61.8%	1.3%
		2005-06	0.5%	39.9%	58.9%	1.1%
	State	2006-07	0.2%	14.4%	75.5%	10%
		2005-06	0.2%	15.7%	78.2%	6%
The state standard for Adequate Yearly Progress (AYP) for language arts literacy is 75% proficient for the school and each subgroup.						

MATHEMATICS		Year	Percent Not Tested	Proficiency Percentages Partial	Proficient	Advanced
State Standard for AYP = 62% Proficient						
All Students	School	2006-07	0.5%	54.6%	40.5%	4.9%
		2005-06	0%	47.7%	46.2%	6.2%
» details for subgroups for MATH	District	2006-07	0.3%	30.2%	55.7%	14.1%
		2005-06	0.6%	28.9%	57.8%	13.3%
	State	2006-07	0.2%	13%	51.1%	35.9%
		2005-06	0.2%	14.6%	50.5%	34.8%
The state standard for Adequate Yearly Progress (AYP) for mathematics is 62% proficient for the school and each subgroup.						

Taken from: http://education.state.nj.us/rc/nclb07/reports/23/3530/23-3530-100.html

New Jersey's assessment programme is high stakes for schools and teachers, and at the high school level, for students as well. There is no separate high school qualifications programme; this is typical for most states. Comparability of achievement in subject areas from one high school to the next has never been of particular interest in the United States. College entrance examinations, the SATs and the ACTs, provide a level of equivalence across students. On the other hand, the

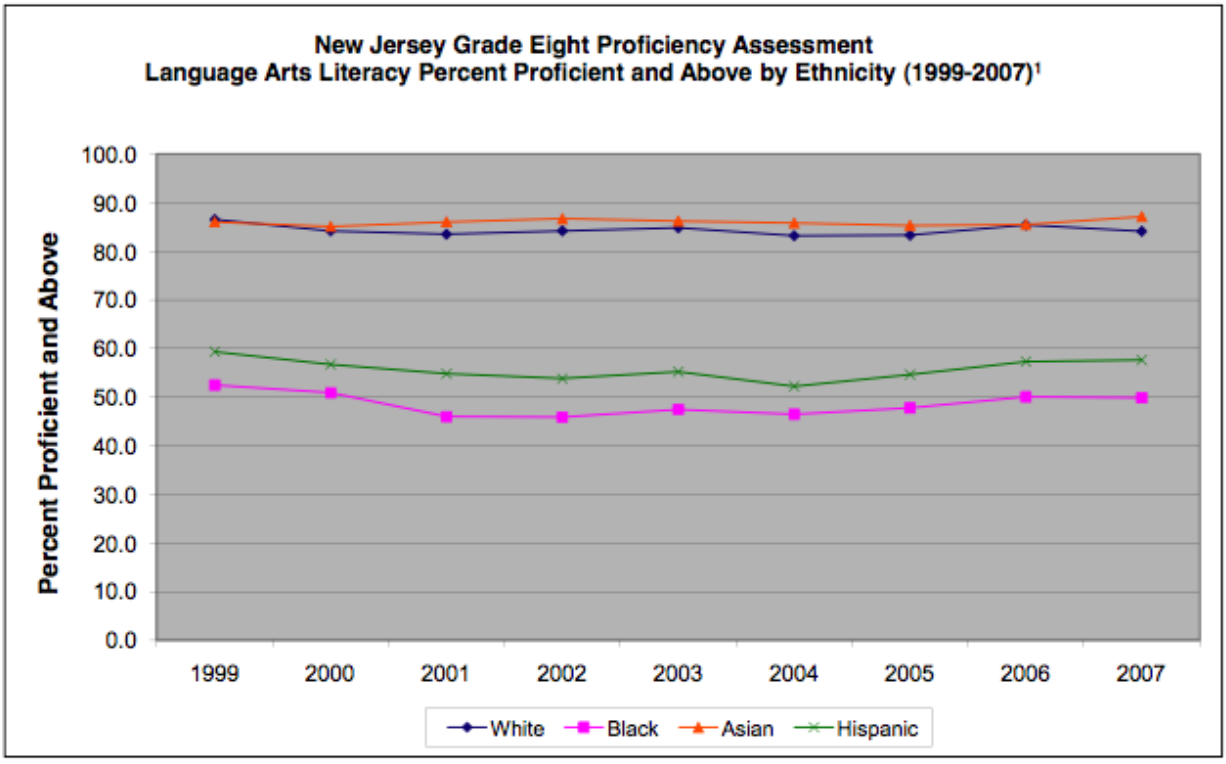
public is very much concerned with comparing the schools themselves; thus, the league tables at the level of the school are published each year.

Another major concern in New Jersey, and in the US in general, has to do with growth over years. In the presentation of scores presented in Figure 3 above, it can be seen that growth is a major concern. In fact, it is the essential basis of the argument for AYP in NCLB. The need to make comparisons on a yearly basis in a high stakes programme is met by developing new assessments each year and linking them statistically via item response theory to the previous year's tests. Previously used forms are not released. The tests are primarily multiple choice (except in the writing assessment), with a few constructed response questions.

Analysis:

New Jersey is no exception in the United States with regard to reactions to NCLB. Although national results in terms of overall student growth are mixed as best, there is no question that NCLB and the testing requirements that accompany it dominate education in the US. As mentioned, New Jersey's approach is a fairly typical one. With more experience in high stakes statewide mandated testing than most states, New Jersey may have adjusted to NCLB mandates more quickly than some states. Indeed, in the National Assessment of Educational Progress testing that serves as a federal level check on the states' programmes, New Jersey is one of the best performing states. This is somewhat ironic in that the state hasn't made much progress in performance over the past decade. Figure 4 shows the change in the Language Arts/Literacy area from 1999 to 2007 broken out by ethnicity.

Figure 4
Progress in New Jersey in Eighth Grade Language Arts by Ethnic Group (1999-2007)



Taken from: http://education.state.nj.us/rc/nclb07/reports/23/3530/23-3530-100.html

New Jersey's assessment programme, both in its current configuration, and over the previous three decades, has had a profound influence on the educational practices of New Jersey educators. Firestone, Monfils, and Schorr (2004) analysed the behaviour of fourth grade teachers in teaching mathematics in urban districts in response to the State's testing mandate. They found that when teachers felt that they understood the ideas and principles behind the standards and goals of mathematics instruction, they used more enquiry-based instruction and tried to integrate their regular instruction with efforts to prepare the children for the tests. But when they felt pressure to obtain good scores and didn't understand the underlying processes well, they engaged in a variety of "teaching to the test" activities. The pressure of the testing programme associated with NCLB nationwide has led not only to teaching to the test, but also to widespread instances of teachers engaging in various forms of activity that can best be described as cheating (Cizek, 2001, 2004).

What might be possible for New Zealand?

This final section of the paper examines what might be possible for New Zealand if knowledge about literacy and numeracy from a theoretical and practical perspective were blended with what we know and can learn from national assessment programmes. The concept of blending substantive knowledge about learning with empirical data from assessment was not something we found as an explicit model in any of the national programmes we looked at, but there are some approaches to assessment that might be seen, at least in part, as similar to this idea. The State of Michigan is exploring a programme that links assessments to growth in learning, but their approach appears to be very focused on making sure items align with standards, and that cut scores have both a substantive and empirical base (Martineau, Pack, Keene, & Hirsch, 2007).

The assessment programmes that we currently use in New Zealand operate independently of one another to a large degree. In particular, asTTle, NEMP, and AtoL are autonomous, but a coordinated programme might utilise all of these and perhaps also PAT, MIDYIS, and other efforts to inform the national picture. The idea here is to explore the feasibility of combining the strong components of assessment that exist in New Zealand today into a more synergistic approach in a national assessment framework. The approach presented here imagines the utilization of a number of extant assessment capabilities in New Zealand. It is primarily focused on assessment in years 1-10, as, at year 11, the NCEA programme becomes the primary approach.

Currently, asTTle provides information for use in schools and classrooms, and to a degree, at the national level, utilizing the strengths of an extensive item bank that has been linked horizontally and vertically throughout much of the fundamental subjects of the curriculum and across a number of years. NEMP provides in-depth information in most areas of the curriculum, along with a highly accurate picture of national performance based on its stratified random light sampling approach. The mission of AtoL is quite different from asTTle and NEMP, as it works directly with teachers to enhance their ability to use assessment effectively, rather than conducting

assessments as part of its mission. But it, too, can rightly be considered to be a national assessment programme.

The question we pose is what might be possible if a vehicle existed that allowed for these programmes to work, at least in part, collaboratively? There is certainly overlap in their activities to a degree in the current configuration, but not really a systematic effort that utilizes the strengths of each group for the good of the whole. This idea of a collaborative effort stems from a concern for being able to communicate more directly and clearly to the general public about progress in education in New Zealand. But it moves beyond that beginning to the potential realisation of an approach to assessment that makes use of an active research and development component with the express goal of helping the nation reach its educational goals.

Learning Progressions

We start with the recent development of "learning progressions" (Ministry of Education, 2007), and in particular, literacy learning progressions as an example. Literacy learning progressions are broad statements of what children who are making appropriate progress in school ought to be able to do at various years in school. An example of a literacy learning progression in reading for the end of year 4 is:

"As they read, students build on their expertise and demonstrate that they can work out the meanings of unfamiliar phrases and expressions (e.g., figures of speech) by using their prior knowledge and the context."

And, in writing at year 8:

"As they read, students build on their expertise and demonstrate that they can plan effectively for writing by selecting an appropriate text form to match the writing purpose and audience and by using strategies such as mind mapping to aid planning."

These learning progressions seem to be an excellent starting point for an effort to blend theory, empirical data, and practice into a coordinated approach to national assessment, one that emphasizes a continual programme of research, national monitoring, and best practice in classrooms. One of the distinct advantages of the learning progressions is that they are broadly stated, yet clear enough to provide solid advice and information to those trying to assess students' abilities in these areas.

Admittedly, the learning progressions are somewhat in their infancy. And that is a good thing. One of the goals of the programme proposed here is to empirically investigate the learning progressions, both from a rigorous statistical perspective, as well as through more qualitatively oriented one-to-one interviews with students, and through field testing in schools. Part of national assessment in this paradigm is that assumptions are continually tested, best ideas are sent out to be trialled in classrooms, and are returned for refinement. Thus, national assessment is not a product, but a process that not only tells us how we have been doing, but also points the way toward improvement.

Researching Progressions and Bringing Them to the Classroom

Learning progressions exist at selected points in time from the beginning of school through Year 10. They are progressions in that one can see how achievements at one level form the basis for the subsequent level. They are bold statements in that they declare what children should be able to do at given developmental points. These bold statements naturally lead to the question: How do we know that? Can Year 4 students really work out meanings of unfamiliar phrases as the progression suggests? Do Year 8 students really understand how to select the appropriate text form? How could we gather information that would confirm these bold statements, or recommend that they need refinement?

It would seem that there are two existing aspects of assessment programmes that could be brought to bear on this issue. Each year, the NEMP programme assesses 1480 randomly selected Year 4 students and an equal number at year 8, often using a labour intensive one-to-one format involving a student and an experienced teacher trained in the assessments that are being given. The NEMP format could be used, at least at Year 4 and Year 8, to provide a rigorous investigation into the learning progressions proposed at those years. Are we on generally on target? Where do we need refinement or revision? Using its nationally representative sample and in-depth, one-to-one assessment format, NEMP has the potential to provide a first level of strong support for the learning progressions (or recommend where changes need to be made). Through collaboration with the developers of the progressions and others with expertise and experience in working with teachers, such as the AtoL team, an assessment approach could be developed for investigating this progression. Consider what is being stated in the reading progression listed above. It is not simply that students will be able to determine the meaning of difficult, figurative language. It has the additional, important aspect that these phrases are ones that they *do not currently understand*. Thus, an assessment of this progression must start by finding phrases that the child does not currently understand. This can be done particularly effectively in a one-to-one setting.

Then, the assessment can turn to what students do when they do not understand a phrase. How do students go about trying to glean the meaning from a phrase they do not understand? What tactics and strategies do they use? Do they turn first to context, or do they work from the words that they *do* know? Are most children successful with this process, or do many have difficulty? And if they do have difficulty, what kinds of scaffolding are most effective in helping their efforts on tasks like these (not simply to get a meaning for a given phrase, but to help them refine their strategies for addressing such situations, and generalising to similar ones). Here again, the advantage of having hundreds of videotaped exemplars can be invaluable.

The research would not end at this point. A next question might be, How can we turn what we have learned here into assessment items that can be used by teachers and analysed via the asTTle IRT framework? Such analysis would allow the progression to essentially be placed on the asTTle learning scale, which would confirm its position in the progression, *and make it available for teachers to use in classrooms when students are ready to take such assessments*. If we could achieve this level of utility, we would have an "off the shelf" assessment that could be used on a "when ready" basis by classroom teachers. These assessment tasks could be

combined with various subsets from the asTTle item bank, allowing for a wide variety of combinations of abilities to be assessed at the desire of the teacher or the school. Teachers could be trained through the AtoL infrastructure, and the assessment of a learning progression would have not only a strong empirical basis, it would have emanated from a strong theoretical basis. Teachers would have a tool that came from a theoretical understanding of reading growth that was tested in extensive one-to-one interviews with veteran teachers who not only were looking at student response, but also at how students make use of prompts and other scaffolding. Finally, the assessment tool could be given to teachers not only in a one-to-one format for in-depth assessment, but also in a more efficient format for the assessment of a whole class.

Using the Approach for Monitoring National Assessment

Let's now look at the question of effective national assessment. How can we get the kind of information we need as a nation with regard to how well our children are doing from such a system? Well, to begin, NEMP data provide a rich, and in-depth picture of Year 4 and Year 8 student progress in any given year. NEMP also covers the breadth of the New Zealand curriculum, utilizing formats and probing for information that is not possible using more efficient forms of assessment. NEMP's approach, utilizing tasks saved from previous administrations, also allows for a strong assessment of growth over time. But it does not currently provide a picture beyond Year 4 and Year 8.

So how can information be garnered of how well children are doing across all years, from Year 1 to Year 10? Here, the already realized utility of asTTle can be augmented by linking the learning progressions, validated by research as being appropriate, into the asTTle scales. This allows for a statistically rigorous located of the learning progressions, and provides ability to use the learning progressions as a basis for communicating to the public how well students are doing.

Finally, the information from the learning progressions, and from asTTle data, can be combined to provide benchmarks of progress that can be rigorously researched. For example, if at a given learning progression benchmark, we see 20% of the students not performing as well as the learning progression indicates they should be, what are the consequences of this lack of achievement? Do these students show learning difficulties later, or do they seem to catch up in later years? We currently do not have a way to ask how performance at a given point in time is related to performance four years later. But all of the pieces for such a system exist; we simply need to use them. Then, when we look at national assessment, we would be able make statements on the order of:

"*We know from our research that success at this learning progression is important for later development. Eighty-seven percent of the students assessed are performing at a level that augurs well for the future, whereas 13% are not."*

And, that statement would be based upon: (1) theory saying that this learning progression represents an important benchmark, and (2) empirical data indicating what has happened with regard to performance on this learning progression in the past.

What Might Be Accomplished

The ultimate goal here would be to have a set of learning progressions that were theoretically determined, empirically validated through individualized assessment, statistical analysis and linking via IRT, and trialling in classroom settings. Accompanying the progressions would be a series of assessments, for both one-to-one usage and for more efficient classroom administration that would place student progress on a meaningful scale of progress in learning. These assessments could be used both for classroom instruction and for monitoring progress at various levels, including a national level.

The development of such a capability would require a model of explicit and ongoing collaboration among the current components of the New Zealand assessment infrastructure. In such a model, we could see:

1. Teams of Ministry of Education (MoE) professionals working collaboratively with AtoL, NEMP, and asTTle teams in determining what progressions would be the next candidates for investigation.

2. NEMP research looking at how well students perform on assessments developed from the learning progressions in intensive assessment settings.

2. asTTle researchers working to incorporate and validate the progressions after initial development and refinement.

3. AtoL, NEMP, MoE professionals, and teachers examining videotapes of students responding to tasks trialled in the previous year to see how the progression under current consideration seems to be working, and seeing what kinds of classroom instructional implications arise from such study.

4. AtoL professionals working with teachers in how to use the most current assessments for "when ready" assessment in the schools.

5. asTTle and NEMP researchers looking at how well students performed on assessments from several years earlier and how well they are doing today. This would allow for the determination of what levels of performance at a given year will predict success four or five years hence. It also would allow for statements such as, "When children score below this point, the research indicates that they could use some assistance to keep them on track with their peers."

Summary and Conclusions

This preliminary sketch of a proposal may seem bold, even pollyannish, but if we do not think boldly, we will be limited in what we accomplish. We have a number of assessment professionals working on a variety of outstanding projects currently, but their efforts are not linked in any serious fashion. In looking at the work on the learning progressions, it occurred to us that this could be a vehicle for just such collaboration. Furthermore, it could be tested on a small scale to begin. We could

pick just a small set of progressions, perhaps at Year 4 or 8, where NEMP already conducts assessments. The exact nature of the ideas presented here are not ultimately as important as engendering the collaboration of the assessment expertise and experience in New Zealand to further the goals of our national educational endeavours.

References

Bandelos, D. L. (2004). Introduction to the special issue on Nebraska's alternative approach to statewide testing. *Educational Measurement: Issues and Practice, 23*(2), 6-8.

Cizek, G. J. (2001). An overview of issues concerning cheating on large-scale tests. In J. O'Reilly (Ed.) *National Association of Test Directors 2001 Annual Proceedings.* New York: National Association of Test Directors.

Cizek, G. J. (2003). *Detecting and preventing classroom cheating: promoting integrity in assessment.* Thousand Oaks, CA: Corwin Press.

Darling-Hammond, L. (2006). Securing the right to learn: Policy and practice for powerful teaching and learning. *Educational Researcher, 35*(7), 13-24.

Eglund, N. (2005). Educational assessment in Danish schools. *Assessment in Education: Prinicples, Policies & Practice, 12*, 203-212.

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: Inquiry-oriented and didactic responses. *Assessment in Education: Principles, Policy & Practice, 11*, 67-88.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing. Yearbook of the National Society for the Study of Education, 104*(1), 1-34. Boston, NA: Blackwell Publishing.

Herman, J. L. (2007). *Accountability and assessment: Is public interest in K-12 education being served?* CRESST Report 728. National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles.

Hutchinson, C., & Hayward, L. (2005). The journey so far: Assessment for learning in Scotland. *The Curriculum Journal, 16*, 225-248.

Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment, 7*(1), 29-38.

Linn, R. L. (2006). *Educational accountability systems*. CSE Technical Report 687. National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles.

Looney, A. (2006). Assessment in the Republic of Ireland. *Assessment in Education: Prinicples, Policies & Practice, 13*, 345-358.

Martineau, J., Paek, P., Keene, J., & Hirsch, T. (2007). Intergrated, comprehensive alignment as a foundation for measuring student progress. *Educational Measurement: Issues and Practice, 26*(1), 28-35.

Ministry of Education (2007). *Literacy learning progressions: Meeting the reading and writing demands of the curriculum*. Wellington, NZ: New Zealand Ministry of Education.

Phelps, R. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice, 19*(1), 11-21.

Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). A strategy for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 17-25.

State of New Jersey Department of Education (2008a). *Historical context: Overview of statewide testing program.* Retrieved 8/1/08 from: http://www.state.nj.us/education/assessment/history.shtml

State of New Jersey Department of Education (2008b*). New Jersey core curriculum content for mathematics.* Trenton, NJ: New Jersey State Department of Education.

State of New Jersey Department of Education (2008c). *New Jersey Statewide Assessment Reports.* Trenton, NJ: New Jersey State Department of Education. Retrieved 8/2/08 from: http://www.nj.gov/education/schools/achievement/2008/gepa/

The Scottish Government (2008). *A consultation on the next generation of national qualifications in Scotland.* Edinburgh, Scotland: The Scottish Government. Available online, August 2, 2008: http://www.scotland.gov.uk/Resource/Doc/226233/0061255.pdf

Wolf, A. (2002). Ships in the American night? Assessment paradigms and political imperatives. *Assessment in Education: Prinicples, Policies & Practice, 9,* 387-400.