

# **The 2009 Year 4 and Year 6 English QCATs**

An analysis of data collected by Queensland Studies Authority

A report prepared for the Queensland Studies Authority

Dr Sandy Muspratt  
Adjunct Research Fellow  
School of Education  
The University of Queensland

24 July 2009

## Table of Contents

Executive Summary .....	iii
Report.....	1
Introduction.....	1
Data collections.....	1
Major questions asked of each data collection.....	3
Analyses .....	4
State-wide data.....	5
250-schools data.....	8
Double marking of QCATs from 100 schools.....	13
Focus group sessions.....	19
Survey .....	22
Conclusion .....	27
Appendix 1: Focus group questions.....	28
Appendix 2: Survey .....	29
Appendix 3: Summaries of Statistical Tests .....	311

## Table of Tables

Table 1: Crosstabulations showing the number of students in the State-wide data collection at each level by gender, by Indigenous status, and by ESL status .....	5
Table 2: Patterns of missing data for Overall Letter Grade for the 250-schools data collection.....	9
Table 3: Patterns of missing data for letter grade for Assessable Elements for the 250-schools data collection .....	9
Table 4: Statistical test for equality of number of returns across the Overall Letter Grades for the 250-schools data collection .....	10
Table 5: Relative importance assigned to each Assessable Element by teachers when deciding the Overall Letter Grade.....	13
Table 6: Patterns of missing data for the letter grades for Assessable Elements for the 100-schools data collection .....	15

## Table of Figures

Figure 1: Distribution of responses across Overall Letter Grades for the State-wide data collection for Year 4 and Year 6 .....	6
Figure 2: Distribution of responses across Overall Letter Grades for Year 4 and Year 6 separated by Gender .....	7
Figure 3: Distribution of responses across Overall Letter Grades for Year 4 and Year 6 separated by Indigenous status .....	7
Figure 4: Distribution of responses across Overall Letter Grades for Year 4 and Year 6 separated by ESL status.....	7
Figure 5: Distribution of responses across Overall Letter Grades for each year level for the 250-schools data collection .....	10
Figure 6: Pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade – Year 4 .....	11
Figure 7: Pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade – Year 6 .....	11
Figure 8: Agreement between (a) pairs of markers; and (b) between markers and the schools when awarding Overall Letter Grades – Year 4.....	14
Figure 9: Agreement between (a) pairs of markers; and (b) between markers and the schools when awarding Overall Letter Grades – Year 6.....	14
Figure 10: Agreement between pairs of markers; and between markers and the schools for each Assessable Element – Year 4 .....	16
Figure 11: Agreement between pairs of markers; and between markers and the schools for each Assessable Element – Year 6 .....	17
Figure 12: Coefficient of agreement (Cohen's $\kappa$ ) between the two makers and between teachers and markers when awarding Overall Letter Grades and the letter grade for Assessable Elements for each year level .....	19
Figure 13: Time spent preparing students for the QCAT .....	23
Figure 14: Time spent contextualising the QCAT with students.....	23
Figure 15: Time taken by students to complete the QCAT .....	24
Figure 16: Number of sessions taken to implement the QCAT .....	24
Figure 17: Mean ratings for four items dealing with teachers' perceptions of the <i>Teacher Guidelines</i> .....	25
Figure 18: Mean ratings for six items dealing with teachers' perceptions of the <i>Student booklet</i> .....	26
Figure 19: Mean ratings for four items dealing with teachers' perceptions of the <i>Guide to making judgements</i> .....	26
Figure 20: Mean ratings for four items dealing with teachers' perceptions of the <i>Sample responses</i> .....	26
Figure 21: Mean ratings for six items dealing with teachers' beliefs about the way in which the QCAT data will inform their teaching, planning and programming.....	27

## Executive Summary

- This report is concerned with the Year 4 and Year 6 English QCATs for 2009
- Five data collections inform the analyses presented in the report:
  - State-wide data - Overall Letter Grades plus students' gender, Indigenous status and ESL status from schools across the State;
  - Data from 250-schools – The schools returned completed *Student booklets* that represented a typical response for each Overall Letter Grade;
  - Double marking for 100 schools – The *Student booklets* from 100 schools (selected from the 250-schools data collection) were double marked by trained markers;
  - Summaries of focus group discussion with the markers at the conclusion of the double marking process;
  - A survey completed by teachers.
- The questions asked of the data collections include:
  - What are the shapes of the distributions across the letter grades, and do the distributions separate according to gender, Indigenous status and ESL status;
  - Are there discernible relationships between the Overall Letter Grades and the Letter Grades for Assessable Elements;
  - Were the markers and teachers consistent when awarding Overall Letter Grades and letter grades for Assessable Elements;
  - What aspects of the QCAT process made it difficult for markers to be consistent;
  - What were teachers' opinions and beliefs concerning the QCAT process.
- For the State-wide data, the distributions follow a typical Normal distribution – small proportions at the extremes (letter grades A and E) with larger proportions in the middle (letter grade C).
- In general, girls did better than boys; non-Indigenous students did better than Indigenous students; and non-ESL students did better than ESL students.
- Year 4 teachers assigned relatively less importance to the 1st Assessable Element (Knowledge and understanding) when making their on-balance judgement for the Overall Letter Grade.
- Year 6 teachers assigned relatively less importance the 3rd Assessable Element (Constructing Texts) when making their on-balance judgement for the Overall Letter Grade.
- The markers achieved satisfactory levels of agreement when awarding Overall Letter Grades and awarding letter grades for the Assessable Elements.
- The levels of agreement between the Overall Letter Grades awarded by the markers and the Overall Letter Grades awarded by the schools were also satisfactory or not far from it; although

somewhat less than the levels of agreement achieved by the pairs of markers. It is with respect to only one Assessable Elements that the two groups were not achieving satisfactory agreement.

- The markers found it difficult to award letter grades when:
  - Distinguishing between borderline grades;
  - Descriptors were appeared to be vague, not specific, not discrete, were misplaced;
  - Assessable Elements drew on information from a number of questions;
  - Deciding how to weight differing letter grades for Assessable Elements when determining an Overall Letter Grade.
- The majority of teachers who responded to the survey claimed that: they took more than one hour preparing students for the QCAT, they took 30 minutes contextualising the QCAT,; students completed the QCAT in about the recommended time; and that it took two sessions to implement the QCAT.
- Teachers' perceptions of the *Teacher Guidelines*, the *Student Booklet*, the *Sample Responses* were on the whole positive. They were more critical of the *Guide to making judgement* than any other documents making up the QCAT.
- Teachers agreed that the data gathered from the QCAT implementation will help to inform programs, planning and teaching.

# Report

## Introduction

This report is concerned with the 2009 trail of the QCATs. At the time of writing (July 2009), the QCATs were administered in Year 4 and Year 6 only - the Year 9 trail is to take place later in 2009. Furthermore, this report is concerned with the English KLA only.

Schools taking part in the administration of QCATs received a package of materials for each QCAT that contained:

- *Teacher guidelines* – containing information about QCATs in general; how teachers prepare themselves and their students for the QCAT; online resources relevant to the assessment; a list of the Essential Learnings that form the basis of the assessment; and models for achieving consistency of teacher judgements;
- *Student booklet* – containing the assessment task to be completed by the students;
- In addition, the *Teacher guidelines* and the *Student booklet* contain the *Guide to making judgements*.
- In addition, *Sample responses* – containing annotated responses - were available on the QSA website.

Teachers are asked to "make a judgement" (award a letter grade on the 5-point scale) related to each Assessable Element according to a set of descriptors, then "make an overall on-balance judgement" (award an Overall Letter Grade on the 5-point scale for the QCAT). On the 5-point scale, "A" represents the highest level of achievement and "E" represents the lowest level.

This report is concerned with the awarding of letter grades; problems that were experienced when letter grades were being awarded, and teachers' perceptions of the usefulness or otherwise of the documents that comprise the QCAT package. The sections to follow provide details of the data collections that inform this report and the major questions asked of the data. These are followed by details of the analyses applied to each data collection.

## Data collections

Five data collections inform the analyses contained in the following section. Three focus on the letter grades awarded for students' responses contained in the *Student booklets*. As well, focus group sessions and surveys were used. The data collections are described below.

### **State-wide data**

This data collection is concerned with the Overall Letter Grades awarded by teachers across the State for two QCATs: Year 4 English and Year 6 English. The schools returned the Overall Letter Grade (i.e., the *Student booklets* were not returned, nor were letter grades for the Assessable Elements available) along with indications of students' gender, Indigenous status and ESL status.

### **For 250 schools**

Across approximately 250 schools, the typical or mid-range *Student booklet* for each Overall Letter Grade was selected and returned. Thus for each QCAT, approximately 1250 *Student booklets* should have been returned (250 schools X 5 Overall Letter Grades). The data for this collection comprised the Overall Letter Grade plus the letter grade for each Assessable Element. It should be noted that if a school could not provide a mid-range QCAT for each of the five Overall Letter Grades, the school was nevertheless asked to return five *Student booklets*, and as a consequence, they would have doubled up on an Overall Letter Grade.

### **Double marking of QCATs from 100 schools**

From the 250 schools, a subset of 100 schools were selected and the QCATs from these schools were assessed by two trained markers. A total of 20 markers took part at each year level. The two makers for each QCAT awarded an Overall Letter Grade and a letter grade for each Assessable Element. From time to time, the two markers met to check for consensus. If, for any *Student booklet*, they failed to reach consensus for either the Overall Letter Grade or the letter grade for an Assessable Element, they were asked to reach consensus, possibly after some discussion. Thus there were four sets of letter grades available in this data collection: one set for each marker when awarding letter grades independently; the consensus set; and the set of letter grades awarded at the schools.

The markers kept brief records of the *Student booklets* (Student ID, QCAT, Year Level, and Assessable Element or question) that were difficult to assess, including an indication of what it was that made the responses difficult to assess. Also, the markers kept similar records of the *Student booklets* for which they failed to reach consensus, including brief comments about why they failed to reach consensus in the first place and how consensus was eventually achieved. These records served as memory prods for the focus group session (see below).

## **Focus group sessions**

At the conclusion of the marking, the markers attend focus group sessions. The markers formed four groups - two groups per year level. Each group comprised the pairs of markers who marked the QCATs for a given year level and a group leader. A semi-structured schedule was prepared (see Appendix 1) to serve as a guide for the discussions, but the group leaders were encouraged to move beyond the schedule to seek points of clarification and elaboration during the discussions. The sessions were recorded and summaries of the recordings were prepared.

## **Survey**

A survey seeking teachers' opinions of the implementation of the QCATs in their schools was available for to teachers on the QSA website. Appendix 2 contains the survey.

## **Major questions asked of each data collection**

### **For State-wide data**

The questions asked of the state-wide data focussed on the shapes of the distributions across the Overall Letter Grades:

- Are the shapes of the distributions for the two year levels comparable?
- Are the shapes and the locations of the distributions comparable across: gender groupings; Indigenous status groupings; and ESL status groupings?

### **For 250 schools**

The questions asked of the 250-schools data again focussed on the shapes of the distributions, but unlike the state-wide data collection where the distributions were expected to follow roughly a Normal distribution, the distributions for the 250-schools data collection were expected to be flat. This is because each school was asked to select a typical example of each Overall Letter Grade. The 250-schools data collection also included the letter grades for Assessable Elements, and so it was possible to investigate the ways in which letter grades for Assessable Elements were awarded within Overall Letter Grades. Thus, questions asked of the 250-schools data collection included:

- Are the distributions at each year level flat?
- What is the pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade?
- Were the teachers assigning roughly equal importance to the Assessable Elements when assigning an Overall Letter Grade?



## **Double marking of QCATs from 100 schools**

The questions asked of the 100-schools data collection were concerned with the consistency with which Overall Letter Grades and letter grades for Assessable Elements were awarded:

- Initially, were there discrepancies between the two markers?
- Were there discrepancies between the consensus letter grades awarded by the markers and the letter grades awarded at the schools?
- Are there discernible patterns associated with discrepancies within year levels and within Assessable Elements?

## **Focus group sessions**

In the focus group sessions, the markers were asked to consider aspects of the marking process that made it difficult for the markers to be consistent:

- Were there problems with the descriptors, the Assessable Elements, or the tasks that contribute towards inconsistencies?
- How did the markers overcome these problems and reach agreement?
- Were there discernible patterns associated with discrepancies?

In addition, the markers were asked to move beyond the direct evidence available to them in the *Student booklets*, and to speculate about:

- The extent to which teachers might or might not be attending to particular curriculum domains;
- The extent to which teachers might be using schemes in addition to or as an alternative to the QSA descriptors when awarding letter grades.

## **Survey**

The survey contained questions concerned with the time taken to implement the QCATs, and the documentation accompanying the QCATs (*Teacher guidelines*, *Student booklet*, *Guide to making judgements*, and *Sample responses*). A copy of the survey is contained in Appendix 2.

## **Analyses**

The analyses are presented for each of the data collections in turn. Where appropriate, the analyses will be supplemented with discussions of technical aspects of the analysis.

### *State-wide data*

Table 1 shows that Overall Letter Grades were obtained for a little more than 40,000 students at each year level from across the State. Table 1 also shows the number of students according to gender, Indigenous status, and ESL status. As expected, the administration included roughly equal numbers of male and female students, but non-Indigenous students and non-ESL students far outnumber Indigenous and ESL students. As shown under the "Unknown" heading in the table, the Indigenous status and the ESL status for a small number of students (less than 0.5%) were not known.

*Table 1: Crosstabulations showing the number of students in the State-wide data collection at each level by gender, by Indigenous status, and by ESL status*

	<b>Male</b>	<b>Female</b>			<b>Total</b>
<b>Year 4</b>	20900	19859			40759
<b>Year 6</b>	20306	20075			40381

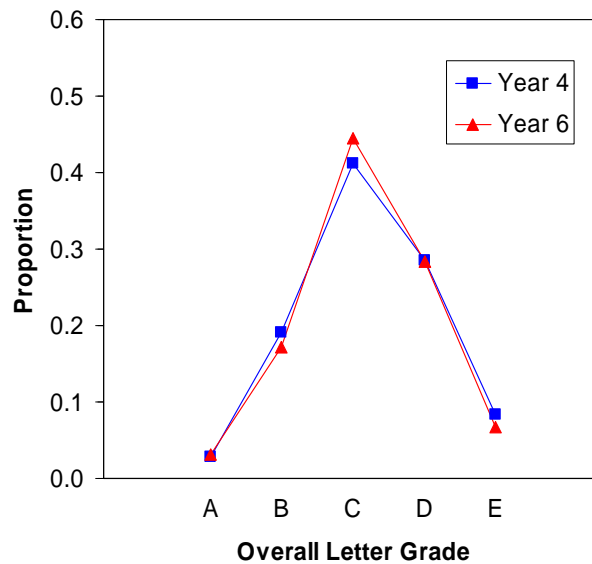
  

	<b>Indigenous</b>	<b>Non-Indigenous</b>	<b>Unknown</b>	<b>Total</b>
<b>Year 4</b>	3026	37635	98	40759
<b>Year 6</b>	2957	37295	129	40381

	<b>ESL</b>	<b>Not ESL</b>	<b>Unknown</b>	<b>Total</b>
<b>Year 4</b>	1967	38715	77	40759
<b>Year 6</b>	1723	38567	91	40381

Figure 1 shows the shape of the distribution across Overall Letter Grades at each year level. The figure shows the proportion of students of the total number of students awarded each Overall Letter Grade. For instance, considering the distribution for Year 4, only small proportions of students were awarded the letter grade A – the letter grade awarded to students achieving at the highest level - less than 10% of students (0.1 of students) received an A grade. The proportions tend to rise for letter grades B and C, then decrease for letter grades D and E. That is, the pattern is roughly a Normal distribution – smaller proportions of students at the extremes of the distribution, with larger proportions of students receiving mid-range letter grades. The shape of the distribution for Year 6 is roughly the same as the Year 4 distribution.



*Figure 1: Distribution of responses across Overall Letter Grades for the State-wide data collection for Year 4 and Year 6*

Figures 2, 3 and 4 show the extent to which the distributions separate according to gender, Indigenous status and ESL status respectively. The left-hand chart in Figure 2 shows that, for Year 4, girls achieve at slightly higher levels than boys. This effect is represented in the Figure by the boys' distribution being displaced to the right compared to the girls' distribution. This shifting of the distributions is the result of larger proportions of girls than boys receiving the higher letter grades (A, B and C), and larger proportions of boys than girls receiving the lower letter grade (D and E). The pattern for Year 6 is similar to the Year 4 pattern; that is, in Year 6, girls achieve at higher levels than boys.

When comparing Indigenous students to non-Indigenous students (Figure 3), the separation of the Year 4 and the Year 6 distributions are generally large. Smaller proportions of Indigenous than non-Indigenous students receive letter grades A, B and C; and larger proportions of Indigenous students than non-Indigenous students receive letter grades D and E. Indeed, the proportion of Indigenous students receiving letter grade E is approximately twice that of non-Indigenous students. It is noted, however, that not much separates the proportions of Indigenous and non-Indigenous students receiving an A grade.

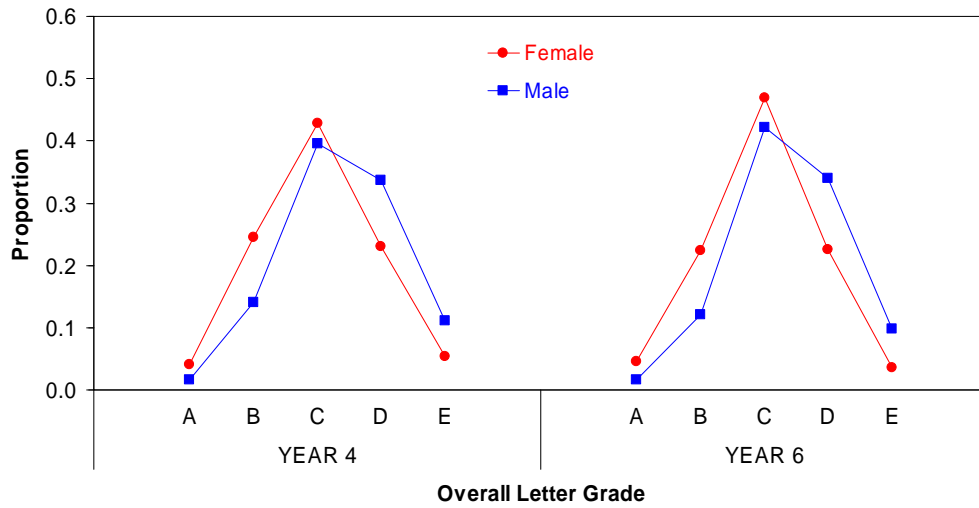


Figure 2: Distribution of responses across Overall Letter Grades for Year 4 and Year 6 separated by Gender

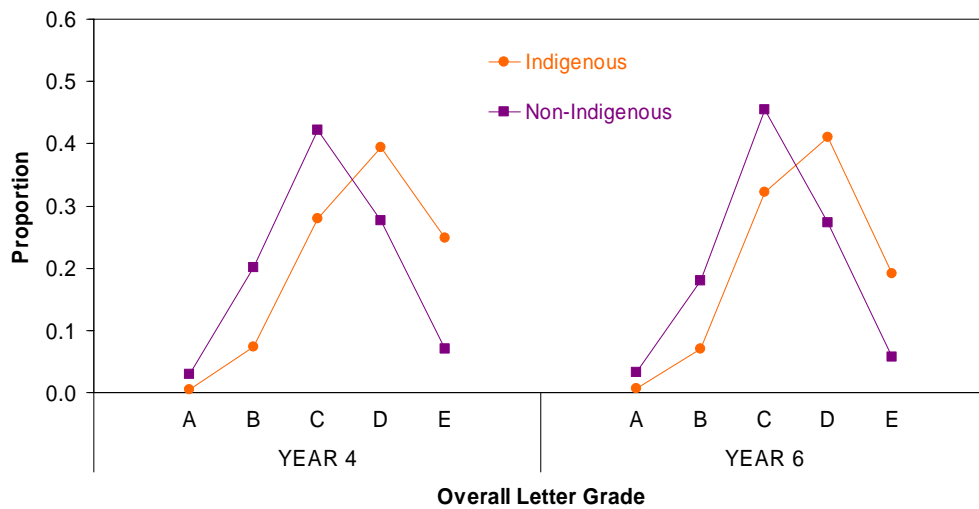


Figure 3: Distribution of responses across Overall Letter Grades for Year 4 and Year 6 separated by Indigenous status

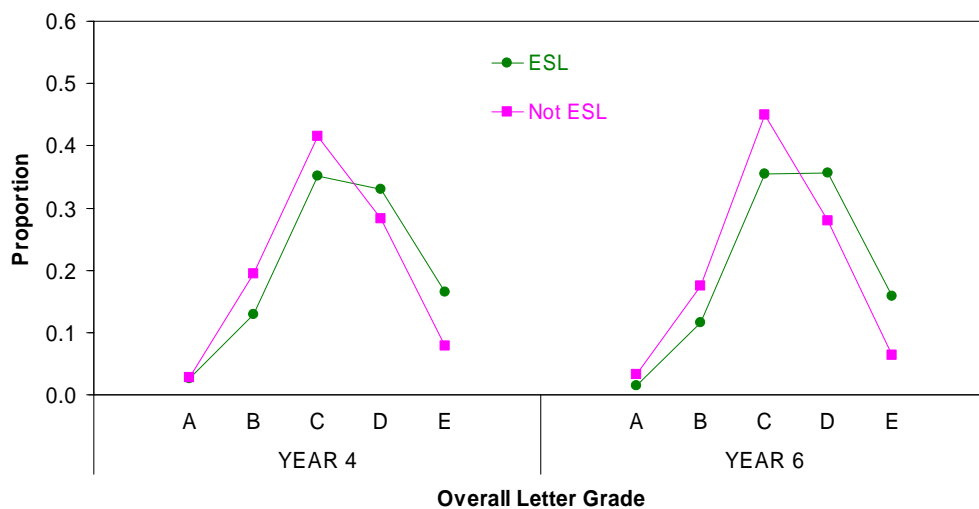


Figure 4: Distribution of responses across Overall Letter Grades for Year 4 and Year 6 separated by ESL status

The separation of the distributions according to ESL status is not as marked as for Indigenous status, but the overall patterns are similar. At each year level, not much separates the proportions of ESL and non-ESL students receiving an A grade, larger proportions of non-ESL students than ESL students receive letter grades of B and C, then larger proportions of non-ESL students than ESL students receive letter grades of D and E.

In summary, girls do better than boys; non-Indigenous students do better than Indigenous students; and non-ESL students do better than ESL students at Year 4 and Year 6.

### ***250-schools data***

Schools were asked to select a typical or mid-range QCAT for each Overall Letter Grade (i.e., one QCAT that was typical of an Overall Letter Grade A; one that was typical of a B, one that was typical of a C, a D; and an E). The number of schools that returned Year 4 and Year 6 QCATs were 233 and 228 respectively. Thus it was expected that there be a total of 1165 returns (5 returns X 233 schools) for Year 4 and a total of 1140 returns (5 returns X 228 schools) for Year 6. However, there were missing data. Table 2 shows where the missing data occurred at each year level. The rates are not large - of the order of 10%. Most of the missing data is a consequence of schools returning fewer than five *Student booklets*.

Table 3 shows the pattern of missing data for the Assessable Elements. There are different reasons why letter grades could be missing for Assessable Elements, including: in some schools, there were no letter grades awarded for any Assessable Element; there were instances of letter grades missing for one, two or three Assessable Elements; and there were responses for which teachers could not distinguish between letter grades. Also shown in Table 3 is the number of times letter grades could not be distinguished for each Assessable Element. For Year 4, no Assessable Element appears to be any more difficult than any other Assessable Element; but in Year 6, it appears that teachers had difficulty with Assessable Element 3.

*Table 2: Patterns of missing data for Overall Letter Grade for the 250-schools data collection*

<p><b>Year 4</b></p> <p>97 booklets not returned (from a total of 52 schools);          15 Could not distinguish between letter grades;          7 No Overall Letter Grade even though AEs had a letter grade;  <b>Total 119 (10.2%)</b></p> <p><b>Year 6</b></p> <p>87 booklets not returned (from a total of 47 schools);          8 Could not distinguish between letter grades;          18 No Overall Letter Grade even though AEs had a letter grade;  <b>Total 104 (9.1%)</b></p>
--

*Table 3: Patterns of missing data for letter grade for Assessable Elements for the 250-schools data collection*

<p><b>Year 4</b></p> <p>76 No letter grades for any AEs;          79 instances of one, two, three or four letter grades missing for AEs mostly because letter grades could not be distinguished.          The number of times letter grades could not be distinguished by AEs:</p> <table border="1"> <tr> <td>AE1</td> <td>AE2</td> <td>AE3</td> <td>AE4</td> </tr> <tr> <td>26</td> <td>27</td> <td>28</td> <td>29</td> </tr> </table> <p><b>Year 6</b></p> <p>80 No letter grades for any AEs;          99 instances of one, two, or three letter grades missing for AEs mostly because letter grades could not be distinguished.          The number of times letter grades could not be distinguished by AEs:</p> <table border="1"> <tr> <td>AE1</td> <td>AE2</td> <td>AE3</td> </tr> <tr> <td>29</td> <td>29</td> <td>44</td> </tr> </table>	AE1	AE2	AE3	AE4	26	27	28	29	AE1	AE2	AE3	29	29	44
AE1	AE2	AE3	AE4											
26	27	28	29											
AE1	AE2	AE3												
29	29	44												

For this data collection, schools were asked to select a typical or mid-range QCAT for each Overall Letter Grade. Thus, it is expected that the distributions across the Overall Letter Grades at the two year levels will be flat, but as can be seen in Figure 5, the distributions are not perfectly flat. Deviations from 'perfect flatness' were tested for statistical significance. The results, shown in Table 4, confirm the indications of Figure 5, that the distributions deviate significantly for 'perfect flatness'. The number of A grades is less than expected, and also possibly the number of E grades is

less than expected. That is, not all schools could find *Student booklets* with Overall Letter Grades at the extremes.

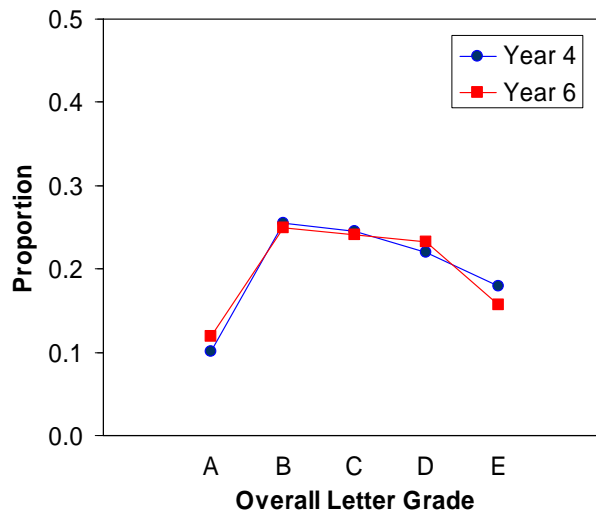


Figure 5: Distribution of responses across Overall Letter Grades for each year level for the 250-schools data collection

Table 4: Statistical test for equality of number of returns across the Overall Letter Grades for the 250-schools data collection

Year Level	$\chi^2_{df=4}$	p
Year 4	75.8	< 0.001
Year 6	60.5	< 0.001

If schools could not provide mid-range A, B, C, D and E responses, they nevertheless submitted five *Student booklets* thus doubling up on an Overall Letter Grade. It might be reasonable to assume that the doubling-up of Overall Letter Grades accounts for the deviations from 'perfect flatness' noted above – as noted in the 1300 schools data collection, letter grades A and E occur less frequently than the other letter grades; therefore in smaller schools, letter grades A and E might be more difficult to find; and as a consequence, letter grades A and E occur less often in the 250-schools data collection while letter grades B, C, and D occur more often.

Figure 6 shows the pattern of letter grades awarded for Assessable Elements within an Overall Letter Grade for Year 4. Consider the patterns for Assessable Elements when an Overall Letter Grade of A was awarded. The most likely letter grade for any Assessable Element was an A.

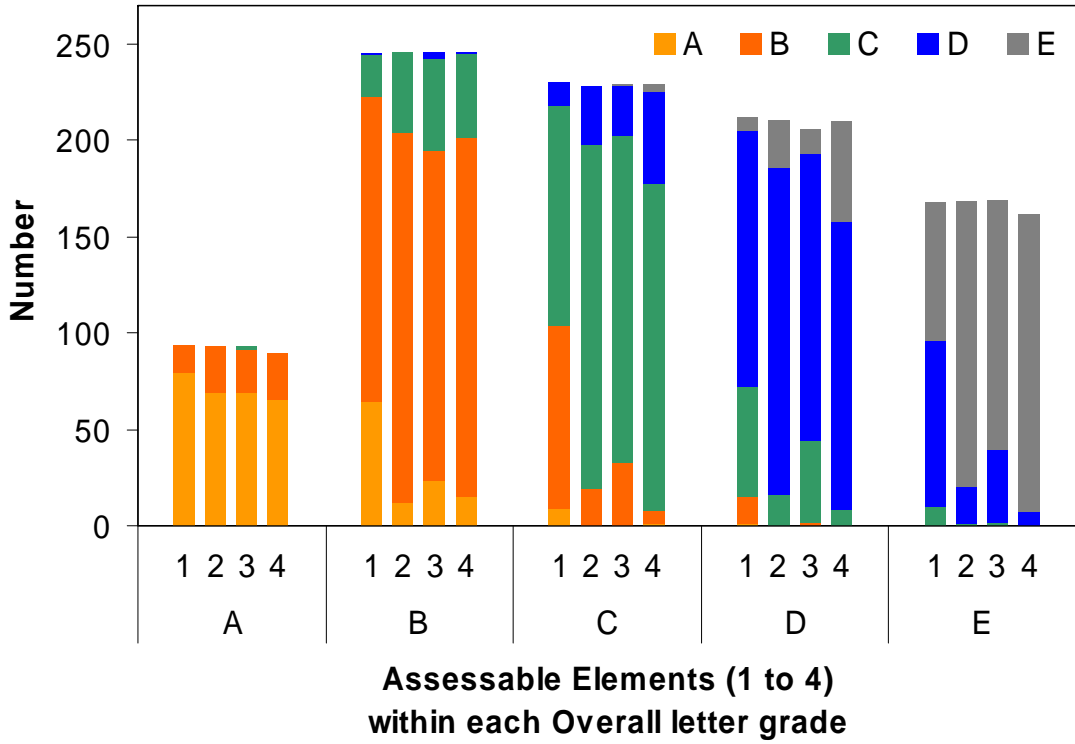


Figure 6: Pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade – Year 4

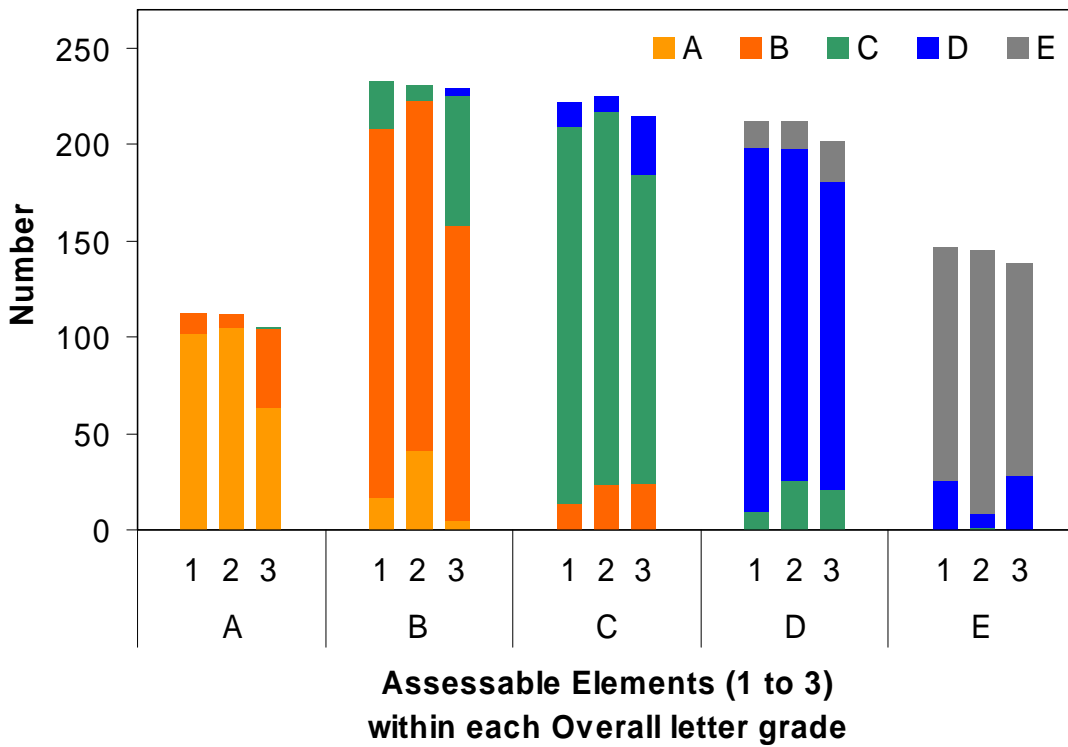


Figure 7: Pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade – Year 6



Similarly, when an Overall Letter Grade of B was awarded, the most likely letter grade for any Assessable Element was a B. There are similar patterns for letter grades C, D and E. That is, the letter grade for the Assessable Element aligns mostly with the Overall Letter Grade. A similar patterns applies for Year 6 (Figure 7).

There is one possible exception to this pattern. It concerns the pattern of letter grades awarded for the 1st Assessable Element in Year 4. When an Overall Letter Grade of B was awarded, most teachers awarded a B for the 1st Assessable Element, but also a substantial proportion awarded an A. Similarly, when a C, D or E was awarded for the Overall Letter Grade, large proportions of teachers awarded C, D or E respectively for the 1st Assessable Element; but also, substantial proportions awarded B, C and D respectively.

Another element that might play a role when awarding an Overall Letter Grade is the importance that teachers attach to each of the Assessable Elements, albeit implicitly. One way to assess 'relative importance' is to examine standardised regression coefficients obtained from multiple regression analyses. For the data at hand, the regressions were set up so that the letter grades for Assessable Elements were used to predict the Overall Letter Grade (the letter grades having first been converted to numeric grades: A = 0, B = 1, and so on through to E = 4).

To read the importance of Assessable Elements, consider the standardised coefficients for Year 4 in Table 5. An increase of one standard deviation for the 1st Assessable Element leads, on average, to an increase in the Overall Letter Grade of 0.17 standard deviations; an increase of one standard deviation for 2nd Assessable Element leads to an increase in the Overall Letter Grade of 0.29 standard deviations; an increase of one standard deviation for the 3rd Assessable Element leads to an increase of 0.29 standard deviations for the Overall Letter Grade; and an increase of one standard deviation in 4th Assessable Element leads to an increase of 0.29 standard deviations for the Overall Letter Grade. Thus, somewhat less importance is assigned to the 1st Assessable Element than to the other three Assessable Elements. For Year 6, less importance is assigned to the 3rd Assessable Element than to the other two Assessable Elements.

Table 5: Relative importance assigned to each Assessable Element by teachers when deciding the Overall Letter Grade

Year 4		Year 6	
Assessable Element	Standardised coefficient	Assessable Element	Standardised Coefficient
Knowledge & understanding	.167	Knowledge & understanding	.379
Interpreting texts	.287	Constructing texts	.396
Constructing texts	.292	Constructing texts	.244
Reflecting	.294		

### *Double marking of QCATs from 100 schools*

In this section, the analyses are concerned with the agreement achieved by pairs of markers when awarding the Overall Letter Grade and the letter grade for each Assessable Element. Also, the analyses are concerned with the agreement between the grade awarded by the school and the consensus grade of the two markers for both the Overall Letter Grade and the letter grade for each Assessable Element. These analyses apply to five *Student booklets* from 100 schools, a sub-sample of the 250-schools data collection.

Figures 8 and 9 give a visual representation of the consistency achieved by pairs of markers when awarding Overall Letter Grades. To read Figure 8 (a), for instance, note that each point is represented by a cloud of points. Consider the point represented by the coordinates (B, B) in the scatterplot (a). There are 78 booklets represented by (B, B), which means that for 78 *Student booklets*, the two markers agreed when awarding the B grade. If the 78 booklets were instead to be represented by a single point, information would be lost – the information about there being 78 booklets. In the scatterplot, each point has been jittered. Jittering means adding a small random element to each data point so that the data points are spread out a little. Jittering generates a cloud of points but it is clear that the cloud for (B, B) is associated with (B, B). Most of the time, interest is focussed not so much on the specific number of points in a cloud but rather on an overall impression of the density of points within a cloud. Thus, it is clear that there is a clustering along the diagonal points: (A, A), (B, B), (C, C), (D, D) and (E, E); with a few points displaced one space off the diagonal. That is, the pairs of markers were fairly consistent. For Year 6 (Figure 9 (a)), the pairs or markers are again fairly consistent, but it is noted that occasionally there are points appearing two spaces off the diagonal.

The scatterplots on the right in Figures 8 and 9 show the consistency between the mark awarded at the schools and the consensus mark of the pairs of markers. (There was minimal data missing for the school awarded Overall Letter Grade - there were only nine and four instances of the Overall Letter Grade being missing at Year 4 and Year 6 respectively.) It is clear that there is a dense cloud of points along the diagonal, but, compared to the scatterplots on the left in each Figure, there are more points displaced one and two points off the diagonal. That is, teachers and markers did not achieve the same level of consistency as achieved by the pairs of markers.

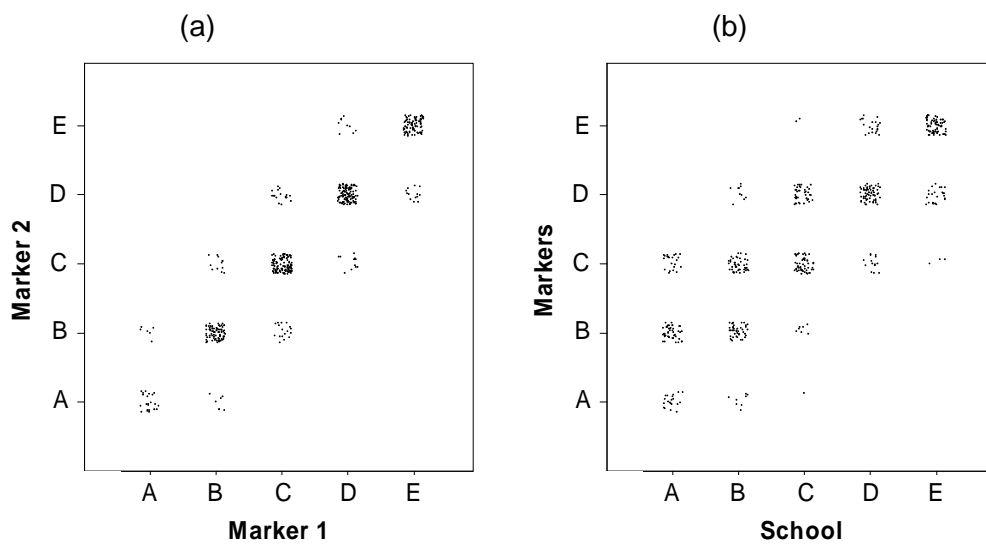


Figure 8: Agreement between (a) pairs of markers; and (b) between markers and the schools when awarding Overall Letter Grades – Year 4

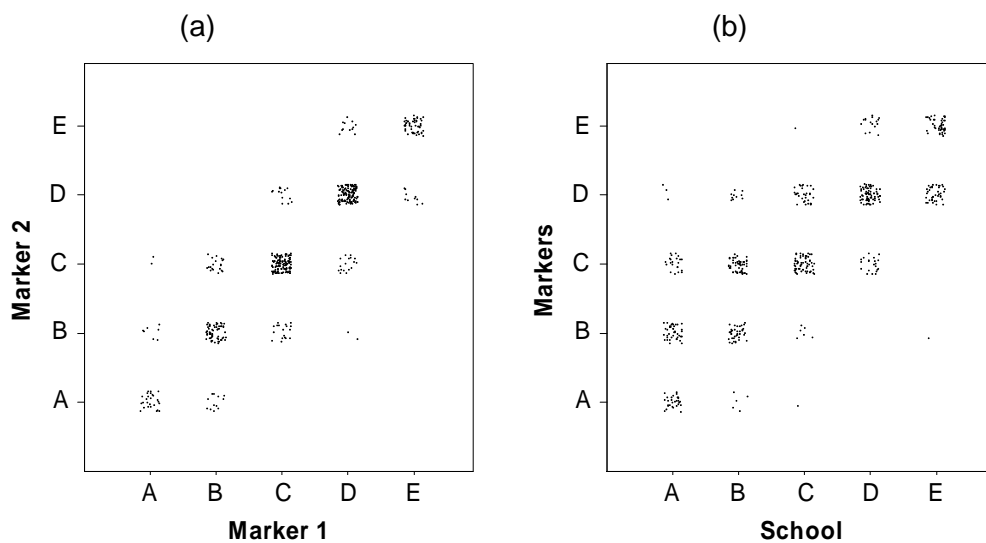


Figure 9: Agreement between (a) pairs of markers; and (b) between markers and the schools when awarding Overall Letter Grades – Year 6

Before turning to the question of consistency when awarding letter grades for the Assessable Elements, it is noted that there is missing data among the letter grades for the Assessable Elements. This should not be surprising given that the 100 schools that comprise this data collection are a subsample of the 250-school data collection. Table 6 shows where the missing data occurred for each QCAT. As was the case with the 250-schools data collection, there were instances of letter grades missing for one or more Assessable Elements, and there were responses for which teachers could not distinguish between letter grade. Also shown in Table 6 is the number of times letter grades could not be distinguished for each Assessable Element.

*Table 6: Patterns of missing data for the letter grades for Assessable Elements for the 100-schools data collection*

<b>Year 4</b>				
37 No letter grades for any AEs;				
36 instances of one, two, three or four letter grades missing for AEs mostly because letter grades could not be distinguished.				
The number of times letter grade could not be distinguished by AEs:				
AE1	AE2	AE3	AE4	
11	10	5	13	
 <b>Year 6</b>				
35 No letter grades for any AEs;				
42 instances of one, two or three letter grades missing for AEs mostly because letter grades could not be distinguished.				
The number of times letter grade could not be distinguished by AEs:				
AE1	AE2	AE3		
14	10	23		

Figures 10 and 11 show consistency in the same way as shown in Figures 8 and 9, except that Figures 10 and 11 show consistency when awarding letter grades for the Assessable Elements. Clearly, the level of consistency declines for the pairs of markers when dealing with the Assessable Elements (scatterplots on the left in each figure). With respect to consistency between the consensus grade and the school grade (scatterplots on the right in each Figure), there does not appear to be a further decline in consistency, except perhaps for the 3rd Assessable Element for Year 6.

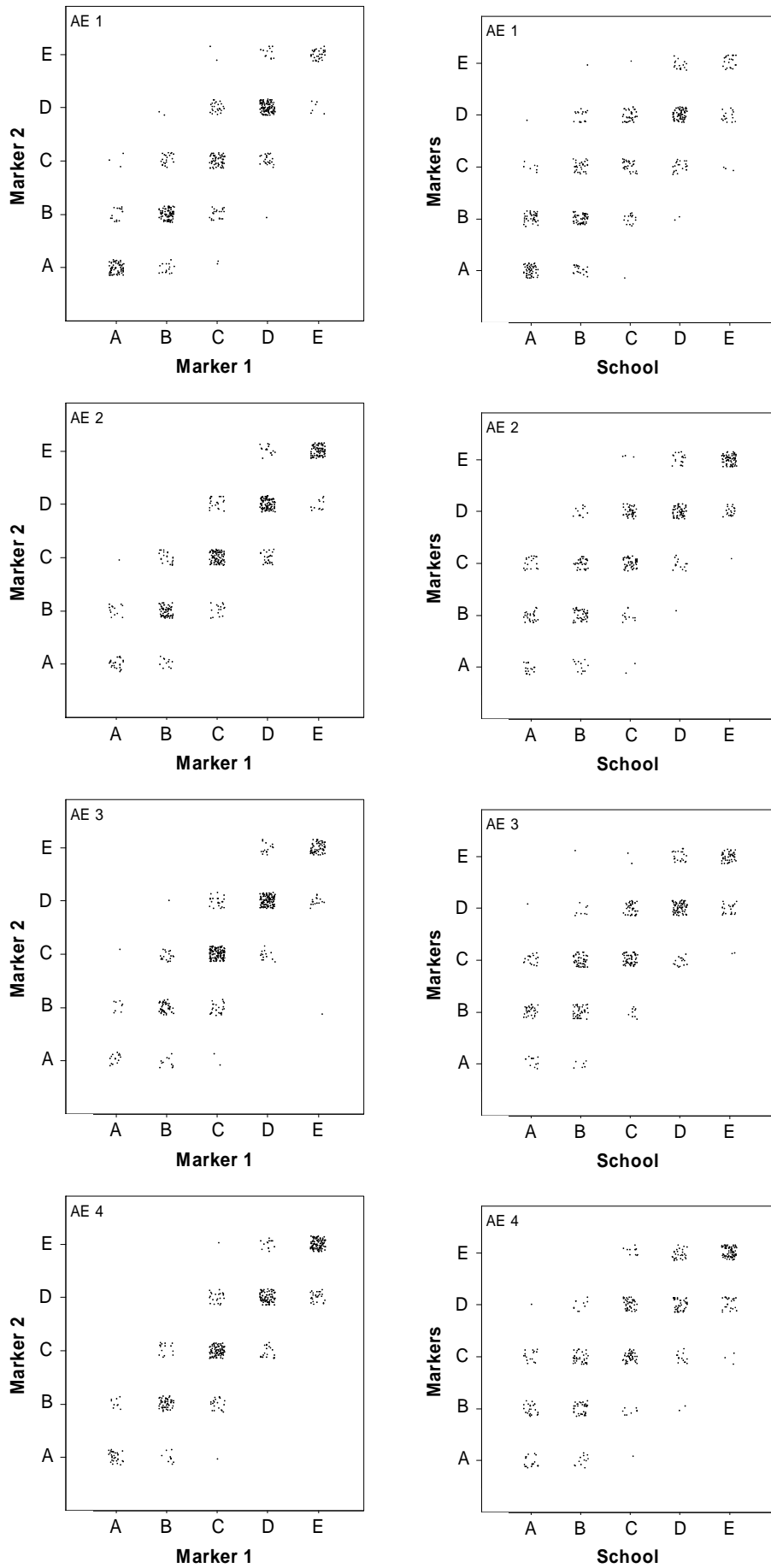


Figure 10: Agreement between pairs of markers; and between markers and the schools for each Assessable Element – Year 4

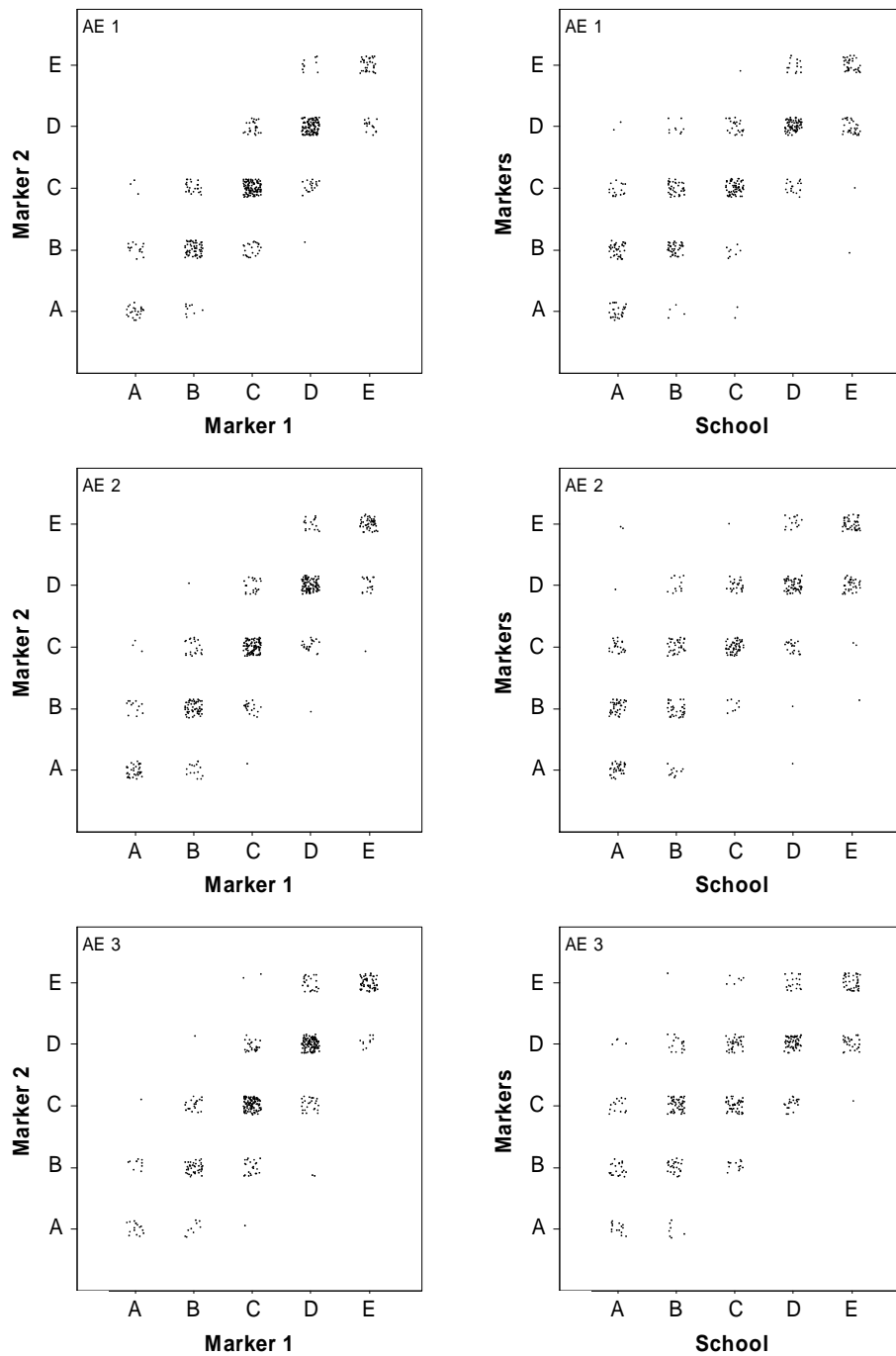


Figure 11: Agreement between pairs of markers; and between markers and the schools for each Assessable Element – Year 6

The consistency between the two markers can be quantified. Cohen's  $\kappa$  is a measure of inter-rater agreement when two raters are rating objects. Usually, Cohen's  $\kappa$  is calculated when the raters are rating objects on a nominal scale (i.e., when there is no order built into the scale), but it can be modified to take account of ordering on an ordinal scale<sup>1</sup>, like the scale used here - A, B, C, D and E. Furthermore, there are two methods for weighting the objects when raters differ in their

<sup>1</sup> Fleiss, J., Levin, B. & Paik, M. (2003). *Statistical methods for rates and proportions*. (3rd ed.) Hoboken, N.J.: John Wiley and Sons.

assessments. The method used here is linear weighting. Cohen's  $\kappa$  ranges between 0 (no agreement other than what would be expected by chance) through to 1 (perfect agreement). A set of descriptors for Cohen's  $\kappa$  is<sup>2</sup>:

< 0.2	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very good

Figure 12 shows the values for Cohen's  $\kappa$  for the two sets of comparisons (between the pairs of markers, and between the consensus grade and the grade awarded at the schools) for the Overall Letter Grade and for the letter grades awarded for Assessable Elements. It is noted that the  $\kappa$  values for pairs of markers are in the "Good" to "Very good" range, but the  $\kappa$  values for the Assessable Elements are slightly less than the Overall values.

The  $\kappa$  values for assessing agreement between the consensus grade and the grade awarded at the schools for the Overall Letter Grades are either at the top end of the "Moderate" range (Year 6) or bottom end of the "Good" range (Year 4). Thus the markers and the teachers could not achieve the same levels of agreement as was achieved by the pairs of markers. Figure 12 also shows that for most Assessable Elements, markers and teachers achieved roughly the same level of consistency as was achieved for the Overall Letter Grade. The 3rd Assessable Element in Year 6 was the exception, where Cohen's  $\kappa$  was 0.5 – well below the "Good" range.

In summary, the markers were achieving satisfactory agreement when awarding Overall Letter Grades and when awarding letter grades for the Assessable Elements. The levels of agreement between the Overall Letter Grades awarded by the markers and the Overall Letter Grades awarded by the schools were also satisfactory or not far from it, although the level of agreement was somewhat less than that achieved by the pairs markers. Similarly, the levels of agreement between the markers and the schools were mostly satisfactory or close to it when awarding letter grade for the Assessable elements, but again, that levels were less than the levels achieved by the pairs of markers. except the teachers had difficulty with the third Assessable Element for Year 6 English.

---

<sup>2</sup> Altman, D. (1991). *Practical statistic for medical research*. London: Chapman & Hall.

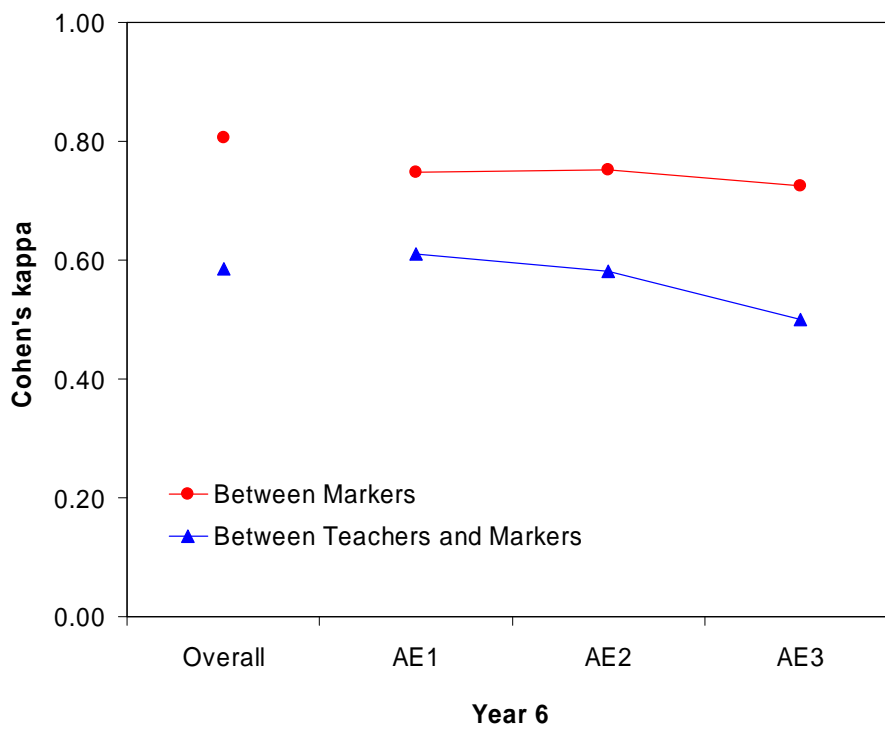
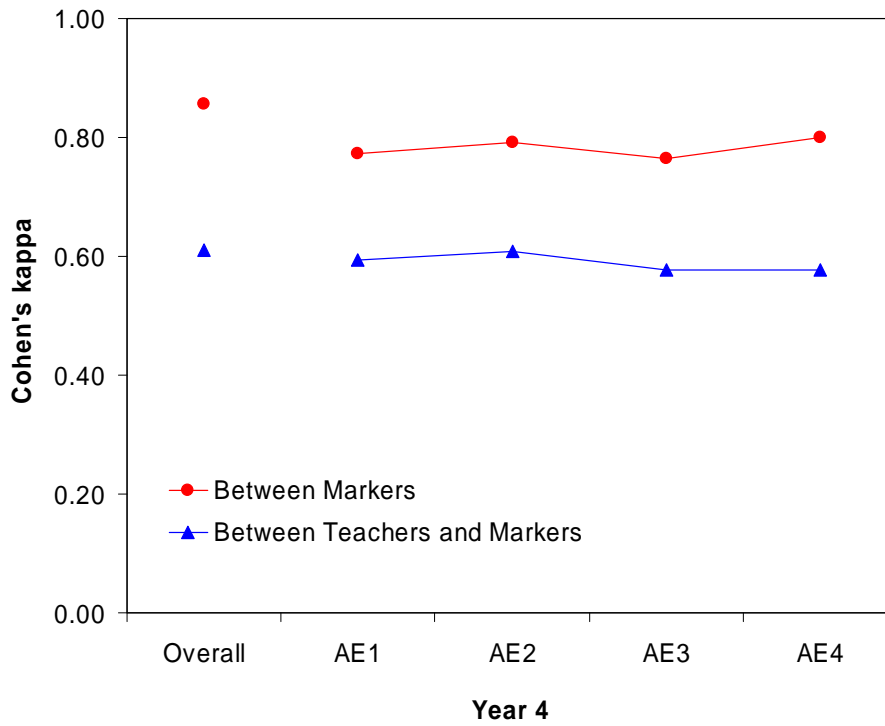


Figure 12: Coefficient of agreement (Cohen's  $\kappa$ ) between the two makers and between teachers and markers when awarding Overall Letter Grades and the letter grade for Assessable Elements for each year level

### ***Focus group sessions***

At the conclusion of their marking, the markers attended focus group sessions to discuss any difficulties that arose during the marking and their perceptions of the consistency the achieved.



The markers claimed to be fairly consistent. On the whole, their assessment of their consistency aligns with the assessments of agreement presented in the previous section (see Figures 8 to 11, and, in particular, Figure 12). Some markers claimed that in addition they were more consistent when awarding letter grades for Assessable Elements than when awarding Overall Letter Grades. Given that the letter grades for Assessable Elements were awarded according to specific descriptors whereas Overall Letter Grades were awarded on the basis of an "overall on-balance judgement" (according to the *Teacher Guidelines*, p. 6), their perceptions might seem reasonable. However, on comparing the clouds of points in Figures 8 and 9 with those in Figures 10 and 11, or on inspecting the agreement values in Figure 12, clearly the markers, on the whole, were more consistent when awarding Overall Letter Grades.

When discrepancies did occur, they were not large, and, with few exceptions, disagreements were not displaced more than one letter grades apart. The markers claimed that disagreements were concerned mostly with borderline grades. Markers claimed to depend heavily upon the *Guide to making judgements* to resolve differences. Some focussed on the purpose; other focussed on the task specific assessable elements; while other returned to the descriptors, even in some situations highlighting key words in the descriptors. Other strategies included:

- Discussion;
- Return to the *Sample Responses*;
- Consulted with the writers about the wording or the intent of the descriptors; and
- Ask other markers to look at particular student's work.

For the markers, a major difficulty when assessing students' work occurred where the descriptors were not sufficiently specific. The first Assessable Elements in the Year 6 QCAT was mentioned as an example. The task-specific Assessable Element contained a list of language and textual features. The difficulty for some markers lay with deciding to what extent all or some of the features needed to be displayed in students' work. The first Assessable Element in the Year 4 QCAT was mentioned as a similar example, where one of the descriptors referred to "key facts". The question for the markers was to what extent do all or some of the "key facts" need to be present in students' work.

Some markers claimed difficulty with awarding letter grades for Assessable Elements that drew on multiple questions, especially when each question required substantial pieces of writing. The difficulty lay with weighting the responses to the different questions when assigning the letter grade. For instance, with respect to the 4th Assessable Element in the Year 4 QCAT, some markers

claimed difficult with deciding on letter grade when, say, only two of the three questions were answered well. Similarly, some markers claimed difficulty with assigning weights to the Assessable Elements when awarding an Overall Letter Grade.

Other areas of concern raised by the markers included:

- Markers, at times, found it difficult to distinguish between descriptors. Some instances that were cited include: "credible" and "convincing" in the 2nd Assessable Element in Year 6; and "evocative" and "emotive" in the 1st Assessable Element in Year 6.
- Descriptors that appear to be out of order: Markers cited descriptors B, C and D for the 1st Assessable element in Year 4 English as an example.
- Descriptors that appear to be misplaced; for example, one marker complained of the use of the term "imagery" in the Knowledge and Understanding Assessable Element in Year 6, arguing that "imagery" would be better placed as a descriptor for text construction.
- Having to overlook or disregard spelling and punctuation errors.
- Assessable elements that were spread across a number of question, or probably a number of pages – "There was a lot of flicking."

The markers were asked if they thought teachers were using schemes in addition to or as alternatives to the QSA descriptors, and if they thought that there were curriculum areas that the teachers were attending to particularly well or areas that teachers were not attending to well. The markers comments here should be treated as highly speculative because they are based on just five booklets from each school. As a consequence, any conclusions drawn from these comments have to be treated with a degree of caution.

With respect to alternative schemes, the markers claimed that teachers had used highlighting, underlining and ticks, but possibly more as a reminder of features used in students' work rather than as an alternative to the descriptors. Also, markers noted that teachers had used various schemes to highlight spelling and grammatical errors. Nevertheless, in some booklets, markers noted that teachers had used methods other than or in addition to QSA's descriptors to award letter grades; including the use of letter or numeric grades in sub-questions or in elements smaller than the Assessable Element.

With respect to curriculum domains that might or might not have been attended to well, the markers' impressions were that while the content might have been well attended to, some students were not well prepared to display specific skills such as dealing with time in an itinerary,

justification and interpretation.. Generally, questions that depended on knowledge, recall and understanding were answered better than questions that depended upon justification and interpretation.

### *Survey*

A total of 182 surveys were completed, but after discarding surveys from teachers responding with respect to Mathematics and Science, 140 remained in the data file: 74 from Year 4 teachers, and 66 from Year 6 teachers. The majority of surveys (89%) were received from State schools, with smaller numbers received from Catholic schools (7%) and Independent schools (4%). As expected, most surveys were received from Primary schools (90%) with the remainder received from P-to-10 or P-to-12 schools (10%). A small number of returns were received from teachers in schools located in remote areas (6%), with the remainder more or less evenly spread across rural (34%), provincial (32%) and Brisbane metropolitan (29%) areas.

The survey contained four questions concerned with the amount of time spent preparing, contextualising and implementing the QCAT:

- How much time did you spend preparing students for the QCAT?
- How much time did you spend setting the scene of the QCAT with students?
- How long did the students take to complete the QCAT?
- In how many sessions was the QCAT implemented?

A series of tests were conducted to determine whether or not responses differed according to year level (Mann-Whitney tests testing for differences across the two groups – Year 4 and Year 6), and according to the location of teachers' schools (Kruskal-Wallis tests testing for differences across four groups – remote, rural, provincial cities, Brisbane metropolitan area). After adjusting the alpha level for each test to take account of the fact that in each case four tests were being conducted, there were no statistically significant differences.

Figures 13, 14, 15 and 16 show the distribution of responses for each question in turn. Each figure shows the response pattern for each year level as well as the overall pattern. Consider Figure 13. It shows the distribution of responses for "Time spent preparing students for the QCAT" at each year level and the "Overall" response pattern. The bars show the proportion of teachers who ticked each time category (30 minutes, 1 hour, more than 1 hour). It can be seen that, overall, a large proportion of teachers ticked the "More than 1 hour" category, with smaller proportions ticking the "1 hour" and "30 minutes" categories. There is a similar pattern for each year level, and indeed, the results of

the significance test (mentioned in the previous paragraph) indicate that, on the whole, there were no differences in the response patterns across the year levels.

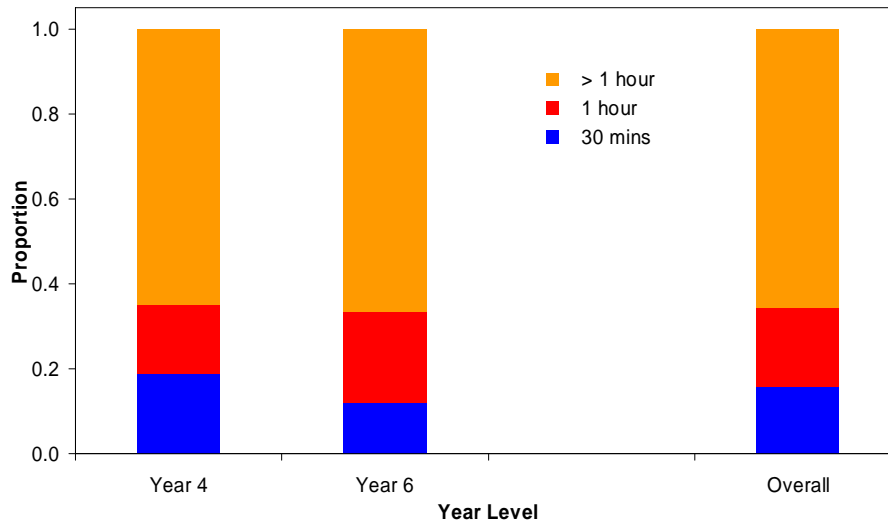


Figure 13: Time spent preparing students for the QCAT

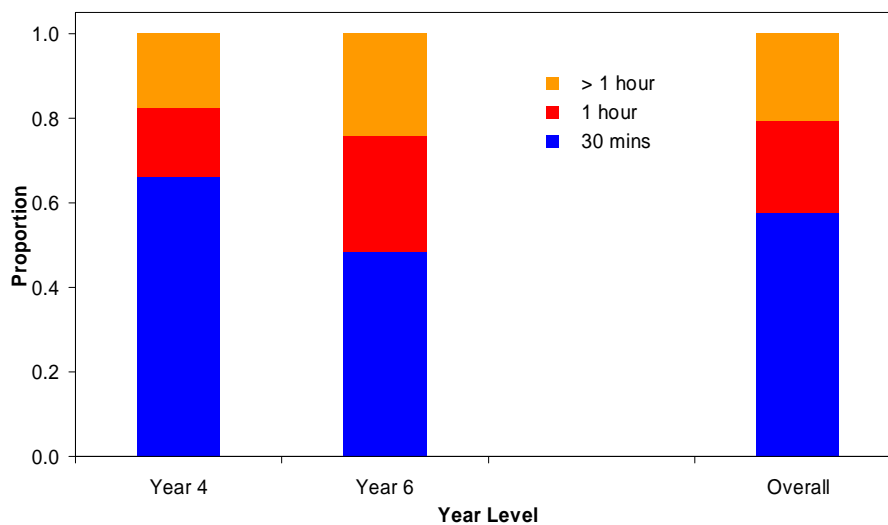


Figure 14: Time spent contextualising the QCAT with students

Figure 14 shows the distribution of responses for "Time spent contextualising". The figure is structured the same way as Figure 13. Overall, a large proportion of teachers ticked "30 minutes", with smaller proportions ticking the "1 hour" and "more than 1 hour" categories. Figure 15 shows the distribution of responses for the question concerning "Time students took to complete the QCAT". Most teachers claimed that students took about the recommended times, with a substantial proportion claiming that students took more than the recommended time. Very few teachers claim

that students took less than the recommended time. Finally, Figure 16 shows the pattern of responses for the question concerning "Number of sessions to implement QCAT". The majority of teachers claimed that the QCAT was implemented in two sessions.

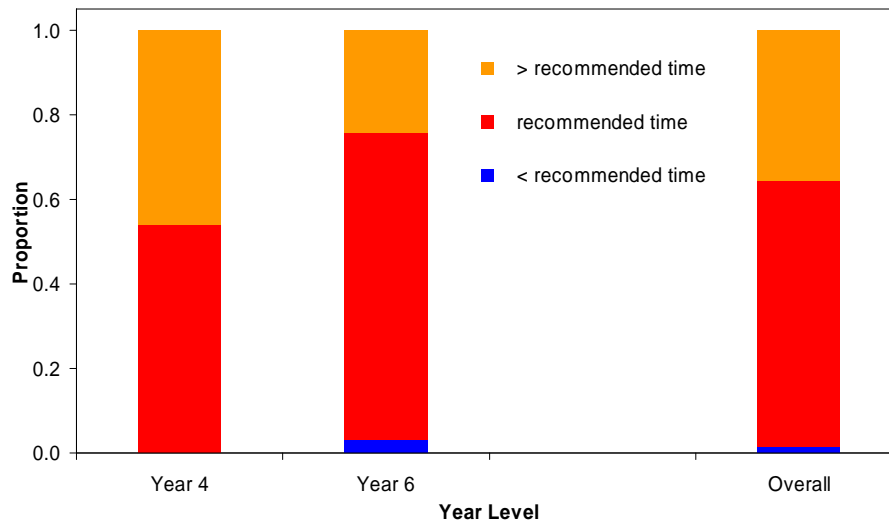


Figure 15: Time taken by students to complete the QCAT

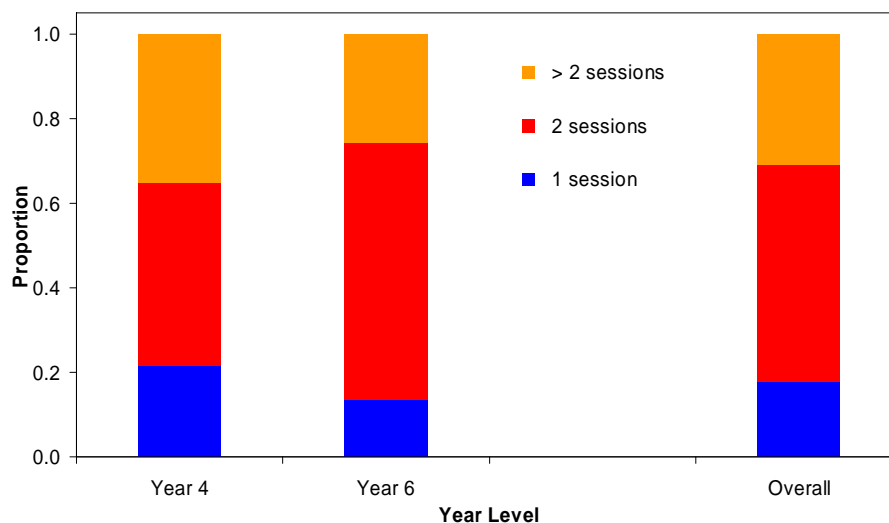


Figure 16: Number of sessions taken to implement the QCAT

There was a series of questions asking teachers their opinions of the QCAT documents: *Teacher guidelines*, *Student booklet*, *Guide to making judgements*, and *Sample responses*. Figures 17, 18, 19, and 20 give the average ratings for each statement about each document in turn. Each figure shows the ratings separated by year level. Note that the scale used in the figures is the reverse of that used

in the survey so that in the figures "stronger agreement" is represented by larger numbers. For instance, for the *Teacher guidelines* (Figure 17), teachers on the whole agreed that the document provided the information that was required, that the instructions were clear, that the suggested level of support to students was appropriate, and that the model response was helpful.

A series of MANOVAs (Multivariate ANalyses Of Variance) showed that there are no statistically significant differences in the way Year 4 teachers responded compared to the way Year 6 teachers responded. Also, a series of MANOVAs showed that there were no statistically significant differences in the response patterns according to the location of teachers' schools (remote, rural, provincial cities, Brisbane metropolitan area). Summaries of results of the MANOVAs are presented in Appendix 3.

With respect to the *Student booklet*, Figure 18 shows that the mean rating for each statement are above the neutral midpoint on the scale, indicating that teachers, on the whole, agreed with the propositions. Figure 19 shows that teachers were more critical of the *Guide to making judgements* than the other documents; but nevertheless, the means remain at or close to the neutral midpoint of the scale. Finally, Figure 20 shows that the teachers were on the whole in agreement with three of the four propositions. The exception was the statement, "Two Sample responses per overall grade was sufficient". The mean rating was at the neutral midpoint of the scale, indicating that teacher, on the whole, were undecided about the proposition.

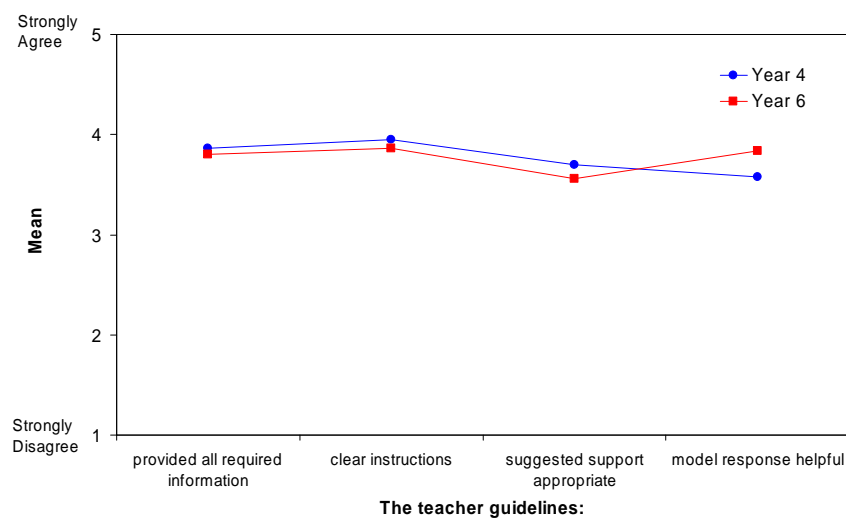


Figure 17: Mean ratings for four items dealing with teachers' perceptions of the Teacher Guidelines

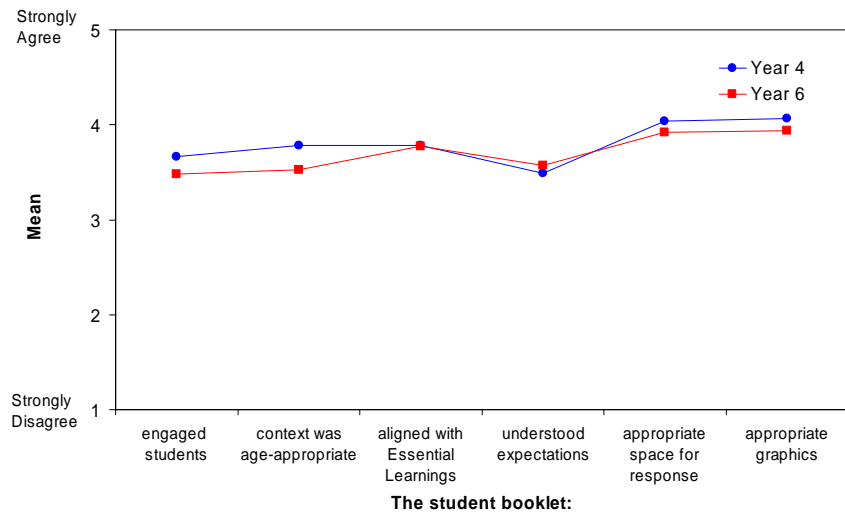


Figure 18: Mean ratings for six items dealing with teachers' perceptions of the Student booklet

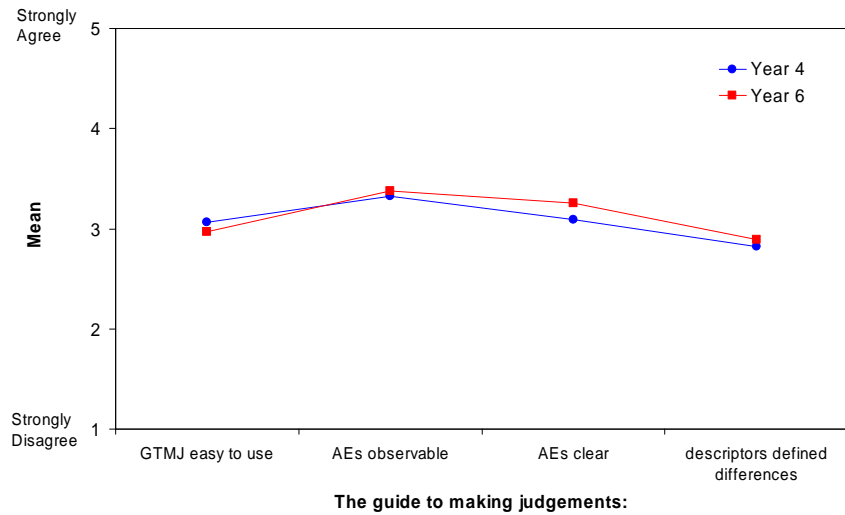


Figure 19: Mean ratings for four items dealing with teachers' perceptions of the Guide to making judgements

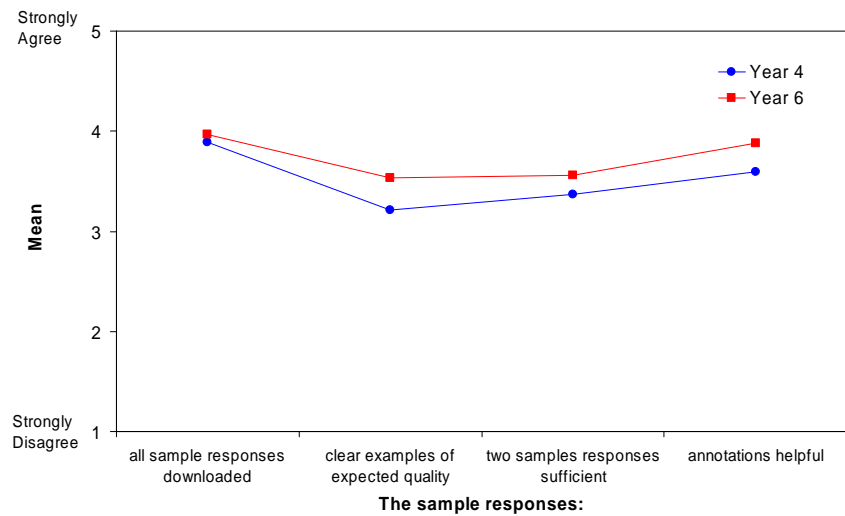
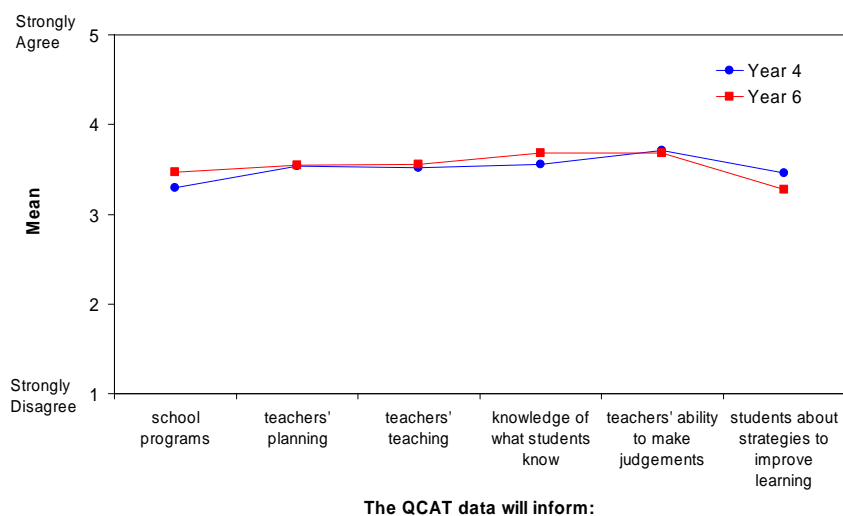


Figure 20: Mean ratings for four items dealing with teachers' perceptions of the Sample responses

The last set of questions concerned teachers' perceptions of the way in which the data gathered during the QCAT implementation would inform teaching, planning and programming. MANOVAs showed that there were no statistically significant differences in response patterns according to year level nor according to the location of the teachers' schools (summaries for the MANOVAs are shown in Appendix 3). Figure 21 shows that teachers were on the whole in agreement with each statement.



*Figure 21: Mean ratings for six items dealing with teachers' beliefs about the way in which the QCAT data will inform their teaching, planning and programming*

## Conclusion

The markers demonstrated that satisfactory levels of agreement can be achieved when awarding Overall Letter Grades and Letter Grades for the Assessable Elements. Thus it might be argued that what the markers achieved, the teachers too should be able to achieve – the markers after all are themselves teachers. But it must be remembered that the markers were brought into a central location to complete the double marking, they had received training before commencing the double marking, they were marking "typical" student responses, they were not having to complete the marking during an already crowded teaching day or at the end of the teaching day, and they could consult with each other whenever difficulties arose. Nevertheless, the teachers were able to achieve satisfactory levels of agreement with the consensus grade when awarding the Overall Letter Grade. It was only with respect to the one Assessable Element that levels of agreement dropped.



# Appendix 1: Focus group questions

## FOCUS GROUP QUESTIONS FOR MARKERS

### Focus area 1: Think about how the students answered the questions.

- How did the students go about answering the questions?
- Were there Assessable Elements or questions that the students answered particularly well?
- Were there Assessable Elements or questions that the students were struggling with?
- Can you say where the students' difficulties might lie – interpreting the question, not knowing the content, ...?
- Are there Assessable Elements or questions that were regularly omitted?

### Focus area 2: Think about where you had difficulty assessing students' work.

- Were there elements that you, individually, had difficulty assessing?
- Where in your opinion did the difficulty lie - the question, the descriptors...?
- How did you overcome the difficulty?

### Focus area 3: Think about the discrepancies between you and your second marker.

- Do you think you and your second marker were on the whole consistent?
- Were there Assessable Elements or overall grades for which you and your second marker had difficulty reaching consensus?
- Where in your opinion did the difficulty lie - the question, the descriptors. ...? How did you reach consensus?
- Were there overall grades for which you and your second marker had difficulty reaching consensus?
- Where in your opinion did the difficulty lie? How did you reach consensus?
- Were there any instances where consensus could not be reached. What did you do in those circumstances?



### Focus area 4: Think back to any notes or marks or ticks that the teachers might have left on the QCATs.

- Was there any evidence that teachers might have been applying numeric methods or some other method (e.g., counting ticks) in making judgements of the quality of students' work?
- Did it appear that they were using these instead of or as well as the QSA descriptors?
- How often did it happen? Were there any discernible clumping patterns (e.g., within schools, curriculum areas, year levels, etc.)?

### Focus area 5: We want you to go beyond the direct evidence contained in the QCAT that you've been marking, and to speculate somewhat.

- Do you think that there are curriculum areas that teachers seem to be attending to particularly well, and/or some that they are not attending to so well?

## Appendix 2: Survey

 <b>Queensland Government</b>	 <b>QSA</b> Queensland Studies Authority <i>Partnership and innovation</i>	<p><b>2009 QCATs</b></p> <p><b>This online survey should be completed and submitted by the teacher/s who implemented the QCATs</b></p> <p><b>(We welcome multiple responses, if more than one teacher implemented the QCAT in the school.)</b></p>
---	---	--

**Please complete and submit a survey for each QCAT implemented.**

<b>1. Which QCAT did you implement?</b>					
<input type="checkbox"/> 4 English	<input type="checkbox"/> 4 Mathematics	<input type="checkbox"/> 4 Science			
<input type="checkbox"/> 6 English	<input type="checkbox"/> 6 Mathematics	<input type="checkbox"/> 6 Science			
<b>2. To which education authority does your school belong?</b>					
<input type="checkbox"/> State (EQ)	<input type="checkbox"/> Catholic (QCEC)	<input type="checkbox"/> Independent (ISQ)	<input type="checkbox"/> Other .....		
<b>3. What type of school?</b>					
<input type="checkbox"/> Primary	<input type="checkbox"/> Secondary	<input type="checkbox"/> P-10/ P-12	<input type="checkbox"/> Special	<input type="checkbox"/> Other .....	
<b>4. What is the location of your school?</b>					
<input type="checkbox"/> Remote	<input type="checkbox"/> Rural	<input type="checkbox"/> Provincial	<input type="checkbox"/> Brisbane		
<b>5. How much time did you spend preparing students for the QCAT?</b>					
<input type="checkbox"/> 30 mins	<input type="checkbox"/> 1 hour	<input type="checkbox"/> more than 1 hour			
<b>6. How much time did you spend setting the scene of the QCAT with students?</b>					
<input type="checkbox"/> 30 mins	<input type="checkbox"/> 1 hour	<input type="checkbox"/> more than 1 hour			
<b>7. How long did the students take to complete the QCAT?</b>					
<input type="checkbox"/> About the recommended amount of time.	<input type="checkbox"/> More than the recommended amount of time	<input type="checkbox"/> Less than the recommended amount of time.			
<b>8. In how many sessions was the QCAT implemented?</b>					
<input type="checkbox"/> 1 session	<input type="checkbox"/> 2 sessions	<input type="checkbox"/> More than 2 sessions			
<b>9. If any students did not undertake the QCAT, give the reason(s)</b>					
<input type="checkbox"/> Absent	<input type="checkbox"/> Special consideration	<input type="checkbox"/> Other			
<b>10. Comment on the <i>Teacher guidelines</i>:</b>					
	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
The Teacher guidelines provided all the information I required	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The instructions were clear	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The suggested level of support to students was appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The model response was helpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>11. Comment on the <i>Student booklet</i>:</b>					
	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
The QCAT engaged students	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The context was age-appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The QCAT was aligned with the targeted Essential Learnings	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Students understood what they were expected to do	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
There was an appropriate amount of space for students to respond	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The graphics were appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

12. <b>Comment on Guide to making judgments (GTMJ):</b>	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
The GTMJ was easy to use to make judgments about the overall quality of student responses	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The Task-specific assessable elements were observable in student responses	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The Task-specific assessable elements were clear	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The Task-specific descriptors clearly defined the discernible differences in student responses	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

13. <b>Comment on the Sample responses:</b>	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
We downloaded all of the <i>Sample responses</i> from the Assessment Bank	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The <i>Sample responses</i> provided clear examples of the quality expected in student work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Provision of two sample responses per grade was sufficient	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The annotations were helpful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

14. <b>The data gathered from the QCAT implementation will help to inform:</b>	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
Our school programs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
My planning	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
My teaching	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
My knowledge of what students know and can do	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
My ability to make consistent judgments	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
My students about strategies to improve their learning	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

15. What processes did teachers put into place to establish consistency of teacher judgments?

Conference/consensus (reaching agreement after grading)     
 Calibration (reaching agreement before grading)     
 Expert (one marker, no conferencing)     
 Other .....  
.....

16. Did teachers from your school work with teachers from other schools to help develop consistency of teacher judgments?       Yes       No

19. General comments:

---

## Appendix 3: Summaries of Statistical Tests

### Mann-Whitney and Kruskal-Wallis tests – Survey questions 5, 6, 7 & 8

The response variable for these questions are at best ordered categorical variables. As a consequence, non-parametric tests were conducted. The appropriate non-parametric analysis when testing for differences between two groups is the Mann-Whitney test, and the appropriate non-parametric analysis when testing for differences among more than two groups is the Kruskal-Wallis test. In the summaries below, the Mann-Whitney analyses test for differences for Year Level (Year 4, Year 6), and the Kruskal-Wallis analysis tests for differences for Location (remote, rural, provincial, metropolitan).

#### Time spent preparing the students for the QCAT

By Year Level – Mann-Whitney  $U = 2348$ ,  $z = 0.47$ ,  $p = 0.640$

By Location -  $\chi^2 = 0.39$ ,  $df = 3$ ,  $p = 0.941$

#### Time spent setting the scene of the QCAT with students

By Year Level - Mann-Whitney  $U = 2030$ ,  $z = 1.94$ ,  $p = 0.053$

By Location -  $\chi^2 = 7.17$ ,  $df = 3$ ,  $p = 0.067$

#### How long did students take to complete the QCAT

(Response categories were ordered from "less than the recommended time" to "more than the recommended time".)

By Year Level - Mann-Whitney  $U = 2020$ ,  $z = 2.10$ ,  $p = 0.036$

(After adjusting the alpha level for multiple testing (across the four questions), the z-value is not statistically significant.)

By Location -  $\chi^2 = 3.81$ ,  $df = 3$ ,  $p = 0.283$

#### How many session did it take to implement the QCAT

By Year Level - Mann-Whitney  $U = 2389$ ,  $z = 0.243$ ,  $p = 0.808$

By Location -  $\chi^2 = 3.08$ ,  $df = 3$ ,  $p = 0.379$

## **MANOVAs – Survey questions 12, 11, 12, 13 & 14**

Summary of MANOVAs testing for significant differences for Year Level (Year 4, Year 6) and testing for significant differences for Location (remote, rural, provincial, metropolitan) on teachers' perceptions of:

### **Teacher Guidelines**

By Year Level - Wilks'  $\Lambda = 0.97$ , MV F(4, 135) = 1.05, p = 0.385

By Location - Wilks'  $\Lambda = 0.90$ , MV F(12, 352) = 1.20, p = 0.284

### **Student Booklet**

By Year Level - Wilks'  $\Lambda = 0.96$ , MV F(6, 133) = 0.89, p = 0.507

By Location - Wilks'  $\Lambda = 0.82$ , MV F(18, 371) = 1.43, p = 0.114

### **Guide to making judgements**

By Year Level - Wilks'  $\Lambda = 0.97$ , MV F(4, 135) = 0.97, p = 0.428

By Location - Wilks'  $\Lambda = 0.92$ , MV F(12, 352) = 0.99, p = 0.458

### **Sample response**

By Year Level - Wilks'  $\Lambda = 0.97$ , MV F(4, 135) = 1.21, p = 0.310

By Location - Wilks'  $\Lambda = 0.95$ , MV F(12, 352) = 0.56, p = 0.873

### **Data will inform**

By Year Level - Wilks'  $\Lambda = 0.95$ , MV F(6, 133) = 1.26, p = 0.282

By Location - Wilks'  $\Lambda = 0.83$ , MV F(18, 371) = 1.38, p = 0.138