

A new procedure by which NZQA monitors secondary exam results to detect anomalies and, if need be, remark is not a return to scaling by another name. That's the key point of the following article prepared in response to a feature article by Professor Warwick Elley in *Education Review* on 21 April 2006.

30 April 2006

Feature article for *Education Review*

In his article entitled “New Year, Same Mistakes” (*Education Review* 21 April), Professor Warwick Elley expresses the broad concern that standards-based assessment, as used in NCEA external examinations, cannot be a consistent method of assessing students’ performance, despite every effort to make it so. The issue is important, and warrants careful consideration.

Any examination system produces some variability or fluctuation in the results from year to year. This would be so even if the distribution of student ability and question difficulty were identical each year. In terms of fairness to students, the critical question is whether or not the variability is such that the same level of student achievement is consistently awarded the same level of recognition (ie grade level).

For example, no one would deny that the very large shifts in the success rate in the Level 1 English standard on “interpreting unfamiliar texts” between the first and second years of NCEA indicated a problem. This problem has now been corrected, as evident in the consistent results in this standard over the last two years.

However, there has been misunderstanding of the way in which inconsistencies evident in the initial implementation of the NCEA examinations have been put right. The claim has been made that the more consistent results of the last two years are the result of concessions being made to norm-referenced assessment. The setting of bands of expected performance for each externally assessed standard and the subsequent remarking of some standards have been cited as the evidence of a change of direction.

In fact, neither the bands of expected performance, nor the remarking, were in any sense used to reintroduce normative criteria, either covertly or explicitly. Rather, they were used as tools for ensuring that the marking process was well calibrated to the standards, thereby making sure that standards were consistently applied during remarking.

The bands of expected performance were set by drawing on information including historical results, cohort changes, and, most importantly, the professional judgment of National Assessment Facilitators, moderators, leaders of marking panels, and examiners. The aim was to identify the range of success that might be expected if the standards were consistently applied.

The bands were not treated as targets, and much less as normative criteria. Rather they were a mechanism to identify instances in which marking might not have been consistent with standards. During the marking of the 2005 NCEA exams the results for each standard were monitored. Results falling outside the expected bands constituted a warning signal that the marking process for that standard needed to be scrutinised.

In 17 cases, following dialogue between National Assessment Facilitators and the leaders of marking panels, it was determined that the calibration of the marking to the standard could be improved, either by modifying the marking schedule or the sufficiency required. Once that was done, remarking then occurred. For others, the evidence from the scripts already marked indicated that the actual results were a better match to the standard than the expected band. In these cases marking continued with no attempt to match results to the expected band.

This does not mean the bands for these standards were not useful – they were valuable monitoring tools, which is all that they were intended to be. With the benefit of more experience with each standard, further years of historical data, and more sophisticated statistical analyses of results, it will be possible over time to produce bands of expected performance with greater confidence, but they will nonetheless remain monitoring tools only. They will never be normative targets, and ultimately those people with the most expertise in assessing the standards – the assessment facilitators, moderators, examiners, and panel leaders – will have the final call with respect to any remarking that takes place.

As a brief digression, it is worth considering exactly what a standard is, and where it resides. Standards are expressed in written form, both in terms of brief descriptions, and more detailed criteria, but, in reality, these written specifications are just *descriptions* of the standards, and not the standards themselves. The standards are actually knowledge shared by the community of educators and assessment experts who use and implement them. The results are not “a lottery”, because, with good communication between members of the education community, shared knowledge of the standard forms a strong basis for assessment with high levels of confidence.

Consistency in results can be thought of in a number ways, for example as consistency with respect to the profiles of expected performance, consistency with respect to previous years’ examinations, and consistency *with respect to the standards themselves*. It is the last of these three that matters most.

While consistency with past results can be an indicator of consistency with respect to the standards, it is not a foolproof indicator, and in some cases consistency with the past would mean *inconsistency* with the standard. This is so, for example, when a past examination round has not accurately reflected the standard – in such a case, consistency with the past would entrench inconsistency with the standard. A less obvious case is when there has been a large shift in the cohort of students undertaking a particular standard. This may result from a change in the profile of the cohort with or without any substantial change in overall numbers.

In some standards there may be an influx or efflux of students with greater or lesser ability than the average for that standard in the past, or an increase or decrease in the number of students whose first language is not English. Indeed, such shifts in cohort profiles are more likely under NCEA than they were under the old system, because students now have more choices.

Other possible reasons for a shift in performance not indicating inconsistency in the application of a standard include improved resourcing of a standard by schools, better training of teachers with specific regard to certain standards, or simply greater familiarity with the requirements of a standard, particularly where a large shift is evident between the first and second years of implementation.

The suggestion has been made that any change in failure rates greater than 5 percentage points warrants investigation. In fact a 5 percentage point margin *was* the rule of thumb that NZQA adopted for setting the expected performance bands for each standard. However, any given margin needs to be a rule of thumb only, because results for standards with small numbers of candidates vary more than those with large numbers of candidates. Latin for example attracts fewer than 250 students at Level 1, and fewer than 100 at Levels 2 and 3. Reo Māori has just over 1,000 candidates in each standard at Level 2, and around 650 per standard at Level 3. In these cases it is not realistic to expect year on year variability in results to remain within a 10 percentage point band.

That said, NZQA has publicly reported that the results in three subjects - Accounting, Agriculture/Horticulture and Technology - are of concern, and that steps need to be taken to ensure that in future marking in these subjects better reflects the standards.

Even so, it is worth getting this in perspective. At the subject-aggregated level, in its biggest fluctuation, Level 1 Accounting between 2002 and 2004 showed a 12 percentage points variation. It was much the same in the days of School Certificate after scaling to meet normative criteria was abandoned in the second half of the 1990s. For example, between 1996 and 1998 the pass rate in this subject changed by 11 percentage points. And Accounting shows more year-on-year variability than most subjects under NCEA. In many cases in 2005, there was considerably *less* year-on-year variability than under School Certificate.

Of course in most pre-NCEA examinations, scaling was used to mask variability, and to condemn a large proportion of students to fail. It is hard to imagine a system more manifestly unjust. Being a fifth form student in the lower academic quartile was profoundly demoralizing and de-motivating. Students know where they sit in relation to other students, and those students would have known that, no matter how hard they worked, they could not get themselves sufficiently high up the rank-ordering to pass even one subject in School Certificate. Consequently there was little incentive for them to try.

Now, as the results show, the vast majority of students leave school with credit for some standards to their names and thus something positive and tangible to show from their schooling. This is not because standards have been lowered, but because students are now recognised in detail for what they know and can do, not just how well they do relative to others on average in a handful of aggregated subjects.

Justice apart, the wastage of the previous system could arguably have been justified in an economy that needed large numbers of unskilled labourers. However, in a modern economy, where skills and the ability to keep learning are demanded from almost every sector of the workforce, to economically and socially disenfranchise so many people would be intolerable.

Karen Sewell
Acting Chief Executive
New Zealand Qualifications Authority