

Getting the most out of NQF statistics: A guide for users

Part 3 – Different ways of comparing data

This section describes different ways of comparing data. Table 3.1 below summarises some of the main kinds of analyses to consider. The focus is on the uses and advantages of each kind of comparison, as well as factors that potentially complicate interpretation.

Please note that:

- the analyses presented here do not constitute an exhaustive list
- the final analysis chosen is dependent on the question/s to be answered
- some analyses may contain elements of more than one of the types described here e.g., an intra-subject analysis might also include a longitudinal component.

Table 3.1 Examples of questions and analyses that can be used with NCEA/NQF data and statistics

Type of question	Type of analysis
How are patterns of results changing over time?	Longitudinal analysis
What are the particular areas of strength and/or weaknesses in a teaching programme?	Intra-subject analysis
How can the particular strengths of different teachers in a department or school be used to assist one another to improve teaching practice?	Inter-staff analysis
What are the particular areas of academic strength and/or weakness within a school?	Inter-department analysis
How does a school compare with other similar schools in the performance of its students on the NQF?	Inter-school analysis

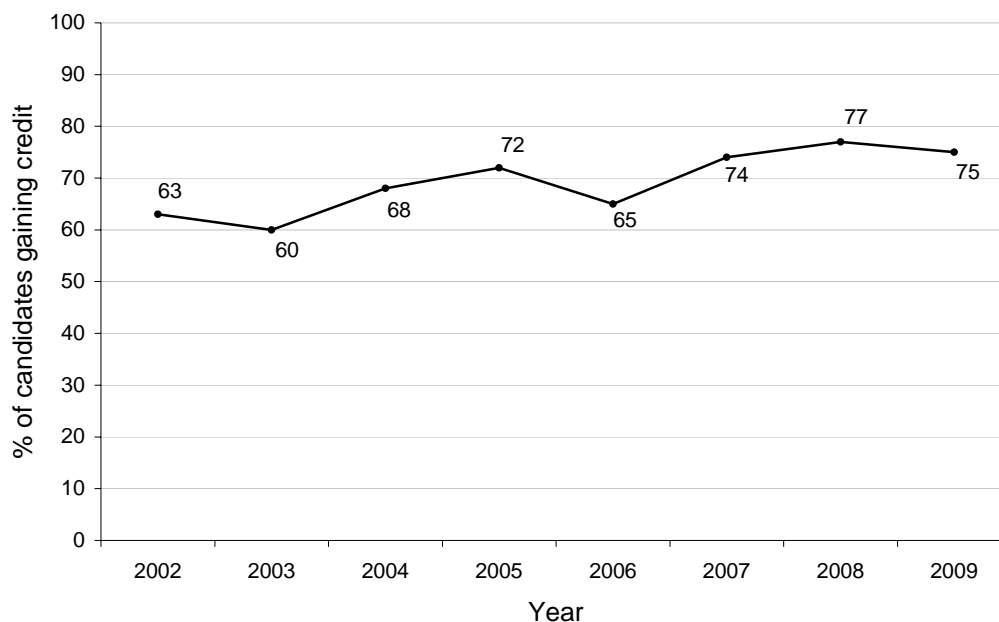
Longitudinal analyses

One of the most useful analyses for tracking results is *longitudinal*; that is, an analysis of how assessment results change over time. Currently, there are not enough years of NCEA data available to make such an analysis particularly meaningful. Longitudinal analyses will become more useful, however, as time goes on. The reason that a number of years of data are needed to make sense of a longitudinal analysis is that, particularly in analyses with relatively small numbers of results (e.g., at the level of a single department), numbers can be expected to fluctuate because of statistical measurement error (see part 5), which has nothing to do with the quality of the teaching or educational programmes. Even for analyses involving large numbers of students, some variability due to measurement error is to be expected.

Longitudinal analysis example

Figure 3a shows the proportions of candidates at a hypothetical school achieving credit in an externally assessed standard over eight successive years.

Figure 3a Hypothetical data showing variability in the percentages of candidates achieving credit in a standard over eight successive years.



Note that while the results shown in Figure 3a fluctuate from one year to the next, over the full time period, an upward trend of an increase in success is evident. For this reason, changes in a results profile in a single year, in either direction, should not be a cause for either concern or celebration. A particularly *large* change in a single year, however, may be a signal that something affecting assessment results in an important way has changed e.g., a high turnover of teaching staff in a particular department, or the revision of one or more standards.

The appropriate basis for longitudinal analysis is the performance of the same entity (e.g., department or school) in the past. This is one of the advantages of longitudinal analyses; it is self-referenced and does not require a comparison with data from other departments or schools. Interpreting inter-school or inter-subject comparisons can be problematic because the entities being compared are often not comparable. While changes in teaching staff or the assessments themselves can still create some interpretative complication, a comparison of a department or school with itself over time is generally more interpretable and more useful than a comparison of one department or school with another.

A candidate's performance in a norm-referenced assessment is always measured relative to the performance of other candidates. So, if there were a change in the strength of a cohort from one year to the next under normative assessment, it would be unlikely to show up in the final results data, because results would be scaled to ensure the distributions matched for the two years; that is, any inter-year variability would be masked by scaling procedures. Under standards based assessment, however, candidates are assessed relative to a standard and not relative to one another. For this

reason, real changes in overall performance can be reflected in results data, and a longitudinal analysis can highlight these changes.

Intra-subject (between standards) analyses

It can be useful, particularly for developing teaching programmes, to track differences in results between standards within a particular subject area. Such analyses can shed light on specific strengths and weaknesses within a particular department. When performing analyses of this kind, however, be aware of the following:

- 1 Analyses of differences between standards within a school typically involve comparatively few students' results. Statistics based on low numbers of data are prone to high variability, and so are not reliable estimates of an actual situation.

The most common, everyday example that illustrates this is a political poll. Any political poll has a margin of error associated with it, which means that it is highly likely (usually 95% likely) that the actual profile of the population will be within the margin of error relative to the poll result. So, if 42% of respondents to a poll say that they will vote for a particular party, and the margin of error for the poll is 3%, then the actual proportion of the population intending to vote for that party is 95% likely to be between 39% and 45%.

Note that a margin of error is actually the same thing as a confidence interval. A confidence interval is relative to the number of results analysed, with low numbers producing a wider confidence interval. An individual class can therefore be expected to have a high margin of error associated with its data, so that a single year of results data from a particular class may not accurately reflect the quality of teaching in that class. A handful of especially strong or weak students in a particular year will make a substantial difference to the results profile of the class – in a class of 25 students, for example, a single student constitutes 4% of the class, so if three students were to move from not achieving to achieving, the results profile would shift by 12 percentage points.

- 2 Results for unit and achievement standards (particularly externally assessed achievement standards) cannot be compared for two reasons. First, unit standards do not have differential grades; there are no Merit or Excellence grades available for them. Second and more seriously, results for internally assessed standards are only recorded in NZQA databases for students who have *achieved* a standard. There is no record of those who have attempted a standard but not achieved it. Because of this limitation in the data, unit standard statistics from the NZQA website cannot be used meaningfully in any analysis except longitudinal comparisons within schools, unless schools keep their own records for unit standards that include students who have attempted but not achieved in these standards.
- 3 If there appears to be a gap in performance at a school or in a department between two standards, this does not necessarily reflect a weakness in the learning programme with respect to the standard in which performance is poorer (nor, necessarily, a strength with respect to the standard in which

performance is stronger). It may simply be that students find the standard with the highest rate of success easier than the standard with the lower rate of success. To address the question of whether the performance profiles for a given two standards reflect any noteworthy issues for a school with respect to its teaching programmes, it is necessary to reference those profiles to the similar profiles in the national statistics. This can be done by referring to the NQF statistics, *results distributions by learning area* on the NZQA statistics website, on which the overall (national) percentages of students receiving grades of *Not Achieved*, *Achieved*, *Merit*, and *Excellence*, for every externally assessed achievement standard are shown.

Intra-subject analyses example

Consider the Level 1 English standards 90053 (*Produce formal writing*) and 90054 (*Read, study and show understanding of extended written text*). The national results profiles for these standards in 2004 are shown in Table 2. Note that the percentages of students receiving *Merit* and *Excellence* grades in each standard are very similar, whereas there is a substantially greater percentage of *Achieved* results in 90054 than in 90053, and a commensurately lesser percentage of *Not Achieved* results. These differences between standards need to be taken into consideration in any comparison of results from these standards within a school.

Table 3.2 National percentages of students in each grade category for English standards 90053 and 90054 in 2004

	Not Achieved	Achieved	Merit	Excellence
Standard 90053 <i>Produce formal writing</i>	44.7%	38.8%	13.9%	2.5%
Standard 90054 <i>Read, study and show understanding of extended written text(s)</i>	37.6%	47.0%	12.5%	2.9%

In a hypothetical Year 11 English class, 79% of the students receive a grade of *Achieved* or better in standard 90053, compared with 68% for standard 90054. Nationally, around 55% of students received a grade of *Achieved* or better in 90053, compared with 63% in standard 90054. What can the teacher of this class conclude from this? First, the students seem to be achieving these standards at a somewhat higher rate than the national average. However, before celebrating this apparent success, two things need to be considered.

Remember that high variability is to be expected in statistics based on small numbers of cases. While the national data are based on a very large number of cases, class data are not. So, the apparently high rate of success for this class may actually be a chance effect. The class may just happen to be particularly academically strong in that year, or the exam may have happened to coincide especially well with work that they had

completed close to the exam time. One way to test these possibilities is to compare data for a number of successive years (longitudinal data), and if this shows that the teacher's classes perform *consistently* above the national average, it would be unlikely to be a statistical anomaly. Another way to test the possibilities is to calculate confidence intervals for the estimates under comparison. If the confidence intervals do not overlap, the observed difference is unlikely to be attributable to measurement error.

Teachers' professional judgment can also shed light on these issues, and consultation with colleagues over these kinds of analyses is encouraged. (In fact, professional judgment should always be used in interpreting any statistical analysis). Another consideration is the demographic circumstances of the school. Setting aside the issues of interpreting the *absolute* difference between the national results and the performance of the hypothetical English class, the analysis of the data shown in figure 1 was actually prompted by the question of what was learnt from the *difference in the results* profiles for the two standards about the strengths/weaknesses in a teaching programme. The national statistics suggest that 90053 is a slightly more difficult standard to meet than 90054, with grades of Achieved or better around eight percentage points more common for the latter. One issue to take into account here is whether the national cohorts undertaking these two standards are likely to be significantly different. Again, professional judgment is probably the best measure of this, but it is perhaps unlikely to be the case for two Level 1 English achievement standards.

Whereas in the national data, 90053 seems to be the more difficult of the two standards, in the hypothetical class results, there was a considerably higher success rate for 90054. As for the absolute comparison, the small size of the data set for the hypothetical class needs to be considered, and similar data from consecutive years would strengthen this conclusion. But on the face of it, the data for the hypothetical class suggest that the teacher has a particularly effective programme for reading and reading comprehension.

In the above example, just two standards were compared. In a real analysis, teachers would probably want to compare performance across all of the standards that they had taught in a particular course, or to a particular class. All of the same caveats would apply when making these comparisons, although an analysis of multiple standards would be likely to be more informative in many ways, because it would be possible to determine whether there is a common theme in standards that stand out as having particularly strong or weak rates of success.

Inter-staff analyses

Heads of department may wish to track the performance of individual teachers in order to improve the overall performance in their departments. It goes without saying that such analyses need to be approached with sensitivity and it is suggested that, if analyses such as this are carried out, results are communicated with a clear emphasis on teachers assisting one another to improve in a collegial environment. Furthermore, it is likely that some teachers will excel with some student groups or with particular curriculum material, and other teachers will excel with different groups of students or other material.

To have any validity, a comparison of teaching performance really needs to be carried out over time. As noted above, a single year's data from a single class does not give a reliable indication of the true nature of the teaching in that class.

Where classes are to some extent streamed, or grouped according to ability, a comparison of teachers' performance becomes next to impossible. Although it would be expected that a strong class would achieve a higher success rate than a weaker class, regardless of any differences in teaching effectiveness, to confirm this would require a reliable way of determining the expected results for each class, taking into account the difference in ability and there is no straightforward way of doing this. Furthermore, it might often be the case that classes of students with differing levels of ability undertake a somewhat different mix of standards in a given subject area. Standards, even within a subject and level, are not homogeneous in difficulty and, again, there is no straightforward way to correct for any differences in standard difficulty in any comparison of data for different classes.

Inter-department analyses

It may be of interest to a school to identify subject areas or departments of particular strength or weakness. As with comparisons of individual standards, a departmental comparison needs to take into account the different rates of achievement that are expected in the different subject areas. There are two reasons for this:

- 1 Any differences in standard difficulty are more likely to be evident across subjects than within subjects (although a range of difficulty is to be expected within subject as well). It is also likely that some subjects are intrinsically more difficult than others for the majority of students.
- 2 Some departments and subject areas tend to attract students of higher ability than others. Clearly these two factors will interact; a difficult standard undertaken by a strong cohort might have similar results profile to a less difficult standard undertaken by an average cohort.

One way to test this hypothesis is to compare differences in standard-aggregated results for two departments, and to determine whether or not this difference is greater than, less than, or about equal to the difference expected on the basis of the national data, or preferably a subset of national data drawn from schools with similar demographic characteristics to your own (e.g., mid-decile co-ed). The following example compares just two subjects. In an actual analysis, it may be more appropriate to take into account a broad range of departments.

Inter-department example

In this example, the relative performance in Level 3 standards of two departments, Geography and Mathematics, are compared for a hypothetical school. Table 3 shows the numbers of results in each grade category for each Level 3 standard in Geography, and each Level 3 standard in Mathematics (excluding probability and statistics standards). The *Total* rows in the table show all of the results in each grade category across all standards in both subjects. The *subject aggregate* rows show the percentage of the total results for each subject falling into each grade category. It seems clear

from these data that the success rate in Level 3 Mathematics at this school is higher than that in Level 3 Geography, with 71% of results in mathematics gaining credit compared with 55% for Geography. What is not clear from these data, is why this is so. Is it because the teaching in Mathematics is superior at this school? Is it because the Level 3 Geography standards are more challenging than the Level 3 Mathematics standards? Is it because the Mathematics cohort is more able?

Table 3.3 Results for a hypothetical school in Level 3 Geography and Mathematics

Geography					
Standard	Total number of results	Number of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90701	45	21	14	7	3
90702	30	18	8	3	1
90704	57	20	25	10	2
Total	132	59	47	20	6
Subject aggregate (%)		45%	36%	15%	4%
Mathematics					
Standard	Total number of results	Number of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90638	84	23	41	14	6
90639	90	30	42	10	8
90644	57	15	22	14	6
90635	69	12	31	18	8
90636	75	27	31	14	3
Total	375	107	167	70	31
Subject aggregate (%)		29%	44%	19%	8%

To answer the question of why the success rate in Level 3 Mathematics is higher than that in Level 3 Geography at this school, it is necessary to compare the school's results in these subjects with those of similar schools nationally. Assume that this hypothetical school is in the mid-decile range. Table 4 shows the numbers of results for each Level 3 Geography and Mathematics standard gained by students at Decile 4–7 schools nationally, as well as the percentages of these results in each of the four grade categories. These are real data on the NZQA website from the 2004 external assessment round¹.

Before it is possible to directly compare these data with the data for the hypothetical school, it is necessary to aggregate the percentages in each grade category across the

¹ <http://www.nzqa.govt.nz/qualifications/ssq/statistics/natl-nqf.do?statsYear=2004#t>

standards in each subject. The aggregate percentages are shown in the bottom row for each subject. Note that the aggregate figures are weighted by the size of each cohort – they are not simple averages.

Table 3.4 Numbers of results and percentage of results in each grade category for each externally assessed standard at Level 3 in Geography and Mathematics, across all Decile 4–7 secondary schools

Geography					
Standard	Total number of results	Percentage of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90701	2,256	55%	34%	8%	3%
90702	2,306	57%	30%	11%	2%
90704	2,335	30%	43%	19%	8%
Subject aggregate (%)		47%	36%	13%	4%
Mathematics					
Standard	Total number of results	Percentage of results in each grade category			
		Not Achieved	Achieved	Merit	Excellence
90638	3,011	40%	41%	16%	3%
90639	2,882	45%	43%	11%	1%
90644	4,369	30%	52%	14%	4%
90635	3,050	33%	43%	20%	4%
90636	3,039	47%	39%	14%	0%
Subject aggregate (%)		38%	44%	15%	3%

Compare the subject aggregate percentages of results for the hypothetical school (Table 3.3) with those for all Decile 4–7 schools (Table 3.4). For Geography, the percentages are very similar – identical for the *Achieved* and *Excellence* categories, and differing by just three percentage points for each of the *Not Achieved* and *Merit* categories. These small differences are well within the margin of error (95% confidence interval) for the estimates of the performance in Geography at the hypothetical school; the estimate of the *Not Achieved* rate, for example, has a margin of error of +/-8% (rounded to the nearest percentage point; check this using the formula for confidence intervals, given in Part 5), so that the confidence interval for this estimate is approximately 37% – 53%. Thus the results here give no reason to suspect that the geography department at the hypothetical school is any stronger or weaker than average.

For Mathematics, the situation is a bit different. The percentage of students gaining credit at the hypothetical school is 10 points higher than it is for the Decile 4–7

aggregate. The margin of error for this estimate is narrower, because the sample is larger; there are 375 mathematics students at the hypothetical school, compared with 132 geography students. The 95% confidence interval for the success rate in mathematics turns out to be 24% – 34%, the upper limit of which is still below the *Not Achieved* rate for the Decile 4–7 aggregate (39%). So we can be over 95% confident that the success rate in Mathematics standards at the hypothetical school is higher than would be expected based on the school’s demographics. The conclusion of this analysis is probably that the school has an excellent Mathematics department, with the only real alternative being that, for some reason, the students at this school are particularly mathematically gifted.

Inter-school analyses

Many schools use their NQF/NCEA statistics to compare themselves with other schools in their local area, with the national statistics, or with a subset of the national statistics such as other schools of the same decile rating. Conclusions from such comparisons should be drawn with a great deal of care. Wrong conclusions can be serious and damaging in terms of the prestige and public understanding of a school. The league tables produced by the media each year when the NQF/NCEA results are published are essentially meaningless when it comes to comparing schools’ performance, because there are a great many factors beyond the control of a school that affect outcomes in assessment.

Inter-school example

Consider the data presented in Table 5, which show that success rates in Level 3 NCEA are very much affected by the demographic characteristics of a school. (In fact, similar disparities are evident in almost all secondary school assessments.) The range of success in Level 3 NCEA makes clear just how profound the effects of demographics are, from low decile, single-sex boys’ schools with a Year 13 achievement rate of around 20%, to high decile, single-sex girls’ schools, with a Year 13 achievement rate above 75%.

Table 3.5 Percentages of Year 13 students at schools with various demographic classifications achieving Level 3 NCEA in 2004

	Co-ed schools	Single sex boys’ schools	Single sex girls’ schools
Decile 1 – 3	28.8	19.7	34.5
Decile 4 – 7	44.2	47.3	56.7
Decile 8 – 10	57.5	55.9	75.5

The point clearly made by the data presented in Table 3.5, is that any comparison between schools at least needs to take into account the decile rating of the schools, as well as their co-educational or single sex status. A comparison with average statistics for schools with similar demographics will be far more meaningful than a comparison with national averages. Consider, for example, two schools, one a Decile 10 single-

sex girls' school with a Year 13 success rate in Level 3 NCEA of 67%, and a Decile 1 single-sex boys' school with a success rate of 30%. In a league table, the former school would look like it is doing a much better job for its students than the latter, particularly if the reader was ignorant of the demographic differences. In the context of the data in Table 3.5, however, it is clear that the reverse is true; students at the low decile boy's school are achieving at much higher rates than expected for a school of that type, whereas those at the high decile girl's school have a much poorer rate of success than students at other similar schools.

Even a comparison of schools with similar decile and gender characteristics has its difficulties. Demographic variables do not tell the full story about a school's circumstances, and there are many other important predictors of success in assessment that are outside of a school's control such as its geographic location and its ethnic and cultural mix. The decile statistic itself is problematic because it accounts only for the proportion of students in a school that is likely to be below the poverty line. It does not describe the *distribution* of economic circumstances within a school, so two schools with the same decile ratings might actually have quite different socioeconomic profiles.

Another approach would be for a school to choose a group of other schools with similar demographics with which to compare its results. Often, teachers and principals have a better idea of which other schools have similar circumstances to their own, than can be provided through official decile ratings or other administrative data.