

Analysis of NCEA, 2002-2004

Mary Jane Sneyd¹

BSc, MBChB, PhD

This is a basic analysis of data for NCEA from 2002 to 2004. The data for 2005 were requested on 29 August, 2006 but have not yet been received from the Ministry of Education.

Because of the size of the task to undertake a full analysis and because I do this in my spare time, of necessity I have chosen research results of most interest to me. None of my children have been at secondary school for several years so this is not influenced by their achievements. I have concentrated on science and maths subjects as I am a medical/science researcher and am greatly concerned for the future of science in New Zealand.

The basis of the following report has been taken from a submission to the State Services Commissioner in 2005 but I have done considerably more analysis since then.

¹ 95 Hillary St
Pine Hill
Dunedin

Summary of main results.

1. Variability of NCEA by year.

If we see differences of 5% or more from one large cohort of students to another it is almost certain that it is because the examination has not been of consistent difficulty and the standards have not been applied equally.

Example 1. Fail rate for L1 Biology 90163 **decreased** by 20% from 2002 to 2003, followed by an **increase** of 11% the following year.

Example 2. L2 Biology 90464 had a fail rate of 34% in 2003 and a fail rate of 63% in 2004, an absolute **increase** of **29%**.

This variability is completely unacceptable when students are depending on the result for admission to further study or competing in the job market.

2. Variability among subjects.

In every year and at every level, languages had the lowest failure rates, followed quite closely by arts subjects. Given these results below, why would students choose to do sciences? Pass rates need to be reasonably consistent among subjects to discourage students from choosing subjects just to achieve the highest pass rates.

Example 1. Level 1, 2003: 17 of 91 (19%) externally assessed standards had failure rates $\geq 50\%$. Twelve of the 17 (71%) were in science or maths. 8 of the 10 standards (80%) with the lowest failure rates, i.e. highest pass rates, were languages.

Example 2. Level 2, 2003: Of the highest 7 failure rates (ranging from a massive 86% fail rate to 62%), 6 were in Science subjects.

Example 3. Level 2, 2004: 29 of 78 (37%) externally assessed standards had failure rates of 50% or more. The 3 standards with the highest pass rates were all in languages.

3. Variability of assessment within a subject.

The ratio between the externally and internally assessed fail rates should all be about 1.0. In 2003 the ratio of external to internal failure rates for level 1 standards ranged from 1.4 for Graphics to 8.2 for Reo Maori. Thus, the failure rate for **externally** assessed standards in Reo Maori was **8 times greater** than the failure rate for **internally** assessed standards.

Not one **internally** assessed standard (with more than 500 candidates) in 2002 or 2003 had a failure rate of more than 45%.

4. Disparity in results between sexes.

Results by sex are only available for 2004.

In almost all subjects, girls did better than boys. They even did better in subjects that are more typically 'boys' subjects and which many more boys sat, such as sciences, economics and technology.

Boys appear to have different ways of learning and NCEA seems unable to capture these abilities.

5. Disparities in requirements to achieve pass, merit and excellence.

We have no consistency for one standard from year-to-year and we have no consistency among standards. Furthermore, we have overlapping and, therefore, meaningless grade categories within each achievement standard.

Example 1. L1 Human Biology 90178, 2004. A student can fail with anywhere between 0 and 94% of the paper correct. Excellence can be gained with 88-100%.

Example 2. L1 Human Biology 90178, 2005, i.e. the same standard as example 1, same curriculum content but different year. A student can fail with anywhere between 0 and 61% of the paper correct. Excellence can be gained with anything between 39% and 100% correct.

6. Incorrect answers provided for internal and external assessment.

We are greatly concerned that so many answers provided by NZQA in the assessment reports and the internal assessment resources are naïve, incomplete and often just plain wrong.

Example 1. L1 Human biology 90178. The ‘correct’ answers provided for the circulatory system questions, in general were wrong and showed an extremely worrying lack of knowledge on the part of the examiners.

Q1(c). Hypertension is NOT caused by blocked coronary arteries. Hypertension is NOT caused by an enlarged heart, despite what is written on the markers’ answer sheet.

Example 2. Internal assessment resource Mathematics 2.5F v4, 2005. What is being asked of the student is **not possible** from the information provided. The sample data **cannot** be generalized to the general population.

7. The grade average.

The calculation of the grade average is a nonsense. NZQA makes a very common but very basic error in the handling of ordinal data.

To make it worse, the ‘fails’ are not included in the calculation of the grade average. If a student sits 8 achievement standards worth 3 credits each, gets ‘excellence’ (worth 4 points) for one and fails the rest, he gets a grade average of 100%. Had his ‘fails’ been included, his grade average would have been 12/96 or 12.5%. Surely a more accurate representation of his ability.

Had a student used this method to calculate an average in a maths exam he would rightly have failed this question. Unfortunately, this tells students that the calculation of average as taught in the curriculum is not in fact what the Ministry of Education itself uses!

Main Results

Note on School Certificate pass rates.

It is often stated as one of the many justifications for the introduction of NCEA is that it is better than school certificate (SC) because SC was a pass fail system in which only 50% passed.

The SC pass percentages were published on the NZQA website. In 2001, percentages who passed ranged from 94.1% in Latin and 93.3% in Indonesian (but only 15 people sat this exam) to 45.8% in Text and Info Management and 49.1% in Food and Nutrition. Of 33 subjects, only 4 had fail rates higher than 45%.

The claim about 50% pass rates in SC does not appear to be factual.

1. Variability of NCEA by year.

Reliability of assessment is crucial. It is well known that large cohorts of students vary little in their performance from one year to the next if the assessment remains the same. Otherwise, whether a student passes or fails depends mainly on the year they sat the exam and not on their knowledge base or their performance in the exam. If we see differences of 5% or more from one large cohort to another it is almost certain that it is because the examination has not been consistent and the standards have not been applied equally. This is clearly unfair to candidates relying on NCEA results for employment, tertiary study or admission to other courses.

Of the **externally** assessed science standards, the fail rate from one year to the next varied a great deal.

For examples see Table 1. The fail rate for L1 Biology 90163 **decreased** by 20% from 2002 to 2003, followed by an **increase** of 11% the following year. Similarly with L1 Biology 90168. Over externally assessed standards for all L1 Biology the failure rate varied, from 52% in 2002, to 38% in 2003, to 49% in 2004.

A very variable pattern also occurred in L1 Chemistry 90173. This achievement standard had a fail rate in 2002 of 15%, in 2003 it was 32% (an **increase** of 17%) and in 2004 it was 50% (an **increase** of 18%).

Table 1. Change in fail rates from 2002 to 2004 for some science achievement standards.

	2002	2003	change from previous year	2004	change from previous year
L1 Biology 90163 % NA (fail rate)	53	33	-20	44	+11
L1 Biology 90168 % NA (fail rate)	61	36	-25	46	+10
L1 Biology overall % NA (fail rate)	52	38	-14	49	+11
L1 Chemistry 90173 % NA (fail rate)	15	32	+17	50	+18
L2 Biology 90464 % NA (fail rate)	-	34	-	63	+29

L2 Biology 90464 was no better (Table 1): it had a fail rate of 34% in 2003 and a fail rate of 63% in 2004, an **increase** of **29%**.

For entrance to University, 14 credits must be gained in Maths level 1 or higher. L1 Math 90147 had a fail rate of 36% in 2002, 50% in 2003 and back to 36% in 2004.

This variability is completely unacceptable when students are depending on the result for university admission: success or failure depends much more on the year of the assessment than the performance of the student.

2. Variability of NCEA among subjects.

It is very important in examinations that consistency and comparability among subjects is maintained to the maximum extent possible. Students should never choose subjects just to achieve the highest pass rates.

Level 1, 2002: 91 externally assessed standards with more than 500 candidates. Of these standards, 19 (21%) had failure rates of 50% or more. Over half were in science or maths subjects.

Of the 10 standards with the lowest fail rates, 5 of these were in languages.

Level 1, 2003: 17 of 91 (19%) externally assessed standards had failure rates $\geq 50\%$. Twelve of the 17 (71%) were in science or maths.

8 of the 10 standards (80%) with the lowest failure rates, i.e. highest pass rates, were for languages.

Level 1, 2004: 18 of 95 (19%) externally assessed standards had failure rates $\geq 50\%$. Over half were in science subjects.

Seven of the 10 standards (70%) with the highest pass rates, were languages.

Level 2, 2003: 22 of 73 (30%) externally assessed standards had failure rates of 50% or more (11 of the 20 were in science or maths).

Of the highest 7 failure rates (ranging from a massive 86% fail rate to 62%), 6 were in science subjects.

Of the 10 standards with the lowest fail rates, 5 of these were in languages.

Level 2, 2004: 29 of 78 (37%) externally assessed standards had failure rates of 50% or more.

Again, of 10 standards with the highest pass rates, 5 of these were in languages.

Level 3, 2004: 17 of 69 (25%) externally assessed standards had failure rates of 50% or more. The 3 standards with the highest pass rates were all in languages.

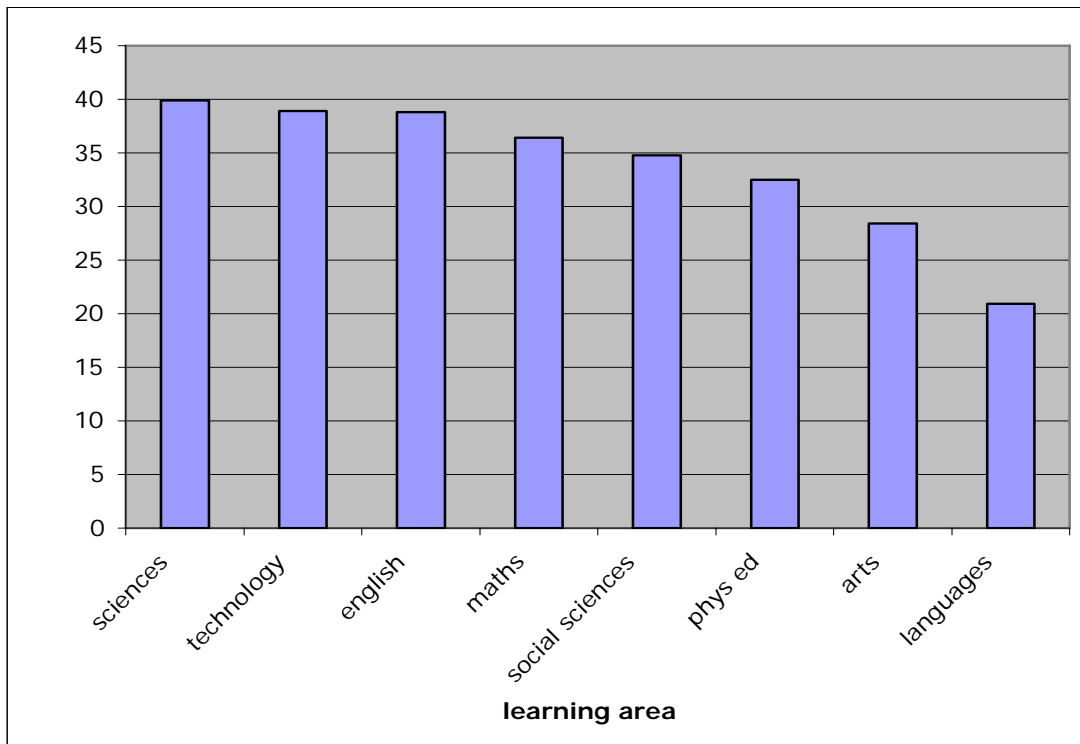
Now look at the externally assessed standards, grouped by learning area.

Level 1 2002: languages (excluding English) had the lowest mean failure rate (20%) followed by phys ed subjects (23%). In contrast, technology subjects had a mean failure rate of 45% and science a weighted mean failure rate of 39.0%. Then in 2003 this swung wildly.

Level 1 2003: Phys Ed subjects went from the 2nd lowest fail rate in 2002 to the highest fail rate in 2003. Technology subjects went from the highest fail rate in 2002 to the 3rd lowest in 2003.

Over all years the highest fail rates for Level 1 learning areas were in sciences and technology and the lowest fail rates were in arts and languages (Figure 1).

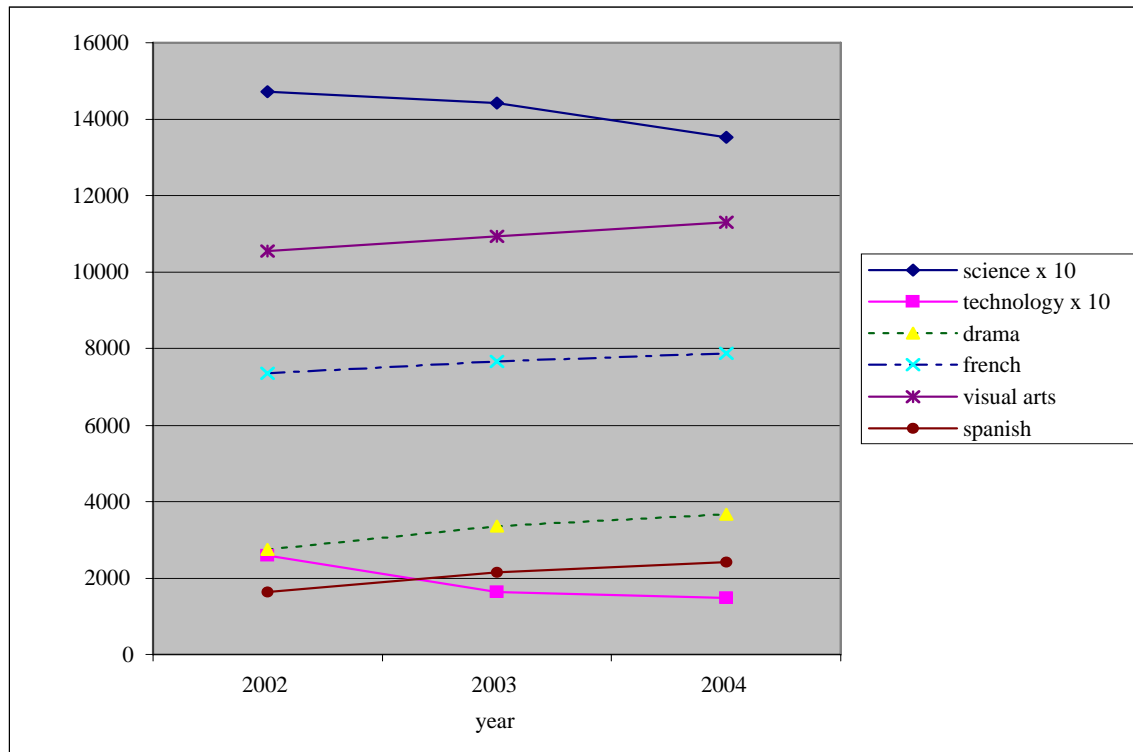
Figure 1. Fail rates for Level 1 learning areas, pooled over all years.



In every year and at every level, languages had the lowest failure rates, followed quite closely by arts subjects. Given these results, why would students choose to do sciences?

The number of students sitting externally assessed L1 science papers has decreased by 8% from 2002 to 2004 and numbers taking technology papers has decreased by 43% (Figure 2). In contrast, numbers doing drama have increased by 34% and both French and visual arts have increased by 7%. It is not difficult to see why. Students and their parents seem to be just doing what they perceive as best for their short-term prospects.

Figure 2. Number of NCEA papers sat for selected subjects. (Numbers in science and technology have been divided by 10 to fit more easily on the graph)



3. Variability of assessment within a subject.

The NCEA is a mixture of internal and external assessment. Unfortunately, in 2002 and 2003 the failure rates for internally assessed standards were not consistently reported by schools. From 2004, instead of insisting that schools report all grades for internally assessed standards, NZQA no longer reports fails for internally assessed standards.

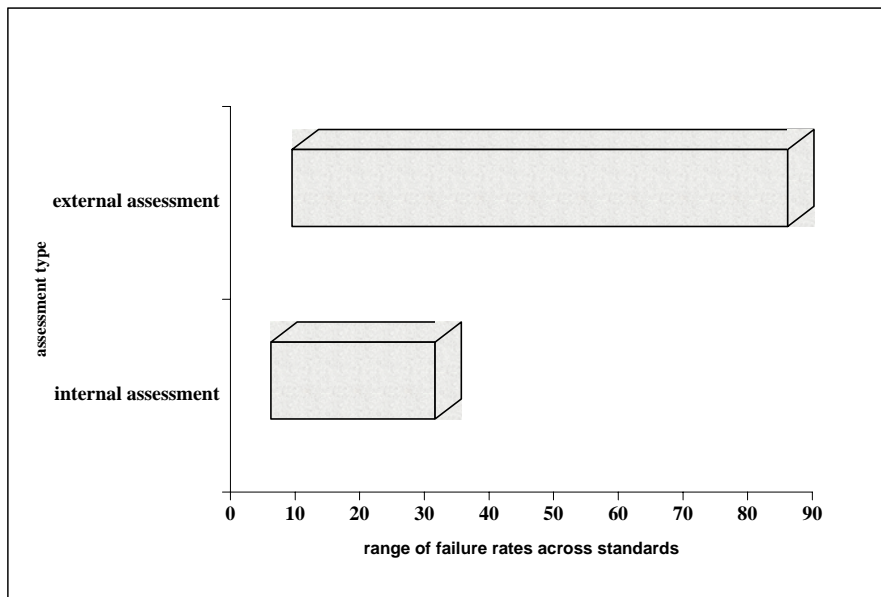
Without the fail rates, the data presented by NZQA for internally assessed standards are meaningless. Therefore, the only useful data available are from 2002 and 2003, which I have used below.

For levels 1 and 2 **internally** assessed learning areas (with 2000 or more assessments submitted over the learning area), the failure rates in 2002 and 2003 ranged from 5.3% in languages to 25% in technology.

Not one **internally** assessed standard (with more than 500 candidates) in 2002 or 2003 had a failure rate of more than 45%.

For level 2 **internally** assessed standards in 2003 (Figure 3) the failure rate ranged from 3.2% in Reo Maori to 28.6% in Home and Life Sciences. In contrast to failure rates for externally assessed level 2 standards in 2003, there were no internally assessed standards with failure rates anywhere near 50% (Figure 3).

Figure 3. Range of failure rates across standards in level 2 assessments in 2003.



If the moderation process was working properly the ratio between the externally assessed and internally assessed fail rates should all be about 1.0, unless it is intended that internal and external assessment should be of a different standard. The ratios (external/internal) of

the L2 fail rates in 2003 ranged from 1.4 for Graphics to 8.2 for Reo Maori. Thus the failure rate for externally assessed standards in Reo Maori was **8 times greater** than the failure rate for its internally assessed standards.

The evidence shows that internal assessment is measuring at a different standard to external assessment - the overall pass rate was much, much lower for externally assessed standards compared with internally assessed standards. Therefore national externally assessed standards are extremely important for establishing national standards. Otherwise, the school a child attended may become more important than their achievement.

From the evidence presented above, students would be daft to choose externally assessed standards if there was a choice of internally assessed ones. For the best results they should choose languages or arts and only choose internally assessed standards.

4. Disparity between sexes.

Results by sex are only available to me for 2004 so I cannot compare with previous years.

In almost all subjects, girls did better than boys. They even did better in subjects that are more typically 'boys' subjects and which many more boys sat, such as sciences, economics and technology.

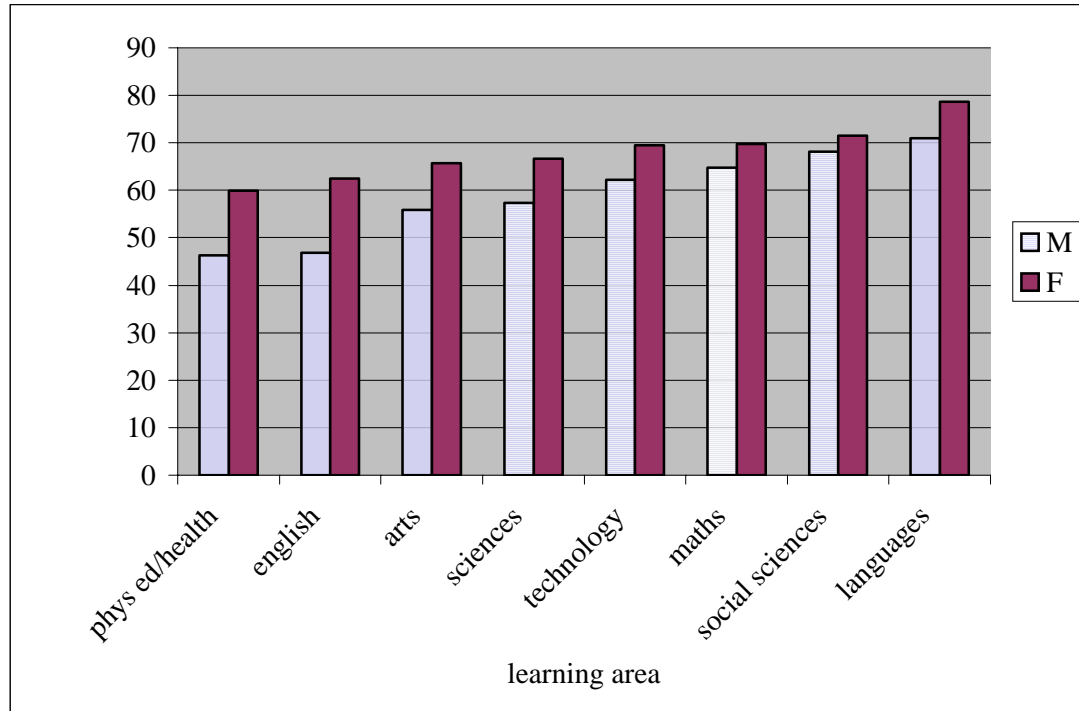
In level 1, more boys passed than girls in only 2 of 51 subjects (with more than 500 candidates): accounting and geography. For both subjects the difference in pass rates was tiny, <1%. When grouped over learning areas, girls had higher pass rates than boys in every area (Figure 4).

In level 2, girls were assessed as more able than the boys in every subject, bar two: accounting and economics.

Again, in level 3, boys only got higher pass rates than girls in 2 subjects: accounting and chemistry.

Boys appear to have different ways of learning and NCEA seems unable to capture these abilities.

Figure 4. Pass rates for learning areas by gender. Level 1, 2004.



5. Disparities in requirements to achieve pass, merit and excellence.

Relying solely on the number, the mark, the grade, or the outcome of assessment misses a large and essential part of the process. It is often more important to closely investigate the procedures and criteria for awarding these numbers, marks or grades.

So let us look in more detail at some of the external assessments and in particular the criteria for awarding 'not achieved' (NA), 'achieved' (A), 'merit' (M) or 'excellence' (E). Percentages have been used for clarity in these examples.

As an example let us take Level 1 Human Biology 90178 (2004). In this achievement standard there are 10 'opportunities' for marks and only in 2 of them can anyone score Excellence.

Not Achieved can be awarded with anything between 0 and 94% of the paper correct (Figure 5).

Achieved can be gained with anything between 35% and 94% correct.

Merit can be gained with anything between 71% and 88% correct.

Excellence can be gained with anything from 88% to 100% correct.

If we now look at the identical achievement standard (Level 1 Human Biology 90178) in 2005, the pattern of grades is completely different. Now there are 18 'opportunities' for marks and only in 3 can anyone score Excellence.

Not Achieved can be awarded with anything between 0 and 61% correct (Figure 6).

Achieved can be gained with anything between 29% and 58% correct. The candidate must get 9 'correct' answers of 18.

Merit can be gained with anything between 35% and 65%. The candidate must get 9 'correct' answers of 18.

Excellence can be gained with anything from 39% to 100% correct. The candidate must get 9 'correct' answers of 18.

In this achievement standard, 4 students may each get exactly 50% of the paper correct and they may each be awarded a completely different grade – Not Achieved, Achieved, Merit or Excellence.

Whatever grade a student gets for this standard, no one can possibly tell what this student knows or how much of the material they have been able to learn in the year because of the overlap in grades. In fact student A in 2005, who achieved 'excellence' (at 39% of the paper correct), knew much **less** than that known by student B (at 61% of the paper correct), who failed. This is obviously ridiculous.

Furthermore the grades awarded change enormously from year to year. The graphs below (Figure 5 and Figure 6) represent the grades and marks for exactly the same standard in consecutive years but they appear completely different.

Figure 5. Marks and grades for Human Biology 90178, 2004.

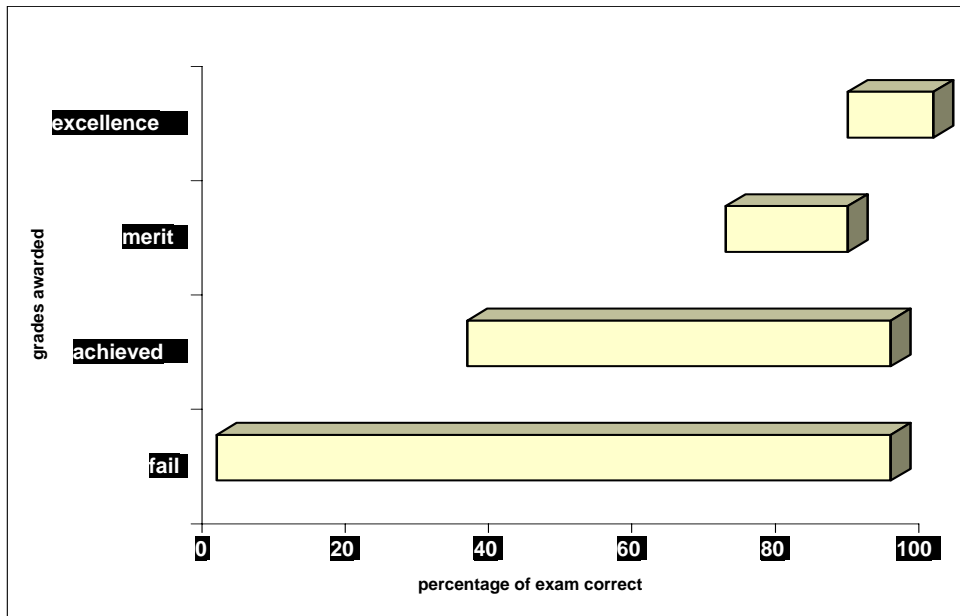
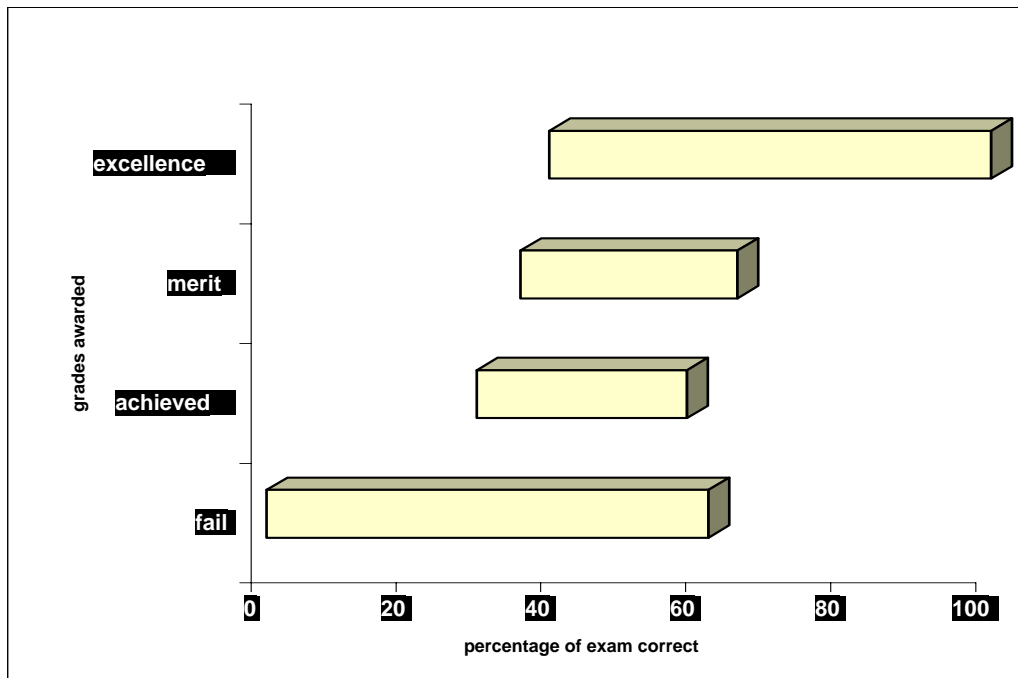


Figure 6. Marks and grades for Human Biology 90178, 2005.



6. Incorrect answers provided for internal and external assessment.

We are greatly concerned that so many answers provided by NZQA in the assessment reports and the internal assessment resources are naïve, incomplete and often just plain wrong.

Let us consider for example (and there are many others to choose from) Level 1 Human biology 90178, 2005. The 'correct' answers provided for the circulatory system questions, in general were wrong and show a worrying lack of knowledge.

Q1.(a) Blood pressure usually increases with age. 110/70 is not a low blood pressure for a lot of people: it would be a normal blood pressure for the people sitting this exam (while they are not stressed). 110/70 does not necessarily indicate a healthy heart. This blood pressure can occur after a heart attack in a person with previous high blood pressure. A wide difference between numbers (pulse pressure) does not indicate a healthy BP. Stiffening of the arteries can cause a wide pulse pressure and this can cause organ damage.

Q1(c). Hypertension is NOT caused by blocked coronary arteries. Hypertension is NOT caused by an enlarged heart, despite what is written on the markers' answer sheets.

As another example L1 Biology 90166 - digestive and skeletal systems.

Q1b. The answer given in the assessment schedule does not answer the question **at all**. Furthermore, some muscles in the leg are under voluntary and involuntary control. How else do we stand up?

Q2c. Why on earth, to get an excellence, do osteoarthritis and rheumatoid arthritis both have to be discussed when the question does not ask for it? How about other forms of arthritis which are more common in the age group sitting this paper, such as septic arthritis or juvenile arthritis? One can give a very good answer without mentioning any specific type because they share features in common. Surely that is the point of the question?

Let us now consider an example from an internal assessment resource Mathematics 2.5F v4, 2005. This requires students to select a sample and from this sample make inferences about the general population. Unfortunately, what is being asked of the student is **not possible** from the information provided. A sample of 75 people is provided, but the selection criteria (or equivalent information), which is absolutely essential information to do this problem, is missing. Students are then asked to take a representative sample of 25-30. They already have all the data for a sample of 75 people so why take another sample? That is ridiculous. The person setting this assessment resource does not appear to have realized that if the initial sample is not representative of the underlying population, neither will a sub-sample be representative. Thus the sample data **cannot** be generalized to the general population without knowing how the original 75 were selected. Many students will understand this and the assessment project will simply confuse them.

7. Grade average.

The results for level 1 to 3 NCEA are categorized as 'excellence', 'merit', 'achieved' and 'not achieved'. This is an example of an ordinal data system whereby the result categories can be placed in an order but do not have specific numerical values.

Arithmetic CANNOT be performed on ordinal data in a meaningful way. However, arbitrary numbers have been attached to the categories and an artificial numerical system generated: 'excellence' = 4, 'merit' = 3, 'achieved' = 2 and 'not achieved' (fail) = 0. To calculate the grade average, these arbitrary numbers are then manipulated as if they were real numbers. However, they are not real numbers and should not be treated as such. This is, unfortunately, a very common but very basic error in the handling of ordinal data. The resultant statistic, the grade average, is a nonsense.

If a numerical answer is required then one should start with numerical data, such as a percentage score? This way a meaningful average can be generated from real numerical data.

However, the deficiencies in the grade average do not stop there. In the interim results notices and the 'record of learning', any standard which has been failed is not listed. Furthermore, the 'fails' are not included in the calculation of the grade average. Failing to

include 'not achieved' results makes an enormous difference to the results. This is deceitful.

Grade average equals the total marks gained (for one level) divided by the total available marks for those achievement standards, converted to a mark out of 100 (percentage). If a student sits 8 achievement standards worth 3 credits each, gets 'excellence' (worth 4 points) for one and fails the rest, he gets a grade average of 100%. Had his 'fails' been included, his grade average would have been $12/96$ or 12.5%. Which score more accurately represents the student's **overall average** achievement at this level?

Had a student used this method to calculate an average in a maths exam he would rightly have failed this question. Unfortunately, this tells students that the calculation of average as taught in the curriculum is not in fact what the Ministry of Education itself uses!