ASSESSMENT FOR SUCCESS IN PRIMARY SCHOOLS

A RESPONSE TO THE NEW ZEALAND GREEN PAPER ON ASSESSMENT

John Hattie

Chair, Professor Department of Educational Research and Methodology,

University of North Carolina, Greensboro, NC, USA

Head of School, University of Auckland, New Zealand

1998

The major issue in this submission is to encourage the New Zealand Government not to tread down the well-worn path of increasing "national assessment", as promulgated in the Green Paper. There are many other successful methods for enhancing the quality of schooling, and rarely does increased external testing feature in these methods.

I am currently the Chair and Professor of one of the largest "Educational Measurement, Research, and Evaluation" departments in the USA, have been involved as an expert witness in the US Federal court, in National testing cases, and am committed to the improvement of assessment in education. I am about to assume the position as Head of the School of Education, University of Auckland (Sept., 1998), have school age children, and thus have multiple reasons to be concerned about the direction suggested in this Green paper.

I categorize my remarks under twelve headings.

1. The Purposes of Schooling

The Green paper commences with the claim that the *single* goal of New Zealand schools is "to ensure that all students are taught the skills, knowledge, attitudes, and values necessary for them to meet their life-long potential". I suggest that there are other outcomes worthy of consideration: affective outcomes, flexibility in processing, creativity, reduction of stress, socialization, economic opportunities, love of learning, surface and deep processing strategies, and willingness to learn. Like Broudy (1988) I contend that education (as practiced in schools) too often is judged by the replicative and applicative uses rather than the more appropriate associative and interpretive uses. The effects of schooling often are assessed with respect to their "screening" functions (identifying the more able, changing the order in the

unemployment queue) rather than with respect to their production functions (increasing the learning skills and love of learning of students). The Green Paper appears to be premised on the facility of schools to "prepare" students for some later life.

A major attraction of the present New Zealand education system is the added goal: "to ensure that all students learn to learn, and participate meaningfully in New Zealand life while they are students". The emphasis on "preparation" in the Green Paper can lead to more utilitarian methods of education to the detriment of methods more conducive to educating present and future effective citizens. The assessment methods proposed in the Green Paper are more related to screening and assessing whether schools are performing similarly, and neglect the measurement of attitudes and values and other important associative and interpretative outcomes of schooling.

Recommendation 1: That the purposes of schooling enunciated in the Green Paper be broadened to include associative and interpretive outcomes as well as replicative and applicative outcomes.

2. Teachers Make the Difference

The emphasis in the Green Paper appears to be on effective school leadership, and this is laudable. I would encourage the Green Paper authors to emphasize the importance of excellent teaching as well as schools lead by excellent school leaders. It is the teacher that makes the differences (Hattie, 1987; Hattie, 1992a, 1992b, 1997) and methods aimed to improve teachers and teaching is what necessarily leads to better outcomes.

As an alternative to spending resources (money, teacher time, student learning missed opportunities, parent confusion, etc., see below) on national testing, I recommend that more appropriate methods for recognizing and rewarding excellent teaching be implemented. The National Board for Professional Teaching Standards has lead the way in such a task, and this could be emulated in New Zealand, to the betterment of schooling—particularly on student outcomes. These methods recognize teaching as a complex activity that involves more than student outcomes on narrow tested domains, but involves higher order thinking, critical decision making, In our recent review of "expert teachers" we identified the following attributes—and these need to be encouraged in New Zealand schools:

Content knowledge. The expert teacher has an appropriate and deep command of the content knowledge that is to be taught.

Pedagogical knowledge. The expert transforms essential aspects of subject matter to connect with students' ways of understanding. These transformations include: identifying essential representations (have deeper representations about teaching and learning; are problem solvers); setting goals for diverse learners (can anticipate, plan, and improvise; are better decision makers, and can identify important decisions); guiding learning through classroom interactions (have excellent class control; have a multidimensional perception of class situation; are more context dependent and have high situated cognition; are diverse in their problem solving; they are flexible), and monitor learning and providing feedback (are more adept at monitoring and provide much feedback; are more able to

check and test-out their hypothesis or strategies; are more adept at recognizing patterns and gathering information; are more automatic).

Affect. Experts have high respect for students, and have much passion about teaching and learning.

Further, we argued that effective teachers have a multiplicity of effects on their students including: enhanced motivation, increased self-efficacy, facility to confront challenging objectives, achievement outcomes—both in curricula coverage and depth of processing, and a conception of learning that transcend the accumulation of information.

All efforts to recognize and encourage these attributes can lead to higher outcomes from schooling. Using National Assessments, as suggested in The Green Paper, too often has the effect of narrowing the concept of teaching, and certainly narrowing the outcomes of schooling.

Recommendation 2: That the efforts and resources to construct new tests for students to adjudge schools, would be better spent on recognizing and rewarding excellent teaching—with the concomitant increased on the student outcomes.

3. The Shift from Inputs to Outputs

I am certainly not averse to models based on outputs rather than inputs (see Hattie, Tognolini, Adams, & Curtis, 1990) but the Green Paper ignores the

intermediary influences—the processes. These processes more under control of the teacher, who are the critical delivers of education. The mantra that follows from the Green Paper emphasis on outputs, self-management, is misleading. The claim is that schools and teachers will have more freedom to decide how they can turn their inputs into outputs is crude, and past experience has demonstrated that the presence of national tests to monitor the outcomes narrows these "freedoms". When the inputs and outputs are controlled (or out of the hands of those who manage" the processes, there is little room left for self-management. The test-driven models that have been introduced in other countries have demonstrated the narrowness that derives for teachers and students. If, for example, the outcomes for all New Zealand students are the same, then alternative schooling (the diverse forms of schooling appropriate to the needs of particular communities" p. 7) are inevitably constrained from being different.

Recommendation 3: That it be recognized that increasing National Testing leads to less innovation in schools, less self-management for teachers, and an attention more on outcomes to the detriment of inputs and processes.

4. Assessment

Assessment dominates this report. The premise is that "assessment and testing should provide the information about children's attainment that schools, teachers, managers, and pupils themselves need in order to improve teaching and learning". Such a model is premised on the following logic:

a. Create a national Standard Course of Study/Curricula.
b. Create excellent National Tests on a limited set of outcomes (numeracy and reading) related to the Standard Course of Study/National Curricula.
c. Report scores to parents and others
d. Students will improve their abilities, learning, and attitudes
e. Desirable outcomes will befall students to the betterment of all.

The logic breaks down most manifestly, between steps c) and d). If there is evidence to support the follow though of the above logic, then I have yet to see it. Such evidence needs to be shared in the Green Paper.

There are few programs in other nations whereby the massive introduction of testing (akin to the Green Paper model) has lead to improvements in outcomes—other than the gains directly attributable to immediate "gaming the test". There are no reports of long term gains. Most programs attain a Lake Wobegon effect (where all school averages are above the national mean!, Cannell, 1989), and the effects on teaching and students are largely negative. Where are the studies, or plan for such studies, to investigate the positive and/or negative effects of the testing program promulgated in the Green Paper? What evidence would the New Zealand Government need to abandon/improve the proposed testing program, and will it seek such evidence?

Yes, many nations have introduced new systems of external tests that involve all the students in the same year of study completing National Tests. The Green Paper glosses over the findings of these testing systems, although this evidence is well documented. There is little debate in the measurement community as to the adverse

effects of these systems, the only question is why is the research ignored—as it is in the Green paper. There is no reason to believe that New Zealand system will go down a different route, the question only is, will it fast or slow.

I am not suggesting that the National system must not seek comparative information on outcomes to influence National Curricula: This does NOT require testing all students (which is totally unnecessary to gauge the effects of the implementation of curricula), but can involve appropriately sampling within schools using a matrix-sampling design.

There is a plethora of evidence demonstrating the effects of high-stakes testing in schools (see Shepard, 1991). The effects are on a narrowing of the curricula, an increased drop off by students interested in such narrow schooling, harder conditions for teachers to turn students onto lifelong learning, decreased confidence by the public in the education system; and greater incentives for excellent teachers to abandon the classroom. Smith and Rottenberg (1991), for example, documented the major implications of testing such as suggested in the Green Paper:

External testing reduces the time available for instruction. This time included: time for testing, for test preparation (while elective, the time reflected the degree of pressure teachers felt to raise the test scores and the importance they perceived the scores to have for administrators and the public), recovery from testing.

Schools neglect material that external tests exclude. Subjects not tested are neglected, reading real books, writing in authentic contexts, solving problems that require more than rote recognition, creative and divergent thinking, longer term integrative projects, and the like were gradually squeezed out of instruction. Concentrating on teaching “basic skills”—they have less to read, write, or think about and fewer avenues of interest such as those provided by science and social

studies. Focusing instruction on test materials also slights - in breadth, complexity, and form—reading, math, and language.

Encourages use of instructional methods that resemble tests. As stakes rise, teaching becomes more like the tests.

Encourages narrow forms of school organization. Test scores influenced decisions to place students in homogeneous groups and tracks, leads to blaming the students.

Negatively affect teachers. Given that teachers are aware of the variability of test scores, and standard error, the effects of other factors than their own teaching, teachers are always kept off balance in the face of the "numbers". They become autonomous professionals. "They redefine problem solving as the operations necessary to solve word problems on tests, they neglect social studies and science, integration of knowledge, production of discourse on novel problems, critical thinking, civic participation, cultural knowledge..."

Recommendation 4: That those who wish to implement National Testing (as outlined in the Green Paper) provide the New Zealand public evidence that the above typically found results of National Testing would not be repeated in New Zealand.

5. Implication for Maori students

I have limited knowledge regarding education for Maori students. I do note the excellent research by Claude Steele (1992) on disidentification. He has demonstrated that the widely reported discrepancies in academic performance between African

American and White students are caused to group differences in identification. Although all students experience anxiety over possible failure in academic settings, individuals who are members of disadvantaged groups also experience the increased anxiety of confirming the negative group stereotype through personal failure. He has provided some excellent empirical studies demonstrating that the mere mention of "test" can cause minority students to perform lower on achievement tests because of their increased personal and group responsibility, and this is not evident among majority students.

Such research leads me to directly oppose the claim that "to identify the effectiveness of school programs for Maori students, schools need to be able to compare their Maori student achievement with that of the rest of the student population" (p. 8). Not only does this highlight the "test stereotype" but puts the Maori students onto the common scale of all students, which is not necessarily the aim of the alternative Maori programs cited on p. 7.

In the following recommendation, I am not suggesting that Maori children may not benefit if their teachers decide to use testing (including national tests of numeracy and reading), but this decision needs to be a local teacher and school decision based on consideration of the information/consequences of the testing at the local level, and not some uniform mass testing for comparative purposes. In the interests of individual Maori students, the government program needs to reduce group stereotypes.

Recommendation 5: That more appropriate assessment be developed uniquely for Maori students, where necessary.

6. Usefulness for Teachers

The Green Paper claims that:

- Teachers use assessment tools to help them to identify systematically specific learning needs for individual students
- Teachers need information to help them to identify whether their judgments about achievement are consistent with national standards
- Teachers use assessment information to help them evaluate the effectiveness of their teaching and learning programs.

These are assumptions and have been well tested. Teachers do NOT use assessment for such comparative, norm related purposes. It is rare, if not impossible, to imagine a class of 30 students that are representative of the New Zealand student population—and thus it is meaningless to presuppose that a teacher could learn much from comparing her/his class to this norm. Teachers *do* ask to compare to some set of curriculum standards, but not to some group that is a meaningless comparison. The National Tests do involve curricula-relative information and this is less questionable, but National tests for school/teacher/student-comparative purposes are invidious.

Consider a germane study: Ask teachers to predict the scores of their students on these National tests, and then ask them “What additional information does the National test scores provide?” Methinks little. In a recent review of 578 studies

investigating the effects of testing on classroom learning, Black and Wiliam (1998) concluded that the tests seldom inform teachers of previously unrecognized student talents and seldom identify deficits in a way that directs remedial instruction.

The report notes the waste of effort that teachers put into "reinventing the wheel".

"Teachers need to be able to moderate their assessment information" – I do not hear teachers asking for such comparisons, and the evidence is that such normative information rarely informs the teacher beyond what they already know about the student, and rarely leads to changes in instructional methods that are more appropriate to the learners.

On p. 13 in The Green Paper the desiradata of school assessment are noted—none of which are served by a national testing program.

- *Helping schools to decide what works well and what does not.* This research is well documented, and few of these research results are derived from testing programs (Hattie, 1987; Walberg, 1994). For example, other than more effective methods for test-taking, there is little educationally to show from the UK or USA testing models.
- *Enabling teachers to identify student learning needs early, before the problems become too big.* Only if the tests are diagnostic can this occur. The development of a diagnostic test that serves the diagnostic needs of all New Zealand students is beyond my comprehension. Such a test has been rarely accomplished elsewhere, and rarely have teachers gained such information from national tests (see Black & Wiliam, 1998).

- *Clearly identifying the effectiveness of a school in improving achievement for Maori or Pacific Islands students.* The research on this topic that has been informative has rarely come from wide scale testing.
- *Ensuring that dialogue between parents and schools is better informed.* The evidence in the US states that have implemented similar testing programs has lead to parents become more jaundiced about the status of schools, and they have been rarely better informed about their own child's status above and beyond what they already know on the basis of teacher information provided from other sources. Further, given the wide distribution of achievement across schools, the media and parents often dwell on those schools that (inevitably given distributions) fall at the bottom end—and there will ALWAYS be such schools given the measurement model. For example, in North Carolina the bottom 15 schools are taken over by the state and extra resources poured in. Surprise, surprise, the next year there are another bottom 15 schools—reinforcing the parent belief of the adverse state of schooling in North Carolina.
- *Giving parents a better picture of the progress of their child and the effectiveness of the school.* Other than relative information about the school, there is little evidence that parents a provided with a better picture (Black & Wiliam, 1998).

I agree that parents are very interested in their child's relative progress. However, there is little information in test scores from national testing, as implemented in most countries, that inform the parents above and beyond what the school has already provided. It is surprising, however, that the research on the best methods for reporting information to parents is woefully lacking (Jaeger, Gorney, &

Johnson, 1994), and I would encourage New Zealand to investigate more effective ways to transmit information to parents. This lack of information has often lead parents to ask for more national tests (and every Gallop and related poll in the USA for the past ten years has noted that parents want more tests—they care less about whether such tests are good, relevant or informative, they just want more tests. The politicians have more than obliged by giving them more tests, and the poverty of the information from these new tests, has lead parents to wanting even more tests. The Green Paper is correct, let us devise more effective methods to inform parents—but national testing is not the answer.

The theme in the Green Paper relates more to devising methods so that parents can evaluate "how effective a school's programs are compared with those of similar schools with the national picture" (p. 15). The common finding from many such comparisons is that 80% of the parents consider the state of education is not desirable, and 80% of parents consider that the school their child attends is desirable. Why is the Green Paper tapping into the prejudices of the first group. We need to increase not decrease the *efficacy* and confidence in our National system of schooling.

It is stated that "The government needs information on the achievement needs of specific groups of students in order to formulate policy and monitor its effectiveness" (p. 15). I definitely agree, but have rarely participated in policy forums where national testing has provided this information—other than confirming deficits already well known.

Recommendation 6: The Green Paper authors need to provide evidence that teachers/schools will make more informed decisions about students and curricula development as a consequence of introducing National tests.

Recommendation 7: That the resources that were to be expended in developing National tests be spent devising a series of effective diagnostic tests (if they do not already exist) that teachers and schools can use if they consider that such diagnostic information provided assists them to better inform parents of the progress of their students.

Recommendation 8: That alternative method for "School Reports" are assessed as to the value and accuracy of information provided to parents.

7. The Green Paper National Testing Model

I recognize that more is proffered in the Green Paper than National Tests. The new assessment package includes: diagnostic tools, exemplars of student work, nationally standardized tests, and modification of the National Education Monitoring project. I applaud all efforts to provide information that can assist teachers in their tasks, and the Government to monitor progress. Given the present battery of available tests (including PAT, TOSCA), which can provide such national benchmarks, I am concerned that more resources will be placed into tests—which

will lead to little change in schools and lead many to believe that they are undertaking a positive step to enhance schooling—while the schools languish under the increased assessment barrage.

Teachers do value diagnostic procedures, and all efforts to provide these procedures so that teachers can choose and appropriately use them is valued. The measurement community is awash with such measures and the issue surely relates to providing access to these available procedures rather than inventing more. Rarely, can national standardized tests serve the diagnostic purposes (at the class or student level) and provide national benchmarking—in a reliable and valid manner. I have no concerns if the Government wishes to encourage the use of more diagnostic tools, as long as teachers are informed about the validity and appropriateness of using these tools. It is rare to imagine such diagnostic tools becoming compulsory, given the needs of teachers for differential diagnostic information, depending on their needs.

I have seen many such volumes of “National Exemplars”, and they clutter many principals back offices. For example, I am currently involved in the evaluation of the one US states’ Assessment model, and this morning received their tests—all 4 large cartons. The problem is that, while they may serve the purposes of assessing schools/teachers/students, there is little diagnostic information, they omit massive topics of concern to teachers, and they are all multiple choice (for cost reasons, amongst others). There are many excellent examples of National Exemplars, and it is laudable that the Green Paper proposal also includes examples of how teachers marked the work and the judgments they made in relation to the learning demonstrated by the student (p. 21). Given the difficulties of ensuring adequate reliability for such ratings, these exemplars can be expensive to develop. They certainly must be defensible psychometrically if they are to have credibility among

teachers. Otherwise, they will not use them and the Exemplars will be brought into disrepute.

Recommendation 9: Increase efforts to provide schools/teachers/ parents with dependable diagnostic assessment tools, and ensure that they are appropriately chosen and that the consequences of using them are defensible and meaningful.

8. Publicity and Externally Referenced Tests

Here is where I have grave concerns. On the one hand I see the benefits of national benchmarks (such as illustrated in the Figure on p. 23). Such information can be most useful in targeting curriculum and teaching enhancements, provide information on return for resources, and guide educational discussions. They are less useful for diagnostic purposes, and rarely provide teachers extra information. On the other hand, as soon as this nationally comparable information becomes public then negative consequences abound.

I would argue that the many benefits cited in the Green Paper derived from National Tests can be achieved without leading to such invidious comparison—by appropriately "sampling" students from schools (as many countries/states currently do). Then it is meaningless and often impossible to compare schools, but the information can provide the National Education policy makers the relative benchmark information—and schools could use this information to demonstrate excellence, and to revise their curricula and teaching.

Under the proposed Green Paper National Testing model there is no question that the issues of "comparability" will dominate the discussion. It is naive and blind to assume otherwise.

Recommendation 10: That appropriate matrix sampling of curricula and students is used to provide national benchmark information, and avoid all possibility of school comparisons being made public on a narrow range of outcomes.

Initially tests would be developed in literacy (English) and numeracy, would be pencil-and-paper—and thus they will "provide an indicator only of student achievement". I missed something here. All test information is indicators. The issue is whether they are valid indicators—particularly of school performance.

Under the Green Paper proposals, you may end up with valid indicators of a limited range of skills; much energy will be expended on testing that is unrelated to the success of many schools, and schools will be judged on a narrow range of attributes. Moreover, national education policy makers and schools can purchase such indicators *now*. The PAT, TOSCA, and so many other dependable tests exist (see Buros, etc.) I would predict the correlation between the scores on these already available tests of numeracy and reading and the "limited indicators" to be developed as a consequence of the Green Paper proposals would be very high. So, why are these alternatives not used, not sufficient, and not advocated. They are cheaper, have known psychometric properties, and available.

If all the benefits that the Green Paper advocates for these tests are to realized, the fundamental question is why are schools and teachers not clamoring to use such tests now—probably because they are not that informative to the teaching process, to student learning, and for schools to change their practices.

Recommendation 11: That information be provided as to: What diagnostic information are schools and teachers requesting; Why are schools and teachers not using the currently available alternative diagnostic tests; and "What diagnostic information would the National Tests provide that is not currently available"?

I applaud the claim that "It will be important to ensure that the test items are appropriate for children of all cultures" (p. 24). I assume, therefore, that a major effort will be to ensure that there is no item bias, and no adverse impact from these tests. As has been demonstrated elsewhere, this can increase the costs of developing the tests, and certainly increases the costs of maintaining an appropriately unbiased set of test items. I am particularly concerned with the potential for adverse impact, and encourage the Green Paper promulgators to pay particular attention to this issue.

Recommendation 12: That the adverse impact of implementing the National Tests be a major focus of attention.

Similarly, I support the claim that: "Tests of other learning, including authentic or performance-based activities, will still need to be developed". The cost of this,

particularly relative to the return is enormous. There are many US states that have embarked on the development of such activities, and New Zealand could well consider the consequences for these states (e.g.., Kentucky, Indiana, Vermont). They are very expensive, they are notoriously unreliable, they sample very limited information, they occupy an inordinate amount of teaching time for little information returned, and they are constantly being modified in light of experience hence increasing the costs. Even in a state, like North Carolina, which emphasizes multiple-choice items, the costs of $170m per year to maintain the testing program seems an absurd waste of time.

Near the end of the Green Paper (p. 26) it is noted that these tests would commence in Year 6 and Year 8—partly because secondary (and I presume intermediate) schools could better plan programs suited to the needs of there prospective students. Where is the evidence that these schools want this information that would be provided on these tests? What do such schools do to change their curricula in light of this information? Why not invite such clamoring school to administer some of the myriad of tests already available to the feeder schools (or when the students first enter their school), and Why insist all students undertake this testing when their prospective schools may not wish this information?

I assume, although I am surprised it is not stated, that these tests of numeracy and reading would be criterion-referenced to the National Curricula, and not norm-referenced to some distribution.

Recommendation 13: Provide evidence that testing on only numeracy and reading would lead to enhance educational outcomes, when the evidence in other National Testing implementations is to the contrary.

Recommendation 14: That the appropriate referencing attributes (the curricula or some norm group) be enunciated, and that the appropriate psychometric models will be used to optimize best test design.

9. The Box-And-Whisker Graph

In the Ethical Principles of Psychologists (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) it is stated that the clients have the right to know the result of tests, the interpretations made of the results, and the bases for conclusions and recommendations made about them. In one state in the US, a study was undertaken of the value of such graphs (based on state-wide testing) were for parents. Hebbler (1997) demonstrated that the majority of parents failed to understand the meaning of the scores, and the standard errors were so large that they probably are not interpreting the scores correctly. He noted that 23% to 39% of administrators and teachers “said the reports were not useful” (p. 20) and felt that the students did NOT understood their scores (a further 42% of administrators and 21% of teachers were “not sure”). Further, 21% of administrators and 47% of teachers did not believe the reports were useful to parents. Parents either could not meaningful

interpret the graphs, or claimed that the gained no information beyond what they already had known from the school. This study has lead to a major and costly program to improve the graphs—so far, with no success.

It is fascinating to note the constant claims in the Green Paper that these tests will assist in providing diagnostic information". I challenge a school or teacher to inform me of the diagnostic information that they would derive from this graph.

The other consequences of the testing are cited on p. 24ff:

- *A report to Government on national and group levels of achievement.* This could be accomplished by appropriately sampling schools and students and thus avoiding the deleterious consequences of national testing and inappropriate comparisons.
- *A report to each school compared to national levels and to similar groups of students.* I would be less concerned if there was a statement that said that if schools did not find this information useful they could opt out of the program; or that there would be research to ascertain the returned value from this exercise. Other than driving schools to be more "test oriented" the effects in other systems have been negligible.
- *A report for schools to send to parents.* What use are group levels of achievement" not individual assessment. or parents. This is not what they are asking for. They want to know the progress of their student, not some group to which their student may or may not represent! Note the above comments on the limited value of commencing this enormous resource hungry method. There are far more effective methods to send information to parents (Jaeger et al., 1994).

Recommendation 15: That evidence be provided that schools/teachers/ parents are appropriately informed about their child's/students progress when using the box-and whisker graph.

Recommendation 16: That alternative methods to National Testing be investigated to meet the aims cited in the Green Paper on p. 24.

Let me cite from Dorn's (1998) analyses of "The Political Legacy of School Accountability Systems":

> Technocratic models of school reform threaten to turn accountability into a narrow, mechanistic discussion based on numbers far removed from the gritty reality of classrooms. Over the past twenty years, the dominant method of discussing the worth of schools in general has been the public reporting of aggregate standardized test score results. Popular news sources typically distort and oversimplify such findings (Berliner & Biddle 1995; Darling-Hammond 1992; Koretz 1992b; Koretz & Diebert 1993; Shepard 1991). The recent public debate over schools is not rich, reliant on multiple sources. (Dorn, 1998, p.1)

Despite the weaknesses of high-stakes testing, the short-term consequence of more standardized testing may be intensified criticism of public schooling and cynicism about the purposes of public educational systems. Schools need to be "public" in the sense of public involvement and political commitment (Fine, 1991, Chapter 9; Katz, 1997). However, the ranking of schools and teachers is inherently a zero-sum game, and not everyone can be above-average. Seeing school performance in such terms, divorced from classroom practice and public policy, makes both meaningful praise and criticism of schools very difficult. Moreover, the constant reinforcement of the myth of declining school performance will continue the erosion of support for the good schools that exist and make intense discussion of the needs of children more difficult.

Accountability should encourage deeper discussion of educational problems. Student performance should be the starting point of educational politics, not an occasion for political opportunism or crude comparisons. Statistical accountability, with the centralization of statistical production and dissemination through popular news sources, encourages oversimplification rather than a more extensive public discussion. Accountability should connect student performance with classroom practice. Statistical accountability, with the abstraction of student performance into numbers without context, removes classroom practices from the discussion of educational reform. Accountability should make the interests of all children common. This sense of commonality is the best meaning of "public" in public schooling. Statistical accountability systems intensify educational triage, encouraging schools to isolate and devote fewer resources to students whom schools judge as difficult to teach. Politically, statistical accountability systems divide the interests of schools and communities through competition for prestige and resources.

10. The Cohort Problem

A constant claim throughout the Green Paper is that schools would be related to "similar groups of students". Such normative emphases are not consistent with the emphasis on producing diagnostic information. Surely, a child, class, or school can obtain diagnostic information if the tests are not related to curricula outcomes. Comparing to norms provides limited diagnostic information.

What are the appropriate comparisons? It is well known that the resources a child brings to school can impact dramatically on that students' achievement. I believe that schools exist to neutralize the negative resourcing that can be brought from the home—the school provides all students an equal opportunity to excel and must provide every incentive and learning opportunity to assist the student to excel (across multiple possible outcomes). By providing comparisons on the basis of home resources leads to cementing many teachers and schools in a cycle of low expectations—and condemns the students to be perceived as "lowly resourced". I am sure the New Zealand educational system already knows which schools have a predominance of students from low resourced backgrounds, so why test to re-discover this. The system can already provide appropriate resources to these schools. Just because School x in a low resourced area has "outcomes" higher or lower than School y is a similarly low resourced area does not mean that the teachers are performing better or worse. Such a comparison model assumes that by merely switching the teachers from School x to School y would reverse the differences—this has never been demonstrated.

The Green Paper claim that, "if all Year 6 students were achieving at level 2 and similar students nationwide were achieving at level 4, the school would clearly have cause to evaluate its programs". This depends on ensuring that the students are

similar—which is notoriously difficult. Does this mean they were “similar” when they enter Year 1; “similar” in the resources provided to the school, “similar” in home background, “similar” in teacher experience and expertise—what are “similar” students? No system has yet provided a “value added” statistical model that has lead to comparing “similar” students, and the Green Paper provides no indication that New Zealand would be the first to achieve this.

The problems of schools with small numbers of children in a cohort are acknowledged in the Green Paper on p. 25. It appears that 40% of schools (p. 26) are deemed “too small to provide statistically reliable comparisons” and the national testing is “unlikely to provide information that is reliable enough to make definitive judgments about a student’s progress” (p. 25). So, for the purposes of comparability constantly echoed throughout the Green Paper, for 40% of the schools the National Tests are close to useless. Many other states and countries also have found that solving the problem of considering “small schools” has been intractable. But no, the Green Paper then suggests these schools ignore this issue as the “national comparative data will provide valuable information to benchmark schools’ expectations of students”.

I am delighted to read that the government will not publish “league tables” of schools. It seems that only the school receives the information. Do not also parents receive the information as promised earlier in the report? It would not take much for an enterprising reporter, parent, researcher, fellow principal, etc. to compile comparative information. Further, such comparisons is what parents most often demand—usually because they consider the only informative purpose of such testing (which tells them very little or nothing about their own child beyond what the school has already provided). A purpose of these assessments are for comparative purposes

and now the Green Paper says that only Government will have access to these data—and not teachers, principals, or parents. It is acknowledged that the information may be released as a consequence of the freedom of information provisions, but that is unlikely that schools with small numbers (40% of all schools) may be able to withhold information. I would wish that legal advice on this matter be stated prior to implementing this scheme. I cannot imagine the New Zealand voters agreeing to finance an assessment scheme based on providing comparable information and then being denied access to that very information.

An issue not raised in the Green Paper that needs addressing is the implications of provided schools/parents/teachers with this "comparable' information across years. Is there any expectation that schools should remain constant, or "grow" across years. The manner in which the National Tests are to be introduced appear to mitigate against any such comparisons in a meaningful way, although I can guarantee that such comparisons will be made. The fundamental issue is that of differences in cohorts. It does not take much experience in schools to realize that cohorts have "personalities" and can differ as a group from year to year. To then compare the mean performance from a cohort this year with the cohort next year is fraught with dangers. As the Green Paper notes, "like" students need to be compared to "like students, and there is no guarantee that students across years within the same school are alike. This cohort problem has plagued many testing programs and has lead to the demise of many such programs—but usually not until after deleterious effects accrue relating to judging the effectiveness of a school.

Further, I would be seriously concerned if any comparisons across years lead to the use of "gain scores". Although seemingly appealing, these gain scores have well known measurement problems, which have not been resolved by the very best

measurement experts. I am confident that if Recommendation 19 is listened to, there will be no use of gain scores in New Zealand.

Recommendation 17: That any testing model that leads to public comparing schools on a few test indicators, regardless of efforts to attend to comparability issues, not be implemented.

Recommendation 18: That the issues of comparing school performance across years be addressed, and all efforts taken to ensure that comparisons on a few indicators under a National Testing model are meaningful.

Recommendation 19: That any use of gain scores, in any form, be fully investigated prior to their recommended use.

11. The Use of a Single Score

The report acknowledges one limitation: "Students must not be labeled on the basis of a one-off snapshot of their achievements in a restricted part of the curricula" (p. 25). But it seems it is permissible for a school to be so labeled, and a teacher, and the national education programs.

The APA, AERA, and NCME Standards for Educational Tests (1999) states:

> In elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score. Other relevant information for the decision should also be taken into account by the professionals making the decision. *(Primary)*

The Standards further claims that test scores are to be presented as one source of information about a student or group of students and should not be used for placement, referral and other consequential decisions on the education of a student.

Recommendation 20: That no test information based on a single score be released.

The Green Paper National tests are to be administered within a two-week period. As the stakes of these Green Paper tests increase the efforts by teachers' etc. to optimize their performances increase—recall, this is the point of introducing these tests, to improve teaching. This is not suggesting that teachers are law-breakers, but that there are many, many instances of some teachers and principals using nefarious methods to enhance their schools' performance. The National Tests must all be administered at the same time.

An external agency will be contracted to administer the tests. The administration (p. 27) does not include the external agency provided a public report on the psychometrics of these tests: the degree of reliability, the reliability of the scorers,

the validity, the bias against sub-groups, the factor structure, the dimensionality, etc. Who provides this critical and necessary information?

A minimum piece of information that must accompany ALL releases of information from these tests, is the standard error. All users of these tests must be educated in the meaning of standard error, as it is fundamental to defensibly interpret sores, when comparing scores between students and schools, and in all testing situations.

Recommendation 21: That the appropriate standard error of measurement always accompany scores on the National tests; and that all users be informed of the interpretation and criticalness of the standard error.

12. How would the Government know that its aims are being realized? (p. 27). How do teachers, principals, and boards know whether their school's programs are effective in improving learning or need improvement?

I believe that it is incumbent on a Government when promulgating national tests, to build into the model a research program to address the fundamental claims on which the introduction of these tests is based. Too often, tests are introduced, causes massive disruption and destruction to learning and teaching, and there is no way to rid the system of the tests.

Recommendation 22: That any implementation of National Testing be contingent on funding a research program that addresses the major issues/consequences underlying the testing program:

I suggest the following questions could guide this research

School

- The extent to which schools move toward school-based management that leads to within-school improvements.
- How does the Green Paper plan effect what is happening in schools?
- How does the Green Paper decisions relate to factors controlled by schools?
- Does the Green Paper plan lead to narrow forms of school organization?

Teachers

- Do teachers seek different and desirable professional development as a consequence of the National Tests—particularly teachers in high gain classes?
- Is there evidence that the quality of teachers and teaching improves as a consequence of the Green Paper proposals?
- Are we attracting better quality teachers attracted compared to pre-Green Paper days?
- Are we retaining quality teachers and removing low quality teachers?
- Are teachers learning how to successfully "game" the Green Paper proposals model, as opposed to enhancing the educational achievements

of the students (Note the Cannell Lake Wobegon effect well documented in the US: every state that uses nationally comparable tests has a state mean score higher than the national average!

- What do those who educate (i.e. the teachers) learn from the results of the Green Paper proposals that they did not already know?
- Is the enhancement of the test scores under the control of the teachers and schools?

Instruction

- How does the Green Paper implementation affect instruction time?
- How does the Green Paper implementation affect instructional methods?
- Are teachers more in "control" of their classes, such that they are, and/or believe they are implementing decisions to more effectively teaching the students?

Curricula

- What are the effects on the curricula?
- What are the effects of the overemphasis on testing?
- How does the Green Paper implementation affect subjects not tested?
- How will the Green Paper plan lead to higher quality questions in the National tests that tap deep as well as surface learning?
- What are the consequences on the curricula and teaching time of having but one score for math, and one for reading?

Tests

- Are the tests being used for the purpose they were validated?

- How will the standard error be incorporated into test score interpretation?
- Given that Year 6 and Year 8 comparisons are being made, what is the effect of using a single year as the "outcome" cohort?

Students

- How will the Green Paper plan affect the learning processes of the students?
- How does the Green Paper affect student learning?
- What pressures are teachers, as a consequence of the Green Paper, placing on students?
- Are students more engaged in school because of the Green Paper emphasis on testing?
- Are students better prepared as a consequence of this testing?

Parents

- How much more information do the parents (and students) derive from the Green Paper test information?
- Does public confidence in public education increase?

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Broudy, H. S. (1988). *The uses of schooling*: Routledge New York.

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests: The "Lake Wobegon" report*. Albuquerque, NM: Friends for Education.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives, 6*(1), 1-32.

Fine, M. (1991). *Framing dropouts: Notes on the politics of an urban public high school*. New York: State University of New York Press.

Hattie, J. A. (1987). Identifying the salient facets of a model of student learning: A synthesis of meta-analyses. *International Journal of Educational Research, 11*(2), 187-212.

Hattie, J. A. (1992a). Do teachers count? *Australian Association for Pastoral Care in Education, 2*, 9-11.

Hattie, J. A. (1992b). Towards a model of schooling: A synthesis of meta-analyses. *Australian Journal of Education, 36*, 5-13.

Hattie, J. A. (1997). *Setting standards for beginning teachers: A discussion paper*. Washington DC: National Council for Accreditation of Teaching Standards.

Hattie, J. A., Tognolini, J., Adams, K., & Curtis, P. (1990). *An evaluation of a model for allocating research funds across departments within a university using*

selected indicators of performance. Canberra, Australia: Department of Employment, Education, and Training.

Hebbler, K. (Ed.). (1997). *Improving the quality of early intervention personnel by enhancing faculty expertise: Findings and recommendations for the regional faculty institutes*. Chapel Hill, NC: Frank Porter Graham Child Development Center.

Jaeger, R. M., Gorney, B. E., & Johnson, R. L. (1994). The other kind of report card: When schools are graded. *Educational Leadership, 52*(2), 42-45.

Katz, L. G. (1997). *A developmental approach to assessment of young children* (No. EDO-PS-97-18 RR93002007). Champaign, IL: ERIC Clearinghouse on Elementary and Early Childhood Education.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73*, 232-238.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11.

Steele, C. (1992). Race and the schooling of Black Americans. *The Atlantic Monthly, 269*(4), 68-78.

Walberg, H. J. (1994). Educational productivity: Urgent needs and new remedies. *Theory into Practice, 33*(2), 75-82.