MINISTERIAL COUNCIL ON EDUCATION,
EMPLOYMENT, TRAINING AND YOUTH AFFAIRS

3579

Analysis of Queensland students' NAPLAN performance

2008 National Assessment Program
Literacy and Numeracy

Queensland Government

Queensland Studies Authority
Partnership and innovation

Contents

Executive summary

2008 was the first year of the National Assessment of Literacy and Numeracy Program (NAPLAN). For the first time students across Australia sat for exactly the same test at the same time of year and the results for all states and territories were released by MCEETYA on Friday 12 September 2008. The full report will be released on 19 December 2008.

The figures released on 19 December will vary slightly from the earlier figures as time has been available to match data from different test booklets that had been attributed to different students. The processing of over a million test booklets caused considerable difficulty in ensuring that the correct information was attributed to the correct child.

Even with the newer data the average results for Queensland students in nearly all aspects of the tests were still below that of every other state and above only those for students in the Northern Territory.

Some of these results could be attributed to the nature of the test in the way it was constructed and the way in which it engaged the students. Other possible reasons for the poor performance have to do with curriculum coverage and the lack of diversity within the range of items on the test.

In the absence of detailed guidelines for test constructers it was obvious that there was no clear direction on what aspects of literacy and numeracy were to be tested. There are issues about mathematical knowledge as opposed to numeric reasoning, spelling as proofreading rather than as necessary for the production of writing, and a lack of understanding of how to measure reading comprehension without relying on pattern recognition.

Reading showed different patterns from other strands, in that the Year 9 results were closer to the national percentage that were correct, whereas in Years 3, 5 and 7 they were further away. After Year 9, Year 5 was closest to the national percentage.

While Queensland students did not perform well on the writing task, Queensland representatives who were involved in national selection of exemplar scripts did not believe that the work they saw of students in other states was necessarily of a higher standard.

Numeracy exhibits the same pattern of students being behind national percentages at the younger year levels, but getting closer to the national average at older year levels, with perhaps the strongest comparative performance at Year 7 and Year 9.

Although the average performances are not encouraging, it is useful to consider the performance of different groups of students. Comparisons of sub-population groups such as boys, girls, remote students and Indigenous students can only be done for Queensland students until all national data are available. It is not possible to determine if these differences are the same for other states and territories. However, it appears that girls outperformed boys in most areas of literacy and that boys performed better in numeracy. Similarly students in non-government schools generally did better than government school students and non-A&TSI students achieved higher averages than A&TSI students.

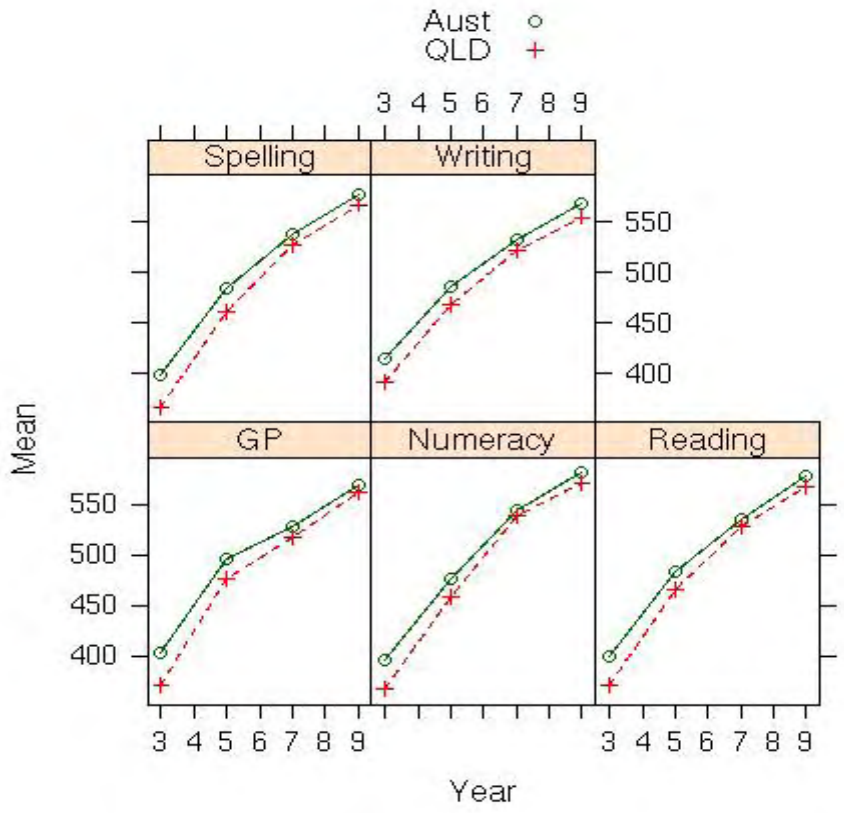
Overall, it is difficult to know what effect engagement has, without knowing the frequency with which interstate students omitted items. However, within Queensland the rate of omitting items is of a concern that suggests there are some overall problems with engagement, particularly in Spelling and overall at Year 3 and Year 9.

While it would not be appropriate to label the tests as diagnostic there are messages in the data for teachers and parents. Schools need to know what to look for in the data and how to interpret it. Teachers need information on how best to prepare students for the test and how to talk about the test with parents.

Results of the tests

Although direct comparisons cannot be made between the results of the former state-based test and 2008 NAPLAN, the patterns in Queensland's results and Queensland's relative performance in NAPLAN are broadly similar to the patterns and relative performance under the state-based testing regime.

As reported earlier, the data show that in many aspects of the test, and particularly at Years 3, 5 and 7, Queensland students on average performed at a lower level than all other states but better than students in the Northern Territory. The following graph represents performance of Queensland students compared with the national average. By Year 9, students have almost reached the national average but the performance at Year 3 is considerably below the national average in all aspects of the test. Data show that the average age of Queensland Year 3 students is 8 years and 3 months and they have been at school for 2 years and 4 months. This is nine months less than the national average.



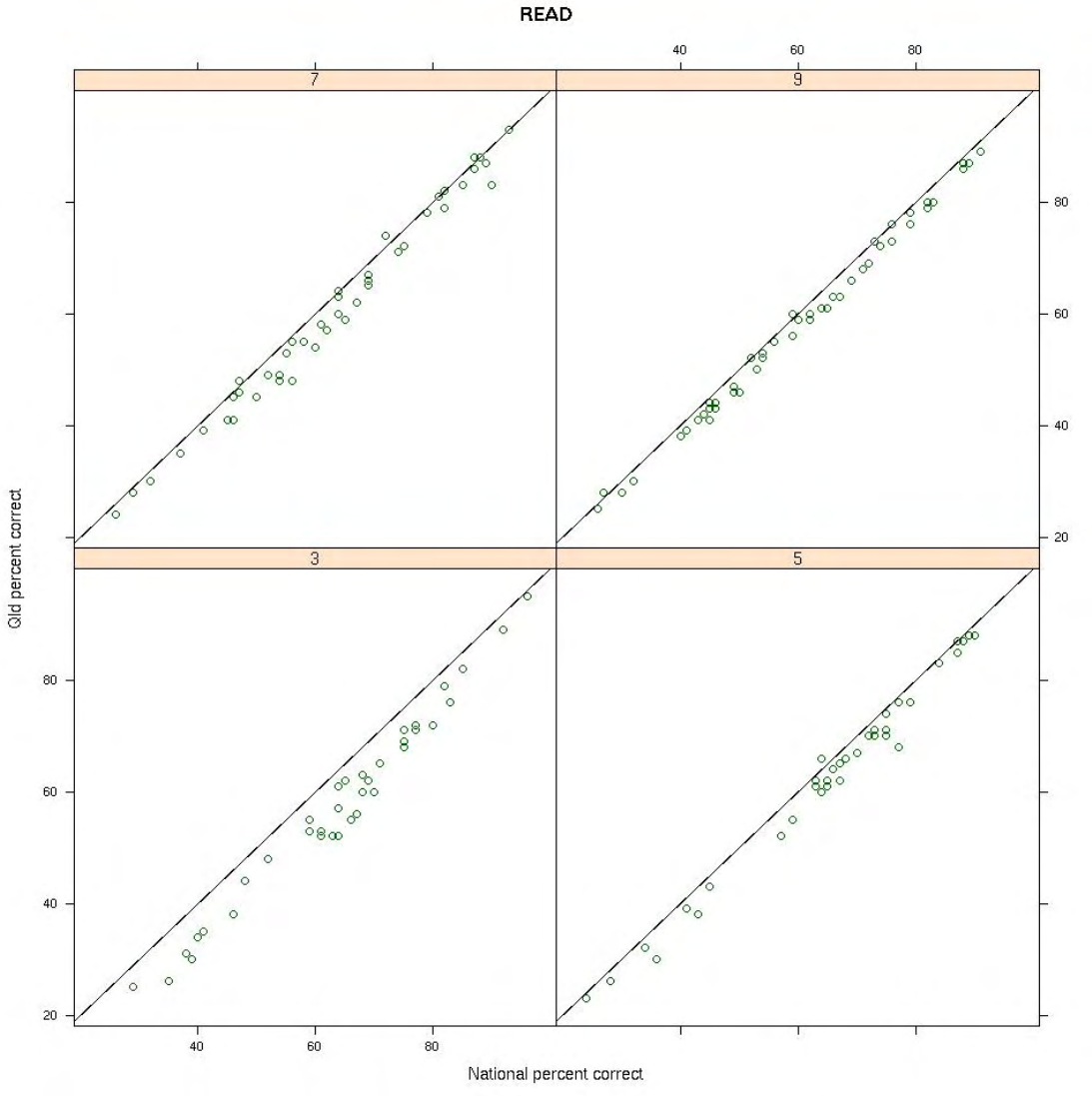
In comparing the results of Queensland students it is possible to map what percentage of Queensland students had particular items correct and then measure if this was more or less than the national performance. Mapping onto a graph how many students provided the correct answer for an item is referred to as the *facility* of the item. High facility rates mean that the majority of students got the item correct and conversely, low facility rates mean that very few students got the item correct.

Comparative results across all the domains tested show a similar pattern: Students at Year 3 were considerably behind the national mean, students at Year 5 were closer to the mean and Year 7 students were closer again to national means. Students at the Year 9 level were generally at about the same point relative to national means as the Year 7 students, which in most cases was at about the national mean.

This pattern was also evident in the performance on individual items. Rather than being closer to the national mean on some items and further away on others, students generally performed at a consistent level below the national mean. This indicates a general problem rather than specific curriculum effects for specific items; that is, when students did less successfully, they did less successfully on all items.

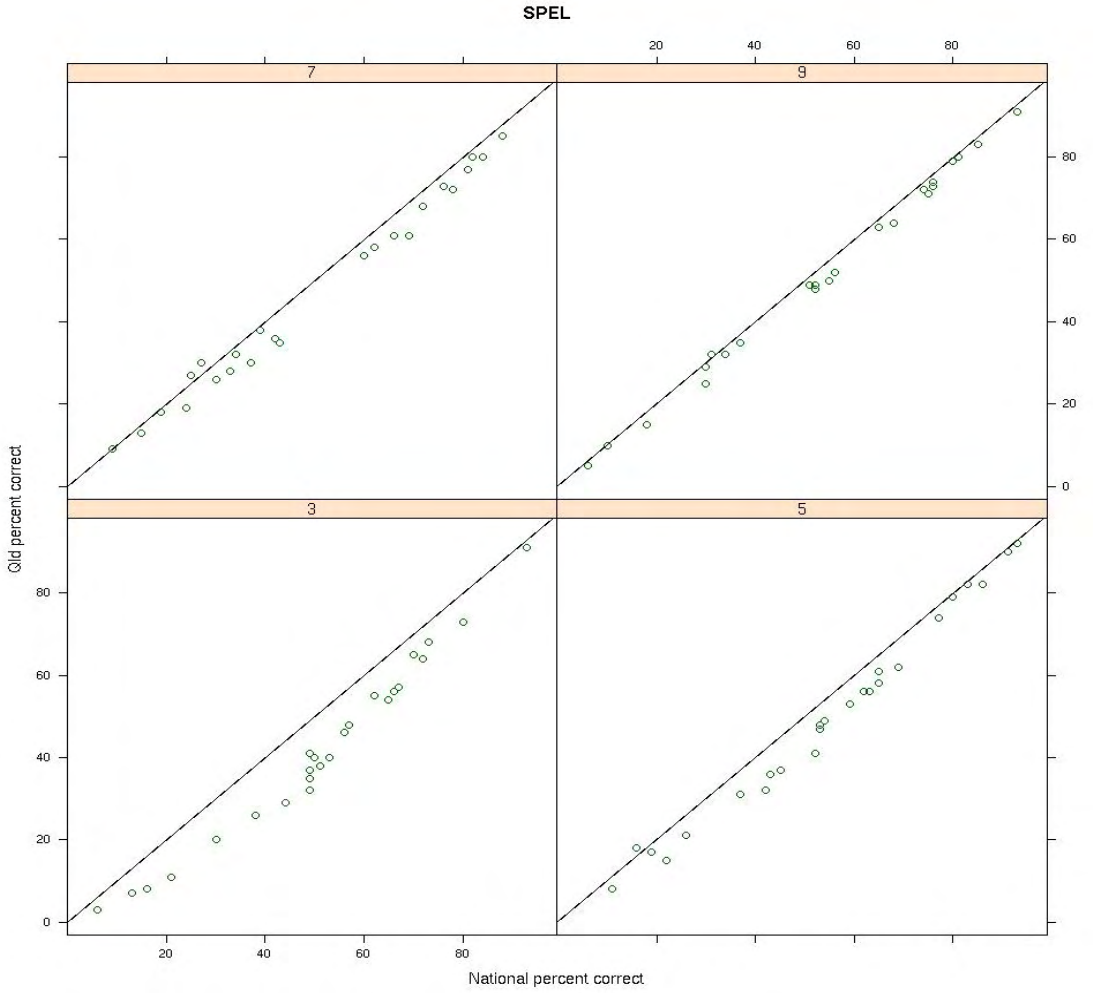
When examining facility rates on items, the expectation is that differences between Queensland students and the national mean should be smaller for items rated as either very difficult or very easy. That is, if 95% of students have an item right or wrong, the difference in facility rate will be little different from one jurisdiction to the next. The greatest differences are expected in items in the middle of the facility range. This is evident in the following graphs which, when comparing national and state facilities, show that at each end of the range Queensland students performed close to the national average but in the middle of the range students were a considerable distance below the national average.

Reading



Reading showed different patterns from other strands, in that the Year 9 results were closer to the national percentage correct, whereas in Years 3, 5 and 7 they were further away. Year 9 students performed closest to the national mean, with the Year 5 cohort the next closest.

Spelling



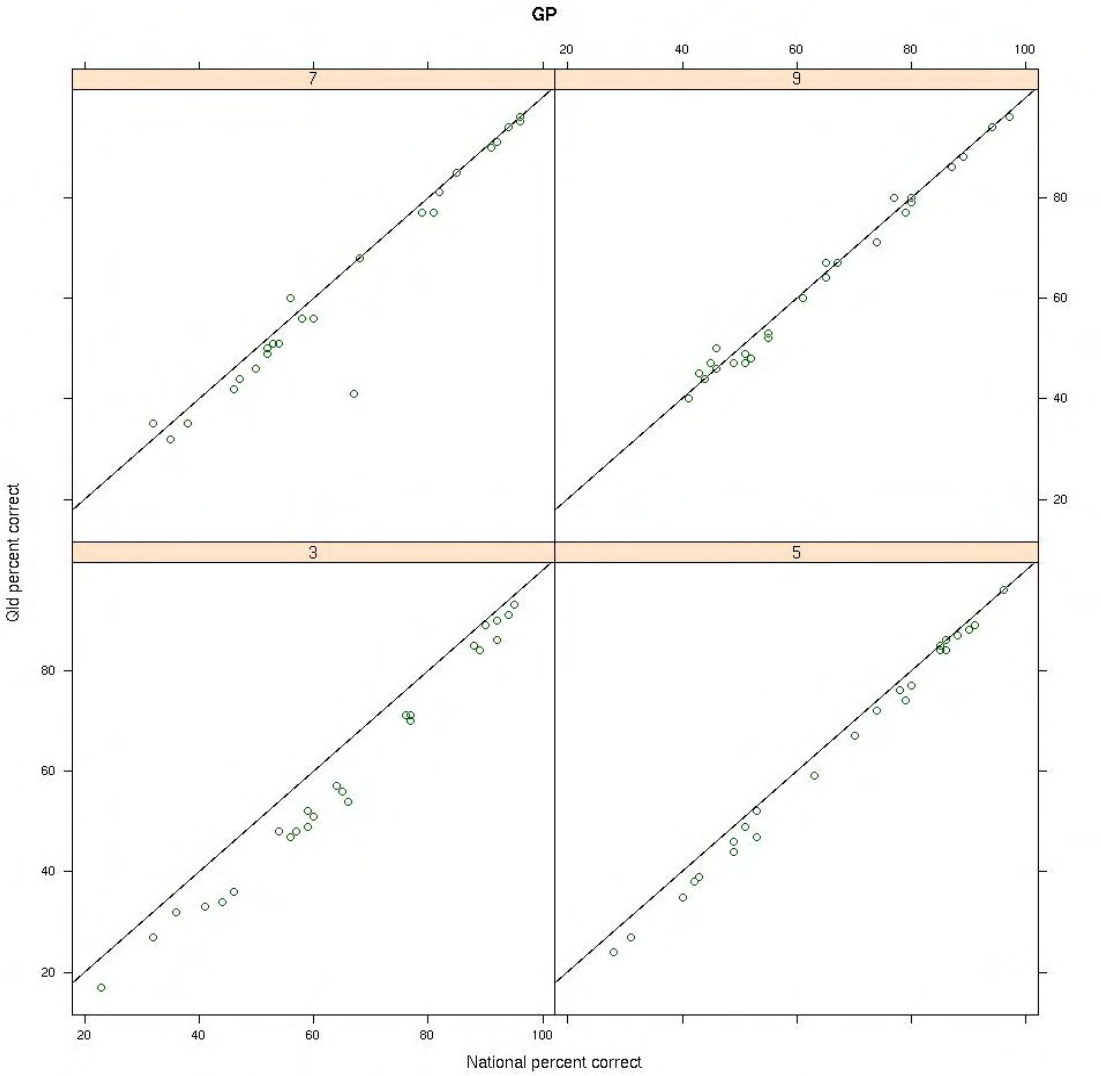
In spelling, Year 3 students demonstrated the most obvious differences in performance. Of the 18 instances where the difference in the facility rate between Queensland students and the national average was greater than 10%, eight occurred in Year 3 spelling. It is interesting to note that Year 7 and Year 9 students were both closer to the national averages than Year 3 and Year 5 students.

While in previous years, the Queensland tests had had multiple measures of spelling which measured both the recognition and production dimensions of spelling, the NAPLAN spelling test had only one — a form of recognition. Rather than spell a word that was presented to them orally, students had to respond to a misspelling. Teachers observed that the spelling test was different from the previous Queensland testing as students had to read as well as focus on spelling. Because the measure was dependent on reading, and the reading levels of the texts was often above that of the year level, students who were good spellers but only satisfactory readers were disadvantaged. They had to read the text, identify the incorrect word and then write it correctly. The measure is confounded by involving more than one process.

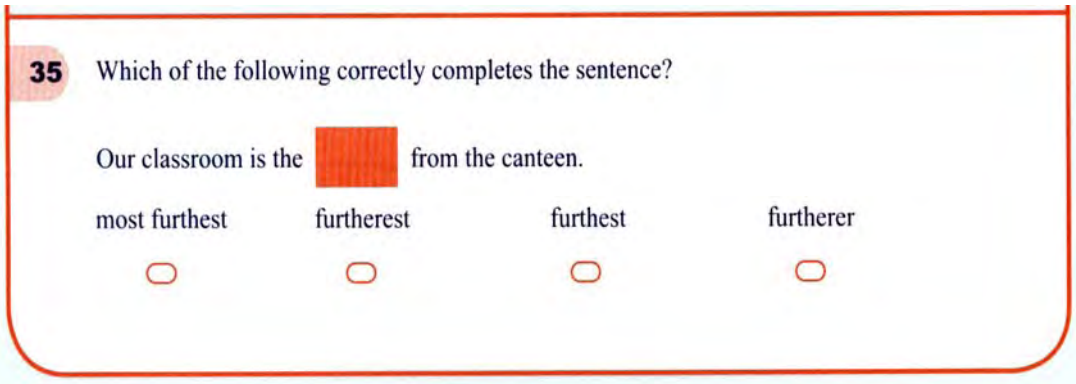
After the test, some teachers were asked to re-administer the spelling words to students but to do so as a dictation rather than proofreading. The results indicated that students could spell more of the words using this technique. Even though there had been consensus during the test construction period that students in Years 3 and 5 would be given the words orally, this decision was overridden before the test was finalised. This difference in performance has an obvious implication for the way in which the test is developed and administered.

Anecdotal discussion and observation by Queensland officers suggest that there were differences in the way in which spelling variations were treated by different jurisdictions during marking. Officially only one variation was allowed but other variations were seen to be credited.

Grammar and Punctuation



Grammar and punctuation is the only strand where one item shows significant difference for Queensland from the national results. Question 35 from the Year 7 Language Conventions test showed the most difference of any item.



Students identified that the superlative ending “est” was required and circled the first one they came to, which was, however, incorrect. In some communities the word “furtherest” is commonly heard as a dialectic variation. This item, along with others, tests membership of one class or another where language usage and pronunciation is less formal.

At Year 9, the relative results were closest to the national average. The Year 9 test can also be seen to be relatively easy, with almost all state and national facilities over 40%.

Writing

It is not possible to do an item analysis for the writing task, but it is evident again that Queensland students did not perform as well as the national average and in some cases were significantly below some states. There is a suggestion that this may have been a result of a variable application of the marking rubric. While similar in approach to the Queensland rubric, it had a different emphasis from that used in Queensland. This rubric favoured performance in the surface features of grammar, spelling and punctuation over the ability of scripts to respond to the stimulus and for students to produce a cohesive and engaging text.

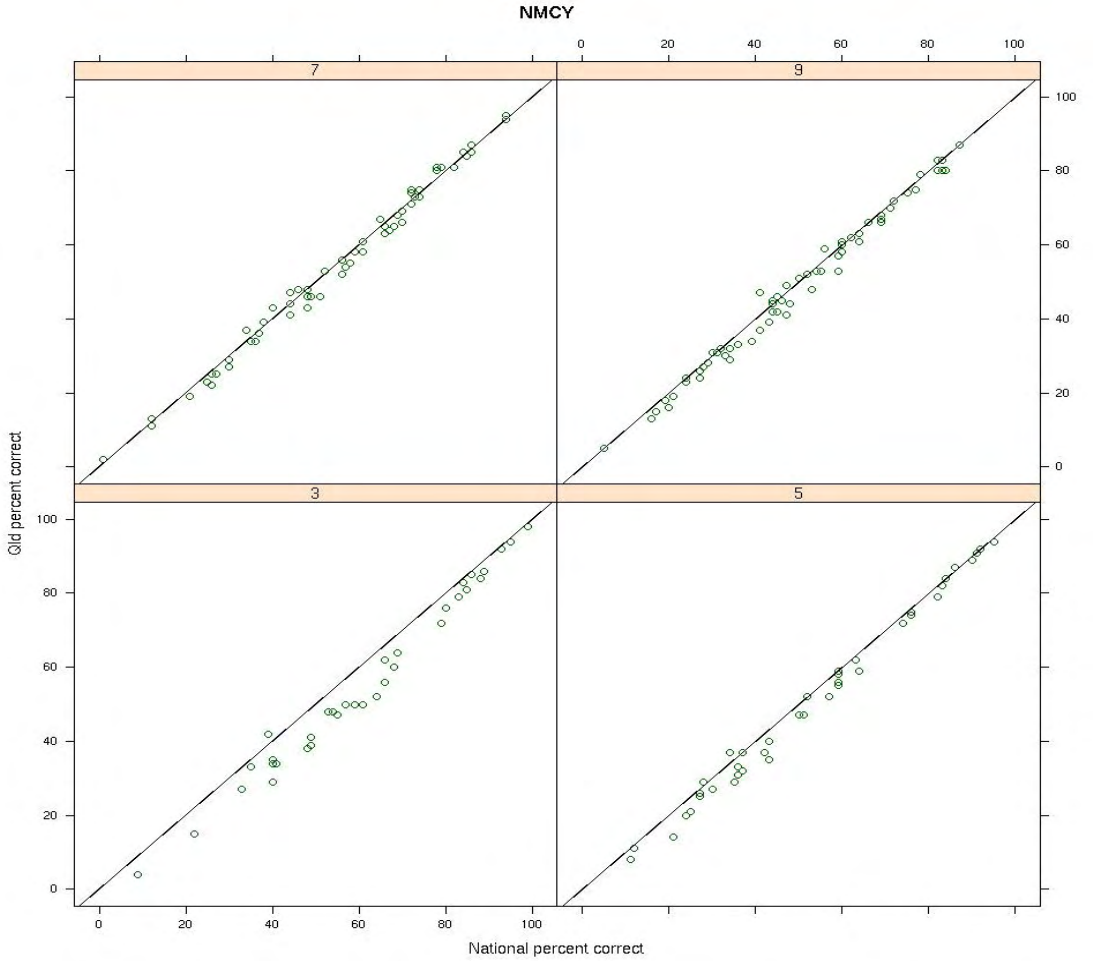
Although Queensland students did not perform well on the writing task, Queensland representatives who were involved in national selection of exemplar scripts did not believe that the work they saw of students in other states was necessarily of a higher standard.

In each session all markers mark the same script to check that they are still marking in accordance with the rubric. These scripts, presented randomly to the marker pool, are called control scripts. During the marking operation, data was available to jurisdictions about performance on common scripts that were marked by all markers for control purposes. Each jurisdiction received information about their data as well as the overall national data on that item.

The agreement between Queensland markers and the agreed marks on the control scripts was good. In addition, on any one script there was approximately the same number of markers who marked higher than the agreed grade on that script as those who marked below. However, evidence from the national data showed that nationally more markers marked the control scripts easier rather than harder. If this also occurred in individual jurisdictions, during regular marking, this would create an overall negative effect on Queensland's results. No information is available to Queensland regarding which jurisdictions were responsible for the overall higher marking nationally.

There is evidence in the data that students in New South Wales and Victoria performed significantly better in writing than their results in reading, spelling and grammar would predict. Without double marking of student scripts to provide greater inter-marker reliability, the control scripts perform an important role in maintaining the accuracy and quality of the marking operation so there should be greater use of these to ensure that all markers apply the marking scheme in the same way. National processes for responding to identified discrepancies in marking approaches need to be developed.

Numeracy



The same pattern appeared in Numeracy, with students being behind national percentages at the younger year levels, but getting closer to the national average at older year levels, with perhaps the strongest comparative performances at Year 7 and Year 9.

There was concern that the readability level of some of the numeracy items was beyond the year level being tested. The readability of the test papers, according to the Flesch-Kincaid Readability scale, is summarised as follows:

- 18 items above grade reading level on the Year 3 Numeracy test
- 12 items above grade reading level on the Year 5 Numeracy test
- 1 item above grade reading level on the Year 7 Numeracy tests
- 5 items above grade reading level on the Year 9 Numeracy tests.

Administration of the tests

The tests took place over three days. Each student in Years 3 and 5 had four booklets to complete while students in Years 7 and 9 had five booklets each. With five separate booklets, matching the information for individual students was particularly challenging. Often teachers did not realise the critical importance of being accurate when completing the covers or the significance of the overprinted covers. In spite of overprinting booklets with the names of students there were instances where not all of the booklets could be matched to provide a reliable result for each student. These problems were a result of:

- students being marked absent when in fact they were present and had filled in the book
- students being marked as present when in fact they were absent
- students being marked exempt but still completing the test
- incorrect names — slightly different names when students filled them in by hand.

The result of the inaccuracies can and frequently did impact adversely on the data.

Empty booklets

After considering handwritten and pre-printed booklets, we calculated that some students had over 12 booklets attributed to them in the data. Inevitably, with over one million testbooks to match, a small percentage could not be matched or processed correctly.

One of the most obvious administration problems that had an adverse effect on schools occurred when students were absent, withdrawn or no longer at a school, yet the school returned empty test booklets with no indication that this was the case. In many of these instances, the students were given zero on every item. This could have affected the results adversely, as one student would have appeared as two separate students in the data, with incomplete booklets, papers with zero scores if a matching book could be found, and the student marked absent if no matching book could be found.

In the analysis by the national contractor, students who were absent were still included in the estimates of parameters such as jurisdictional means, but with a score imputed from their other results. Because students with mismatched booklets tended to be, on average, lower achieving students, this effectively resulted in their lower performance being counted twice. That would have slightly lowered estimates of the mean for Queensland. Since the submission of data for the national release, the QSA has been able to match many student booklets that were previously unmatched and for which two records, or more, had been created. This has had the effect of decreasing the overall population in each cohort by at least 1000 students, and resulted in a slightly increased mean. There are more such students who still exist in the data.

Other jurisdictions had more resources to identify and follow up potential mismatches with schools, so this was a source of relative difference between jurisdictions. Familiarity with procedures and understanding of the consequences of this, as well as the potential to follow up some of these problems when they are identified, may result in fewer problems in future years.

Construction of the test

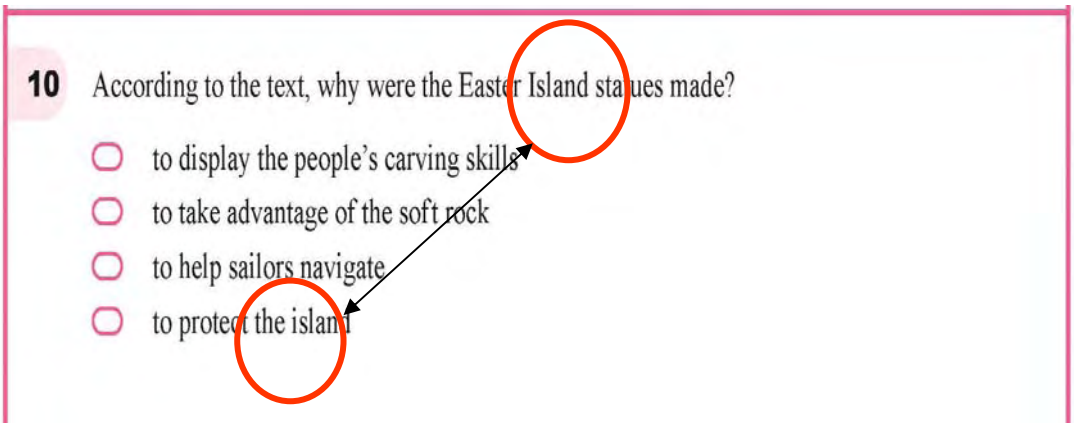
The tests have been developed by Curriculum Corporation on behalf of Commonwealth and State ministers. As this is the first year, the development processes have been those of consultation and consensus rather than of applying the best practices of accepted test development. While it is understandable that development should happen this way, the resulting decisions have had a marked and often detrimental effect on the outcomes.

Familiarity

The national tests very closely resembled the tests administered in some other states. They are most like the tests administered in New South Wales.

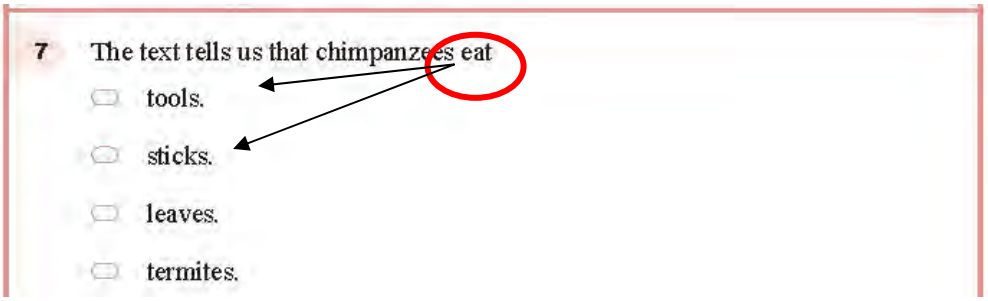
This meant that students in other states would have had considerable practice in answering particular questions, with a focus on how to answer "tricky" questions. This is the kind of knowledge teachers in other states would have had experience with, and thus taught. While this is an obvious response to the situation, it is not desirable test practice as it narrows the curriculum.

In the following reading question students from other jurisdictions would have been trained in reading comprehension to look for patterns. The primary pattern involves matching an item word with a text passage word. Another technique would involve matching the words in the question and the correct response. This is an example of familiarity with the way in which the test works but is not necessarily a good question to judge reading comprehension. This type of item would not be familiar to Queensland students as pattern reading was previously avoided in test construction.



A strategy they would be taught is to look for words that look similar in the text and the question. This compromises the validity of many items.

Another reading question has a similar flaw in that there are grammatical links between the stem and the key while at least two of the distracters are implausible and non-performing.



Items with non-working distracters function as true/false items and contribute to the tests differently. More seriously, they do not comply with the Rasch model which requires the distracters to discriminate between the groups. As these items do not have distracters that tap into plausible errors associated with the construct being assessed, they provide little if any information about student learning or teaching practice.

In spelling, the items are written to a readily identifiable formula.

Drop a letter — *compeated*
Add a letter — *vanila, swiming*
Reverse letters — *muscel*
Trade a letter/s — *butten, taiste*

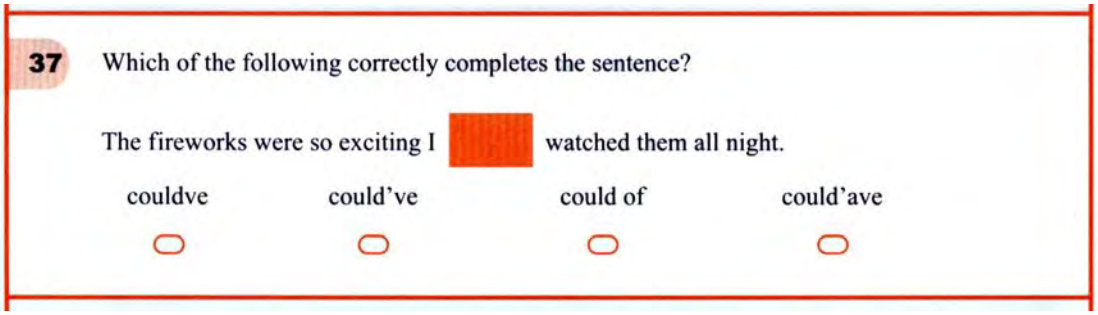
Because this pattern is idiosyncratic to the tests, students from the states where these kinds of items had been used in the past would have been schooled in how to change spelling errors.

The impact of the bad item construction is that the practice effect has an undue influence on the way students interact with the test. Because the tests are badly constructed, the practice effect becomes more significant as students need to know what is meant rather than what is said.

Not only do flawed items increase the impact of a “practice effect” on the results, but they also encourage teachers to spend valuable teaching time focusing on the worst possible aspects of test-wiseness.

Curriculum coverage

Despite the public perception, the best use of the data is to inform teachers and systems about student learning. As the reliability of class, school and cohort data is strongest, the tests do have the potential to operate as an effective tool to evaluate different approaches to curriculum development and learning. For this to be effective the construct needs to be defined for testing so that the items can be assessed for validity. Failure to match the test items to curricula can have unintended consequences. For example, because some states thought it important to test knowledge about grammar as a separate entity instead of making judgments about student knowledge and use of grammar to create meaning in writing and reading, no agreement could be reached about what was meant by grammar. The consensus position was that usage, rather than the taught curriculum, would be tested. While this has several effects on the construction of items, the major implication was that what was assessed by these items was underlying dialect, often an aspect of social class. The consequences of these decisions for Queensland students can be demonstrated by such questions. In the following item, many Queensland students opted for “could of”, which is phonetically close to what they hear but of course is not correct. Only these two options were attractive to the students, so this too is an example of an item functioning as a true/false item rather than as a multiple choice able to provide curriculum information.


37 Which of the following correctly completes the sentence?

The fireworks were so exciting I _____ watched them all night.

couldve could've could of could'ave

With respect to Numeracy there were concepts tested that are not covered by the *Statements of Learning for Mathematics* (SOLS).

Year 5 students were expected to find two lines of symmetry (no lines drawn) in shapes with colours and patterns whereas the students in Years 7 and 9 were required to recognise symmetry in simple shapes with the line drawn in for them. For Year 5, finding only an obvious line of symmetry is described in the *Statements of Learning for Mathematics* for this age group.

The Year 7 papers tested the same content repeatedly. For example, eight questions relating to the calculation of rates and proportions was an over-representation of this aspect of numeracy. Because of the design of the Queensland curriculum, it is likely our Year 7 students would have had limited experience with this content so early in the year. This means that they may have been disadvantaged up to eight times because of the poor test design.

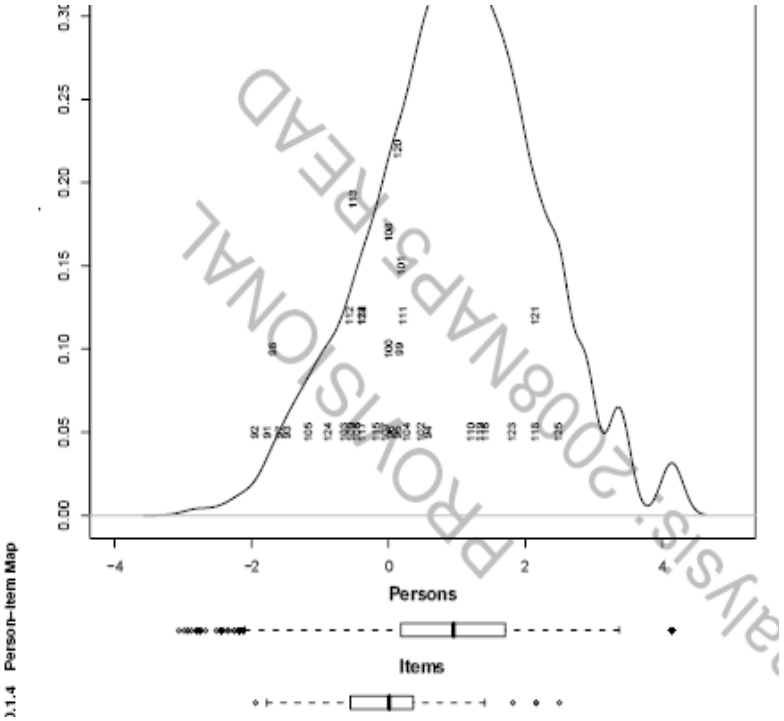
In the Year 9 Numeracy paper, six questions tested the same conceptual understandings of the mathematical sub-strand of Space. These items required students to calculate angles using knowledge of triangles or squares. Therefore, students who were particularly skilled in solving these types of questions were advantaged six times. They could repeatedly draw on the same understandings to answer the questions, thereby skewing their results. Conversely, students who were not as skilled, or who had less experience with this aspect of spatial understanding, were disadvantaged. A better balance of space items is imperative to ensure test validity.

Other areas that were also heavily weighted included Algebra and, in particular, the strategy of substituting a value into an equation. This skill was assessed four times overtly and twice more if students chose to use this strategy to help select a correct equation from an incorrect one. However, students with strong algebraic understandings could not demonstrate higher level understandings as there were no items testing these skills. Those students who were not strong in this area were disadvantaged up to six times. Once again, a better balance of items is needed.

The tests were heavily weighted towards mathematics and could not be considered a balanced test of numeracy skills.

Ability to discriminate across the range of students

Meeting one jurisdiction's needs to have a number of questions easy enough for their students to answer has resulted in a fundamental imbalance in the papers. An extreme demonstration of this effect can be seen in the grammar strand of the Year 7 where six of the 25 questions have facility rates of more than 90%. Questions such as these that are too easy provide little or no information about student performance. Having so many of them reduces the number of questions on which to discriminate the performance of the population. In order to see the relationship between the level of difficulty of the items and the ability of the test-takers, an item–person map can be constructed. Such a map shows the range of the items and the range of the candidates. That the tests had a preponderance of easy items and did not function across the range of students is demonstrated in the item–person map for the Year 5 Reading.



In a well-constructed range test, a close relationship between the bars representing the items and those representing the students would be expected. But in the case of the Year 5 reading test, the relative position of the bars suggests a mismatch between the range of items on the test and the suggested ability of the students.

Range and balance

The range of items did not meet the specifications set down for the test construction. This was true of both item difficulty and the range of curriculum aspects covered.

The lack of difficult items not only fails to provide information on a large proportion of the cohort, it means that no data are collected that can be used to inform teaching and learning. There were very, very, few items that tapped into higher-order comprehension. A quick perusal of the item demands would substantiate this. If data are going to be created from this testing program for scrutiny of schools and systems, then in all fairness the range of items must be present in the tests.

This problem of the range is often exacerbated by the fact that it was not uncommon for the same literacy skill to be tested more than once. This not only restricts the range further, it also means that some students are advantaged and others are disadvantaged. In addition to the grouping mentioned above, the Year 7 Grammar consisted of:

- 4 items dealing with strong verbs
- 10 items dealing with punctuation, 11 if you count the contraction question as a punctuation question rather than a spelling one or even, given the distracters, a test of language usage (an example of trickiness).

The decision on the inclusion or exclusion of terms for the test was made on a consensus basis. This meant that clusters of items were developed around points of agreement only. The result of this was that instead of testing a particular educational concept once, the same concept was tested several times. Because they were testing the same concept, often in the same way, these items performed in much the same way. It can be assumed that only one of the items was needed on the test to provide information. The presence of the others simply reduced the potential for other concepts or levels of performance to be assessed. For students and for states, the effect of this could be severe. The performance of one student, who happened to be strong in this one area, might well be enhanced, while the student who happened not to have been taught this concept would be seriously disadvantaged and may appear to be weaker in literacy or numeracy than they really are.

If the cluster of items assessed a concept not taught at that level by a particular jurisdiction, as was the case in Year 7 Numeracy where ratio was tested six times, the effect upon jurisdictional data could be severe.

Representation of the different sub-strands of mathematics was not spread consistently throughout the paper. For example, of the items identified as being in band 10 of the achievement band scale, 83% were Number and Algebra and no Space items were included. In bands 8 and 9, only 40% of items were from the Measurement and Space sub-strands. While the spread of items in the tests may have matched the specifications overall, the distribution across the bands was unbalanced.

Students with strong computational skills and experience with Algebra were able to demonstrate higher levels of understanding than similarly high performing students with a strong sense of spatial reasoning. More difficult questions that allow students to demonstrate thinking, reasoning and problem solving with content other than Algebra need to be included on the tests.

Areas that were under-represented included financial literacy, distinguishing variation in data collections and solving authentic problems that require students to demonstrate a range of strategies. There were also no items that really tested students' abilities to use calculators to solve problems. That is, there were no items with large numbers that required complex computations.

Engagement

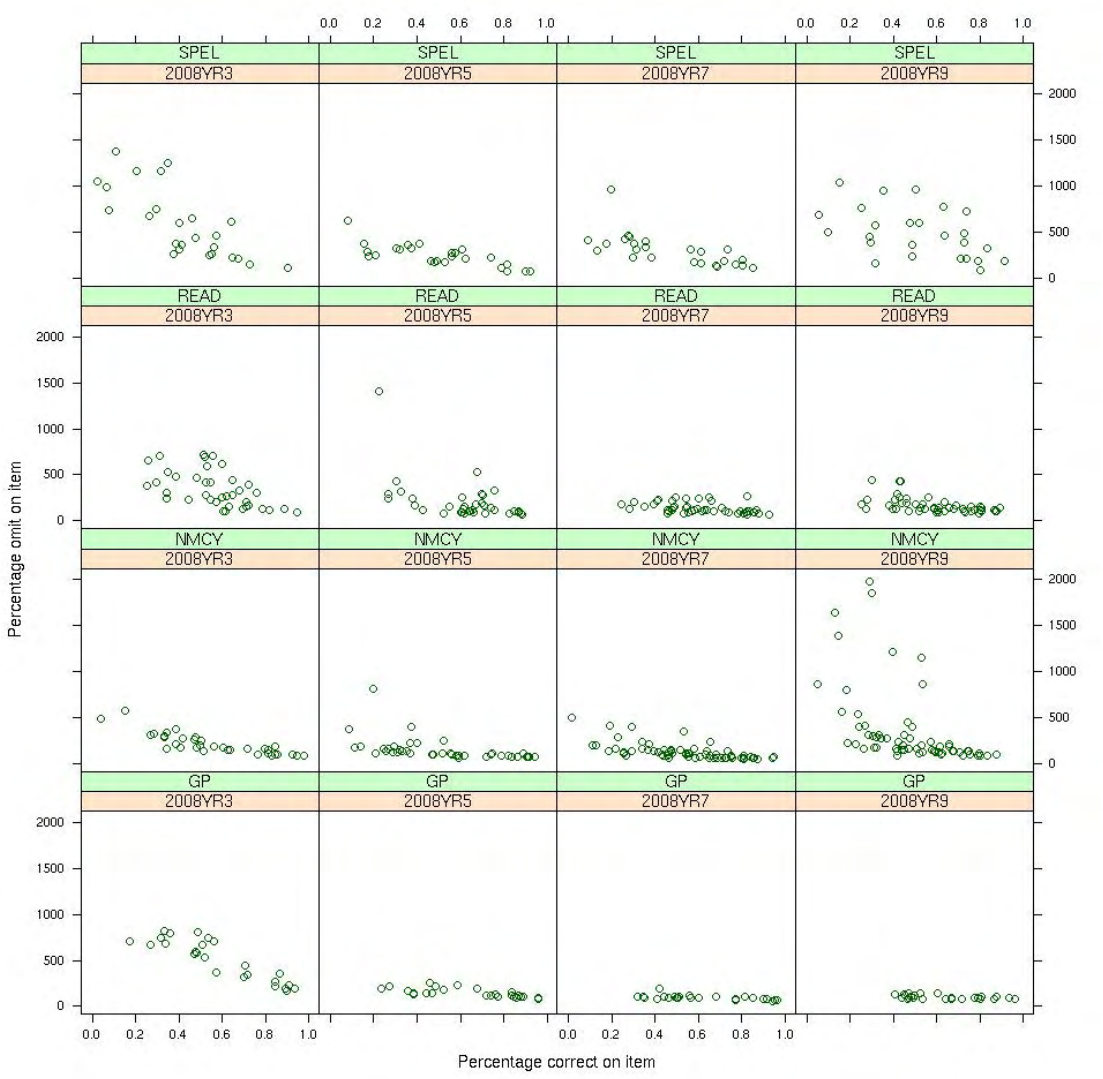
There is evidence that the tests failed to engage Queensland students — a significant percentage of students (30% in some cases) omitted items that were considered to be simple and were not at the end of the papers. There is evidence in spelling that many Year 9 students did not complete any of the open-ended items.

On almost all of the tests, there was a lack of difficult items to challenge good students. In easy tests, if there is low student engagement and students are more likely to omit easy items, this could have had a larger impact of the test scores for Queensland. Without evidence of omit rates from other jurisdictions, it is hard to know if this was a significant reason for Queensland's relatively poor results.

Although the response patterns of other jurisdictions to items is not known at this point, an examination of the omit rates on items within Queensland can indicate important information about student engagement.

Overall, higher omit rates tend to indicate lower engagement. Omit rates usually tend to be naturally higher on harder items, but if student engagement is low students may also not be attempting items that they should be able to get correct, and may omit rather than guess when they do not know the answer, both of which punish these students relative to other groups.

The omit rates are shown below. Each panel contains the items for a year level and a strand. Harder items are towards the left, easier items towards the right. The higher up the panel a data point appears means that more students omitted the items. The overall pattern that should be expected is a line that slopes down towards the right (with easier items having lower omit rates). Generally, the older that students are (that is, in panels to the right), the lower the omit rates should be.



There are some interesting patterns in this data. In Spelling (the panels in the first row), the overall engagement increases from Year 3 to 5 to 7, but at Year 9 not only do omit rates dramatically increase but the students did not systematically omit the harder items. This indicates quite low engagement. It is interesting to contrast Spelling with Grammar and Punctuation (bottom right panel) for Year 9. The omit rates on these items were almost nil, although they were part of the same test paper as the Spelling. It has to be noted that all of the Grammar and Punctuation items on this test were easy. So this relationship shows that students were actually present for the test and the fact that these items appear blank for some students could not be attributed to the fact that they might have been absent for the test.

The pattern is repeated for Reading, but in a less dramatic fashion, with omit rates lowest at Year 7 and slightly higher at Year 9. Predominantly, the questions for Reading ask students to locate a piece of information — 22 of the 38 Year 3 questions are of this kind. The difficulty of these questions often comes from the difficulty of tracking through the text rather than a failure to apply any comprehension knowledge or strategy. This perseverance with the text could account for the lack of engagement by boys in reading at the older year levels.

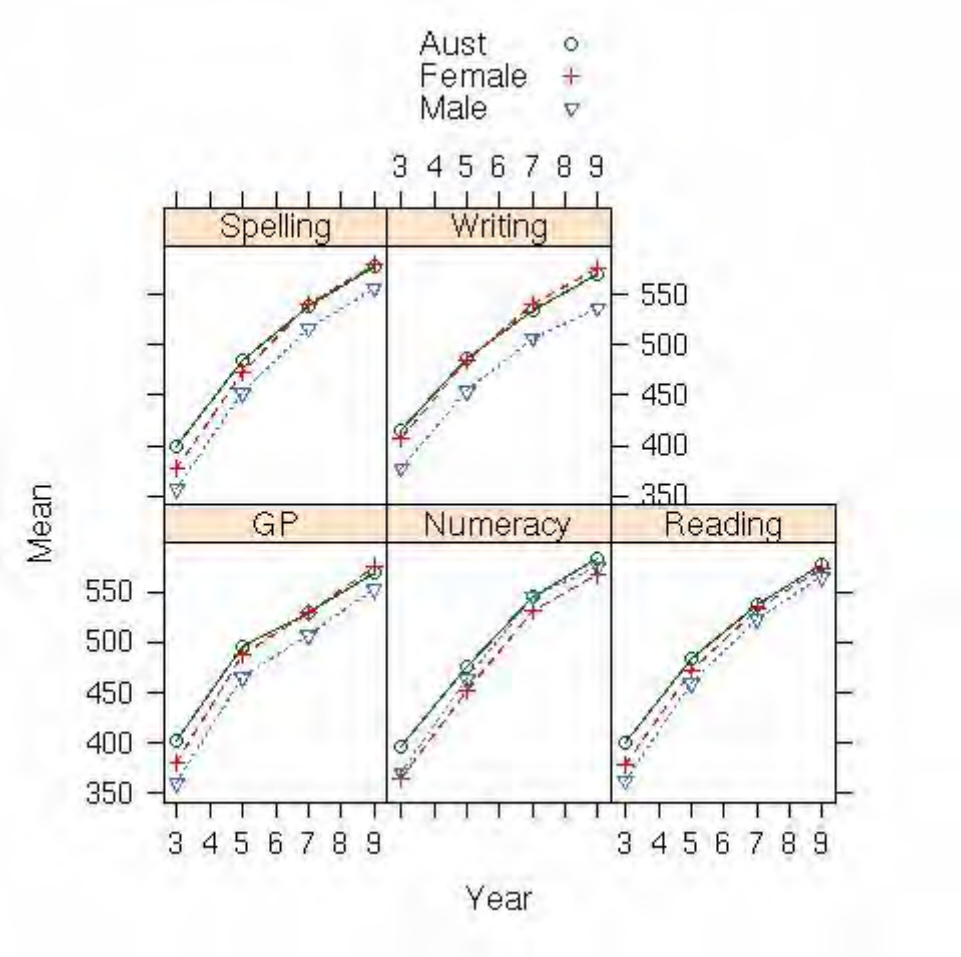
Numeracy has relatively low omit rates at Year 3 compared to other strands, however rates remain about the same in Year 5 and Year 7. Again, at Year 9 the omit rates are dramatically higher. Analysis of these items reveal most of the items with high omit rates are the algebra items. This suggests that there is a mismatch in the timing of curriculum implementation and that many students had not yet encountered this aspect of numeracy. Items with high omit rates are spread across the two test papers, indicating less effect from the length of testing and the use of two test booklets than from the effect of curriculum.

Omit rates are high at Year 3 for the Grammar and Punctuation items, indicating some problem with engagement at this level, but the omit rates were virtually negligible for items at other year levels. In the Year 3 Language Conventions test, much of the omit rate can be attributed to the poor pacing of items through the test. This aspect of test construction is critical in keeping students engaged.

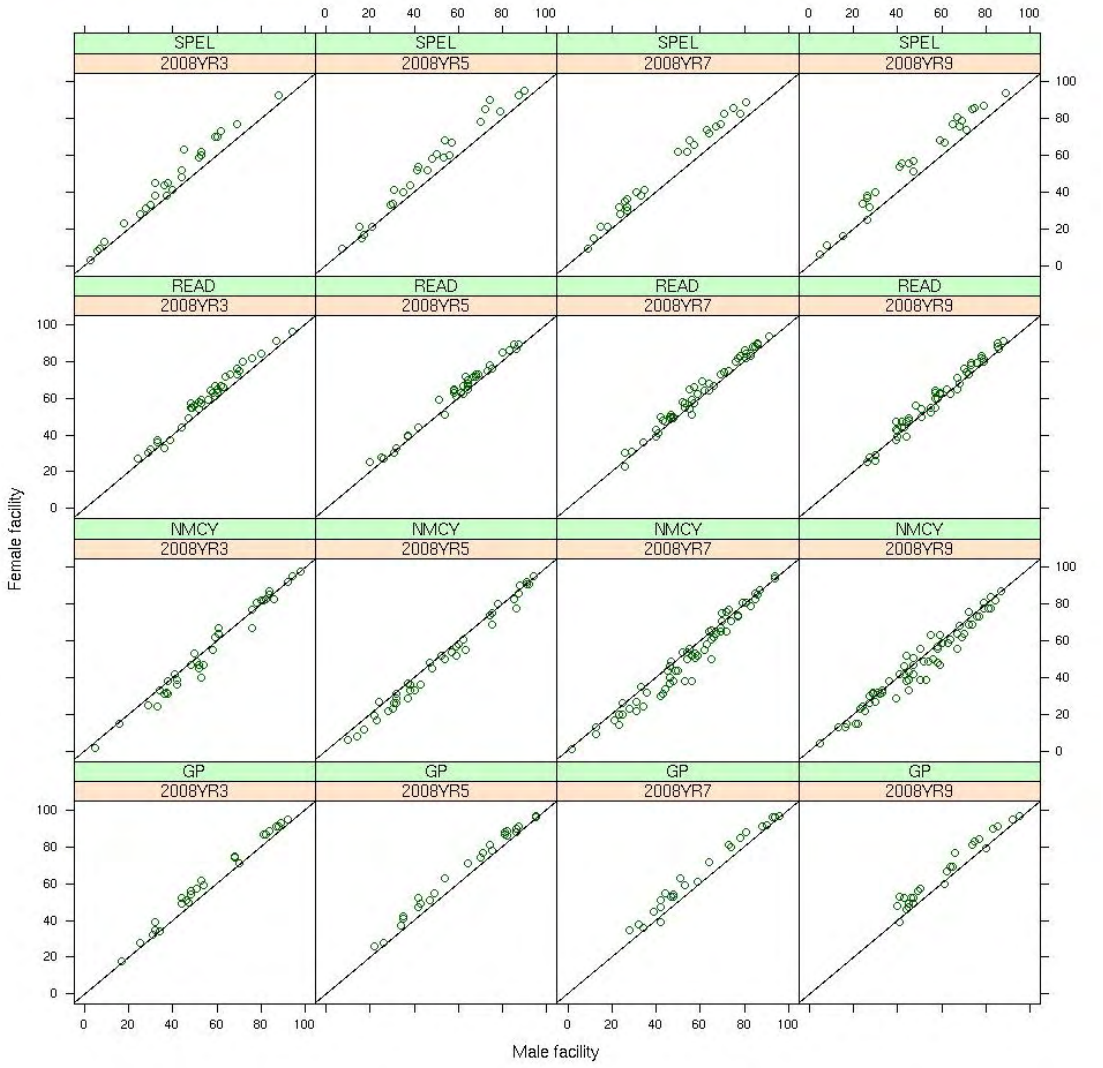
Overall, it is difficult to know what effect engagement has without knowing omit rates for other jurisdictions, but within Queensland this suggests that there are some overall problems with engagement, particularly in Spelling and overall at Year 3 and Year 9.

Gender differences

The following graph shows that in Spelling, Writing and Grammar and Punctuation the girls' results were better than the boys' with the girls generally performing at the national average except at the Year 3 level. The gap in reading is not nearly as significant, while the boys have a slight edge in numeracy overall and are close to the national average at Year 9.



We can expand on this information by looking at the relative performances of boys and girls on individual tests items. Each data point is again an item, and data points above the line indicate girls did better on items, data points below the line indicate that boys did better on items.



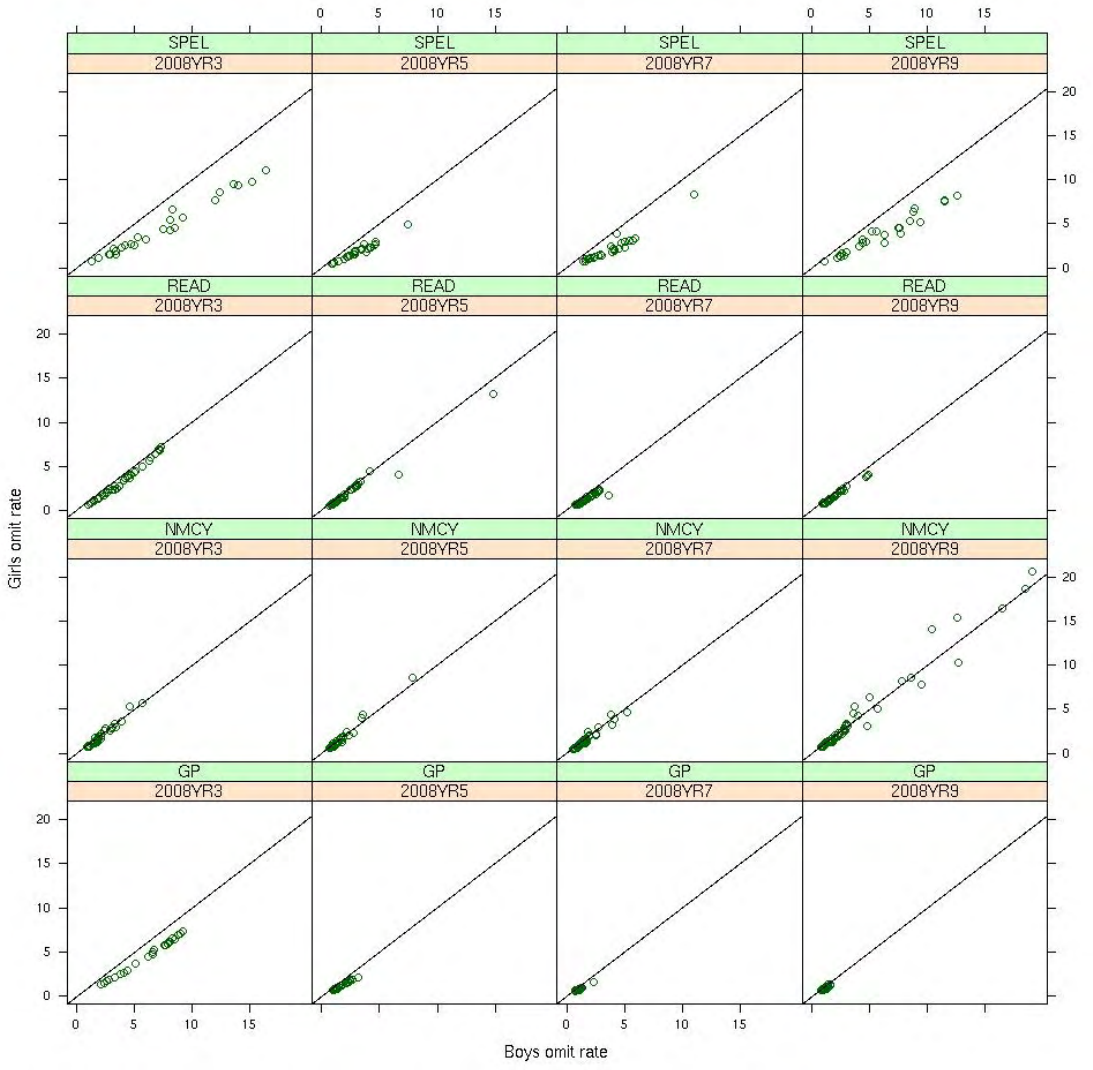
Spelling shows an overall marked difference between boys and girls, with girls performing better across almost all items.

In Reading, the performance of girls was significantly better at the Year 3 level, but the difference was smaller in higher year levels, and was quite small by Year 9.

In Numeracy, boys performed better overall at all year levels, but in Year 9 there are a number of items where girls performed better.

In Grammar and Punctuation, girls outperformed boys consistently at all year levels.

The following graph shows relative omit rates on items for boys and girls. Items that are below the line have higher omit rates for boys.

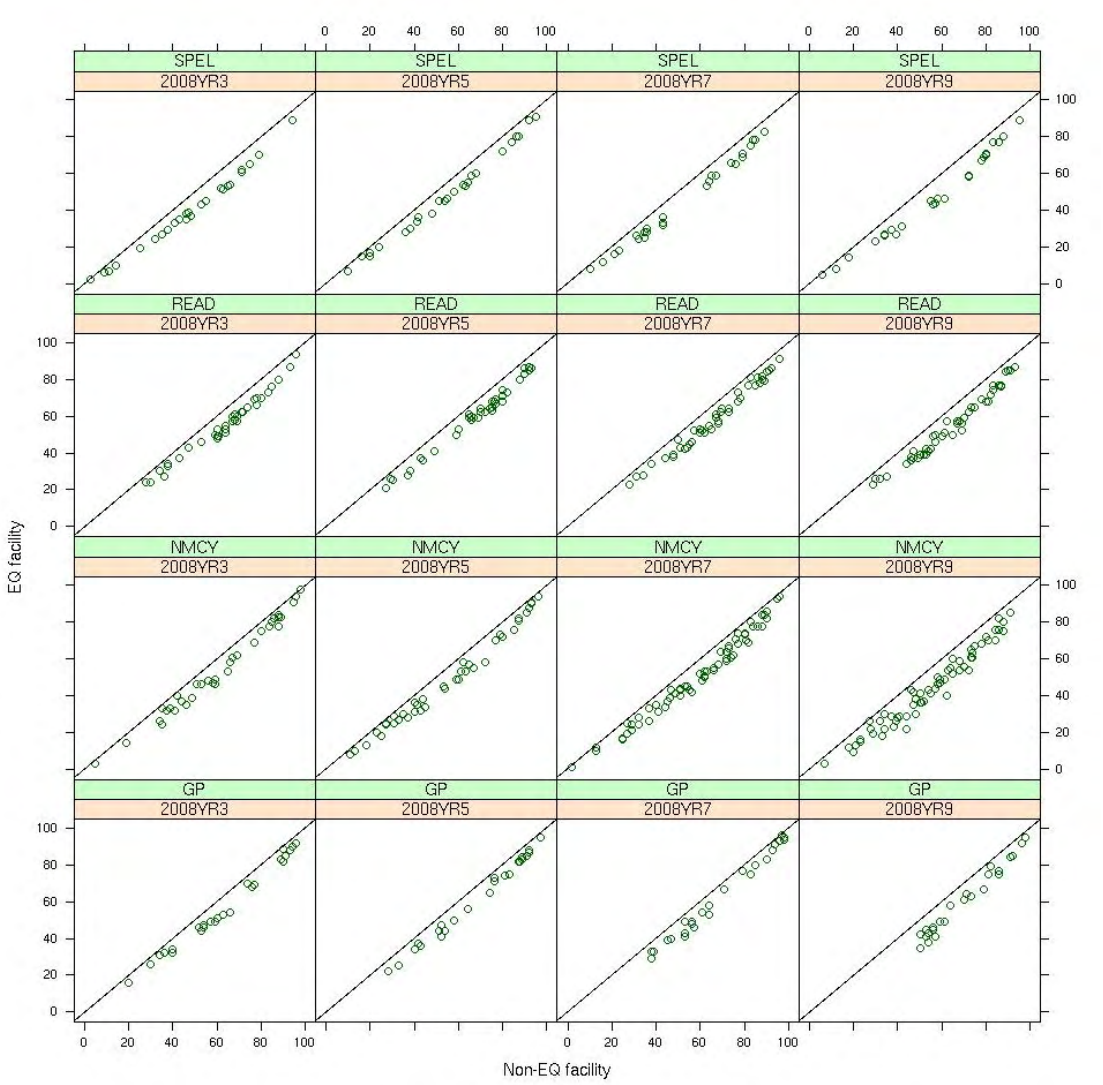


This shows that there are very different patterns in omit rates between the strands. In Spelling, the omit rate (and by inference, lack of engagement) is much higher for boys regardless of year level or overall omit rate. In Reading and Numeracy, on the other hand, omit rates are very similar between boys and girls at all year levels. In Grammar and Punctuation there were only significant omit rates at Year 3, and these tended to be more the boys than the girls.

When read in conjunction with the overall omit rates, this indicates fairly clearly that the engagement problems identified with Spelling, and at Year 9 particularly, affected boys most profoundly. The other areas of relative lack of engagement (Year 9 Numeracy, Year 3 Reading, and to an extent Year 3 Grammar and Punctuation) had less gender difference. This may suggest some focus for test-taking strategies that might have greatest influence in future years.

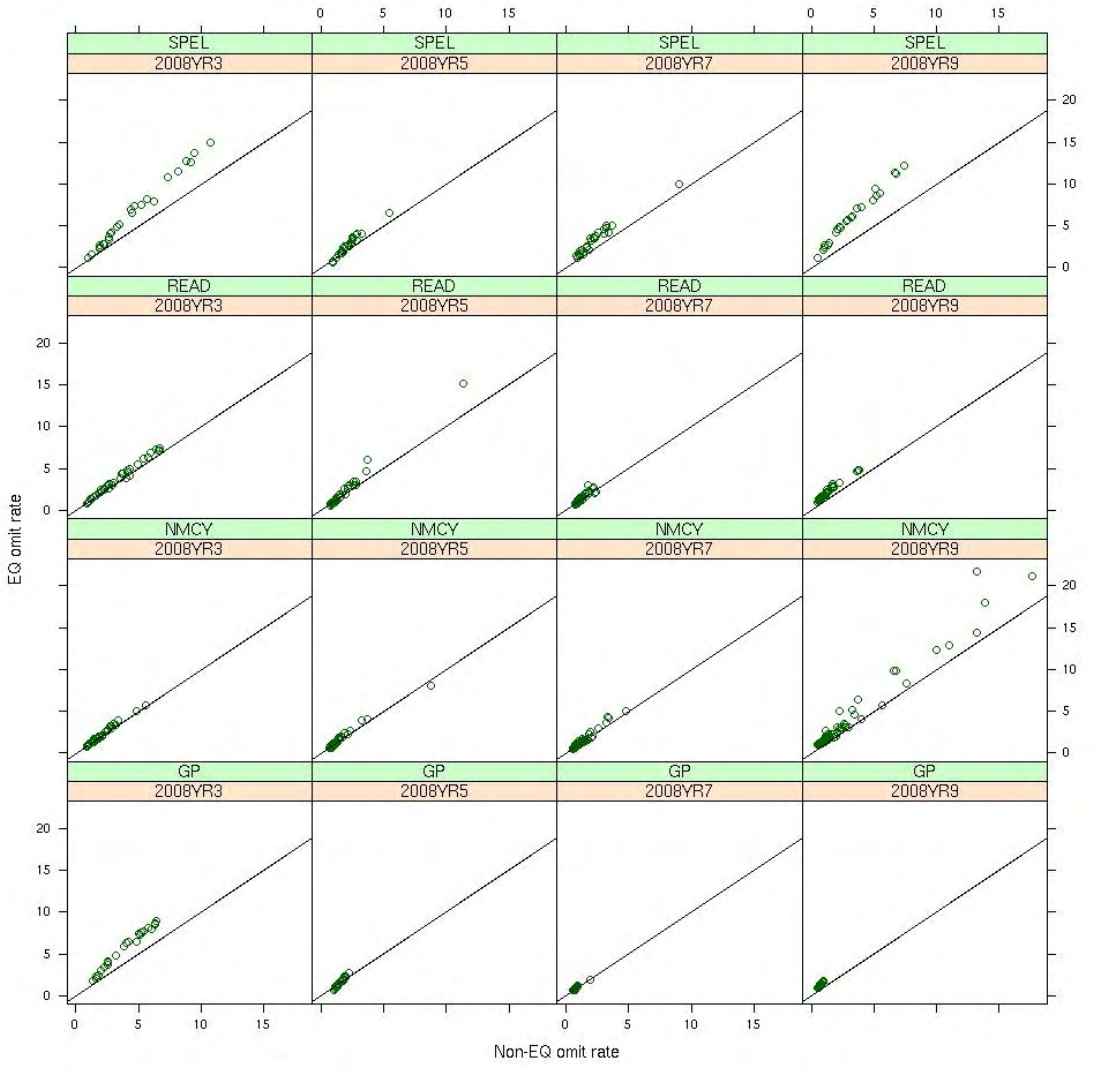
Sectors

The following graph shows relative facility rates for government and non-government schools on test items.



Overall, data points are below the line, indicating government school students performed at a lower level than their non-government school counterparts. The pattern is relatively consistent, with little difference at any year level or in any strand. The greatest difference is perhaps in numeracy at Year 9, with overall lower performance in government schools. Generally, the differences are slightly larger at Year 9, possibly reflecting the more selective nature of private schools at the high school level rather than any educational effect.

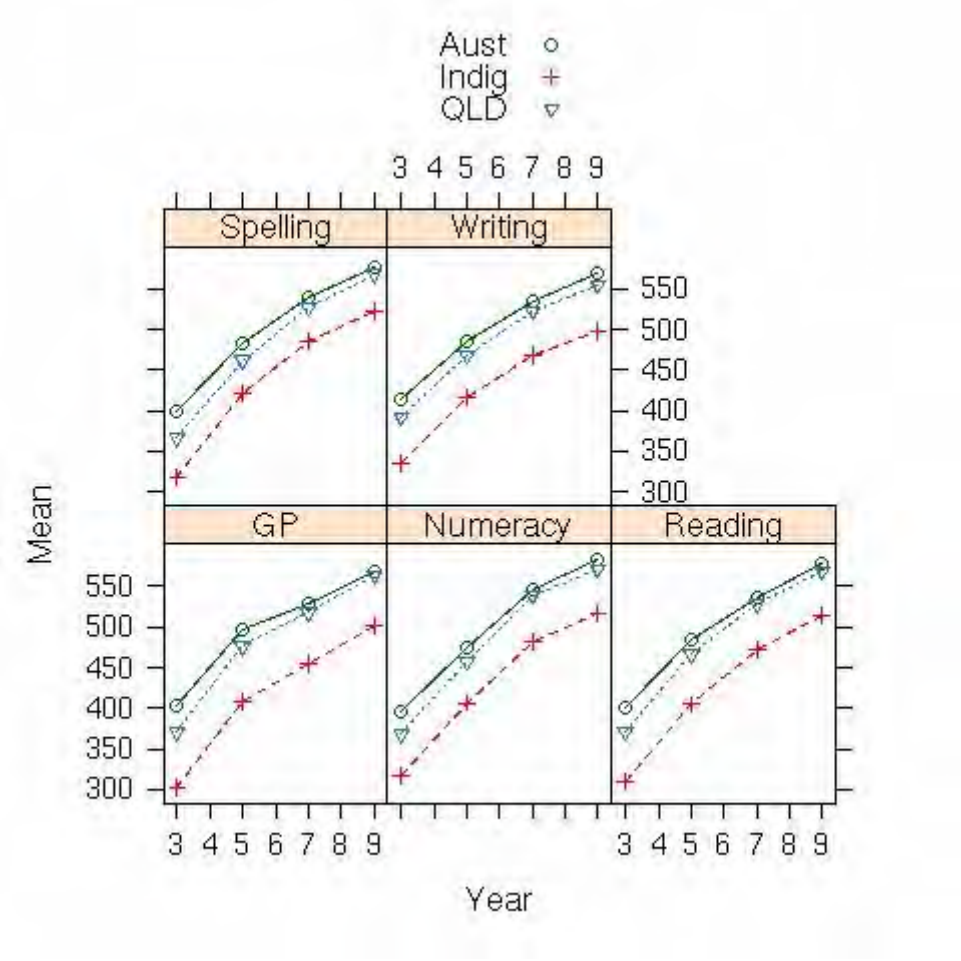
The following graph compares omit rates on items between government and non-government schools.



It can be seen that the omit rates are generally higher across strands and year levels, indicating overall lower engagement at government schools. The difference is particularly marked for Spelling. For reading at Year 3, the differences are relatively small, and they are also less marked for Year 9 numeracy than perhaps expected. If we hypothesise that overall lower engagement of students at government schools explains most of the differences, this means that the relative omit rates on the Year 3 reading test are particularly interesting. This shows no relative differences in omit rates by gender or by government/non-government, which suggests that differences seen here are not a function of overall engagement but are rather a particular feature of that test.

A&TSI students

The performance of Queensland Aboriginal and Torres Strait Islander students is compared with the rest of the Queensland students and the national cohort in the following graphs.



There is no more information to be gained by looking at the detailed item analysis, because the same pattern emerges.

Overall, this shows relative performance of A&TSI students to be well below that of non-A&TSI students, but with little differences between year levels or strands. There is perhaps a larger difference in performance for younger students, probably indicating engagement problems for younger A&TSI students — with school generally and particularly with the tests.

Strategies and advice

In attempting to analyse the test results we are hampered by the number of variables that impact on performance. In the preceding analysis we have tried to concentrate on possible explanations for the apparent poor performance of Queensland students. Two issues that seem important are the results as measured against national minimum standards and average performance compared to the average performance of students in other states.

References to national minimum standards are credible when there has been a definite attempt to use test items that deliberately assess the ability of students to carry out particular aspects of the curriculum. The calculation of the performance levels, and in particular the national minimum standard, has been a numeric cut-off on raw scores that automatically ensures that a certain percentage of students will fall below the mark. There can be no guarantee that the questions that these students got wrong were actually testing the skills described as being the benchmark required for ongoing learning. We need to lobby very strongly for standards to be accurately described and tested so that when statements are made about students' skills they are truthful and reliable indicators of a lack in certain areas of the curriculum.

While there is a simplistic notion that if all students sat the same test the results would represent the same for all students, this is not necessarily the case. We have found that results can be influenced by familiarity with the test type, variations in marking of answers and students' access to aspects of the curriculum. We need to ensure that Queensland students have the test-taking skills that allow them to produce their best performance. There need to be stringent controls on marking operations so that there is more confidence in the grades given, particularly for the writing task.

Response to the test results needs to be both a short-term dealing with aspects of curriculum and test-wiseness that will have an immediate effect and long-term systemic reform.

System response

The greatest influence we can have however, is in determining that students are taught the appropriate curriculum. It is expected that the provision of a National Curriculum on which national literacy and numeracy tests will be soundly based will ensure that all students are given the opportunity to learn what is required. This means we need a rigorous curriculum, excellent teaching and a test which reliably assesses the curriculum.

In making comparisons between the state averages we must acknowledge that there will always be those above and below average. It is self-evident that the ACT because of its population composition could expect to be above average. There is no reason however, why Queensland should be lower than Western Australia, Tasmania or South Australia. There are areas around engagement of boys with the test, curriculum coverage and the performance of rural students that need to be attended to.

Messages for schools from NAPLAN

School climate

The research into high performing schools identifies the key role professional leadership plays in successful schools. These include:

- a clear school mission
- effective instructional leadership and practices
- high expectations
- a safe, orderly, and positive environment
- ongoing curriculum improvement
- maximum use of instructional time
- frequent monitoring of student progress.

Success is planned for in a school where values are respected and upheld. It is important that all members of the school community are working towards the same goals with practices in place that support performance. Every child in the school needs the attention of their teachers so that self-esteem and high expectations are promoted. All participants need to show a passion for learning and achieving.

Role of the Principal

It is time for the Principal to resume the curriculum leadership in a school. The Principal needs to be at the centre of the curriculum planning and implementation. However, this is not something that can be done in isolation. Principals need to learn from each other. Their networks should promote collaboration and mutual learning, not competition. Programs for intervention need to be coordinated so that a child is not being dragged in different directions because of the need to comply with a program.

Curriculum

Too often the curriculum agenda is distracted by identified initiatives that erode the time for true learning. If we analyse the content of staff meetings and school newsletters it will be evident where the energies of the school are focused. This focus needs to be on the curriculum. Coordination of the curriculum needs to be organised so that there is seamless provision of learning for all students. The quality of the curriculum on offer needs to reflect the priorities that children need to learn.

Classes

There is considerable debate around the need to stream classes. Some teachers will argue that it is the only way to cater for all levels of learners. Streaming could occur by ability irrespective of age. Others will offer stories of despair where the early labelling of children has the result of reducing performance outcomes for students who are deemed to be in the lower ability groups. Another strong debate is around class size. While some would promote a notion of no more than 15 students per class group others would argue that class sizes can be bigger if there is sufficient aide time to ensure each child receives the individual attention required.

What is not an option is an assumption that one model will fit all and that it is enough to teach to the middle level of achievement.

Role of the teacher

Literacy and Numeracy are not the sole responsibility of the teachers in Years 3, 5 and 7. Nor are they the responsibility of only the English and Mathematics teachers for Year 9 students. The school climate and the leadership of the principal should encourage teachers of all subjects and all levels to engage with the need to promote literacy and numeracy for all students.

Timetable

To achieve outcomes teachers need quality teaching time with limited interruptions. The distractions of innumerable initiatives and the daily routine of schools need to be minimised to provide a focus on learning.

Early intervention

It has been suggested that schools should not get any surprises when the data from the tests are provided. Even by the beginning of Year 3 teachers should have identified students who need attention or who are having difficulty. Dialogues between teachers and collegial planning should ensure that curriculum programs are capable of catering for the needs of all students.

Professional development

Teachers need time and space to reflect on their practice and to have role models that will help them provide the best pedagogy for the particular children in their care. Additional studies by teachers need to be recognised and rewarded. Any professional development needs to be carefully planned, based on research and aimed at providing teachers with the skills to improve performance. However, the messages in professional development need to be aligned so that teachers are not confused by conflicting messages.

There needs to be professional development around the interpretation of test data so teachers understand the implications and what they can learn about the areas of learning that are not being achieved by their students. A significant aspect of data analysis is knowing what the next step is in terms of the skills acquisition of learners.

Another source of professional development is the marking operation where teachers get first-hand involvement with student scripts and the marking rubric.

Partnerships

Success depends strongly on the partnerships that are established. Of prominence are the partnerships established with parents. It is very obvious that where parents are interested and play their role through reading to children even before they get to school that good habits are formed.

Schools need to establish partnerships with universities that can assist with action-based research studies and strategies. The independent eye might find problems and solutions that people too close to the scene may not be aware of. There are already a number of programs in existence on which future relationships can be built.

Preparing for the test

There is significant pressure on teachers when a national test is in place to spend considerable time teaching the specific content of the test and concentrating on test-taking strategies which are intended to improve performance.

> *Research shows that high achievement and high test scores result when what is tested is woven into daily teaching and challenging curriculum in a relevant manner.*
>
> Routman, R. 2005, *Writing Essentials*, Heinemann, Portsmouth, New Hampshire, p. 245.

This quote suggests that the best thing that teachers can do is teach well. However, without specifically teaching to the test and overwhelming children with practice tests there are some ways in which children can be supported in their test taking. It was obvious that in 2008 many children suffered considerable stress during the testing time. Anxiety needs to be quelled so that children complete their testpapers confidently. Certainly children need to be familiar with what the test will look like and what they will be expected to do. Fatigue or boredom must be resisted and students encouraged to participate until the end. While the stakes may not be high for some individuals the test should be seen as a chance to show off performance and all children need to believe that it is worth doing. Children should be taught to read carefully and to reread questions and their answers. The published writing marking rubric is a convenient guide to help understand what is valued when trained markers judge student writing.

There are specific aspects of literacy and numeracy that teachers need to be aware of and to promote positively through their teaching.

Using the data

Each student answers relatively few questions but each question is answered by the full cohort of students. Thus the information the tests provide about individual performance is not as reliable as what the cohort data provides about learning sequences, trends or general gaps in learning and/or curriculum. The focus of the Queensland Literacy tests was on these very aspects and so the test reporting handbooks from previous years contain helpful information about what can be done to improve performance.

To make the best use of the information provided by the data, schools need to analyse their data not just for the percentage correct but also for the percentage of students who chose each of the options. They then need to make a judgment about what the data particular to the school means for their classroom curriculum and pedagogy. This is the critical difference between the *prescription*, which is the most common response to external accountability measures such as a testing regime, and the *precision* that comes from data-driven focused instruction. It is the latter that makes the critical difference in improving student outcomes and supporting system reforms.

In conclusion

Around the world, those systems that have invoked testing as a tool of public policy and an instrument of change have been disappointed. Failure to engage with the classroom implications and use of the data have meant that reform has become stalled and accountability repressive. If the data are to be used constructively to improve educational standards in Queensland, it is imperative that we advocate for better test construction. At the same time we need to support schools to develop the internal accountability measures that help them to develop their own professional learning. This is a proven road to continuous systemic improvement.