[Hall, C. (2000). National Certificate of Educational Achievement: Issues of Reliability, Validity and Manageability. *New Zealand Annual Review of Education, 9,* 173-196]

National Certificate of Educational Achievement: Issues of Reliability, Validity and Manageability

CEDRIC HALL

Abstract:

This paper examines the implications of the NCEA approach to standards-based assessment, in particular the reliability of assessment against separate achievement standards, and the pedagogical implications of the policy of non-aggregation. The paper argues that assessment against separate standards is unlikely to yield sufficiently reliable results to satisfy public credibility, and that the same focus is likely to foster a "bricks without mortar" approach to course design, delivery and assessment. The paper also argues that the NCEA involves manageability issues similar to that of unit standards. The recommendation is made that the designers of the NCEA rethink the basis by which internal and external assessments could be blended within a standards-based system so that the strengths of each approach to assessment are emphasised.

In November 1998, as part of a policy initiative called "Achievement 2001", the Minister of Education announced that a new qualification, the National Certificate in Educational Achievement (NCEA), would be introduced at the senior secondary level in 2001. The introduction date has subsequently been deferred to 2002 to allow more time for implementation issues to be worked through and to give teachers and schools the opportunity to prepare for the new system (eg. undertake professional development). Implementation of the policy is under the control of the Ministry's Qualifications Development Group (QDG). Key features include the following:

- the NCEA will replace existing qualifications – School Certificate, Sixth Form Certificate and University Bursary;
- broad-based achievement standards will be created for most conventional school subjects at 5th form, 6th form, 7th form and scholarship levels (5th form provisions commence in 2002);
- each standard will specify the learning outcomes and criteria for the award of credit, and each standard will include criteria for pass, merit and excellence;
- for each school subject at each level, at least half of the credits (12 out of 24) will be assessed externally, and external assessment will be by examinations, except for skills which cannot be properly assessed in that form (e.g., art, where examinations are inappropriate);
- credits will be able to be accumulated from achievement standards, unit standards in non-conventional subjects, other approved examinations or qualifications, and industry-developed unit standards;
- in order to gain an NCEA at a particular level, 80 credits would need to be achieved, of which 60 must be at the level concerned (i.e., students may study credits at different levels in any one year).

Policy Drivers

During the 90s, attempts were made by the New Zealand Qualifications Authority (NZQA) to replace the existing norm-referenced examination system with standards-based unit standards. Currently, the examination system remains dominant, although some schools use unit standards, in both conventional and non-conventional subjects. The NCEA policy seeks to address the following perceived problems/issues with the current mixed system of assessment:

- the micro-definition of outcomes in unit standards;
- the workloads for teachers and students and general manageability of standards-based assessment;
- the manageability of the moderation of standards;
- the difficulty of applying unit standards to conceptual learning;
- the lack of recognition for excellence in unit standards;
- the need to establish a credible standards-based system to replace the norm-referenced focus of current external examinations;

- public disquiet about comparability and fairness in internal assessment;
- the lack of recognition on the National Qualifications Framework of traditional examinations.

Expert Panels

In pursuance of this policy, the QDG has established expert panels in English, Te Reo Maori, mathematics, accounting, social sciences, arts, science, languages, technology, and health/physical education/food and nutrition. These panels met three times over the period May-September 1999. Their role has been to develop level 1 standards and related materials for the areas of their expertise. According to a pamphlet released by the Ministry of Education (1999a), panel members were selected through rigorous application of criteria which included "high professional standing", "an ability to see the wider picture", and an ability to "contribute to a group development process". A process of refereeing was used. The QDG has also appointed six project facilitators whose job it is to manage consultations with stakeholders and ensure that the NCEA is trialed and implemented properly. Most of these facilitators have work experience with NZQA or with the unit standards system.

Secondary Schools Sector Forum

In 1999, a secondary schools sector forum, comprising a range of school principals, teachers, and tertiary and industry sector representatives, was established to advise the Secretary for Education on "macro" issues relating to the introduction of the NCEA. At its initial meeting, the forum made recommendations which included several basic design principles which the NCEA should follow. These principles emphasised notions of validity, fairness, inclusiveness in coverage of the school curriculum, manageability for teachers and schools, external credibility, clarity for parents and communities, and cost effectiveness (Ministry of Education, 1999b, p. 4). The forum recommended that the Ministry and Post Primary Teachers Association jointly undertake consultation with schools. This was done through 25 regional meetings with school representatives and a questionnaire which was responded to by over 11,000 teachers. The consultation process covered matters such as the standard required to obtain credit at level 1, minimum requirements for literacy and numeracy, the award of "scholarship", the balance between

internal and external assessment, reporting of students' achievements, and moderation of internal assessments. In its summary, the report of the forum identified several factors critical for successful implementation. These included (Ministry of Education, 1999b, pp. 20-21):

- *quality of assessment materials:* the documentation of achievement standards, exemplars, assessment schedules and associated guidelines should be of the highest quality;
- *support of teachers:* teachers would not "buy in" to the system without access to clear information, on-going consultation and assurance of credible moderation systems;
- *adequate resourcing:* teachers would need access to banks of assessment tasks and schedules, the opportunity to undertake professional development, and the time to carry out all the tasks asked of them;
- *on-going communications:* a system of on-going communications would be needed which would encourage two-way consultation between the Ministry and schools and teachers.

Issues Not Dealt with in the Forum Report

The rest of this paper concentrates on two issues which are not directly addressed in the forum report but require in-depth treatment if the NCEA is to progress on a sound footing. These issues cover:

- the reliability of the assessment of individual achievement standards;
- the pedagogical and practical implications of assessing students against separate (non-aggregated) achievement standards.

These issues focus very much on notions of validity and reliability, which are central to good assessment practice. Validity is concerned with "fitness for purpose", and generally includes a consideration of the extent to which an assessment task samples fairly the knowledge, skills or values that it is intended to cover. Underpinning this focus on the content of what is being assessed are some important considerations. Are tasks relevant and non-trivial in their coverage of the course content? Have expectations of students been reasonably communicated to them before they undertake an assessment task? Are teachers and external examiners well versed in the standards being assessed? Are there unintended consequences or side-effects of the assessment process which have a negative impact on teaching and learning? These are all important matters which need to be considered when an assessment system or particular assessment procedure are being put in place. Indeed, most of these appear to be well recognised in the thinking behind the NCEA to date. However, a later section of this paper deals with some potentially negative pedagogical consequences of the NCEA. The consequences are serious enough to warrant strong reflection on the current NCEA policy of non-aggregation across standards within a subject.

On a similar tack, this paper addresses a serious problem concerning the reliability of assessment for the NCEA. Whereas validity focuses on fitness for purpose, reliability focuses on "accuracy" of measurement. Most readers will be familiar with the notion of "measurement error". For example, political polls in New Zealand typically report a margin of error of ± 3% in their analyses of the standing of political parties. Research in the 1970s on British secondary examinations (e.g., O-level) reported estimates of error typically around ± 7% (Willmott & Hall, 1975). Over recent years, reliability has tended to receive a "bad press" in the assessment literature because it has too often taken precedence over validity in both the design of assessment tasks and research on assessment. However, while validity should rightly take precedence – assessing what should be assessed is the first consideration – little is gained if the assessments made of student work are of insufficient accuracy to place credence on the results. The next section looks at this issue in respect of the NCEA policy of assessing students against each separate standard rather than drawing out an overall (aggregated) result for the subject.

The Reliability of Individual Achievement Standards

The NCEA is to be a standards-based system of assessment. Standards-based assessment requires that students are evaluated against written specifications which define what students should know or be able to do in order to meet the standard being assessed. This sounds relatively straight-forward in theory, but is actually quite difficult to carry out in practice. It may be helpful first to consider the nature of an "educational standard". The following brief explanation gets to the heart of the problem.

One useful distinction when thinking about educational standards is that between "student" standards and "system" standards. Student standards focus on the particular knowledge, skills or values required

of learners, and are typically embedded in the design, delivery and assessment requirements of the individual courses or training that make up a qualification. Such standards may be subdivided as, for example, by the distinction between competency standards (pass/fail) and achievement standards (eg. A, B, C, D etc., or distinction, merit, pass). System standards, in contrast, focus on what educational providers, teachers and trainers have to do in assuring the quality of their educational activities. Such standards exist to support learner achievement of student standards and may include what Harvey et al. (1992) call "service" standards (e.g., "...teachers will return all marked work to students within two weeks"). Most commonly, system standards focus on requirements for accreditation and approval, along with ongoing monitoring of the quality of the education being provided.

Typically an educational standard has two main components which can be described in advance. In the case of a student standard, the first component focuses on the "content" of what has to be achieved and identifies the knowledge, skills and values that are being addressed. The content is often written in the form of intended learning outcomes. A system standard has a parallel structure – the content focuses on what has to be done to support the education being provided.

The second component deals with the "level" of performance which is acceptable for meeting the standard. Unfortunately, the clarity with which this may be stated varies considerably with the type of content being addressed. For example, a student standard dealing with typing speeds is easier to define with precision than a standard which focuses on the level for acceptable writing skills.

Unfortunately, the notion of a standard is more complex than simply specifying in words the content and level. Most educational standards (student and system standards) require subjective interpretation – the specified words are not enough. In the case of student standards, every marker or examiner carries his or her own interpretation of the standard into their assessment of student work. Furthermore, their application of that standard interacts with, and is influenced by, the particular representation of the standard in the work of the student. Two students may express their understanding in different words, actions or behaviours. Both students may meet the standard but the marker will need to interpret their different performances to determine (judge) whether or not each has met the standard. The standard thus lies not only in the words that are specified, but in the mind of the marker, the representation of the standard in each student's work, and the interaction of the student's representation and the marker's interpretation. ***A student standard, therefore, represents a judgement exercised in relation to interpretations of student work, taking account of the content being assessed and the specified assessment criteria.*** A parallel definition can be given for a system standard.

From the perspective of reliability, the above definition and discussion highlights the fact that all assessment of student work is prone to variation. Two examiners may think they have the same understanding of a standard, but in all likelihood variations (often subtle) exist between the interpretations of the different markers.

However, the problem of reliability does not simply lie with the interpretation and application of a standard. We also know that students do not behave consistently when being assessed. Their performances fluctuate because of factors which may have nothing to do with their understanding of the content being assessed. The literature has shown that students can be affected by health, home factors which coincide with an important assessment, simple misunderstandings of the instructions associated with a test or a task, the particular choice of questions which are answered, and so on. Students sometimes have "off" days; sometimes they are lucky and strike an assessment task which they are particularly well suited to in terms of their preparation or personal background. These particular problems are accentuated in contexts which depend on a "one-off" assessment of a student's achievement. External examinations, which appear to have high credibility with the public, fall into this category.

Earlier reference was made to research on British examinations which yielded estimates of error of the order of ± 7%. Technically, these estimates are referred to as the Standard Error of Measurement (SEM). In order to calculate the SEM for examination marks, it is first necessary to estimate the level of reliability of the results. Operationally, this is usually calculated by measuring the extent to which each student has produced consistent results, either on different assessments (e.g., by taking a test twice or taking parallel versions of the same test) or across the different components (questions) of a single assessment.[1] If a student demonstrates consistency, then the interpretation is made that the results are reliable.

The scale on which estimates of reliability are reported usually ranges from zero (no reliability) to 1.0 (perfect reliability). In the context of educational assessments, estimates that reach 0.9 are generally considered to be "high" while values over 0.8 are considered to be satisfactory or "workable". Most of the estimates for the British examinations referred to above (eleven subjects were studied) exceeded 0.8 but only two (for French and mathematics) reached 0.9. There have been no published reports of a similar kind of the reliability of New Zealand public examinations. However, during the mid-1980s the writer was supplied with information by the then New Zealand Department of Education concerning School Certificate English. This information enabled the writer to estimate the reliability of the marks concerned to be close to 0.9. This value had an associated SEM of approximately 5.5%.[2]

An important consideration is how one interprets the SEM for an examination. If we take the value of 5.5% given above for School Certificate English, and imagine that a particular student achieves a score of exactly 50%, then there is a reasonable chance that the range 50 ± 5.5 (ie. 44.5 to 55.5) would capture the student's typical performance level in the subject. We would be even more certain that the range 50 ± 11.0 (twice the size of the SEM) would cover the variations you might get in the student's score as a result of the many chance factors that could influence his/her performance. The reason for taking twice the SEM as a guide is that it approximates closely what is known as the "95% confidence interval" for a given score.[3] That is, we would be about 95% confident that the range 39% to 61% captures the achievement level of the student in the subject.

Returning to the information (albeit limited) that we have on external public examinations, we have evidence that the reliability in most cases is likely to exceed 0.8, and in some cases 0.9. Table 1 below sets out a range of "hypothetical" estimates of reliability for external examinations (0.95 down to 0.70) and provides the associated SEMs and 95% confidence intervals for these reliability estimates. It will be noted that each reliability estimate is paired with three different values for the standard deviation of the scores of students. The standard deviation gives an indication of how much students' scores are spread over the scale being used to assess performance. We know from available evidence (e.g., Willmott & Hall, 1975) that standard deviations typically vary between 14% and 20% in public examinations. The actual size of an SEM is dependent not only on the reliability of the marks of students, but also on the amount of the scale being used.[2]

Table 1 Theoretical Estimates of Examination Reliability and Associated Estimates of the Standard Error of the Measurement (SEM) and 95% Confidence Interval

Estimate of Reliability		Standard Deviation 14%	17%	20%
0.95	SEM	3.1	3.8	4.5
	95% Conf. Int.	± 6.1%	± 7.4%	± 8.8%
0.90	SEM	4.4	5.4	6.3
	95% Conf. Int.	± 8.6%	± 10.6%	± 12.3%
0.85	SEM	5.4	6.6	7.7
	95% Conf. Int.	± 10.6%	± 12.9%	± 15.1%
0.80	SEM	6.3	7.6	8.9
	95% Conf. Int.	± 12.3%	± 14.9%	± 17.4%
0.75	SEM	7.0	8.5	10.0
	95% Conf. Int.	± 13.7%	± 16.7%	± 19.6%
0.70	SEM	7.7	9.3	10.9
	95% Conf. Int.	± 15.1%	± 18.2%	± 21.4%

Looking at Table 1, we can see that if an examination achieves a reliability of 0.90, with a standard deviation of 17%, the SEM is 5.4 and the associated 95% confidence interval is ± 10.6. In other words, if a student scores at the cut-off point for a particular grade boundary, we can say that we are 95% confident that the student's underlying level of achievement in the subject is somewhere between 10.6% below the cut-off point and 10.6% above the cut-off point. Another way of looking at this is to think of a ten point grade scale and imagine the margin of error associated with this scale. Based on Table 1, we would be 95% confident that a student who scored, say 7, has an achievement level somewhere in the range 6 to 8. Realistically speaking, this level of accuracy is about as good as could be expected in most assessment situations, although one might aim to improve on this, as long as it did not compromise validity.

Now consider the worst case scenario in Table 1: a reliability estimate of 0.70 associated with a standard deviation of 20%. This produces a SEM of 10.9 and a 95% confidence interval of ± 21.4%. In other words, to the nearest whole number, someone who scores 50 can be thought of as lying somewhere between 29 and 71. If the example of the ten point grade scale is used again, the 95% confidence interval covers a range of ± 2 grades to the nearest whole number. That is, someone who scores 7 can be thought of has having an achievement level somewhere between 5 and 9.

There is a reason for choosing reliability estimates of 0.9 and 0.7 from Table 1. These represent the writer's best guess as to what the designers of the NCEA might expect of their external assessments of students. The higher reliability value, with some variation across subjects, would be obtained if an aggregated examination total was used to describe a student's level of achievement in a subject; the lower value would be obtained for the reliability of assessments against individual standards within the subject. This "guess" requires further explanation.

First, we need to acknowledge that reliability is an issue. It may not be as important as validity, but it must be considered if NCEA is to achieve public credibility. The two aspects of reliability mentioned earlier – variations in the interpretation of a standard and inconsistency in student behaviour – will place a ceiling on reliability which no assessment system will overcome. What we should hope for is that the NCEA will produce improvements on what is currently obtained in public examinations. However, from the perspective of reliability, this is very unlikely if the unit of measurement shifts from the overall performance of students in a subject – an aggregated total – to assessment and reporting on the basis of performance in individual standards.

Let us imagine that we are dealing with a subject which, at level 1 of the National Qualifications Framework, is to be tested through eight standards, four of which will be internally assessed by teachers and four by an external examination. This, in fact, is similar to the draft standards developed for Year 11 (fifth form) English. Now let's imagine we decide to administer the assessment for just one of these standards as a full three hour examination. Based on what we know about external examinations, and taking account of the improvements that have been made to examination practice since the mid-70s, we would expect the reliability of this examination to be around 0.9, possibly higher. We need not be concerned about the method of estimating reliability at this point – let's just suppose that a suitable procedure exists for dealing with standards-based assessment. Note that the figure of 0.9 relates to a three hour assessment. However, under NCEA, the assessment of this standard might well be undertaken along with the assessment of three other standards. In other words, the standard concerned must share the examining period with three others. This suggests that each standard will receive about 45 minutes of attention on average (some may get more time, others less). One of the things we know about reliability is that the more information collected (e.g., by setting a longer test rather than a shorter one), the higher the level of reliability. However, the relationship of test length to reliability mirrors the "law of diminishing returns" – if a test is increased from 30 minutes to 60 minutes, the incremental increase in reliability will be greater than if the test is increased in length from 60 minutes to 90 minutes. This relationship is represented in what is known as the "Spearman-Brown Formula".[4]

If the Spearman-Brown Formula is applied to the situation described in the previous paragraph, the reliability estimate of 0.9 for the full three hours of assessment reduces to 0.69 for a 45 minute assessment. Furthermore, if the SEM for the three hour examination is around ± 5%, then the SEM for the 45 minute assessment is likely to be at least doubled (i.e., greater than ± 10%). If the cut-off point for a particular grade (e.g., merit) corresponds to a score of, say, 70%, then in the three hour context the 95% confidence interval covers 60-80% while in the 45 minute context, the interval covers 50-90%. ***The argument made here is that the latter situation will challenge public credibility in the system of standards-based assessment being developed for the NCEA.***

The developers of the NCEA may choose to dismiss the figures presented in this paper as speculation. However, the scenario described here is not unrealistic, if evidence from the evaluation of an achievement-based (standards-based) programme in Year 12 (sixth form) English is taken as a guide (Hall, 2000a). The programme is known as the "English Study Design" (ESD) and has been implemented in ten secondary schools in Auckland, Bay of Plenty, Hawkes Bay, Christchurch, Otago and Southland. It was first introduced in 1998 as an alternative to unit standards. Underpinning the ESD are the processing strands of the national curriculum in English: critical thinking, exploring language and processing information. These are not treated as discrete divisions of the programme but as related elements which are blended through the course activities undertaken by

students. This therefore requires that assessment activities, and the reporting of results, should include a strong focus on both the "whole" and the "parts" of the course. This is a pedagogical decision (see next section for further discussion).

In grading student work for the ESD, teachers make their assessments initially in respect of criteria specified at five levels of achievement; they then make a further discrimination and decide whether a student's work falls within either the upper or lower half of the level. Separate grade-related criteria exist for the different components of the programme, but each assessment component allows students to be graded 1 to 10. The overall (aggregated) course grade for a student is achieved by identifying the typical level of performance of the student across the course. This is close to the average grade of the student but allows for the most recent work of the student to be given more weight in the final grading (i.e., there is a focus on where the student has reached). Further information on the programme is provided by Locke (1998, 1999a).

The evaluation of the course has included surveys of teachers and students about various aspects of the programme (e.g., manageability, coherence of the programme, and usefulness of the grade criteria for assessment). It is sufficient to say here that teachers have been very positive about all aspects of the programme, while students have given mixed evaluations. Some components of the course (e.g., assessment, including the grade-related criteria) have received good feedback from students while other features (e.g., the preparation of a workfile which emphasises self-evaluation) have been less well received (Locke & Hall, 1999; Hall, 2000b).

However, of particular significance to this paper is the reliability analysis of the internal assessments given by teachers, and the comparable reliability of a moderation test which was constructed to assess the extent to which schools graded students according to the same standard (Hall, 2000a). It should be noted that the reliability analysis was based on four components making up 60% of the course total – the focus was on various aspects of *close* reading and writing. The reliability of teachers' assessments (separately for each school) ranged from 0.96 to 0.83 with a median of 0.92. The SEMs for these schools ranged from 0.24 to 0.52 of a grade (i.e., 2.4% to 5.2%) with a median of 0.45 (4.5%).[5] The comparable reliability of the moderation test (a two hour test) was 0.81 with SEM of 0.66 (6.6%). Based on the Spearman-Brown Formula, these latter figures adjust to 0.86 for reliability and 0.56 (5.6%) for the SEM if the length of the test were extended to three hours.

On the surface, all of these values can be thought of as at least satisfactory, if not very good. The higher estimates for the internal assessments of teachers (in comparison to the moderation test) are not unexpected. Teachers base their grading on a much larger body of information collected during the school year whereas the moderation test is a one-off measurement. As already indicated, the literature on testing indicates that reliability will increase given more, rather than less, information upon which to base estimates. A further point is that the moderation test was used for the first time in 1999. As a result of the analysis of the test, improvements to content and structure have been indicated. These should enhance both reliability and validity. The test is also positioned so that it could act both as an external examination and a moderation device should the schools choose to incorporate both internal and external assessments into their grading of students (this possibility is being considered).

Also of significance to this paper is the method used to estimate the reliabilities of teachers' assessments and the moderation test. Because the performance of each student is represented by a single grade in Sixth Form Certificate rather than by a profile of different scores across components, the study investigated the extent to which the overall standards-based grade of students represented a reliable aggregation of student performance. The measurement principle underpinning this notion is that aggregation should only be undertaken if the components being combined contribute positively to the same overall trait being measured. One method of checking this is provided by Cronbach's Coefficient Alpha – a technique for estimating reliability based on the notion that all components should work together in producing a reliable total score (Cronbach, 1947). The Alpha values obtained above suggest that the practice of aggregation is justified.[6]

The point can be made that the use of Cronbach's Alpha is usually associated with norm-referenced assessment where the focus is on discriminating between students in terms of their performance. The application of Alpha to the ESD was made possible because students' work was classified on a 10-point grade scale. Although no attempt was made to force students' grades to fit a pre-defined distribution (e.g., bell-curve), the scale allowed for differences between students to be recorded in terms of their meeting the standards-based criteria associated with each task. An alternative approach to reliability – one

which may be incorporated into future analyses of the ESD – is to assess the extent to which accurate or consistent judgements are made at key grade boundaries. In respect of the NCEA, the developers need to conduct research which focuses on the pass/fail, pass/merit and merit/excellence boundaries. The developers should not be too hopeful that they will obtain significantly better results in respect of external examinations than those suggested above. Not only will each standard receive a limited coverage, the time allocated to each will also be split between the three grade boundaries under focus. This will put considerable pressure on examiners to come up with new kinds of assessment tasks which will give the focus that is needed for reducing the margin of doubt at each grade boundary. However, they will still not overcome the ceiling effect associated with variations in the interpretation of a standard and inconsistency in student behaviour.

The developers of the NCEA might also wish to consider one more factor which places strain on attempts to assess students through standards-based assessment in the context of a one-off examination. The research of Willmott and Hall (1975) identified that students' performances towards the end of an examination tailed-off significantly. Those questions attempted last were those that were least well done. One interpretation is that students picked their best questions and answered them first. However, the most common pattern was for students to work through the examination in the order of questions as presented. There was also evidence to show that many students were unable to allocate efficiently their examination time or suffered from fatigue towards the end of the examination period. Under traditional (aggregated) examination conditions, performance on the earlier questions can legitimately compensate for problems such as those just described; under the NCEA system, a student who misjudges his/her examination time or tires towards the end of the period, will receive no compensation. The implications are clear: the performances of students on a particular standard will be affected by the location of that standard in the order of questions as they appear in the examination paper and by factors related to each students' exam-taking techniques and level of fatigue. Students who are affected may will receive eight or nine credits where their true understanding might otherwise merit the full twelve credits.

It must be stressed that the arguments presented here on reliability are not a challenge to standards-based assessment. The ESD is a standards-based model: it emphasises both the whole and the parts and reports on both. The important point is that assessment information on the "parts" is simply not strong enough to allow each component to stand alone for measurement purposes. It is difficult to see how the NCEA will do any better, particularly for those standards which are to be given 30-60 minutes of attention in an external examination.

Pedagogical and Practical Implications of Non-aggregation

Validity and reliability are inevitably connected. In some circumstances, a preoccupation with one can have detrimental effects on the other. This occurs, for example, when "closed" type assessment tasks (e.g., multiple choice tests) are used inappropriately to assess open-ended knowledge contexts. The inappropriate choice of assessment task in this case may well increase precision in measurement, but it could also be to the detriment of content validity – important knowledge and skills may be ignored. It is evident that the designers of the NCEA are well aware of the issue of content validity in their decision to move away from the "micro" definition of standards to the specification of broadly based standards, which are intended to better reflect the content and intended outcomes being assessed. However, a number of questions require consideration if the NCEA is to provide a pedagogically defensible approach to assessment. Assessment must be coherent with course design, teaching and learning. The remainder of this paper addresses some questions which impact on the validity, coherence and manageability of assessment for the NCEA.

1. To what extent is the design of the NCEA being influenced by political pressure to include a strong external assessment focus? If so, how effective is NCEA in blending the best of both worlds – external examinations and teachers' internal assessments?

2. To what extent is the assessment of students against separate standards likely to foster a "bricks without mortar" approach to teaching and learning? Is the policy of non-aggregation in fact creating a new dichotomy in education – standards-based assessment versus course-based assessment?

3. From the perspective of school organisation and manageability of assessment, is the NCEA treading the same ground as that covered by unit standards?

Other questions might also be asked. Indeed other writers (e.g., Irwin, 1999; Locke, 1999b) have posed different questions which the NCEA designers should reflect on if the NCEA is to be built on solid ground.

1. The political pressure for external examinations, and the blending of internal and external assessment

It is very clear that strong political pressure has been exerted upon the Ministry of Education to include a major component of external assessment in secondary level public examinations. This is evident both in the policy statement of the NCEA and the policy drivers which have influenced the directions being taken. For example, the policy requires that at least half of the credits for each subject at each level be assessed externally. In most cases this is to be by examination except for those skills which do not suit this method of assessment. It is clear that the policy has been influenced by public perception of the importance of external assessment for ensuring an independent measure of student achievement – teachers' assessments are not seen to have the same level of objectivity. The pressure to include external assessment is also backed by the survey of teachers' opinions in the report from the Secondary Schools Sector Forum:

> Teachers strongly endorse the forum's position. Over 91% of respondents agreed that external assessment should count for at least half of NCEA credit in conventional school subjects.
> (Ministry of Education, 1999b, p 12)

It is not hard to see why there is a strong backing for external assessment. For teachers, the burden of assessment is reduced, leaving them to focus more on teaching and learning. The experience of unit standards has no doubt influenced this position. External examinations also offer far greater confidence that a student's work is his/her own – authenticity is less of an issue. As already indicated, the public perception is that external examinations are more objective – everyone sits the same examination and everyone can be compared against the same benchmarks. However, it is clear that public perception does not go as far as to include knowledge of the reliability and SEM of external examinations. On this count there is strong evidence to suggest that teachers' assessments are at least the equal of external examinations, for the reasons given earlier in this paper. The problem with teachers' assessments, however, is that unless they are moderated, there is no certainty about the comparability of the assessments from different teachers and schools. For example, the comparability analysis of the

ESD identified two schools that were out of line by approximately two grades on average (out of ten) with the grades awarded by the other schools.

The point to be made from the previous paragraph is that external examinations and teachers' assessments each have their particular strengths and weaknesses. The most appropriate system of assessment is one which blends the two so that the strengths of each are to the fore. It is hard to see how this will be achieved in the NCEA. External assessment and teachers' assessments are not blended: some standards are assessed by one method, some by another. This is not a genuine blending of approaches. In fact the system exposes the weaknesses in each.

Consider the use of external examinations as a vehicle for standards-based assessment. Since an essential feature of standards-based assessment is that student performance should be constantly monitored on an individual basis – teachers, not external examiners, are in the best position to do this – it is clear that a one-off examination is not a suitable mechanism for reaching valid conclusions about each student's performance. In most subjects, there would be a need for at least three, and in some subjects as many as nine or ten, different assessments annually to cover sufficiently the standards being assessed. This simply cannot be done through external assessment. Imagine the effort needed by NZQA if they were to attempt to administer in each subject anything from three to ten separate assessments in a year. The fact is that teachers are in the best position to monitor achievement, to redirect student learning quickly and to implement a testing programme at the appropriate points in the learning process. The place for external assessment in this process is to provide a check on moderation and to contribute an appropriate percentage of marks to the overall grade of students. Such a system would represent a genuine blending of internal and external assessment: teachers would be used to monitor students' performances and to contribute valid and reliable information on students' achievement; the external examination would provide an independent assessment of students' work and act as a moderator of teachers' assessments. Together, the two would enhance both validity and reliability by building on each others' strengths. As mentioned, the proposed NCEA system uses external assessment for some standards, despite problems of validity and reliability, and internal assessment for others with no clear or strong method of moderation at present being mooted. Under this system, it is likely that some standards – those that are externally assessed – will gain higher status because they have the stamp of independent measurement.

Unless genuine blending is to take place between internal and external assessment, there seems little point in proceeding with standards-based assessment. Either do it right, or not at all. Too much damage has been done to standards-based assessment already by inappropriate models of implementation.

2. Bricks and mortar: Is the policy of non-aggregation creating a dichotomy between standards-based assessment and course-based assessment?

The whole-part issue has already been discussed in respect of reliability. However, the problem extends beyond reliability. The first point to note is that a particular division of a subject into achievement standards is but one construction of the underlying standards that could be defined. The division proposed in the ESD, for example, is not the same, nor does it take the same form, as that proposed for the NCEA. Which is right or better? There is no answer to this question. Both are intended to be valid constructions of the National Curriculum. The mistake that appears to have been made by the designers of the NCEA (the same mistake as was made for unit standards) is that because a particular division is suggested, it necessarily follows that the resulting standards should then be treated as independent components of the assessment of the subject. As mentioned, a subject can be analysed in many ways depending on the focus that is to be taken. This does not mean that each standard is separate from others in any pedagogical sense. The standards can be thought of as providing the "bricks" – a particular combination of bricks may be the basis for designing the assessment for a particular course – but the "mortar" for the course, the particular knowledge and skills which connect standards and provide the integration and transfer of knowledge from one part of the course to another, is likely to be de-emphasised or de-contextualised in any scheme which treats standards as separate entities.

We have already seen from the literature, albeit mostly in the context of norm-referenced assessment, that the different parts of a course yield internally consistent assessments of students' performance overall. Standards-based assessment does not change this. The threads and interweaving of the content (knowledge, skills and values) give the "whole" a particular meaning. The division into a particular set of standards no doubt provides a basis for giving students useful feedback on their achievements, but it does not negate the notion of an overall (aggregated) assessment of student performance. Aggregation gives impetus to teaching and learning strategies which encourage students to focus on the "whole" as well as the links between different components of a course. Teachers are likely to design their teaching differently if the focus is only on the parts. The impetus for inappropriate modularisation is strong – the simplest approach to take by teachers is to teach to each standard separately even if this is not the intention of the designers of the NCEA.

The writer has made the same criticisms in the past about unit standards (e.g., Hall, 1994, 1995). We know that in "high stakes" assessment contexts, what is to be assessed quickly dictates what is taught and how it is taught. The achievement standards proposed for the NCEA set out in broad terms the intended learning outcomes that students should meet – these will have a direct influence on what is taught and assessed. The policy to award separate credit for each standard further emphasises the partitioning of the course content and intended outcomes into segregated components. Separation of standards in this way will impact strongly on how teachers design courses and teach and assess their students.

If standards are to make sense they need to be embedded within a teaching and learning structure which ensures that the objectives, content, delivery and assessment are all connected. The danger with the NCEA model is that courses will lose this overall focus on coherence: performance in a course overall will not be recorded, just performance on the parts. This might be acceptable if the parts are uncorrelated and only a profile of students' achievements is needed; but when the parts correlate and integrate, the "whole" is a particularly important piece of information. Perhaps the opposite pole of "standards-based assessment", as it is being implemented in New Zealand, is not "norm-referenced assessment" but "course-based assessment". Under the former, the emphasis is on measuring student achievement in discrete segments; under the latter, recognition is given to the course as a coherent, working unit in which the pedagogy focuses on developing the all-round knowledge and skills of students, including the important links between the parts of a course.

3. *Manageability of assessment: Is the NCEA treading the same ground as that covered by unit standards?*

Underpinning the policy of non-aggregation of standards is the belief that the NCEA will open the door to much greater flexibility in the design of teaching and learning to meet the individual needs of students. The belief is that schools will be able to tailor programmes in a more student-centred way than is possible under existing (course-based) practices. This, in fact, is a direct transportation of thinking from unit standards. However, there are some genuine manageability problems with this thinking.

The first constraint is that the complete tailoring of achievement standards is not possible because most institutions simply do not have the resources to individually tailor courses in this way. The problem occurs at the moment in both schools and polytechnics which offer unit standards. A complete pick-and-mix based around unit standards is not possible without a major increase in resourcing. The same will apply to NCEA achievement standards.

A related problem is that it is not at all clear how schools will handle students who pass some standards in a year but not others. For example, will a student who achieves 12 credits in a subject be allowed to undertake all the standards in the subject at the following level or will s/he be required to do 12 credits at the advanced level and 12 at the earlier level? How will this be influenced by the external assessment process? Will a student be allowed to sit an examination and answer only those questions concerned with standards that s/he has not passed on a previous occasion?

If students are to be required to study some standards at one level and others at the next, how will schools design and timetable courses? Will this not lead to greater emphasis on modularisation (bricks without mortar again) so as to give students freedom to take a particular combination of standards, bearing in mind the particular combination they have already achieved?

To what extent will a school take account of the literature on learning which tends to show that student learning in a subject is uneven? A student may not pass a standard at one level but then operate successfully in the same area at the next level. Will such a student receive the credits at both levels?

The NCEA policy will allow students to carry credits/standards from one institution to another. Unless a course being undertaken in the first institution draws upon the same standards, and in the same order, as a course being offered in the second institution, a transferring student runs the risk of either repeating or missing out altogether some of the standards in a subject. The question is also raised as to what happens over Ministry funding. Will a student be funded for undertaking the same standard twice?

There may well be answers to some of these questions. Clearly, some schools will manage the practicalities of multiple level teaching, learning and assessment with greater ease than others. However, the introduction of such a system is something that is going to need a lot of planning and a considerable level of goodwill from school administrators and teachers. The system being developed is far more complex than the traditional course-based approach that has underpinned senior secondary education until now. It will involve greater effort in course management, assessment and record keeping.

Conclusion

This paper has focused on the implications of the NCEA approach to standards-based assessment, in particular the reliability of assessment against separate achievement standards, and the pedagogical implications of the policy of non-aggregation. The arguments made here are that assessment against separate standards is unlikely to yield sufficiently reliable results to satisfy public credibility, and that the same focus runs the risk of fostering a "bricks without mortar" approach to course design, delivery and assessment. The paper also argues that the NCEA treads the same ground regarding manageability as that of unit standards.

The way forward is for the Ministry to rethink the particular strengths and weaknesses of external examinations and internal assessment, and then to find a way to blend these so that the strengths of each are emphasised. The proposed system does not represent a genuine blending – it exposes the weaknesses of each. The Ministry might like to look seriously at the model of standards-based assessment being implemented for the English Study Design at Year 12. This model has the capacity to maintain overall course coherence but within a framework which allows reporting against the parts as well as the whole. It also has the capacity to genuinely integrate internal and external assessment.

Notes

1. Traditionally, three main approaches have been taken to the estimation of reliability. The first, known as test-retest reliability, involves giving a test or examination twice to the same group and correlating the results. A good correlation is assumed to indicate accurate or reliable information. The second approach involves the administration of two different but equivalent tests/examinations to the same group and correlating the results. Again, a good correlation is assumed to indicate accurate or reliable information. However, because it is often impractical to assess the same group twice, a third approach is often taken which focuses on the extent to which each student has performed consistently on the different parts of the test or examination. Again a high level of consistency in students' performances is inferred to indicate reliable or accurate information. (Most texts on educational testing and measurement provide detailed explanations of the different approaches to the estimation of reliability.)
2. The Standard Error of Measurement (SEM) is represented by the formula: $SEM = SD \times \sqrt{(1 - R)}$, where SD is the standard deviation of the observed (obtained) scores of students, and R is the estimate of reliability.
3. The 95% confidence interval is estimated by: $X \pm (1.96 \times SEM)$, where X is the obtained score of an individual and SEM is the Standard Error of Measurement. This calculation, along with the estimation of the SEM above, assumes that the population is normally distributed in respect of the measurements being made. Where this assumption cannot be met, alternative techniques often exist.
4. The Spearman-Brown Formula is used to predict the reliability of a test of a different length, given the length and reliability of an existing test. The new test may be longer or shorter than the original. The new reliability is represented by: $RR = (N \times R)/(1 + (N-1)R)$, where RR is the new reliability, N is the number of times the length of the test is to be increased or decreased, and R is the original estimate of reliability. (See Guilford & Fruchter, 1973, p 491, for the derivation of the formula.)
5. The analyses were based on eight of the participating schools. One school did not provide complete data and the tenth school has recently joined the ESD.
6. The reliabilities for teachers' assessments of student work in the ESD could well be slight "overestimates" of the true internal consistency of the aggregated data. This could come about if teachers are influenced by halo effects, that is, they tend to base their judgements of students' work on generalised perceptions of each student's performance in the subject rather than on the particular strengths and weaknesses of the student.

However, such an effect is unlikely to influence the moderation test results as each student's performance was centrally marked by an external examiner unaware of the student's general performance in the subject.

References

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12,* 1-16.

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill.

Hall, C. (1994). *Obstacles to the integration of university qualifications and courses into the National Qualifications Framework. Higher education in New Zealand: Occasional paper no. 1.* Wellington: Syndicate of Educational Development Centres of New Zealand.

Hall, C. (1995). Why universities do not want unit standards. *New Zealand Vice-Chancellors' Newsletter, 35*, 5-8.

Hall, C. (2000a). *English Study Design 1999: Evaluation report of a standards-based moderation test. Report to the English Study Design Coordinator.* Wellington: Victoria University of Wellington, School of Education.

Hall, C. (2000b). *Feedback from teachers and students on the Year 12 English Study Design 1999. Report to the English Study Design Coordinator.* Wellington: Victoria University of Wellington, School of Education.

Harvey, L., Burrows, A., & Green, D. (1992). *Criteria of quality.* Quality in Higher Education Project. Birmingham: The University of Central England.

Irwin, M. (1999, October). *Achievement 2001.* Paper delivered to the Examining Assessment Conference, New Zealand Council for Educational Research.

Locke, T. (Ed.) (1998). *English study design: A practical guide for course development and assessment: Year 12* (2nd ed.). Hamilton: University of Waikato.

Locke, T. (1999a). Standards-based assessment in English: Take 3. *Waikato Journal of Education, 5,* 13-31.

Locke, T. (1999b). Lacks of Achievement 2001. *English in Aotearoa, 39,* 62-67.

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.

Ministry of Education. (1999a, June). *Introducing the National Certificate in Educational Achievement: Developing Achievement Standards, Issue 1*, Web site <www.minedu.govt.nz>

Ministry of Education. (1999b). *Achievement 2001: Report from Secondary Schools Sector Forum.* <www.minedu.govt.nz>

Willmott, A. S., & Hall, C. (1975). *O Level examined: The Effect of question choice.* London: Schools Council Research Studies, Macmillan Education.

The author

Cedric Hall is Professor and Head of School, School of Education at Victoria University of Wellington. He was formerly Director of the University Teaching Development Centre at the same university. His research interests include course design, teaching, learning, assessment and evaluation. Some of his recent papers have included a focus on quality assurance in higher education and developments in the National Qualifications Framework.