[Elley, W., Hall, C., & Marsh, R. (2004). Rescuing NCEA: Some Possible Ways Forward. *New Zealand Annual Review of Education*, *14*, 5-25]

Rescuing NCEA: Some Possible Ways Forward[1]

WARWICK ELLEY, CEDRIC HALL AND REG MARSH

Abstract:

This article examines the issue of "variability" that resulted in widespread media criticisms of the 2004 Scholarship examinations in New Zealand. The authors argue that the New Zealand Qualifications Authority (NZQA) and the Ministry of Education have taken an ideological position on the use of standards-based assessment, ignoring the evidence both from the literature and from international experience that recognises the immense difficulties in implementing a "pure" form of standards-based assessment. The article questions the capacity of NZQA to administer the National Certificates of Educational Achievement (NCEA), given the outcomes of the 2002-2004 period of implementation. It concludes by recommending 13 key changes or developments that are needed to redress the problems currently plaguing NCEA and Scholarship.

> Clark says NZQA "on the mat" in exam debacle (*New Zealand Herald*, February 15, 2005, p. A3)
>
> NCEA row threatens senior jobs (*The Press*, February 16, 2005, p. A1)
>
> Ministers "knew of NCEA fiasco" (*Dominion Post*, February 17, 2005, p. A3)

As the above headlines from three daily newspapers testify, results from the National Certificates of Educational Achievement (NCEA) and Scholarship results could hardly have been more controversial. Arguably, no educational topic in the 15 years since the introduction of the education reforms has received such widespread and prominent coverage in daily newspapers. The key theme in these reports was the *variability* in the percentage of students receiving

Scholarship awards across different subjects. For example, while only 1 percent and 3 percent of students gained scholarships in biology and physics, 51 percent and 65 percent gained success in accounting and visual arts. These results prompted the Government to introduce distinction certificates for students who did exceptionally well at NCEA level three in subjects where few scholarships were awarded. Two major reviews were also established by the Government, one comprising a group of practitioners and assessment experts with the task of reshaping the 2005 Scholarship examination (known as the Scholarship Reference Group), the other looking at the setting and management of Scholarship examinations and the performance of the New Zealand Qualifications Authority (NZQA). The latter review was undertaken by a panel appointed by the State Services Commissioner.

However, the problem of variability is not restricted solely to Scholarship results. Statistical analyses of the 2002-2003 results by Elley (2003, 2004) and Elley, Hall and Marsh (2005) point to a major variability problem which also existed in the NCEA results for Years 11 and 12. National data supplied to the authors[2] also identify serious anomalies in the Year 13 results for 2004, as well as major fluctuations between years in the percentage of students gaining ratings of "achieved", "merit" and "excellence". The concerns raised by the analysis of these data is well captured in the following extract from an editorial in the Otago Daily Times under the heading "NCEA's future".

> THE DISGRACEFUL debacle over NCEA level 4 (scholarship) has, at least initially, spared the New Zealand Qualifications Authority (NZQA) the scrutiny it deserves over the other three levels. Many teachers and interested observers were willing to give NCEA the interim benefit of the doubt, despite misgivings over the practicalities of a standards-based system. Many must now be disillusioned, because it was not just for scholarship that the results were erratic. In some subjects, or parts of subjects, it was relatively easy for pupils to manage standards of "achieved", "merit" and excellence". In others, it was extremely difficult. The distributions are all over the place. Glaring anomalies will continue to arise, as they did all last week, and public confidence in NCEA has been shaken if not shattered. (*Otago Daily Times*, February 14, 2005, p. 14)

The furore over the Scholarship and NCEA results should not be underestimated, either in its intensity or in the negativity created towards NZQA and NCEA. The earlier claim that "arguably" no educational topic in the past 15 years has received such widespread and

prominent coverage, is based on a comparison with research reported by Roulston (2005). Using a technique for assessing the "prominence" given to articles or reports in newspapers, Roulston identified that education themes or topics received a very low level of prominence over the 12 year period 1988-1999, during the implementation of the educational reforms. This contrasts markedly with the coverage given to NCEA and Scholarship in most New Zealand dailies over the period February-March, 2005. For example, several articles appeared as front page news, or were given substantial coverage on other high profile pages.[3]

The purpose of this commentary is not to expand further on media coverage of the controversy surrounding scholarship and NCEA. Nor is it the intention to report in detail here on the findings of the two reviews of Scholarship. It is sufficient to say that the Scholarship Reference Group (2005) proposed 26 recommendations, of which 25 were immediately accepted by Cabinet. The thrust of the recommendations was that Scholarship should incorporate assessment processes that identify the top group of students, more or less equally (in percentage terms), in each subject area. In effect, the approach to assessment in the future will include a norm-referenced element. The review team appointed by the State Service Commissioner has presented its first report (of two), focusing on the adequacy of the setting and management of the Scholarship examination. The report on the performance of NZQA has yet to be presented. The first report identified a large number of deficiencies, including:

> Officials were focused on operational risks and lost sight of the higher level implementation risks [e.g., the variability in results across subjects]* which impacted on outcomes. Strategies to mitigate these risks were not identified and put into effect in the approach to the 2004 Scholarship to ensure a fair result for students. (State Services Commission, 2005, para. 19) *[Inserted by authors]

The rest of this paper focuses on what its authors see as the underlying flaws in NCEA and Scholarship which led to the two reviews being commissioned. It also suggests key elements which a revision of NCEA should incorporate in order to place it on a sounder educational footing. The writers are not arguing for a return to the old School Certificate and Bursary examination systems. However, elements of these systems are included in the recommendations. The following discussion draws in part upon submissions made to the State Services Commission review (Elley, Hall, & Marsh, 2005).

Problems Besetting NCEA

The introduction of NCEA was intended to bring about a fairer and more valid system of assessment than existed under previous senior secondary school examinations (School Certificate, Sixth Form Certificate, and Bursary). The key argument in replacing the old system by the new was that students would no longer be compared with each other (norm referenced assessment) but against pre-defined written standards that would enable all students to achieve on the basis of their own merit (standards based assessment). As noted by Hall (in press):

> At a superficial level, the philosophy of the NCEA has a logic that is hard to deny. For too long the achievements of a significant proportion of school leavers have not been recognised either because courses were insufficiently tailored to their needs or because a student's partial knowledge was not recognised. In this sense, the NCEA policy has a social justice dimension to it. Unfortunately, the design of the NCEA places this dimension in direct conflict with important educational principles creating a major design flaw; this flaw has to be corrected or these otherwise harmonious principles and pressures will remain in conflict.

The educational principles referred to above focus on issues related to validity (the fragmentation of learning encouraged by NCEA), reliability (the variability in students' results as a consequence of the purist application of standards based assessment espoused by NZQA), and manageability (the increase in the amount of assessment needed to cover all the standards in a subject).

The difficulties associated with the use of standards based assessment (SBA) on a national scale were well documented by a number of educationists *before* the NCEA was introduced (e.g., Codd, McAlpine & Poskitt, 1995; Elley, 1995; Tuck, 1995). Specific problems related to the design of the NCEA were also made very explicit to the Ministry of Education (which had responsibility for the design) both in person and in writing (Hall, 2000). It also appears that overseas experience and literature were ignored (e.g., Wolf, 1995; Wood, 1991). In a short review of British experience in using criterion-referenced assessment (the original term for SBA), Wood (1991, pp. 91-92) concluded:

> To implement criterion-referenced assessment in large-scale examinations is problematical, as the efforts to date show. No doubt there has been a cargo cult mentality in evidence, believing that

> notions like grade-related criteria and grade descriptions would overnight eradicate the bad practice and inequities associated, sometimes unfairly and erroneously, with norm-referenced examinations. The emphasis on criterion referencing in GCSE will be beneficial insofar as it promotes clearer thinking about what students are expected to be able to do and reduces the effects of capricious question selection, but the gains are bound to be modest.

This conclusion is essentially the position reached by the present authors in respect of their analysis of NCEA. The following points expand on this conclusion and exemplify the troubles that have dogged NZQA in its attempts to apply SBA to the NCEA.

International standards in testing and assessment

The first point to be made is that international standards covering both educational and psychological testing have been established by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). These standards are revised regularly and are widely observed by test developers and assessment agencies throughout the world. While they are generally focused on standardised assessment instruments and psychological tests, several of the standards nevertheless apply to the construction, administration, analysis and reporting of public examinations systems. It is ironic that NZQA, with its purist belief in SBA, has moved progressively away from observing these international standards – we are not even sure that NZQA staff are aware of their existence. For instance, NZQA downplays the importance of reliability in assessment, and never reports standard errors of measurement, or evidence of validity, as the international standards require. The staff do not appear to pretest questions (or alternatively, use pre-tested instruments as moderation devices), and they set arbitrary cut-off points for deciding whether students are successful or not. These policies are not defensible when students' futures are at stake. Based on the evidence from the 2004 Scholarship results, as well as the 2002-2004 NCEA grade distributions, it is clear that NZQA has not observed these standards, even when they are entirely applicable.

The evidence for the variability (unreliability) of NCEA

It is not possible to summarise all the evidence relating to the variability of Scholarship and NCEA grade distributions within the space of this

article. Examples of the variability in Scholarship results have already been presented. Other points worthy of note are:

- In the six academic subjects taken by large cohorts of students in the first two years of NCEA (typically 9000+ students at Year 11), 44 achievement standards were assessed. In 32 of these 44 standards, the discrepancy in pass rates between 2002 and 2003 was more than 5%. Any discrepancy over 5% needs explaining, because there is substantial evidence that adjacent (nation-wide) cohorts *rarely* differ by more than 2% or 3% in achievement levels (Elley, forthcoming). Fourteen of the 44 standards showed discrepancies of more than 10%. It is incredible that NZQA ignored these variations and even claimed that all was well.
- The same level of variability occurred in respect of the percentages of students gaining excellence. For example, of the 39,605 students who attempted Mathematics 1.8 (Achievement Standard 90152), 5069 (12.8%) gained excellence; the following year just 81 students (0.2%) out of a total of 40,494 candidates obtained this grade. Other examples include: Biology (Achievement Standard 90176) which rose from 1.3% to 20.9%; Chemistry (Achievement Standard 90168) which rose from 8.6% to 32.4%; Economics (Achievement Standard 90199) which rose from 5.3% to 17.6%; and Mathematics (Achievement Standard 90148) which fell from 31.0% to 1.3%. Other examples showing similar discrepancies are available.[4] The 2004 results were just as variable.
- One of the most blatant shortcomings of the first two years' results can be seen in the contrast between internally-assessed standards and externally-assessed standards. All the internally-assessed Level 1 standards for academic subjects in 2003 showed a *rise* in pass rates (relative to 2002). Yet, three-quarters of the externally-assessed standards in the same subjects showed a *decline*. This is simply not credible. Clearly, the style of moderation used by NZQA is not working.

The significance of these results is threefold. First, movements of the kind demonstrated here should have triggered alarm bells in NZQA that something was seriously wrong. Secondly, that no action was taken is highly indicative of two interpretations – that SBA has become an ideological position rather than just the application of an assessment approach to school qualifications, and/or that NZQA did not have the

expertise amongst its staff to understand the implications of these results. Thirdly, that other key organisations failed either in their analysis of the situation or in their courage, and did not exert pressure to bring about change. We include here in our criticisms the Ministry of Education (who must be held responsible for the design of the NCEA), the New Zealand Vice Chancellors' Committee, the Post Primary Teachers Association, the Secondary Principals Association of New Zealand, and teacher education bodies. That so many organisations with a vested interest in having a successful secondary level assessment system should show either lack of understanding or lack of resolve in dealing with the very obvious failings of the NCEA, is a sad commentary on New Zealand's educational scene.

Reasons for the variability

It is very clear that in its purist standards-based assessment approach to NCEA, NZQA has been promising far more than it could deliver, and, according to Wood (1991), as quoted earlier, more than the examination systems in other countries have achieved. The attempt to use SBA for Scholarship defies all logic. Scholarship presents a competitive situation, and even tertiary students studying assessment courses for the first time quickly learn that norm referenced assessment is the appropriate method for such situations. However, as already noted, a purist SBA approach to the wider NCEA system also spawns other serious problems. Factors known to impact on the stability of assessment results include:

- the vague and open-ended nature of most educational standards;
- variations in the interpretation of standards by examiners and teachers;
- the difficulty in setting questions and tasks to target the intended level of a standard (this is why professional educational testing agencies pretest questions or apply adjustments after the event);
- variations in assessment practices in different schools or sites (e.g., school policies on "re-sits");
- uneven or inadequate moderation practices;
- the impact of assessment context on student performance (e.g., the time available to students to answer questions in an examination);
- undue weight being given to particular assessment criteria in SBA systems.

All of the above have been problems in NCEA. The difficulties have been compounded by the attempt to capture the whole of a year's study in seven or eight standards. Subjects that have a large complex knowledge base cannot be summed up in this way. Consider Level 1 English as an example. For some unaccountable reason there are *four* standards designed to test students' ability to "read and understand text" – one on short texts, one on extended texts, one on unfamiliar texts, and one on visual texts. (Of course, these skills are all highly correlated, and not worth differentiating, but that is another matter.) Generations of research on such topics as the readability of text (Klare, 1984; Elley & Croft, 1989), on background knowledge effects (Yates & Chandler, 1994), and on the cloze procedure (Bormuth, 1966) make it clear that the difficulty of text is the dominant source of variation between students in their ability to comprehend what they read. None of this would appear to have been considered by NZQA examiners. No wonder there are unacceptable variations from year to year, and no wonder we find the extraordinary result that in 2002 *unfamiliar* text was better understood by candidates than *familiar* text.

Moderation

Moderation is the process of checking teachers' assessments to see whether standards are being applied consistently, from class to class, school to school, and year to year. Moderation was intended to pick up discrepancies in teachers' judgments. It may have helped reduce them somewhat, but the size of the variations in Scholarship and NCEA indicates that it has not been effective. Well-endowed research institutions in other countries have been exploring the problems of moderation for years without breakthroughs. NZQA tried to solve the problems by fiat. They can scarcely use the reference-test approach, which was widely used in New Zealand, and other countries, in the 1980s and 1990s, because there are no *marks* given in NCEA. Indeed the NZQA interpretation of SBA wrongly attributes the use of marks to norm referenced assessment only. Teachers' assessments therefore cannot be adjusted to bring them into line with any reference test results. If the Geography teacher says everyone in the class can read maps, and the reference test says that only half the class deserve to pass in the relevant geography standard, who should be failed?

The situation is complicated when students fail a standard first time round; the new system allows for them to "*re-submit*", or try again. Unfortunately, there is no requirement that schools should follow a

uniform policy in allowing students to re-submit (how many times, on the same or another task, and so on); this is a source of unacceptable variation between schools. There is much research designed to explore ways of setting parallel tests, in order to handle such situations. But NZQA appears to be unaware of such research, and recommends merely that teachers talk to the students to gain further evidence, or look for evidence in earlier work, or allow them to try again on the same task, or try again on another (unspecified) task (New Zealand Qualifications Authority, 2002). Such procedures are naïve, and quite unreliable.

Fragmentation of assessment

NCEA grades are also unreliable for reasons not described above. One major problem lies with the fragmentation of assessment into a large number of discrete assessment standards. Year 11 English, for example, is no longer reported as a single overall grade. Instead, students receive nine separate results, one for each assessment standard. This presents both reliability and validity difficulties (Hall, 2000).

In the traditional scheme, students were judged on a three-hour examination, usually supplemented by some internal assessment. This would typically produce an acceptable level of reliability for a course overall (+0.90 or better) with a margin of error of approximately ±10% if expressed as a 95% confidence interval. (For non-statisticians: this means that if a student scored 60, out of a possible 100, we would be fairly sure that the person's real score in the test or examination lies somewhere between 50 and 70). In NCEA, however, students may be externally assessed on *four* or *five* standards in one three-hour examination, each needing to achieve the same level of acceptability (the equivalent of ±10%, if marks were used). However, the standards must compete with each other for examination time. A single standard might only have 45 minutes (or less) of assessment time. This is quite inadequate and, based on the evidence in the literature, would rarely produce a reliability greater than 0.70, with a margin of error (expressed as a 95% confidence interval) greater than ±20 marks on a 100-point scale, and therefore will also have an unacceptable level of validity. Translated to the NCEA four-point scale ("not achieved", "achieved", "merit", and "excellence"), this means that the only safe conclusion is that students who "fail" are probably not "excellent", and vice versa; no safe conclusions can be drawn between other pairs of grades.

The point that needs to be made from the previous paragraph is this: if a subject is broken down into a large number of separate assessment standards, the total amount of assessment needed to make sure that each separate standard meets an acceptable level of reliability is substantially increased. The greater the fragmentation, the greater the amount of assessment needed. This identifies a fundamental flaw in the NCEA standards based assessment model, at a time when the Minister and NZQA claim that teachers are doing too much assessment (the explanation for the workload that teachers carry). The SBA model adopted by the Government actually demands that more, not less, assessment is needed to achieve acceptable levels of reliability. It is also hard to see how more time being spent on assessment and less time on teaching and learning will improve educational performance!

A further problem, one which has been consistently underestimated by NZQA and the Ministry, despite powerful evidence in the literature and frequent challenges by teachers, is that of the backwash effect of the NCEA model on teaching practices. Because assessment tends to drive curriculum and teaching, the division of a subject into arbitrary components for assessment purposes leads to the consequence that standards may become treated in isolation from each other. Yet one of the Government's main goals is to achieve an education system that fosters life-long learning and people contributing to knowledge generation and change. Educationally, the kind of intellectual and practical skills that are needed to support these goals – the ability to integrate knowledge from different areas of learning, to transfer ideas/concepts across divisions of a course and between subjects, and to create new ways of seeing and doing things – comes not from assessment against separately defined standards, but from an integrated approach to assessment. In short, assessment needs to integrate learning across boundaries, not just measure performance in isolated packages. Paradoxically, NZQA claim that their separate standards model provides the flexibility needed by schools to create tailor-made courses for their students – courses that transit different subjects. But a course is not simply a collection of achievement or unit standards: it must be a coherent whole that provides genuine integration of the parts. The NCEA model in fact fails to achieve genuine integration because it gives absolutely no weight to the notion of a course as a whole; only the parts are assessed and this is done in discrete packages (Hall, 2000; Hall, in press).

Reporting under NCEA

One commonly stated benefit of the NCEA scheme is the increased detail in the information that employers and tertiary institutions would receive about students – not just a mark of 75 percent in English, say, but a grade on as many as nine different aspects of English. This claim has persuaded many. *But what use is such information if it is unreliable – based on chance variations and very difficult to interpret correctly?* Most of the nine aspects are positively correlated, so that any apparent strengths or weaknesses in a student's profile are likely to be due to random errors in the assessment process, rather than real differences in the achievement profile. Moreover, the particular profile of the "standards" chosen for recording will rarely match the profile of skills that any particular employer or organisation is seeking. And skills are not static. What a student "knows and can do" today will be different in 12 months time. Of more use to employers and other users of NCEA results, such as universities and other tertiary institutions, is the far greater stability provided by a whole course result. This is much more likely to indicate the student's potential for learning new skills, whether on the job or in further education, than the fragmented and unreliable profile currently provided by NCEA.

Manageability

The preceding evidence and arguments lead to the conclusion that NZQA has not demonstrated the capacity to undertake its role in implementing and managing the NCEA and Scholarship examinations. In part it can be claimed that NZQA has been given an impossible task; the NCEA has all the hallmarks of being a bottomless pit for draining educational resources. The system is unmanageable: there are too many standards; too much assessment needed to support the standards; too many difficulties with setting examinations and tasks to meet vague, pre-defined standards; too much time spent on cumbersome moderation procedures which do not work properly; too many difficulties in communicating clearly with teachers, students and other stakeholders; and too many manageability issues that require attention to detail beyond the capacity of NZQA to handle. Added to these problems has been NZQA's unwillingness to engage with international literature and experience, an ideological adherence to a purist SBA approach, and a complete failure to monitor properly the variability (unreliability) of NCEA and Scholarship results.

Some proponents of the system claim that it will settle down in due course, that the present set of discrepancies are only "teething troubles". It should be remembered that NZQA staff have been working on the grand SBA plan for nearly 15 years. They have wasted millions of dollars trying to train teachers and develop resources in order to make the system work. Millions have also been spent on setting up the technology support for running and recording information on the system, including each student's record of achievement. This amount of investment itself is likely to become a reason for resisting change – no government is ever likely to admit that it has got it badly wrong when the scale of the operation matches that of NCEA. The current cry from many principals and teachers is that the NCEA needs more resources. As pointed out by Hall (in press), "this is a totally unacceptable solution.... Given the high level of contestability for government funding from all areas of education, there is no case for supporting a system that diverts substantial resources from the core functions of teaching and learning to the administration of a flawed assessment design." The point that needs to be made is that policy on assessment should be grounded in sound educational and measurement principles. These should lead development, which *should never be established on a legacy of mistakes already made.* Having said this, the writers of this article acknowledge that the way forward is not to return to the old system; it is now better to start from the present structure and introduce key changes and modifications that might alleviate the deficiencies. Blending norm referenced assessment with SBA should be part of this revision.

Elements of A Future Design for NCEA

In this section we set out the key elements of what is needed to put NCEA on a strong footing. What we present is *not* a detailed design, but simply the main elements that need to be present in a future system. A detailed design requires a lot more than can be presented here, and also requires participation of stakeholders and people with expertise (principals, teachers, employers, and curriculum and assessment experts). To move ahead, we need a system which:

- is fair and credible;
- encourages strong and coherent course design;
- has a reduced assessment workload for teachers; and
- is acceptable to parents, schools, employers and tertiary institutions.

The following features are recommended in the design of a practicable system that builds on what has gone before, while avoiding the least satisfactory features of the current model.

1. *Students should be assessed across a large enough body of knowledge and skill to produce reliable and valid results.*

Breaking a subject up into eight or nine components and producing a dependable grade for each one imposes a heavy burden on teachers and examiners, and has so far produced very dubious results. Under the present NCEA model, students are often evaluated on a very limited range of tasks, and within a short time period (e.g., 30-45 minutes within an exam). The evidence to date is that the results are questionable in terms of their reliability, their meaning (what exactly is being assessed), and the usefulness of the information obtained. A more stable and useful result would be obtained by basing the assessment on a number of distinct questions or tasks, each sampling in a non-trivial way a significant component of a course. The results for each question should be combined or aggregated to provide a more rounded picture of students' achievement. The award of credit from an external examination should be based on at least two hours of testing (not 30-45 minutes) in order to provide an acceptable level of reliability. As mentioned below, it is also desirable that the results for an external examination be combined with internally assessed marks or grades.

2. *Emphasis should be given to whole-course design and the integration of assessment across the parts of a course.*

This feature follows from the first. As already pointed out, a course is not simply a collection of achievement and/or unit standards. It is a dynamic and complex working unit in which teachers and students interact, and learning takes place. Course coherence is a central requirement for ensuring that the learning is meaningful and integrated. It implies that the parts or components of a course are sensibly connected by their content, and that this is reflected by the course structure and sequence as well as the selection of assessment tasks. Within this framework, assessment tasks should be selected such that there is a focus not only on the mastery of basic knowledge and skills (as might be identified in the course components) but also on the connections between course components. Assessment must therefore include a focus on integrative skills – analysis, synthesis, transfer, problem solving, and other "higher" processes. In effect this means that

the design of the course comes first, while decisions on assessment should arise out of this design. The current focus on *achievement and unit standards* leads teaching and learning in the reverse direction – assessment comes first and course design then follows. This is at the heart of the fragmentation issues surrounding NCEA.

3. *Assessment should include a whole-course grade.*

In line with the above whole-course principle, it is essential that students receive an overall grade for a course, not simply a profile of grades on the parts of a subject. A whole-course result reinforces the notion of course coherence and the need to recognise that the "whole" has meaning. It represents an overall judgment of a student's learning in a subject, and provides a stronger basis than part-course performance for guiding students (and their parents) on the future study and employment directions that they may elect to take.

4. *Where appropriate, assessment should include a part-course performance profile.*

Where a subject is able to be divided into two or three distinctive components, a simple profile of performance should be generated (e.g., English might be divided into reading, writing and oral language). A key point is that the profile should complement, not replace, the whole-course reporting of results. The profile should be standardised nationally – the same components should be used for the same course across the country and should use the same grading scale. However, caution should be shown in using the profile as a basis for awarding separate credit for each component. Unless it can be shown that each component yields data that are reliable and stable across cohorts, credit should only be awarded at the whole-course level.

5. *Assessment should blend internal and external assessment.*

The present system does not blend in any way internal and external assessment: some standards are totally internally assessed, the rest are totally externally assessed. No actual blending takes place. As pointed out by Hall (2000), this exposes the weaknesses in each approach to assessment. External examinations are important because they give an independent measure of each student's performance. They are also useful as a moderation device for judging the comparability of teachers' assessments across schools. They also have high public credibility. However, "one-off" external examinations are limited in that in most

subjects there are usually skills that cannot be effectively assessed in an examination environment (content validity is an issue). One-off assessments are also less reliable than results obtained by sampling student performance over time. Internal assessments are useful because they can cover a wide range of knowledge and skills, and they enable judgments to reflect a student's typical performance in a course over time. While the balance between internal and external assessment should be agreed upon on a course by course basis (e.g., mathematics might be different from English), each should receive a significant weighting to enable the strengths of each to be incorporated.

6. *A systematic procedure should be developed to pre-test examination questions.*

Over time, a bank of valid examination questions should be established in each subject based on a proper pre-testing of questions and their marking schemes. Models for doing this already exist, the Assessment Resource Banks (ARBs) developed and administered by the New Zealand Council for Educational Research, for example. This would have the benefit of ensuring some predictability of how questions will function, thus avoiding or reducing the need to adjust results after the event. If no other way can be found to ensure the security of questions being trialed, then the pre-testing should be undertaken overseas, as was done here in New Zealand for the Australian Commonwealth Scholarship examinations in the 1960s. The pre-testing of questions will also have the benefit of encouraging greater confidence in moderation comparisons based on external examinations. Such questions, once used in an examination, could become the basis for providing schools with exemplars of student work indicative of different levels of performance.

7. *Students' results should be assigned a mark or grade on a 10-point scale, at least, in each subject and component of a subject.*

Locke and Hall (1999) found that teachers are generally comfortable using a mark/grade scale of 10 points or more. With such a scale, employers and tertiary institutions would be able to make better selection decisions, and brighter students would have more motivation to excel. The marks used in assessing should be able to be combined across different tasks and course components (using an appropriate method of aggregation) to produce totals for each subject. Of interest here is the Waikato Certificate of Studies–English (WCS-E), a criterion referenced programme in English at Years 12 and 13 (this programme

counts credit towards NCEA). WCS-E reports results on a 10-point scale, and also includes an aggregate result across assessment components to yield an overall course grade. Reliability estimates for this overall course grade typically reach 0.9 (Hall, 2004). (For further information about the programme and the scale used, see Hall, 2004; Locke, 2001, 2003; and Locke & Hall, 1999.)

8. *While clear SBA criteria could not be reliably set and applied for all 10 grades, they could still be described and published, after the assessments have been made, for (i) the "top" grade and (ii) the "passing" grade.*

This feature acknowledges the potential value for having descriptions of performance available to assist people in their interpretation of NCEA results. The description of these "standards" of performance would necessarily be general, rather than specific, and would attempt to portray a "best fit" description of the typical abilities that these students have demonstrated. Then students, employers and other stakeholders would have a clearer idea of what the assessments meant. Intermediate grades between "top" and "passing", would be inferred by the user.

9. *The percentage of students receiving each grade should be guided by, but not necessarily conform to, those of other years, and other subjects.*

This proposed feature of a modified NCEA system will assist in dealing with the variability problems currently bedevilling NCEA. The important point is that large differences between subjects and between years should be investigated and explained. If they cannot be explained, then some form of adjustment after the event is necessary. For example, if pass rate discrepancies greater than 5% are observed, examiners should be required to defend them or accept adjustment to the grades of students. Clearly this introduces an element of norm referencing to the system, but this is a very practical solution to a critical problem with the current NCEA system.

10. *Any other method of moderation used to align marks and grades should not involve large effort on the part of teachers or assessors, and should ensure similar practices across schools.*

As already indicated, assessment in most subjects should be a blend of internal and external assessment, with the external examination being used to moderate school based assessments. Subjects that warrant a large internally assessed component (e.g., English, Maori, art, music, social studies), should have strict constraints put on the average and range of their internally assessed grades. If the external examination

proves not to be the most appropriate method, a moderation test or common assessment task could be used. In some subjects, sample scripts could be sent to moderators for checking. In addition, exemplars of students' work, with commentaries, should be used to guide teachers' assessments, as at present. However, the procedures that are used should be the same for each school and should be both manageable and applied consistently.

11. *NZQA should employ a significant number of senior staff (at least) with advanced and broad-based qualifications in educational measurement.*

There should be in all major examination authorities, such as NZQA, senior staff of the same calibre that would be found in any national testing organisation overseas. This is essential not only to ensure that valid and reliable assessment procedures are in place, but also to encourage public confidence in the assessment system. It is particularly important for people who who may have to submit their New Zealand qualifications overseas for scrutiny – the credibility of New Zealand qualifications is at stake.

12. *NZQA should adhere to the relevant standards in "Standards for Educational and Psychological Testing".*

NZQA should recognise that SBA is no different from all other forms of assessment in needing to follow established principles and guidelines in all phases of the assessment process. It is of interest that information relating to the validity, reliability and administration of the Waikato Certificate of Study–English is provided in an annual report by the external evaluator (Hall, 2004).

13. *Universities, colleges of education and other teacher education institutions should review their course offerings in assessment to ensure that teachers receive a comprehensive and non-ideological training in educational measurement.*

There is a need for all teachers to be well-versed in the theory and practice of educational assessment and measurement. Teacher education courses should not promote an ideological stand of the kind promoted by NZQA. Teachers need to have the knowledge and skills to assess their own students validly and reliably, and they also need to be critically aware of current issues in assessment so that they can contribute locally and nationally to the development of sound assessment practice.

The Examination Structure: Reducing The Amount of Assessment

The overall structure of the examination system needs reconsideration. NZQA has not impressed with its expertise and its policies; other organisations should be encouraged to develop assessment schemes (e.g., the Waikato Certificate of Studies could be expanded into other subject areas) in order to avoid the monopoly currently enjoyed by NZQA – variety often promotes strength in a system. If these schemes can be shown to meet quality assurance standards (as has been done by the Waikato Certificate of Studies–English at both Years 12 and 13), then they should count towards credit for NCEA as well as contribute to university entrance.

At present, there are examinations at each of the last three years of school (Years 11, 12 and 13), which is unusual, internationally, and totally unnecessary from an education standpoint. As fewer than 15 percent of students leave school from Year 11, it may be possible to phase out any formal exams at this level and provide schools with standardised literacy and numeracy tests to be taken by all students, or by those who expect to leave from this level. The results would be included on the records of such students. They would of course, have school reports on their other skills to show potential employers.

As for the high-stakes qualifications of Years 12 and 13, half-year courses could be designed for all subjects offered by a school, with examinations and internal assessments collated and recorded at the end of each half-year. Over the two years, students could build up a more elaborate profile of results, and would not need to wait 12 months to repeat a failed examination. They could leave at the end of Year 12 with some credible qualifications. Academic subjects like mathematics, science and English could be offered cumulatively in basic and advanced courses, and so allow for the wide range of ability found in the senior school.

Cultural, vocational and practical subjects, which appear to lend themselves more readily to a standards-based philosophy, may continue to use this approach, if the majority of relevant subject teachers favour it, and the results are reliable and defensible. However, the teachers of such subjects should be encouraged to experiment with a policy that describes the standards *after* the assessments are completed, rather than before. Assessing students in relation to pre-set generic standards has proved to be impracticable in many subjects. This is not to say that students should not be given assessment criteria in advance of doing an assessment task or sitting an examination. However, giving students

information that helps them understand what is expected of them is a long way from assuming that clearly defined standards have been set. The above suggestions will have the additional positive impact of reducing the amount of assessment that is demanded from the current system. As already pointed out, there is an urgent need to reduce the number of standards or components in each subject. The more a subject is broken into separate assessment components, the greater the workload for all who participate in the system. We believe the suggestions presented here should have appeal to teachers, students and administrators alike.

Notes

1. This article includes material presented in two submissions made by the writers to the *Review of the Adequacy of the Setting and Management of the 2004 Scholarship Examinations and the Performance of NZQA,* conducted by the State Services Commission (2005).
2.. We thank Simon Peek, Associate Principal, Macleans College, Auckland for supplying these data.
3. A systematic analysis of newspaper coverage of the NCEA and scholarship will be reported in a forthcoming paper. For further information, contact Cedric Hall, Victoria University of Wellington College of Education. Email: cedric.hall@vuw.ac.nz.
4. Data supplied by Simon Peek.

References

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

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*(3), 79-132.

Codd, J., McAlpine, D., & Poskitt, J. (1995). Assessment policies in New Zealand: Educational reform or political agenda? In R. Peddie & B. Tuck (Eds.), *Setting the standards* (pp. 32-58). Palmerston North: Dunmore Press.

Elley, W. B. (1995). What is wrong with standards-based assessment? In R. Peddie & B. Tuck (Eds.), *Setting the standards* (pp. 78-98). Palmerston North: Dunmore Press.

Elley, W. B. (2003, February 19-25). New assessment system does not pass test. *Education Review*, pp. 5-6.

Elley, W. B. (2004, June 16-22). NCEA still does not pass the test. *Education Review*, p. 7.

Elley, W. B. (forthcoming). On the remarkable stability of student achievement standards over time. (To appear in a future issue of the *New Zealand Journal of Educational Studies*)

Elley, W. B. & Croft, A. C. (1989). *Assessing the difficulty of reading materials: The noun frequency method*. Wellington: New Zealand Council for Educational Research.

Elley, W. B., Hall, C., & Marsh, R. (2005a). *Submission to the State Services Commissioner: Review of the adequacy of the setting and management of the 2004 Scholarship examinations and the performance of NZQA.*

Elley, W. B., Hall, C., & Marsh, R. (2005b). *Second submission to the State Services Commissioner: Review of the adequacy of the setting and management of the 2004 Scholarship examinations and the performance of NZQA – Some possible ways forward.*

Hall, C. (2000). National Certificate of Educational Achievement: Issues of reliability, validity and manageability. *New Zealand Annual Review of Education, 9*, 173-196.

Hall, C. (2004). *Statistical analysis of the Waikato Certificate of Studies (English) 2003*. Report to the Chairperson, University of Waikato Certificate of Studies Management Board.

Hall, C. (in press). NCEA: Is there a third way? In J. A. Codd & K. Sullivan (Eds.), *Education policy directions in Aotearoa New Zealand: Is there a third way?* Palmerston North: Dunmore Press.

Klare, G. R. (1984). Readability. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 681-744). New York: Longman.

Locke, T. (2001). English and the NCEA: The impact of an assessment regime on curriculum and practice. *Waikato Journal of Education, 7*, 99-116.

Locke, T. (2003). *English study at years 12 & 13: Quality management handbook.* Hamilton: The University of Waikato, School of Education.

Locke, T., & Hall, C. (1999). The 1998 Year 12 English Study Design Trial: A standards-based alternative to unit standards. *New Zealand Annual Review of Education, 8*, 167-189.

New Zealand Qualifications Authority. (2002). *NCEA Update, 11.*

Roulston, D. (2005). *Educational policy change, newspapers and public opinion in New Zealand, 1988-1999.* Unpublished PhD in Education thesis, Victoria University of Wellington

Scholarship Reference Group. (2005). *A report prepared for the Associate Minister of Education*. Wellington: Ministry of Education.

State Services Commission (SSC). (2005). *Report on the 2004 Scholarship to the Deputy State Services Commissioner*. Wellington: SSC.

Tuck, B. (1995). Issues of objectivity in assessment: A plea for moderation. In R. Peddie & B. Tuck (Eds.), *Setting the standards* (pp. 59-77). Palmerston North: Dunmore Press.

Wolf, A. (1995). *Competence-based assessment.* Buckingham: Open University Press.

Wood, R. (1991). *Assessment and testing: A survey of research.* Cambridge: Cambridge University Press.

Yates, G., & Chandler, M. (1994). Prior knowledge and how it influences classroom learning: What does research tell us? ***set:*** *Research Information for Teachers, 2*, Item 6.

The authors

Warwick Elley is Emeritus Professor of Education of the University of Canterbury, now living in Auckland. He has conducted numerous surveys of achievement in the past, and chaired the steering committee of the IEA surveys in 32 countries. He now works on assessment issues, and studies of literacy in developing countries.

Cedric Hall is Professor of Education and Deputy Dean of the Victoria University of Wellington College of Education His teaching and research interests include assessment and evaluation, professional education and training, educational research methods, learning and teaching, and policy in higher education.

Reg Marsh was Professor of Education at Victoria University of Wellington from 1972 to 1987, before travelling overseas to similar posts in various subjects, including educational assessment. Now retired, he is an honorary Associate Professor at the Auckland Clinical School of Medicine.