# A Change of Direction for NCEA: On Re-marking, Scaling and Norm-referencing

*New Zealand Journal of Teachers' Work, Volume 2, Issue 2, 100-106, 2005*

ROY NASH
*Massey University*

New Zealand's National Certificate of Educational Achievement (NCEA) has been a reality of secondary teachers' work – it must be responsible for a great proportion of it – for several years now and yet the examination remains controversial. The past year, 2005, may prove a turning point in the realignment of the conduct of the qualification, and this paper will take up certain themes, those of re-marking, scaling, and their relationship to norm-referencing, that have attracted recent public attention. There is no need to begin at the beginning and rehearse the entire history of NCEA, for this audience is more than familiar with it, but a little scene setting may be allowed. The crisis of 2005 will merit more than a footnote when some historian of the future produces the definitive account of New Zealand's experiment with standards-based qualifications. The year began, as readers will recall, with the release of Scholarship results that showed wide variability between subjects, provoking two official reports from the State Services Commission and another from the *ad hoc* Scholarship Reference Group set up by the Associate Minister of Education, and ended with public comment on the extent of re-marking in NCEA and its justification. It may also be pertinent to note that I am on record as having advocated standards-based assessment (Nash, 1988), rather than norm-referenced assessment, as a practice to be adopted wherever possible, and that even now I consider NCEA to be, on the whole, a better qualification for secondary schools than School Certificate. But there are realities about the nature of assessment that the New Zealand Qualifications Authority (NZQA) has consistently attempted to deny, which in 2005 at last made their presence felt.

The purpose of an examination is to find out whether candidates know what they are required to know in the domains assessed by the examination. As some candidates invariably know more than others, examination results are usually graded. The process can appear entirely routine and unproblematic. Such appearances are, however, deceptive. The field of assessment and testing is one of the most hotly contested in education and the struggle between the principles of norm-referencing and standards-based (or criterion-referencing) lies at the heart of it. This is not just a local matter. Hamilton (1982), a long-standing critic of normative assessment, reports how marks were allocated in a Scottish school:

> Each child was awarded a grade (A-E) on the basis of the distribution of marks across the department: 10 per cent received grade A, 20 per cent grade B, 40 per cent grade C, 20 per cent grade D and 10 per cent grade E. Had the grades been criterion-referenced it would

> have been possible for a large number of children, if not all, to get grade A (instead of only 10 per cent). In the criterion-referenced case 'success' is based on achieving a certain standard, not on the overall distribution of the marks. (p. 195)

As it happens, a form of standards-based assessment influential in the development of our system was soon to be introduced in Scotland, and our qualification system differs mainly in its adoption of an even purer model. The appeal of a standard, of an unchanging criterion against which to assess achievement, has a compelling allure and yet, as recent New Zealand history shows, the reality is somewhat more complex. The Qualifications Authority has imposed a criterion-referenced assessment regime through the development of unit standards right through the secondary educational system, with only the universities more or less exempt from its reach, and our experience of trial and error learning in this area must be the most extensive in the world. The process may, at last, have reached that point on the learning curve where further reinforcement generates increasingly diminishing returns, and it seems likely that certain lessons have been learned.

The basic difficulty is almost self-evident. What seems to be a precise conceptual distinction between norm-referenced and criterion-referenced attainment is extremely difficult to apply in certain areas of practice. Criteria are much easier to set in some subjects, such as those where a definite motor skill is involved, than they are in others where the performance resists precise definition and requires considerable expert judgement to recognise. If an employer requires applicants for a secretarial position to type a set passage of 300 words in six minutes with no more than three errors then that is the criterion, and an examiner knows exactly how well those who have met that criterion can type with reference to that 'standard'. But an employer who wants to know how well an applicant can read has a more difficult problem. PAT Reading Comprehension scores, although formally linked to standards of achievement, do not really provide information comparable with that to be gained from a typing exercise. The scores indicate that a student with a percentile score of 70 can read better than 69% and worse than 29% of the standardisation population, but what that student can actually read, or comprehend of the questions, is best determined by reference to the text, which is actually not much help in this respect as a given score can be obtained by more than one combination of responses. So why should we not test reading as we do typing, without reference to norms, but by adopting a set text and a given level of performance in reading it as a standard? The idea seems eminently plausible. Taken at face value, criterion-referenced assessment appears to have much to recommend it (the performance demonstrated is a well-specified task open to interpretation) and norm-referencing very little to recommend it (the level of performance must be gauged from the relative position obtained), nevertheless, there are difficulties that make the introduction of criterion-referenced assessment in areas like reading, mathematics, and so on, much less smooth than this view might lead one to anticipate.

Standards of attainment in basic academic subjects are actually very difficult to fix and maintain and the reasons for this must be grasped. If a national curriculum had fixed the standard of English in about 1000, our ten-year-olds would now be faced with texts like, 'On siex dagum waeron geworhte heofonas and eorde, sunne and mona, sae and fiscas' [*Heaven and earth, sun and moon, sea and fishes were made in six days*]. A fixed standard is not

possible in reading for the sufficient reason that languages change. In a millennium a language is likely to change dramatically, but even 20 or 30 years often makes a significant difference. In practice, it seems impossible for examiners using judgement alone to produce equivalent texts, even from one year to the next, within the degree of consistency that has proved essential to the maintenance of public confidence in the examination. In any case, the same text could not be used year after year – as one might use the same passage for a typing exercise – because schools would be unable to resist the temptation to 'teach to the test'. It can be tempting to ride roughshod over these fundamental objections and seek some technical solution to the problems. But there are few available and they all involve standardisation to a norm in one respect or another. For example, in any year a certain number of test questions may be standardised, to ensure that they are equivalent in the sense of being answered correctly by a known proportion of pupils, and used in subsequent years as markers. This procedure, known as 'equating', is used by the National Educational Monitoring Project (NEMP), and is actually recommended by the State Services Commission report on NCEA (Martin, 2005a; 2005b) for consideration by the Qualifications Authority.

The judgements of examination writers are always open to challenge and without a record of such past performance rates comparability over time is impossible to achieve. The content of any standard, moreover, is itself norm-referenced. Indeed, the conceptual purity of the criterion-referenced/ norm-referenced distinction virtually collapses once it is appreciated that the specified attainment targets have been set in full knowledge of the proportion of pupils able to perform such tasks. A level of attainment is specified for each school year and the content for that level is established by what most students within each year actually learn. If students at level 4 of the mathematics syllabus are supposed to know that 43 x 8 = 344 is equivalent to 344 / 8 = 43 (an actual example) that is because it has been determined, either by expert judgement or by actual trial, that students of that age, ten years, are capable within this system of being taught the mathematical skills necessary to perform such computations. In this way, a norm-referenced reality always underpins criterion-referenced standards in the educational system. The NCEA level that replaced School Certificate was designed to allow 70 per cent of students to gain an acceptable standard. There *is* a difference between a testing regime that produces information such as, 'IQ 110-120: interpretation: could succeed at commercial college', which IQ tests offered, and one that states, 'level 5 pass: interpretation: can write a competent job application and read the leading article in an evening newspaper', which criterion-referencing promises, but the distinction in practice is not as sharp as the conceptual distinction.

There is no escape from this logic. It is all too easy to shrug off criticism of this kind as academic pedantry, or to misconstrue it as opposition to the introduction of standards-based assessment practices, but the inadequacy of this impatient response should now be clear. Warwick Elley has repeated these objections over and again in the last five years only to be sidelined as a hostile critic, and he is, indeed, no supporter of NCEA, but there are people who would have been wiser to listen to him. Failure to grasp his elementary point that '[e]xaminers cannot tell in advance how well a cohort will do, on untrialled questions' (Elley, 2005), and to appreciate its political implications, has forced the resignation of more than one great eminence connected with the Qualifications Authority. It is all very well for a State Services Commission review to criticise those officials who 'failed adequately to take stock of the

policy settings and provide the government with an explicit analysis of the implications for the outcome of the 2004 Scholarship' (Martin, 2005a: 3), but those officials had been permitted to run NCEA at a lower level with huge variability; for example, in 2002 more than 5000 candidates were graded 'excellent' in a mathematics standard but in 2003 only 70, and there is no reason why officials should have supposed that such 'flexibility' would prove intolerable at Scholarship level. Elley's analyses of such anomalies were dismissed by the Minister, in a response doubtless prepared by officials, and presented as a necessary consequence of the assessment system (Elley, 2003; Mallard, 2003). If it is true that 'NZQA was aware that there would be variation, but was not aware that this variation would be a surprise to everyone else', and if its officials 'saw the variability of results between subjects as a consequence of their approach to Scholarship, rather than as a significant risk that could undermine the credibility of the examination' (Martin, 2005a: 3), that can only be because the authority had reason to believe that it was acting in accordance with policy. Mallard's response to Elley acknowledged that, 'across subject groups, there is expected variation in the distribution or [sic] results', and that 'the distribution of results for various achievement standards within a subject varies greatly', and gave a variety of reasons why these properties of the system should be expected and accepted as evidence of its superiority to norm-based approaches. Scholarship proved to be different because the results affected access to higher education courses, and perhaps officials should have realised that, but then so should have the politicians. It is not fair if a flawed examination prevents a student from entering a university physics course, but it was not fair to award 5000 'excellent' grades one year and 70 the next, even if the stakes were different, and had the *morality* of that been attended to events might have been different.

It is not surprising in this context that when in early 2005 NZQA published Scholarship results markedly inconsistent between subjects, its initial response was to defend them as an accurate reflection of the standards achieved, and defend the integrity of the system. This position was only abandoned in the face of political and public pressure sustained over a period of several weeks. The situation proved to be so unacceptable, however, that Cabinet itself sanctioned the issue of revised grades and, after official investigations, both the Chief Executive and the Chair of the Board resigned their positions. The most puzzling thing about all of this is that it was entirely predictable. NZQA had no mechanism in place to prevent year-to-year and subject-to-subject variation and seemed not to believe that it was necessary to have such a mechanism. And so we come to the end of 2005.

In December, Karen Sewell, NZQA's Acting Chief Executive, was reported as stating that after last year's problems, the authority had developed profiles of expected results for each exam. If results appeared to be way off the expected profile, the marking guidelines were investigated and changed. A new Minister of Education, Steve Maharey, noted that in previous years: 'This system wasn't in place, and as a result they got these massive variability [sic] and they couldn't do anything about it' (*A question of missing the mark*, 2005). This is finally to admit that there are technical problems in setting and marking standards-based examinations that cannot be solved other than by using comparative information (on year-to-year and subject-to-subjects award rates) in order to control the proportion of candidates allocated to each grades.

By the end of 2005 a dozen or so NCEA papers had been re-marked – a matter that attracted political criticism and press comment – and NZQA began

to refer to cut-off points. In the context of a discussion on re-marking in three English standards, Karen Sewell was quoted as saying that marks given to questions dealing with comprehension of an unfamiliar text showed that the marking schedule was too specific in one area. This, she said, 'had made the cutoff point too high' (Mulroney, 2005). It is clear that the authority became aware that the marking schedule was too specific only because its monitoring of mark allocation indicated that the cut-off point was too high, and this is exactly what such monitoring, newly introduced by political directive, is intended to achieve. We also know how the cut-off point is determined. Karen Sewell, using a phrase from the State Services Commission report, has stated that the cut-off points are within a 'general band of tolerance' based on historical data and the professional expectations of teachers and examiners. Those who appreciate irony will note that the reference to teachers' expectations in this context is somewhat at odds with the Ministry of Education's preferred theory that these sentiments are largely responsible for social and cultural disparities in the educational system. The newspaper also reports her comment that the 'profiles' will be made public after the results come out, but this is hardly necessary as the information may be ascertained from the results, which must now be consistent from year to year and subject to subject. The report advices NZQA to 'define and bring forward normative boundaries to function as a safety net for the four grades' and to allow 'potential comparisons with past examination results', and this advice will no doubt be followed (Martin, 2005b: 6).

But how are these cut-off points to be legitimated within NZQA's standards-based, rather than norm-referenced, theory? That is a much more difficult problem to resolve and has already generated characteristic ambiguities in the discourse. NCEA's supporters, including NZQA, the Ministry of Education, and the PPTA, are reluctant to concede that that re-marking to maintain expected 'profiles' introduces a form of norm-referencing and, being concerned to defend the examination against criticism that threatens to undermine public confidence in it, have no interest in debating the point. This is not the best position from which to acknowledge that a 'profile' of plus or minus 5% (or whatever it might be), is essentially arbitrary, deemed necessary for political rather than educational considerations, and by definition fixed to points on the normal curve, even if it happens to be true. Karen Sewell denies that re-marking to a profile is 'scaling', and if scaling is defined as adjusting given marks to fit the normal curve, that is correct, but re-marking papers achieves the same result by allocating what amounts to a proportional quota for each grade. Re-marking is not scaling, but the technique is used to generate a distribution of grades more or less consistent between subjects and years, and is thus simply an expensive way of accomplishing the same outcome. This practice will, moreover, have the effect of ensuring that any actual improvement in the standards of achievement will be disguised. Should actual achievement in, say, mathematics, improve dramatically over a period of five years, that fact could not be revealed because re-marking to maintain the historically determined 'profile' (such as the Scholarship Reference Group's 2-3% ± 5 target for 'excellence'), would prevent it from being recognised. Compare that situation with reports of an improvement in athletic standards, such as a 20% increase in the proportion of 14-year-old girls running 100 metres in under 12 seconds, and the difference is immediately apparent. There will be no call to re-mark those achievements, because only by the grossest forms of incompetence could examiners fail to measure distances and record times accurately.

Within the pure doctrine of standards-based assessment, the results achieved by a candidate on an examination of that kind are just what they are. If an examination paper is set to the required standard, which is determined by an expert panel, and marked according to the schedule, which is again constructed by experts, then, a candidate who achieves a mark within a specified range has achieved the grade allocated to that range. But under the new rules, that grade cannot be awarded if the effective target cutoff point is under- or overshot by markers. Re-marking allows the authority to argue that it is acting within the spirit of standards-based assessment by responding to information about the properties of the examination paper or the schedule (or both) and effectively revising the examination process itself in a manner compatible with the claim that standards-based procedures are working. Thus, Karen Sewell argues that: 'It may be that a group of students had come up with quite a different way of answering the question, but it showed that they could do it. So we change the schedule' (*A question of missing the mark*, 2005). The problem with this argument, which is a little disingenuous, is that the information the authority collects and acts on is not primarily and definitively of that kind. It actually uses data on the proportion of students at each cut-off point and, armed with this, then modifies whatever sections of the marking schedule do not meet its expectations. If the cut-off points are too far out, there are by definition flaws in the schedule, or its interpretation, to be corrected. In effect, NZQA has introduced real-time item trialling, and is attempting to make a virtue of it. Whether this is an interim feature, to be replaced by 'equating', which is a form of pre-trialling, remains to be seen.

Scaling would produce the same outcome and the reason why it has been rejected is instructive. Let us remind ourselves of how scaling works. There can be, say, no more than 15% of candidates in the 'merit' category, but 20% have marks above the scheduled cutoff point, so we fill the quota with those having the highest marks and allocate the remainder to the 'achieved' category. The candidates shifted from 'merit' to 'achieved' by scaling should be the same (by and large), as those reallocated by re-marking for re-marking designed to remove (as in this case), a quarter of those deemed to have been misallocated as 'meritorious' is likely to regrade those in this set who initially gained the lowest overall marks (this is why the Scholarship Reference Group (2005), thinks that scaling is unnecessary for the purpose of awarding Scholarship because those with the highest marks in each subject can always be identified). Outright scaling of the results, however, would be difficult to present as consistent with standards-based assessment, and that alone is sufficient to make the practice unacceptable. The consequence of this, however, is an expensive bill for candidates whose fees pay for a re-marking that is technically unnecessary, inasmuch that scaling would accomplish the same end with greater efficiency.

The day-to-day experience of NCEA will change little for teachers and students. The modular curriculum, the regime of internal assessment, and so on, will continue in its now accustomed path. Student assignments will receive, as they have for some years, the grades Non-Achieved, Achieved, Achieved with Merit, and Achieved with Excellence, in their more or less constant proportions. And yet, at a deeper level the imposition of fixed bands of success represents a fundamental shift away from the purist concept of standards-based assessment towards a pragmatic acceptance of normative comparisons. As the review into secondary school qualifications points out, the 'climate of polarisation' between standards-based and norm-referenced approaches is

unhelpful, and suggests that 'it is more constructive' to see both as 'complementary and differing largely at the point where norms are applied' (Martin, 2005b: 17). That the Qualifications Authority has been forced to learn this lesson the hard way is unfortunate – and even now it is not clear that it has been fully assimilated as its face-saving ambivalence about re-marking indicates – but in time one may expect to see a more open acceptance of the real difficulties of assessing performance in activities where a technical standard cannot be devised. This better grounding in reality might well, notwithstanding NZQA's current unease, actually lead to an improvement in the conduct of the examination and a better experience for all those involved with it. Elley (2003) noted that the basic problem, there in 1991, was unresolved 12 years later: if it is to be fixed at last, educators might like to reflect on the fact that political rather than educational forces were primarily responsible.

## REFERENCES

A question of missing the mark. (2005). *The Dominion Post*, A6, 24 December.

Elley, W. (2003). New assessment system does not pass test. *New Zealand Education Review*, 19-23 February, pp. 5-6.

Elley, W. (2005). *Facts and fallacies about standards-based assessment*. Paper accessed 24 December 2005 from: www.macleans.school.nz/news/pages/2005/pdfs/warwickelley_address.pd

Hamilton, D. (1982). Handling innovation in the classroom: two Scottish examples. In: T. Horton and P. Raggatt (Eds) *Challenge and change in the curriculum*. London: Hodder and Stoughton, pp. 177-195.

Mallard, T. (2003). Education Minister responds to Elley's comments. *New Zealand Education Review*, 19-23 February, p. 6.

Martin, D. (2005a). *Report on the 2004 Scholarship to the Deputy State Services Commissioner, by the Review Team led by Doug Martin*. Wellington: State Services Commission.

Martin, D. (2005b). *Report on the performance of the New Zealand Qualifications Authority in the delivery of secondary school qualifications, by the Review Team led by Doug Martin*. Wellington: State Services Commission.

Mulroney, P. (2005). Glitches in NCEA not over. *The Dominion Post*, A3, 5 December.

Nash, R. (1988). Towards post-school education for all in New Zealand: a community based curriculum, *Tutor*, *36*, 27-33.

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

## About the Author(s)

*New Zealand Journal of Teachers' Work, Volume 2, Issue 2, 2005*

ROY NASH
*Massey University*

Roy Nash is a professor of education at Massey University College of Education. He is a sociologist of education with a speciality in the explanation of social disparities in educational achievement. Recent papers can be found in *New Zealand Journal of Educational Studies*, *Journal of Curriculum Studies*, *Waikato Journal of Education* and *International Studies in Sociology of Education*.

Professor Roy Nash
Department of Social and Policy Studies in Education
Massey University College of Education
Private Bag 11 222
Palmerston North
r.nash@xtra.co.nz