The 2008 extended trial of the Queensland Comparable Assessment Tasks

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

7 April 2009

Table of Contents

Table of Tables

Table of Figures

Executive Summary

- Five data collections inform the analyses presented in the report:
    - 1300-schools - Overall Letter Grades plus students' gender, Indigenous status and ESL status;
    - 30-schools – The schools returned completed *Student booklets* that represented a typical response for each Overall Letter Grade;
    - Double marking for 10-schools – the *Student booklets* from 10 schools (selected from the 30-schools data collection) were double marked by trained and independent markers;
    - Summaries of focus group discussion with the markers at the conclusion of the double marking process;
    - A survey distributed to 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 1300-schools data, most 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). The exceptions are Year 6 and 9 Mathematics, and Year 9 Science in which large proportions of students were awarded an E grade.
- In general, girls do better than boys at English, but not much separates girls and boys in Mathematics and Science; non-Indigenous students do better than Indigenous students; and not much separates ESL and non-ESL students (with the exception of Year 6 Mathematics.
- Some teachers found it difficult to award a single Letter Grades for Assessable Elements, and instead awarded either multiple Letter Grades or no Letter Grades at all. One reason could be that the descriptors in the *Guides to making judgements* did not always align with Letter Grades; but that was not the complete explanation.

- Teachers who did provide Letter Grades for all Assessable Element assigned relatively less importance to the 'Knowledge and understanding' Assessable Element when making their judgement of the Overall Letter Grade.
- The pairs of markers achieved satisfactory levels of agreement when awarding Overall Letter Grades and mostly satisfactory levels 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. It is with respect to the Assessable Elements that the two groups were not achieving satisfactory agreement.
- The markers found it difficult to award letter grades when:
    - They had to make decisions concerning the extent of a penalty for students' responses that did not fit the instructions or that strayed from the expected context;
    - Assessable Elements drew on information from a number of questions;
    - Assessable Elements were not sufficiently discrete to make an adequate judgement;
    - Deciding how to weight differing letter grades for Assessable Elements when determining an Overall Letter Grade.
- Teachers on the whole agreed that the *Teacher guidelines* provided the information that was required, that the instructions were clear, and that the suggested level of support to students was appropriate, but teachers of Year 6 and Year 9 Mathematics expressed somewhat less agreement for the statement: *Suggested level of support to student*s *was appropriate*.
- With respect to the *Student booklet*, Mathematics and Science teachers were less convinced than English teachers that *Students understood what they were expected to do*, and were more critical of QCATs' capacity to engage students, particularly Year 9 Science teachers.
- Teachers were more critical of the *Guide to making judgements* and *Sample responses* than other documents, Year 9 Science teachers more so.
- In their written comments, teachers touched on many of the same concerns as expressed by the markers: that assessable elements should not draw on information across a number of questions; how to weight assessable elements when determining an overall grade; that descriptors should align with letter grades; that there should be a larger range of responses contained in the *Sample responses*.

Report

Introduction

The alignment of curriculum, assessment and reporting of achievement for students in Years 1 to 9 is a policy objective of the Queensland State government. Answering the policy objective, Queensland Studies Authority (QSA) developed the Queensland Curriculum, Assessment and Reporting (QCAR) Framework. The QCAR Framework has five interlocking components[1]:

- Essential Learnings – what students should know, understand and be able to do;
- Standards - the quality of student achievements described on a 5-point (A to E) scale;
- Assessment bank – an online collection of assessments and resources linked to the Essential Learnings and the Standards;
- Queensland Comparable Assessment Tasks (QCATs) - authentic, performance-based assessment tasks for students in Years 4, 6 and 9 in the Key Learning Areas (KLAs) of English, Mathematics and Science;
- Guidelines for reporting – how schools should provide information about students' learning.

This report is concerned with the QCATs, but there are two models for the development of QCATs: a set of assessment tasks devised centrally by QSA; and assessment tasks devised locally at the school-level. This report is concerned with the 2008 trail of the centrally devised QCATs.

Schools taking part in the trial 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;
- *Sample responses* – containing a model response and five annotated responses, one for each point on the 5-point (A to E) scale;

[1] The descriptions of the core components are taken from the Queensland Studies Authority website: www.qsa.qld.edu.au/assessment/qcar.html

- In addition, the *Teacher guidelines* and the *Student booklet* contain the *Guide to making judgements*.

Teachers are asked to "make a judgement" (award a letter grade on the 5-point scale) related to each Assessable Element within each QCAT, 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

A number of different data formats and methods of data collection were used during the trial. Three of them focussed on the completed *Student booklets* or rather the Letter Grades awarded for students' responses. As well, focus group sessions and surveys were used. In total, five data collections informed the trial. These are described below.

For 1300 schools

Across 1300 schools, teachers awarded Overall Letter Grades on the 5-point scale for a total of nine QCATs: three KLAs (English, Mathematics, and Science) from each of three year levels (Year 4, 6, and 9). The schools returned the Overall Letter Grade (i.e., the *Student booklets* were not returned) as well as students' gender, Indigenous status and ESL status. It should be noted that not all schools implemented all QCATs.

For 30 schools

Across 30 schools, the typical or mid-range *Student booklet* for each Overall Letter Grade was selected and returned. Thus for each of the nine QCATs, 150 *Student booklets* should have been returned (30 schools X 5 Overall Letter Grades). Thus 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*, that is, by doubling up on an Overall Letter Grade.

Double marking of QCATs from 10 schools

From the 30 schools, a subset of 10 schools were selected and the QCATs from these schools were assessed by two independent and trained markers. A total of 18 markers took part (9 QCATs X 2 markers). 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 by 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. The markers brought these records to the focus group session (see below) to serve as a memory prods during the discussions.

Focus group sessions

At the conclusion of the marking, the markers attend focus group sessions. The 18 markers formed three groups of six markers; each group comprised the pairs of markers who marked the QCATs for a given KLA 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 distributed to teachers. Appendix 2 contains the survey.

Major questions asked of each data collection

For 1300 schools

The questions asked of the 1300-schools data collection focussed on the shapes of the distributions across the Overall Letter Grades:

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

For 30 schools

The questions asked of the 30-schools data collection again focussed on the shapes of the distributions, but unlike the 1300-schools data collection where the distributions were expected to follow roughly a Normal distribution, the distributions for the 30-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 30-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 30-schools data collection included:

- Are the distributions for each QCAT 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 grade?

Double marking of QCATs from 10 schools

The questions asked of the 10-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 KLAs, within tear levels, and within Assessable Elements?

Focus group sessions

In the focus group sessions, the markers were asked to consider aspects of the QCAT 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, the documentation accompanying the QCATs (*Teacher guidelines*, *Student booklet*, *Guide to making judgements*, and *Sample responses*), and moderation processes. 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, limitations of the analysis, and limitations of the data.

1300-schools

As Figure 1 shows, the number of students for whom Overall Letter Grades were received varied across the QCATs. The Year 4 Science and Year 9 Mathematics QCATs outnumbered the other QCATs by a factor of about two. One explanation for the uneven distribution could lie with the fact that different regions were selected to implement the different QCATs. If the regions selected to implement Year 4 Science and Year 9 Mathematics QCATs were populous regions, then it is reasonable to find larger numbers of returns for these two QCATs. Figure 1 also shows that the number of returns for Year 9 QCATs were generally larger that the number of returns for the

QCATs at the other two year levels. But given that secondary school classrooms on average tend to be larger than primary school classrooms, it is reasonable to find the number of Year 9 returns being larger than the number of Year 3 or Year 6 returns.

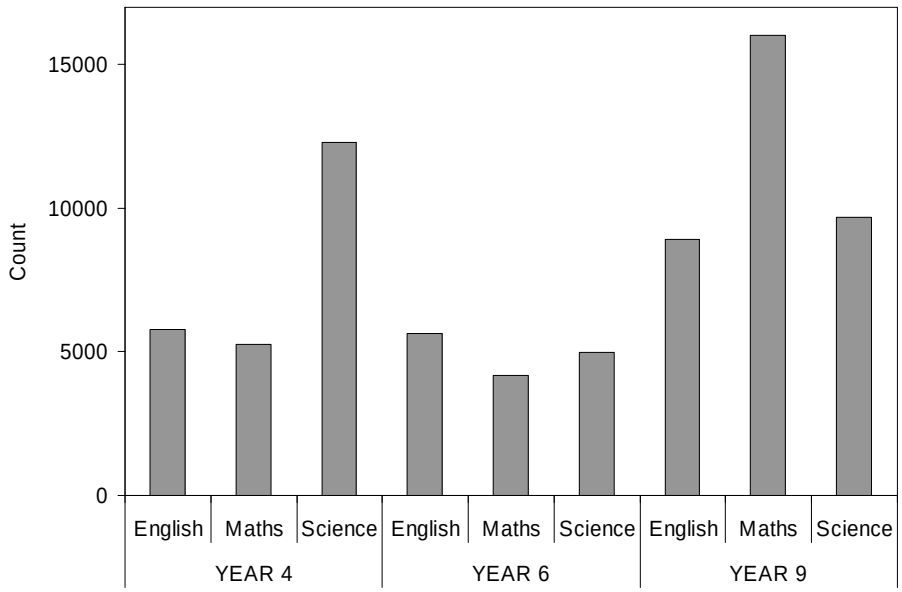


Figure 1: Number of responses for the 1300-schools data collection across KLAs and year levels

Figure 2 shows the shape of the distributions across Overall Letter Grades for each QCAT. The Figure shows the proportion of students of the total number of students for the QCAT awarded each Overall Letter Grade. For instance, considering the three distributions for the Year 4 QCATs, only small proportions of students were awarded the letter grade A – the letter grade awarded to students achieving at the highest level. Typically, less than 10% of students (0.1 of students) receive 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 typical pattern is roughly a Normal distribution – a distribution with smaller proportions of students at the extremes of the distribution, with larger proportions of students receiving mid-range letter grades.

The Figure shows that the distributions for Year 4 Mathematics differs a little from the other two Year 4 distributions. Responses for Mathematics were more evenly spread across letter grades B, C and D than was the case for English and Science. But notice that at Year 4, the shapes of the distributions, even for Mathematics, are roughly the

same: small proportions at the extremes with larger proportions in the middle. For Year 6, however, the shape of the distribution for Mathematics differs markedly from the other two; and at Year 9, both Mathematics and Science differ from the Year 4 pattern. In all three cases, there is no sharp drop-off at letter grade E; that is, the proportions receiving letter grades C, D, and E were relatively constant.

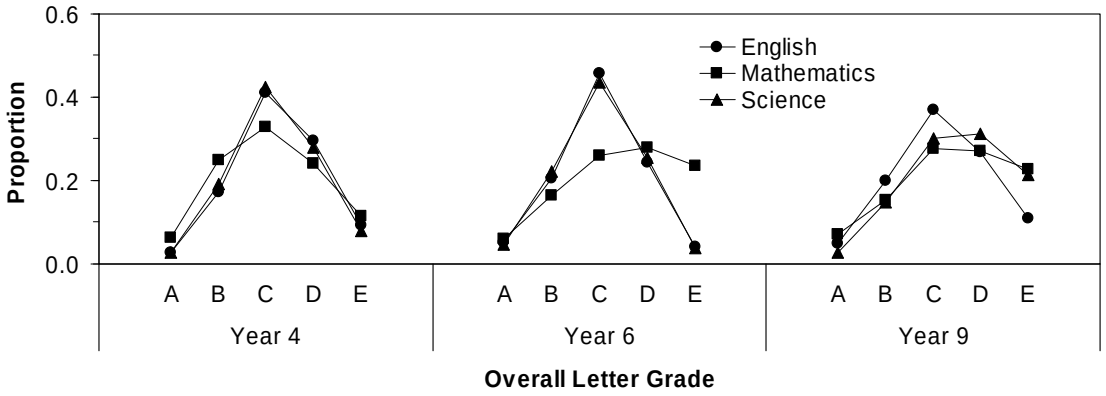


Figure 2: Distribution of responses across Overall Letter Grades for each QCAT

Figures 3, 4 and 5 show the extent to which the distributions separate according to gender, Indigenous status and ESL status respectively. The top left-hand chart in Figure 3 shows that, for Year 4 English, girls were achieving 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 relative 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 patterns for Year 6 and Year 9 English are somewhat similar to the Year 4 pattern. That is, across the three year levels, girls are achieving at higher levels than boys. However, there is little or no separation of the boys' and girls' distributions for Mathematics and Science across the three year levels. That is, boys and girls are achieving at comparable levels for Mathematics and Science.

The separation of the distributions are generally larger when comparing Indigenous students to non-Indigenous students (Figure 4). Indeed, the distributions for Indigenous students for Year 6 and 9 Mathematics are shifted to such an extent that close to 50% of Indigenous students were awarded an Overall Letter Grade of E. For Year 9 Science and Mathematics, large proportions of both Indigenous and non-Indigenous students were

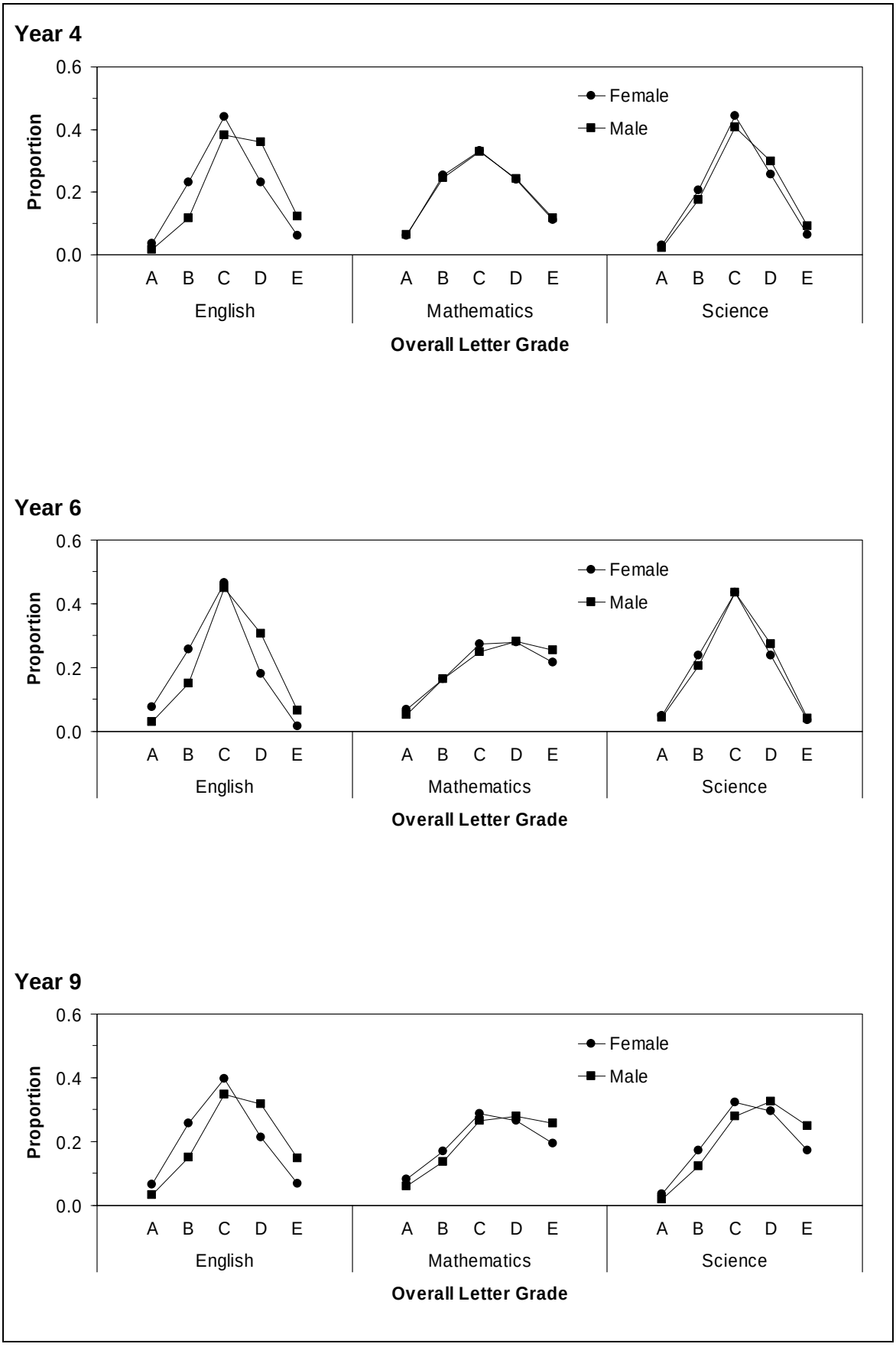


Figure 3: Distribution of responses across Overall Letter Grades for each QCAT separated by Gender

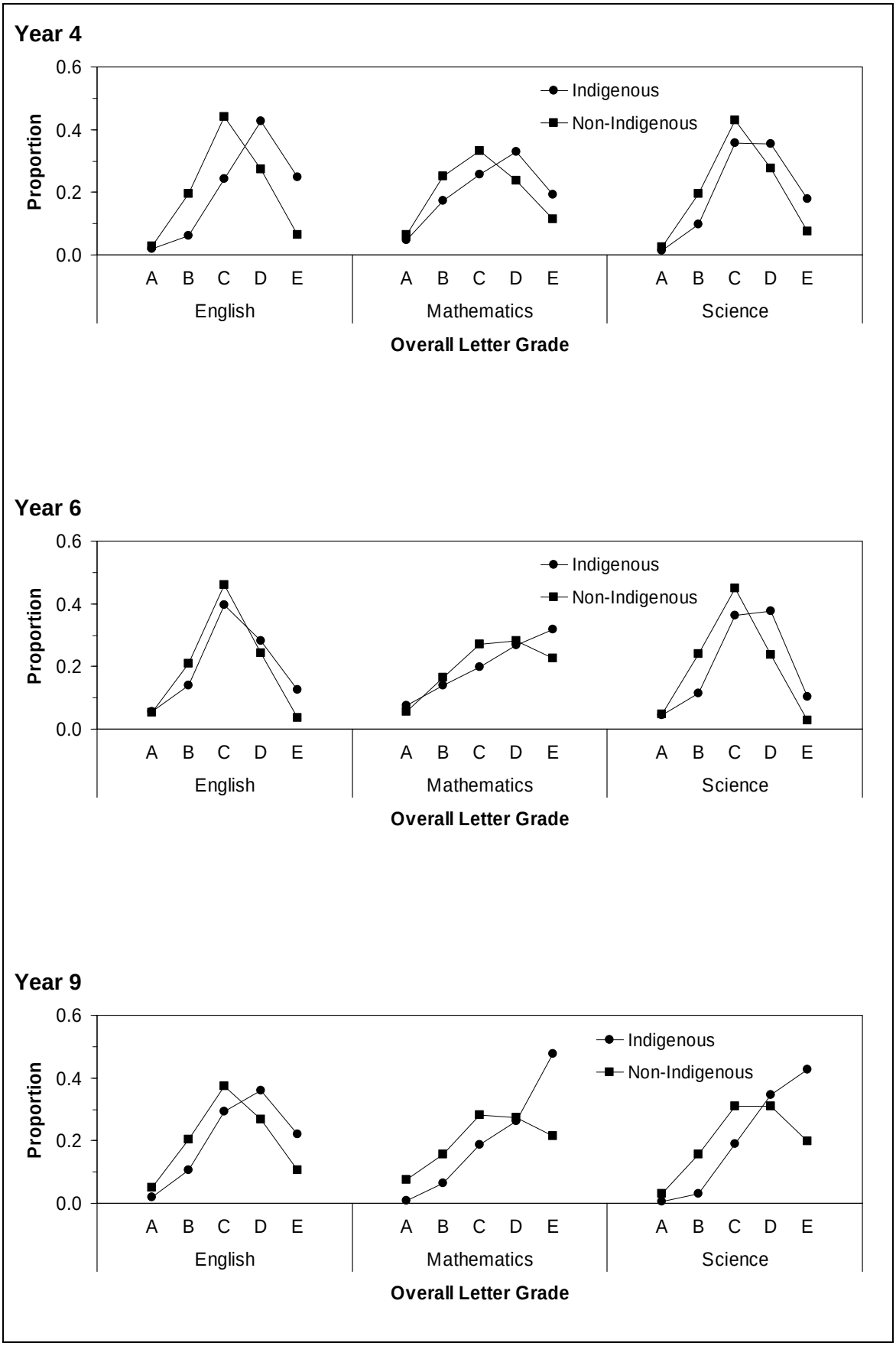


Figure 4: Distribution of responses across Overall Letter Grades for each QCAT separated by Indigenous status

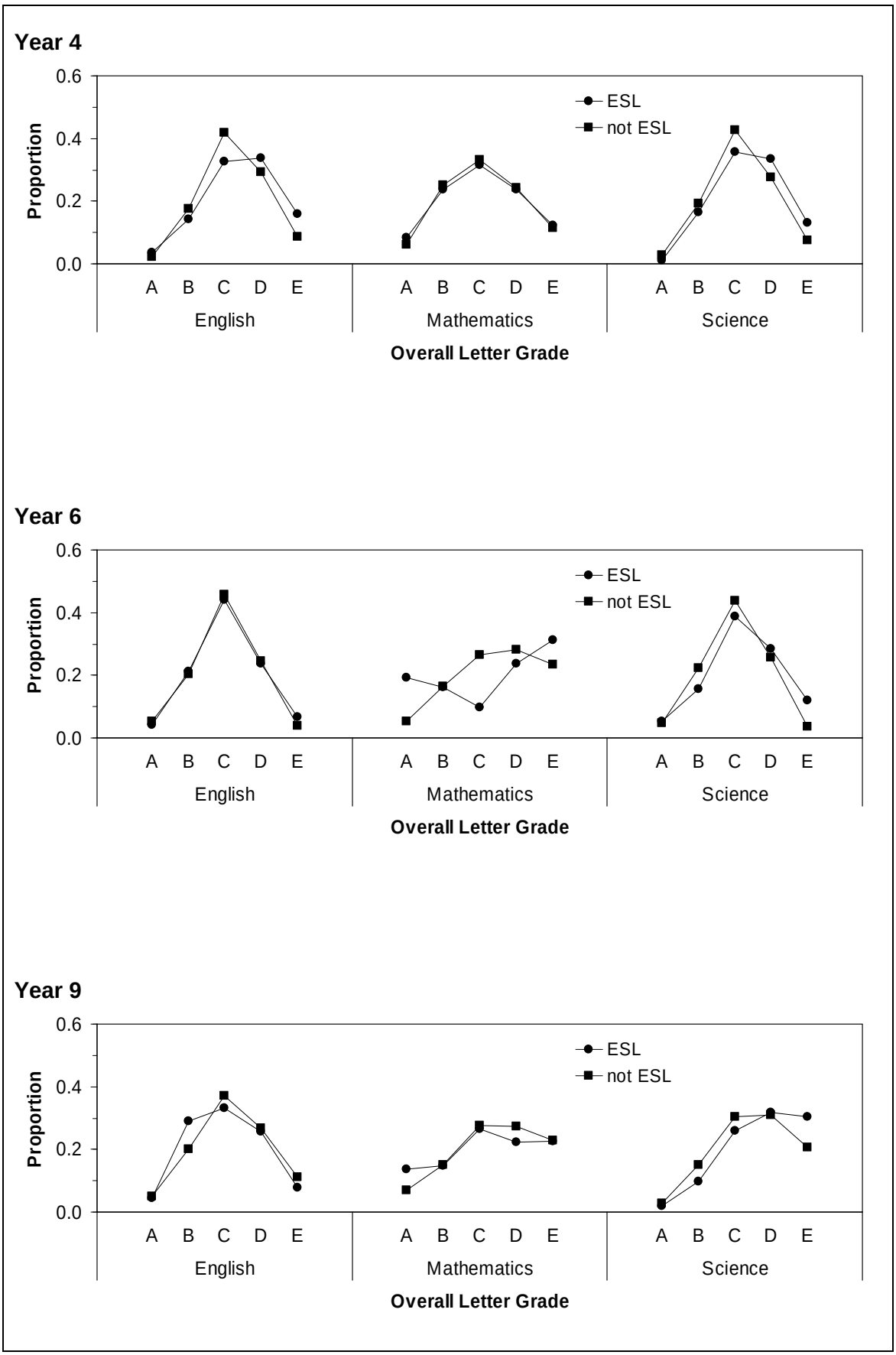


Figure 5: Distribution of responses across Overall Letter Grades for each QCAT separated by ESL status

awarded an E grade, but the proportion of Indigenous students receiving an E grade is even larger. It is noted, however, that with the exception of Year 9 Mathematics, the proportions of Indigenous and non-Indigenous students receiving an A grade are approximately equal.

Students' ESL status appears to have only a small association with students' achievements. With the exception of Year 6 Mathematics, not much separates the distributions for ESL and non-ESL students. For Year 6 Mathematics, the distribution is almost the reverse of a typical Normal distribution; that is, a small proportion of ESL students were awarded a C grade, but larger proportions were awarded letter grades at the extremes. This could mean that being an ESL student is not necessarily an impediment for high achieving students but it is an impediment for low achieving students. Across all the QCATs, the proportions of ESL students awarded an A grade are equal to or greater than the respective proportions of non-ESL students.

In summary, girls do better than boys in English but boys and girls appear to be doing equally well at Mathematics and Science; non-Indigenous students do better than Indigenous students; and ESL and non-ESL students for most QCATs appear to be doing equally well, but there is the exception for Year 6 Mathematics.

There is a methodological issue concerned with the data, or rather the presentation of the data, for this data collection. The data were available only in aggregated formats; that is, as counts of students who received each Overall Letter Grade aggregated for a given QCAT, but the counts could also be separated according to gender, Indigenous status, or ESL status. There are two concerns with data being presented in these formats. First, it might have been possible to determine the extent to which the distributions were separated according to gender, or Indigenous status, or ESL status, but it was not possible to test for significant interactions among these groupings of students.

Second, it is known that the data are hierarchically structured; that is, students are clustered within school. It is reasonable to assume that the schools students attend impact on students' attainments, and moreover, that these effects change from school to school. A consequence of the clustering is that dependencies can arise. Dependencies arise because students within a school share the common school environment, share the same teachers, are in direct communication with each other, come from similar

neighbourhoods, and so forth. Also, dependencies can arise because students in different classrooms might experience the implementation of the QCATs differently. Ignoring these dependencies can led to spurious significant effects, but in this data collection it is not known which students belong to which schools. Rather than run analyses in which there was no choice but to ignore the dependencies, and thus run the risk of claiming spurious significant effects, no analyses were run. Thus all the discussion and the charts presented in this section have to be treated as a presentation or discussion of descriptive statistics only, not discussions of inferential findings.

30-schools

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). Thus it is expected that there be a total of 150 results (5 returns X 30 schools) for each QCAT. However, there was some missing data. Table 1 shows where the missing data occurred for each QCAT. The rates are not large; they are all less than 10%, although the rates for Year 4 and Year 6 Science are close to 10%. One point to note, though, is that a small number of teachers for some QCATs could not distinguish between Overall Letter Grades.

As shown in Table 2, the rates for missing data for the Assessable Elements are considerably larger, in particular for Year 4 QCATs and for Year 6 and 9 Science. 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 in all five *Student booklets*; in some schools, there were no Letter Grades awarded for any Assessable Element in some *Student booklets*; 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 Grade including some instances in which up to four or five Letter Grades were indicated.

To repeat, 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 for all KLAs will be flat, but as can be seen in Figure 6, the distributions are not perfectly flat. Deviations from 'perfect flatness' were tested for statistical significance, and the results of the tests are shown in Table 3. Despite appearances to the contrary (from Figure 6), the Table suggests that only three

Table 1: Patterns of missing data for Overall Letter Grade for the 30-schools data collection

Year 4	**English**	7	No booklet (3 from one school)
		1	Could not distinguish between B and C
		1	No Overall Letter Grade even though each Assessable Element had a Letter Grade
		Total 9 (6.0%)	
Year 4	**Maths**	4	No booklet
		Total 4 (2.7%)	
Year 4	**Science**	6	No booklet (3 from one school)
		1	Could not distinguish between C and D
		3	Could not distinguish between D and E
		3	No Overall Letter Grade even though each Assessable Element had a Letter Grade
		Total 13 (8.7%)	
Year 6	**English**	5	One school did not return booklets (withdrew from the trial)
		Total 5 (3.3%)	
Year 6	**Maths**	3	No booklet
		1	Could not distinguish between D and E
		Total 4 (2.7%)	
Year 6	**Science**	8	No booklet (3 from one school)
		1	Could not distinguish between A and B
		1	Could not distinguish between D and E
		3	No Overall Letter Grade even though each Assessable Element had a Letter Grade
		Total 13 (8.7%)	
Year 9	**English**	5	No booklets
		1	No Overall Letter Grade even though each Assessable Element had a Letter Grade
		1	Could not distinguish between A and B
		Total 7 (4.7%)	
Year 9	**Maths**	1	No booklet
		Total 1 (0.7%)	
Year 9	**Science**	1	No booklet
		6	No Overall Letter Grade even though each Assessable Element had a Letter Grade Included here is 5 returns from one school
		Total 7 (4.7%)	

Table 2: Patterns of missing data for Letter Grade for Assessable Elements for the 30-schools data collection

Year 4	**English**	3 schools provided no Letter Grades for all Assessable Elements
	Maths	6 schools provided no Letter Grades for all Assessable Elements
		plus one school provided only one Letter Grade across its five booklets
	Science	2 schools provided no Letter Grades for all Assessable Elements
	In addition:	
Year 4	**English**	14 booklets were missing all Assessable Elements
		plus 4 instances of one or two Assessable Elements missing
		25 could not distinguish between Letter Grades
	Maths	23 booklets were missing all Assessable Elements
		plus 11 instances of one or two Assessable Elements missing
		35 could not distinguish between Letter Grades
	Science	10 booklets were missing all Assessable Elements
		plus 9 instances of one or two Assessable Elements missing
		16 could not distinguish between Letter Grades
Year 6	**English**	3 schools provided no Letter Grades for all Assessable Elements
	Maths	2 schools provided no Letter Grades for all Assessable Elements
	Science	0 schools provided no Letter Grades for all Assessable Elements
	In addition:	
	English	1 booklet was missing all Assessable Elements
		plus 5 instances of one or two Assessable Elements missing
		8 could not distinguish between letter grade
	Maths	12 booklets were missing all Assessable Elements
		14 could not distinguish between Letter Grades
		plus instances of one or two Assessable Elements missing
	Science	2 booklets were missing all Assessable Elements
		plus 3 instances of one or two Assessable Elements missing
		30 could not distinguish between Letter Grades
Year 9	**English**	0 schools provided no Letter Grades for all Assessable Elements
	Maths	2 schools provided no Letter Grades for all Assessable Elements
	Science	0 schools provided no Letter Grades for all Assessable Elements
	In addition:	
	English	3 booklets were missing all Assessable Elements
		plus 1 instance of one Assessable Element missing
		4 could not distinguish between Letter Grades
	Maths	7 booklets were missing all Assessable Elements
		plus 6 instances of one or two Assessable Elements missing
		16 could not distinguish between Letter Grades
	Science	5 booklets were missing all Assessable Elements
		29 could not distinguish between Letter Grades

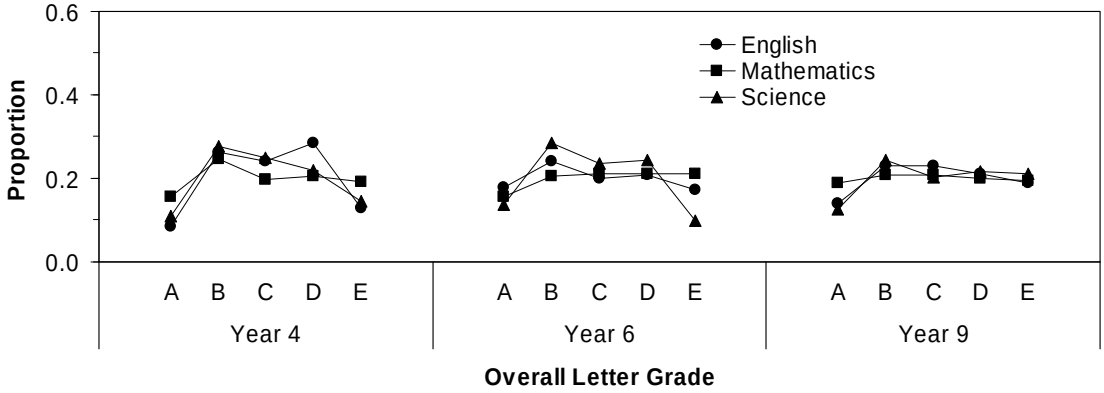


Figure 6: Distribution of responses across Overall Letter Grades for each QCAT for the 30-schools data collection

Table 3: Statistical test for equality of number of returns across the Letter Grades for the 30-schools data collection

Year Level	KLA	$\chi^2_{df=4}$	p
Year 4	English	21.9*	<0.001
	Mathematics	3.0	0.57
	Science	13.5*	0.009
Year 6	English	2.1	0.72
	Mathematics	1.7	0.80
	Science	17.2*	0.002
Year 9	English	4.1	0.4
	Mathematics	<1	
	Science	5.6	0.23

* Statistically significant χ^2

distributions (Year 4 English and Science, and Year 6 Science) deviate significantly for 'perfect flatness'.

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 therefore, Letter Grades A and E will occur less often while Letter Grades B, C, and D occur more often. It is noted that the doubling-up occurs most often with English and

Mathematics booklets, and it is the case that the Year 4 English distribution deviates substantially from 'perfect flatness', but the others that deviate from 'perfect flatness' are Science distributions not Mathematics distributions. even if the doubling up of letter grades could explain deviations from 'perfect flatness', questions remain: Why was it easier to find Letter Grades A and E in Science than in English and Mathematics?; and, Is it reasonable that more than one-third of schools could not find Letter Grades A and E in English and Mathematics?

Figure 7 shows the pattern of Letter Grades awarded for Assessable Elements within an Overall Letter Grade. Consider the top-left graph which shows the pattern for Year 4 English. When an Overall Letter Grade of A was awarded, the most likely Letter Grade for any Assessable Element was an A. Similarly, when an Overall Letter Grade of C was awarded, the most likely Letter Grade for any Assessable Element was a C. The are similar patterns for Letter Grades D and E. That the Letter Grade for the Assessable Element aligns mostly with the Overall Letter Grade is a reasonable pattern. However, the pattern when Overall Letter Grade is B is somewhat different - for the 4th Assessable Element, there were almost equal numbers of Letter Grades B and C. An examination of the *Guide to making judgements* for Year 4 English suggests a possible explanation - there was no descriptor aligning with Letter Grade B. Perhaps some teachers found it difficult to award a Letter Grades in those situations where no descriptor aligned with the Letter Grade.

For other QCATs, it was the case that large numbers for consecutive pairs of Letter Grades occurred when descriptor does not align with one or the other of the Letter Grades (for instance: the 2nd Assessable Element for Year 4 Mathematics; the 1st Assessable Element for Year 6 Mathematics; the 1st and 2nd Assessable Elements for Year 6 Science; the 1st Assessable Element for Year 9 English; and the 1st Assessable Element for Year 9 Mathematics). However, the pattern is not consistent; for instance, no descriptor aligns with B for the 1st, 2nd and 3rd Assessable Elements for Year 9 Science, yet teachers had no difficulty in awarding B for the Assessable Elements. The 3rd Assessable Element provides an example of a pattern that appears to work in reverse – despite there being a descriptor aligning with C, there were few Cs awarded. Another unexpected pattern is shown for the 3rd Assessable Element for Year 4 Science: there were few As when the Overall Letter Grade was B (on its own, not an unusual result), but a large number of As when the Overall Letter Grade was C. That

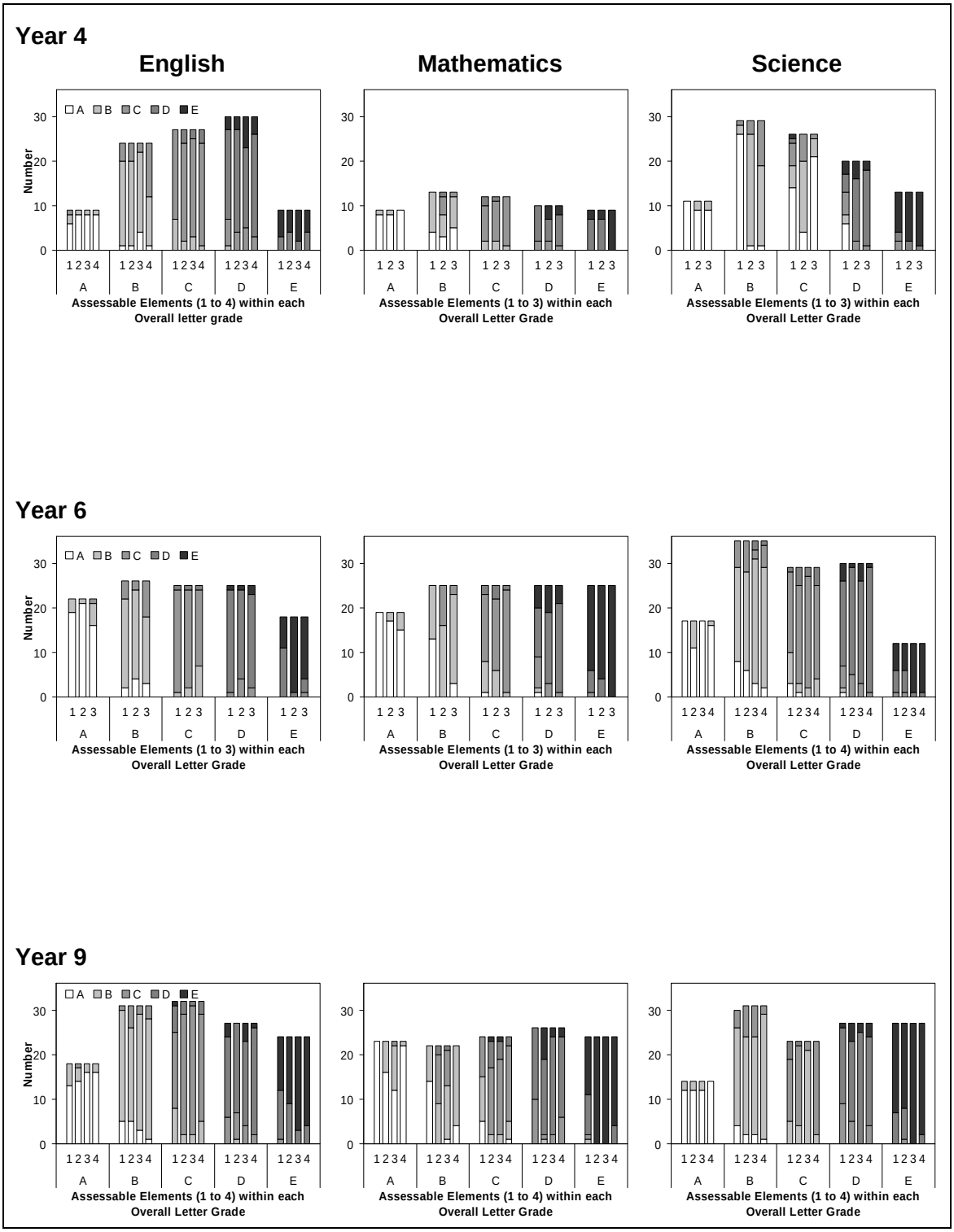


Figure 7: Pattern of Letter Grades awarded for Assessable Elements within each Overall Letter Grade for each QCAT

descriptors do not always align with a Letter Grade provides some explanations for the patterns evident in Figure 7, but it does not provide the complete explanation. For now, all that can be said is that, mostly, the Letter Grades for an Assessable Element usually aligns with the Overall Letter Grade, but there are other patterns evident in Figure 7 that are difficult to explain.

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 = 1, B = 2, and so on through to E = 5). The regressions were multilevel because 'booklets' (i.e., students) were nested within 'teachers'; that is, the regressions take account of the dependencies that can arise when students are clustered with schools. Nine multilevel regressions analyses were run - one for each QCAT. The standardised regression coefficients are shown in Table 4, but interpreting the coefficients has to proceed with some caution. First, the number of teachers in each analysis is small by the standards required by multilevel regression (although this was somewhat compensated for by using what is referred to as MCMC estimation); second, the number of booklets was also small for some schools because of missing data; and third, the correlations among the Assessable Elements were high which means that, together with the small numbers, the coefficients are unstable.

To read the importance of Assessable Elements, consider the standardised coefficients for Year 4 English in Table 4: an increase of one standard deviation for 1st Assessable Element leads to, on average, an increase in the Overall Letter Grade of 0.18 standard deviations; an increase of one standard deviation for 2nd Assessable Element leads to an increase in the Overall Letter Grade of 0.27 standard deviations; an increase of one standard deviation for the 3rd Assessable Element leads to an increase of 0.36 standard deviations for the Overall Letter Grade; and an increase of one standard deviation in 4th Assessable Element leads to an increase of 0.19 standard deviations for the Overall Letter Grade. Thus, more importance is assigned to the 3rd Assessable Element, and less importance is assigned to the 1st and 4th Assessable Element.

A consistent finding across the QCATs is that the 'Knowledge and understanding' Assessable Elements have, if not the least importance assigned to them, a small importance assigned to them. Also, for English at any of the three year levels, the most important Assessable Element is one that has an element of 'Text construction' attached to it; and for Mathematics at any of the three year levels, the most important Assessable Element is 'Communicating'. But the difficulty with labelling the 'most important' Assessable Element is that the labels for the Assessable Elements change across the three KLAs and, to a smaller extent, across the three year levels within a KLA.

Table 4: Relative importance assigned to each Assessable Element by teachers when deciding the Overall Letter Grade

Year	KLA	AE1	AE2	AE3	AE4
4	English	.18	.27	.36	.19
	Mathematics	.23	.26	.57	
	Science	.20	.45	.41	
6	English	.30	.49	.24	
	Mathematics	.13	.29	.59	
	Science	.20	.18	.29	.40
9	English	.15	.21	.33	.35
	Mathematics	.20	.20	.29	.36
	Science	.21	.19	.36	.28

There are two major difficulties for the analyses in the 30-schools data collection. First, the data has a hierarchical structure for the same reasons that the 1300-schools data collection has a hierarchical structure. Unlike the situation with the 1300-schools data collection, the clustering structure of students within schools in 30-schools data collection was known. The difficulty for the multilevel analyses analysis was that the number of schools was small. This was exacerbated by the second difficulty - the missing data for the Assessable Element - which made the already small sample size even smaller. Small sample sizes coupled with high correlations among the Assessable Elements result in unstable estimates (i.e., the uncertainties surrounding the estimates in Table 4 are large). These problems noted, the indications are that some teachers found it difficult to award a single Letter Grades for as Assessable Element, and awarded instead either multiple letter grades or no letter grades at all. One reason for the difficulty could

be that in the *Guide to making judgements* letter lrades did not always align with descriptors; but that was not the complete explanation. For the teachers who did provide Letter Grades for all Assessable Element, they assigned relatively less importance to the 'Knowledge and understanding' Assessable Element when making their judgement of the Overall Letter Grade.

Double marking of QCATs from 10 schools

In this section, the analyses are concerned with the agreement achieved by a pair of independent and trained 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 ten schools, a sub-sample of the 30-schools data collection.

Figure 8 gives a visual representation of the consistency achieved by the two markers when awarding Overall Letter Grades for each QCAT. To read Figure 8, note that each point is represented by a cloud of points. Consider the point represented by the coordinates (B, B) in the scatterplot for Year 4 English. There are five booklets represented by (B, B), which means that for five *Student booklets*, the two markers agreed when awarding the B grade. If the five booklets were instead to be represented by single point, information would be lost– the information about there being five booklets. In the scatterplot, each point has been jittered. Jittering mean 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, interested 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, and even fewer points displaced further off the diagonal. For Year 4 English, it is clear that the two markers were fairly consistent. The other scatterplots in Figure 8 indicate that the pairs of markers for the other QCATs were also fairly consistent.

Figure 9 shows the scatterplots comparing the Overall Letter Grade awarded at the schools with the consensus grade of the two markers. For most scatterplots in Figure 9,

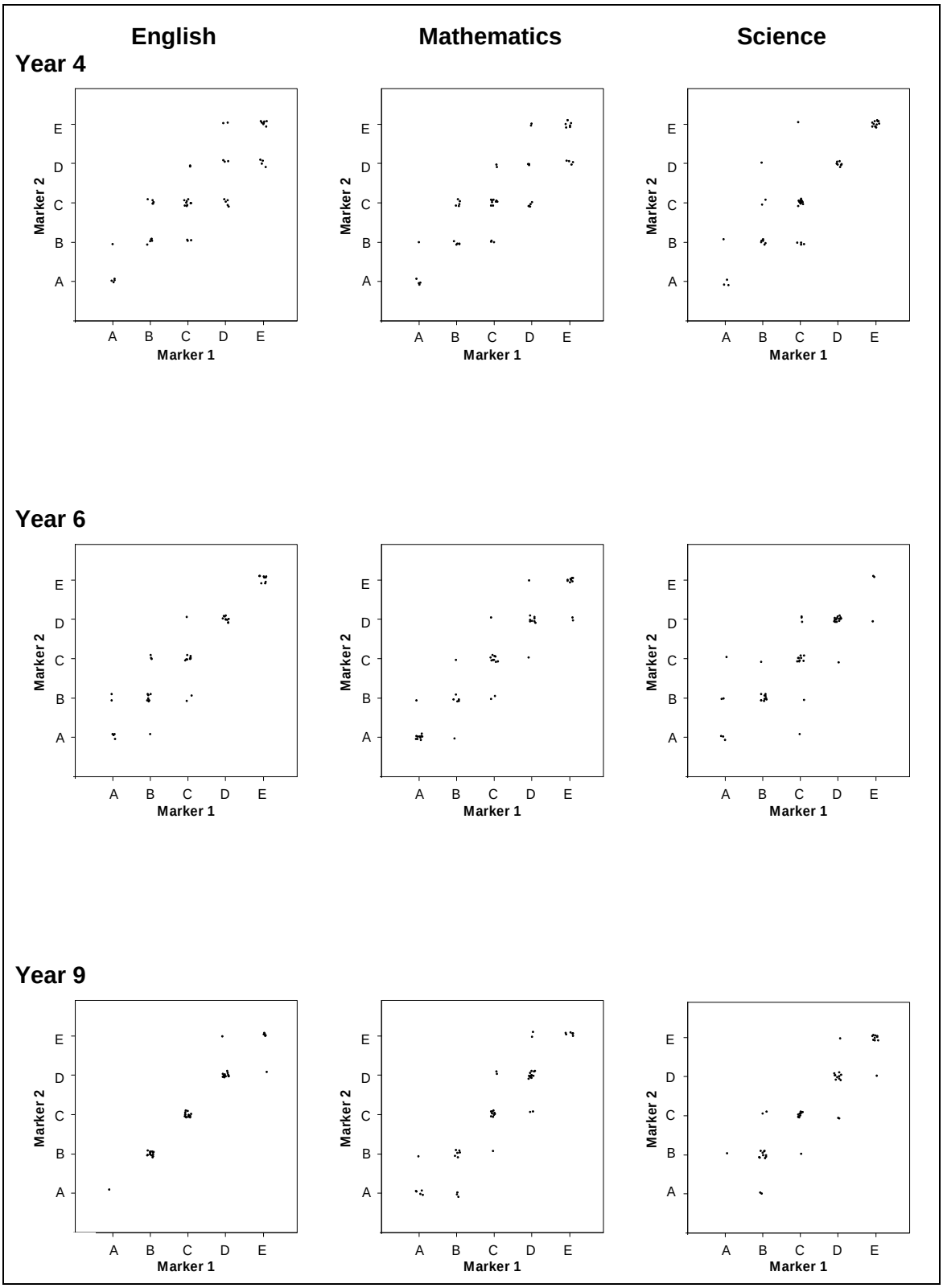


Figure 8: Scatterplots showing the degree of consistency between the two markers for each QCAT

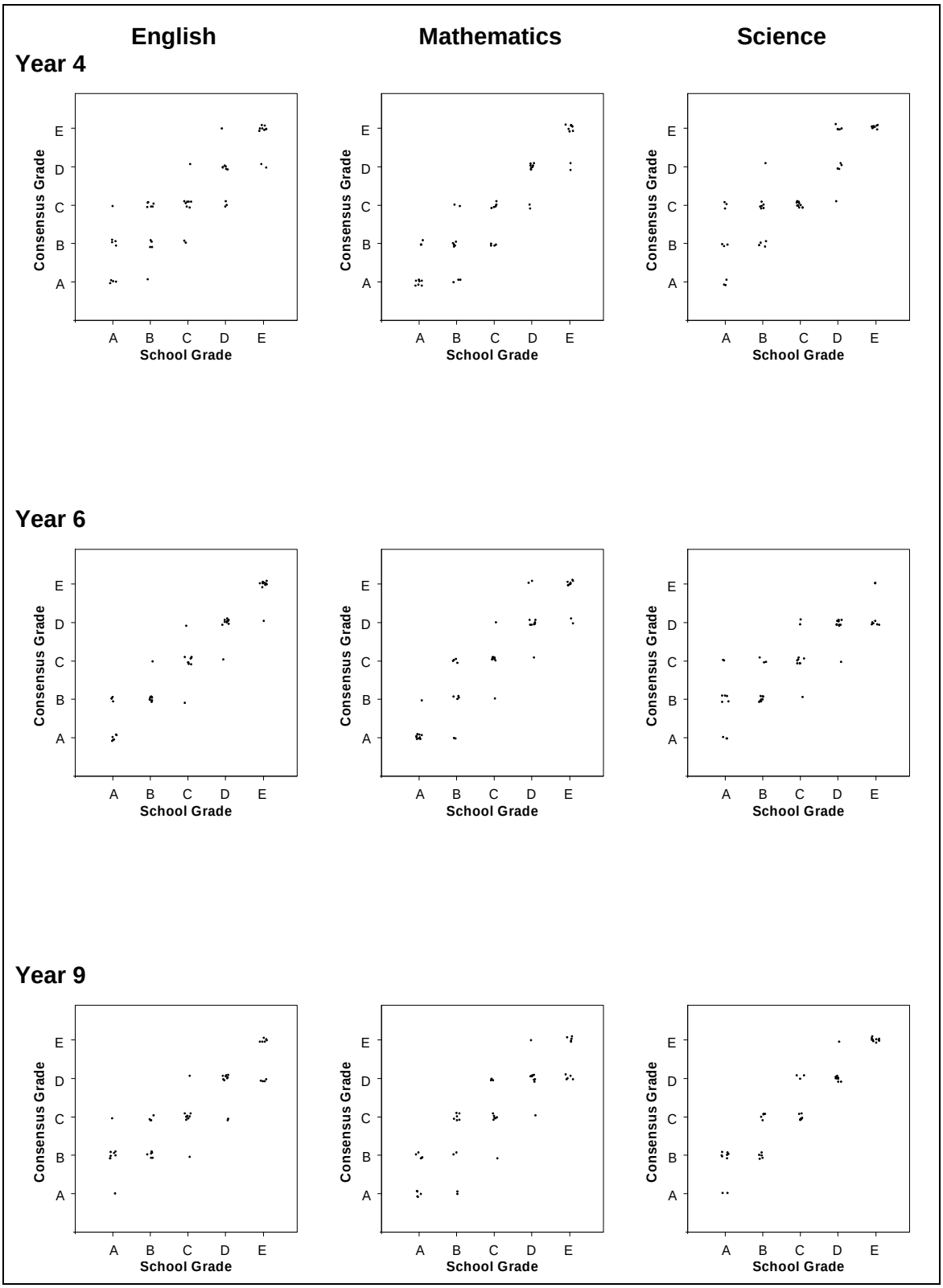


Figure 9: Scatterplots showing the degree of consistency between markers (consensus grade) and teachers (school grade) for each QCAT

there are few large discrepancies between the two sets of Overall Letter Grades. Also, when comparing a scatterplot in Figure 9 with the scatterplot for the corresponding QCAT in Figure 8, it appears that the consistency between teachers and markers is only a little worse, if any at all, than that achieved by the two markers.

The consistency between the two markers can be quantified. Cohen's κ is a measure of inter-rater agreement when two raters are rating objects. Usually, Cohen's κ 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 (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 assessments. The method used here is linear weighting. Cohen's κ 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 κ is[2]:

< 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 10 shows the values for Cohen's κ for all QCATs for the two sets of comparisons (between the two markers, and between the consensus grade and the grade awarded at the schools). It is noted that all values are in the "Good" to "Very good" range. Nevertheless, the agreement between the two markers was less than "Very good" for Year 4 English and Mathematics, and for Year 6 Science. Also, the agreement between the consensus grade and the grade awarded at the schools was less than "Very good" for Year 4 English and Science, for Year 6 Science, and for Year 9 English, Mathematics and Science.

Before turning to the question of consistency when awarding 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 ten schools that comprise this data collection are a sub-sample of the 30-school data collection. Table 5 shows where the missing data occurred for each QCAT. As was the case with the 30-schools data collection, there were instances of Letter Grades missing for one, two or

[2] Altman, D. (1991). *Practical statistic for medical research*. London: Chapman & Hall.

three Assessable Elements, and there were responses for which teachers could not distinguish between Letter Grade including some instances in which up to four or five Letter Grades were indicated. The rates for missing data appear to be largest for Year 4. More Year 4 teachers than teachers of other year levels did not indicate a Letter Grade for any Assessable Element or could not distinguish between Letter Grades. For Year 6, most of the missing data appears to be a consequence of two schools (10 booklets) not providing data for any Assessable Element; otherwise, the rates for Year 6 appear to be small. The rates for missing data for the three KLAs for Year 9 were minimal. The missing data means, particularly for the Year 4 QCATs, that the sample sizes are reduced, increasing the likelihood that the estimations of agreement between markers and teachers are unstable.

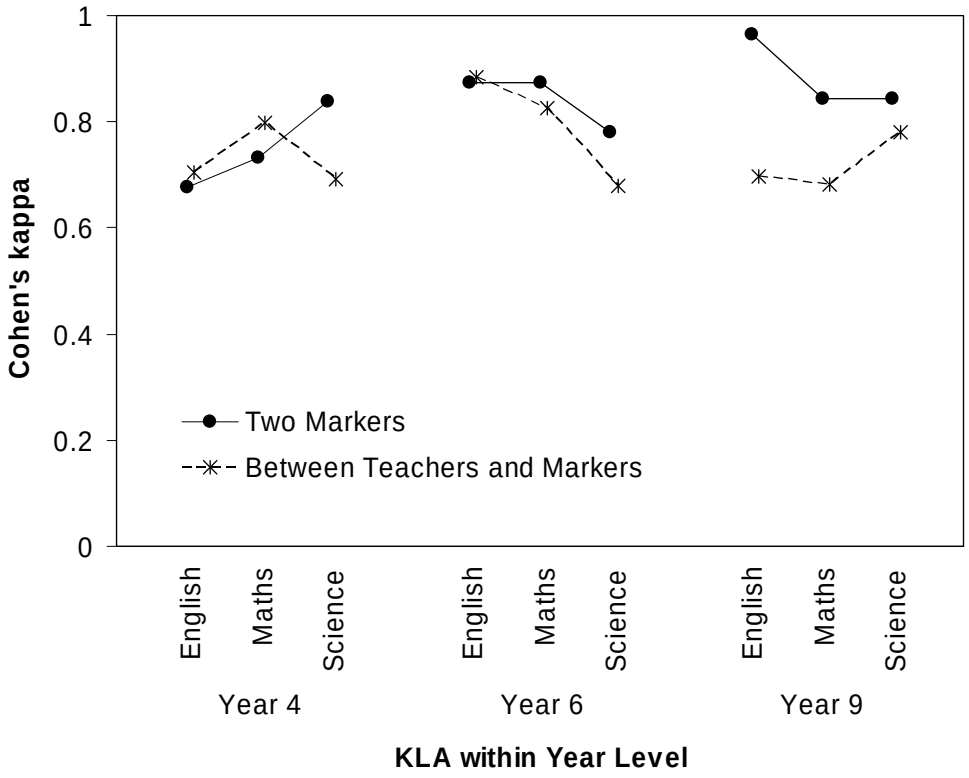


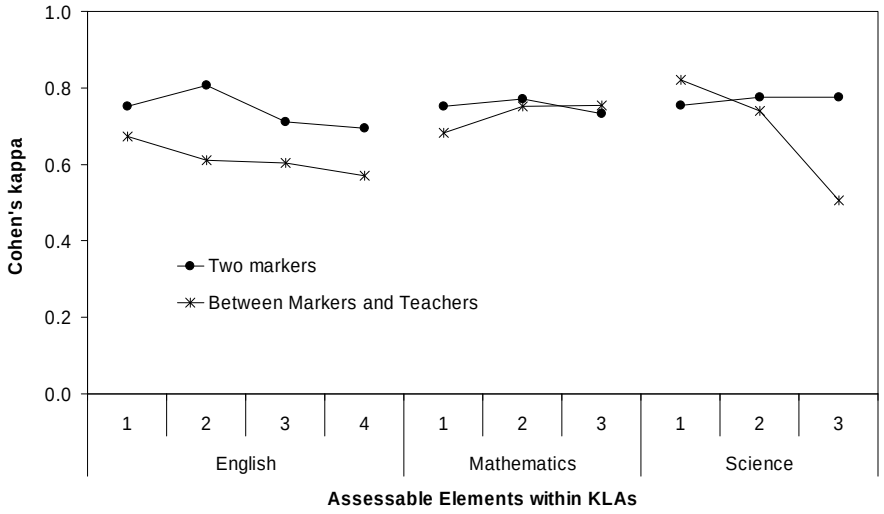
Figure 10: Coefficient of agreement (Cohen's κ) between the two makers and between teachers and markers when awarding the Overall Letter Grade for each QCAT

Figure 11 shows that for most Assessable Elements the two markers achieved "Good" to "Very good" agreement. A Cohen's kappa of 0.7 is usually taken as indicating "satisfactory agreement" between raters. Thus, for the three KLAs at Year 4, the agreement between the two markers was satisfactory or not far from it. For Year 6, there

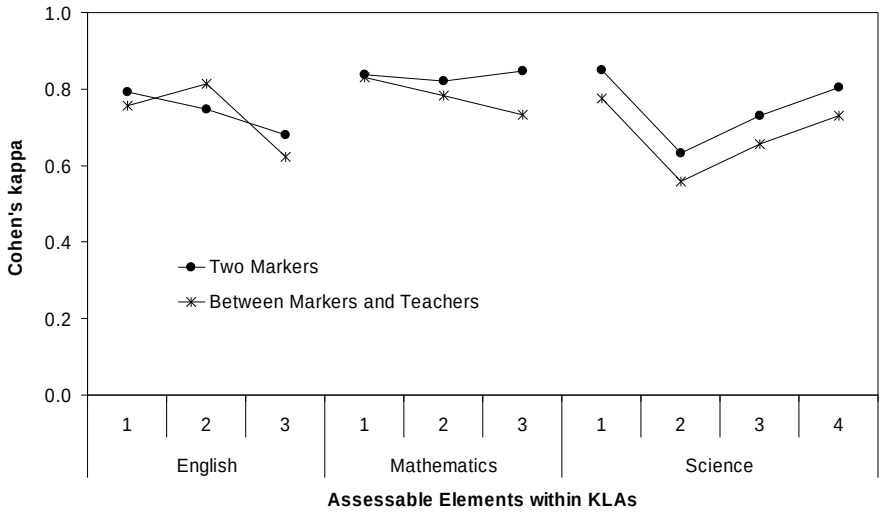
Table 5: Patterns of missing data for Overall Letter Grades and for Letter Grade for Assessable Elements for the 10-schools data collection

Year 4	**English**	1 school provided no letter grades for all Assessable Elements
	Maths	1 school provided no letter grades for all Assessable Elements
	Science	1 school provided no letter grades for all Assessable elements
	In addition:	
	English	4 booklets were missing all Assessable Elements plus 10 instances of one or two Assessable Elements missing 10 could not distinguish between letter grades
	Maths	4 booklets were missing all Assessable Elements plus 5 instances of one or two Assessable Elements missing 21 could not distinguish between letter grades
	Science	5 instances of one or two Assessable Elements missing 9 could not distinguish between letter grades
Year 6	**English**	2 schools provided no letter grades for all Assessable Elements
	In addition:	
	English	1 booklet was missing all Assessable Elements plus 1 instance of one or two Assessable Elements missing 3 could not distinguish between letter grades
	Maths	5 booklets were missing all Assessable Elements plus 1 instance of one or two Assessable Elements missing 7 could not distinguish between letter grades
	Science	1 booklet were missing all Assessable Elements 6 could not distinguish between letter grades.
Year 9 Maths		1 school provided no letter grades for all Assessable Elements
	In addition:	
	English	2 booklets were missing all Assessable Elements 4 could not distinguish between letter grades.
	Maths	1 booklet was missing all Assessable Elements plus 3 instances of one or two Assessable Elements missing 1 could not distinguish between letter grades
	Science	2 could not distinguish between letter grades.

Year 4



Year 6



Year 9

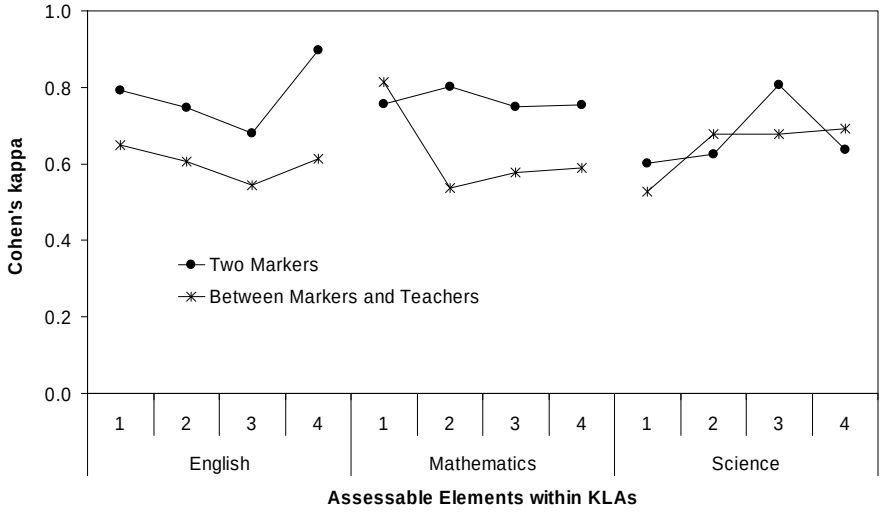


Figure 11: Coefficient of agreement (Cohen's κ) between the two makers and between teachers and markers when awarding the Letter Grade for the Assessable Elements for each QCAT

is a drop in agreement for the 2nd Assessable Element for Year 6 Science, and for the 3rd Assessable Element for Year 6 English, but otherwise, there was "satisfactory" agreement between the two markers. Apart from three of the four Science Assessable Elements (1st, 2nd and 4th), there was "satisfactory" agreement between the two markers for Year 9 English and Mathematics Assessable Elements.

When comparing teachers and markers, and keeping the cautionary note from above in mind, the agreement between teachers and markers is, with few exceptions, less than the agreement between the two markers. Furthermore, the agreement between teachers and markers is less than "satisfactory" for:

- the four Assessable Elements for Year 4 English (especially the 2nd, 3rd and 4th Assessable Elements);
- the 4th Assessable Element for Year 4 Science;
- the 3rd Assessable Element for Year 6 English and the 2nd Assessable Element for Year 6 Science;
- nearly all Year 9 Assessable Elements (the exception being the 1st Mathematics Assessable Element) although one or two others are close

In summary, the markers were achieving satisfactory agreement when awarding Overall Letter Grades and mostly satisfactory agreement 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. It is with respect to the Assessable Elements that the two groups were not achieving satisfactory agreement. If the rates for missing data can be taken as expressions of uncertainty among the teachers (especially those instances when multiple grades were awarded for Assessable Elements) then it is likely that there would be even less agreement between teachers and markers.

Focus group sessions

At the conclusion of their marking, the markers attended focus group sessions to discuss any difficulties that arose during the marking. They claimed that they experienced the greatest difficulty in awarding a Letter Grade when:

- Instructions to students in the *Student booklet* were not explicit which meant that the markers had to make decisions.

- No descriptor aligned with a letter grade (the difficulty had less to do with interpolating between the descriptors and more to do with borderline grades - for instance, when there no descriptor aligning with, say, B, it was difficult to award a Letter Grade when the response was on the borderline between A and B or on the borderline between B and C).
- Responses did not fit with the descriptors or when responses did not quite answer the question.

Markers claimed to depend heavily upon the *Guide to making judgements*. At times, however, the *Guide to making judgements* did not contain sufficient information to make a judgement; or at other times, explicit instructions in the *Student Booklet* were not carried over into the *Guide to making judgements*. In addition, the markers argued that the *Guide to making judgements* and the *Sample responses* needed to cover a wider range of possible answers. They claimed that difficulties lay not so much with straight forward responses; rather, they needed more examples of responses to guide their judgements when the *Student booklets* contained unexpected responses, or when the responses had the potential to be complex (such as Question 5 in Year 6 Science), or when students' responses deviated from the expected context.

Other areas of concern raised by the markers included:

- Assessable Elements that drew on information from a number of questions (for example, in Year 9 Mathematics, markers had to draw upon questions spread across the *Student booklet* when grading the 4th Assessable Element);
- Assessable Elements that were not sufficiently discrete to make an adequate judgement, especially Assessable Elements that draw on information from a number of questions. As an example, if a student made a conceptual mistake with one question, a grade of, say, C was not available because that descriptor refers to "Minor mechanical errors", but the markers believe the remaining questions could be A or B grade;
- Awarding an Overall Letter Grade when the Letter Grades for the Assessable Elements differed (i.e., their concern lay with giving weights to Assessable Elements).
- A concern raised in connection to Science but could potentially be relevant to the other KLAs – the descriptors for Assessable Elements dealing with "Communication" specify the use of correct and appropriate scientific

terminology but descriptors make no mention of the correctness of the content of the communication.

Despite the difficulties they encountered, the markers claimed to be fairly consistent; more so when awarding Overall Letter Grades than when awarding Letter Grades for Assessable Elements. On the whole, their assessment of their consistency aligns with the agreement measures (Cohen's kappa) presented earlier (compare Figure 10, which shows agreement between markers when awarding the Overall Letter Grade, with Figure 11, which shows their agreement when awarding Letter Grade for Assessable Elements). The only exception might apply to the Year 4 English QCAT, where there was less agreement when awarding Overall Letter Grades than was the case when awarding letter grades for some Assessable Elements. When discrepancies did occur, the markers claimed that mostly they were concerned with borderline grades. Some markers attributed their high rate of agreement to the fact that, initially, they worked closely together to reach consensus with potentially difficult responses.

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 ten schools. 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, overall, there was little evidence to indicate widespread use of alternative methods. Nevertheless, in some booklets, teachers used methods other than or in addition to QSA's descriptors to award letter grades; in other booklets, teachers awarded letter or numeric grade 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 higher order skills such as explaining, justifying, and applying knowledge in different domains. Generally, questions that depended on knowledge, recall and understanding were answered better than questions that depended upon justification, application to new areas, and consolidation.

Survey

A total of 480 surveys were returned, representing 281 schools. Figure 12 shows the distribution of returned surveys across KLAs and year levels. The number of surveys returned by teachers responding with respect to Year 3 Science and Year 9 Mathematics was approximately twice that of any other QCAT. The majority of surveys (83%) were received from State schools, with smaller numbers received from Catholic schools (5%) and Independent schools (11%). Most surveys were received from Primary schools (54%) or Secondary schools (28%), with smaller numbers from P-to-10 or P-to-12 schools (17%), and even smaller numbers from Special schools (2%).

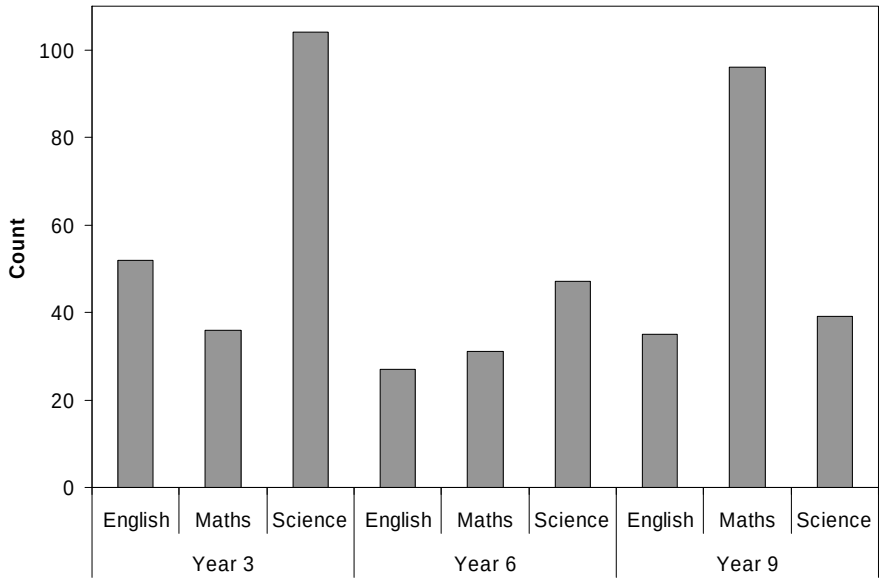


Figure 12: Distribution of returned surveys across KLAs and year levels

Responses to the questions concerning what teachers did to prepare for implementation of the QCAT and what teachers did to prepare students for the QCAT were varied, but included:

- Read the documents;
- Familiarise, discussed with others, meeting with others;
- Worked through the QCAT;
- Attended professional development or information sessions prepared at the cluster level;
- Accessed and \ or obtained resources;

- With students, outlined the purpose of QCAT; discussed time management, how to focus on necessary information and write to a stimulus; and outlined "test" procedures;
- Prepared practice QCATs;
- Modified the teaching program; and /or revised content;
- And there was a number who claimed the did "Nothing", "continued with normal classroom activities", "no preparation", treated it as "another unit of work"; "not doing coaching".

Clearly, some teachers went beyond the suggestions contained in the *Teacher Guidelines*.

The survey contained four questions concerned with the amount of time spent preparing, contextualising and implementing the QCAT. There were statistically significant effects for KLA and year level for three of the questions, and the fourth was not far from statistical significance. With respect to "Time spent preparing students for the QCAT," there was a statistically significant χ^2 obtained from a Kruskal-Wallis test ($\chi^2(8) = 20.0$, $p = 0.010$). Figure 13 shows the distribution of responses for "Time spent preparing students for the QCAT" across the three response categories for each KLA and year level. At the bottom of the Figure, the "Overall" bar show the proportion of teachers, as a percentage, who ticked each time category (30 minutes, 1 hour, more than 1 hour) but note that the "Overall" bar does not take account of KLA or Year level. Above the "Overall" bar, the bars show the proportion of teachers, as a percentage, who ticked each time category within KLAs within year levels. It can be seen that Year 6 Science, and possible Year 6 English are driving the statistically significant effect noted above – overall, a large proportion of teachers ticked the "More than 1 hour" category, but an even larger proportions ticked that category if they were responding with respect to Year 6 Science and possibly Year 6 English. That is, the tendency was for Year 6 Science teachers and possibly Year 6 English teachers to spend more time preparing students for the QCAT.

With respect to the "Time spent contextualising the QCAT with students", the effect did not reach statistical significance (in a Kruskal-Wallis test, $\chi^2(8) = 15.1$, $p = 0.057$), but it was not far from statistical significance. Figure 14 shows the distribution of responses for "Time spent contextualising" across the three response categories for each KLA and year level. The Figure is structured the same way as Figure 13. Overall, a large

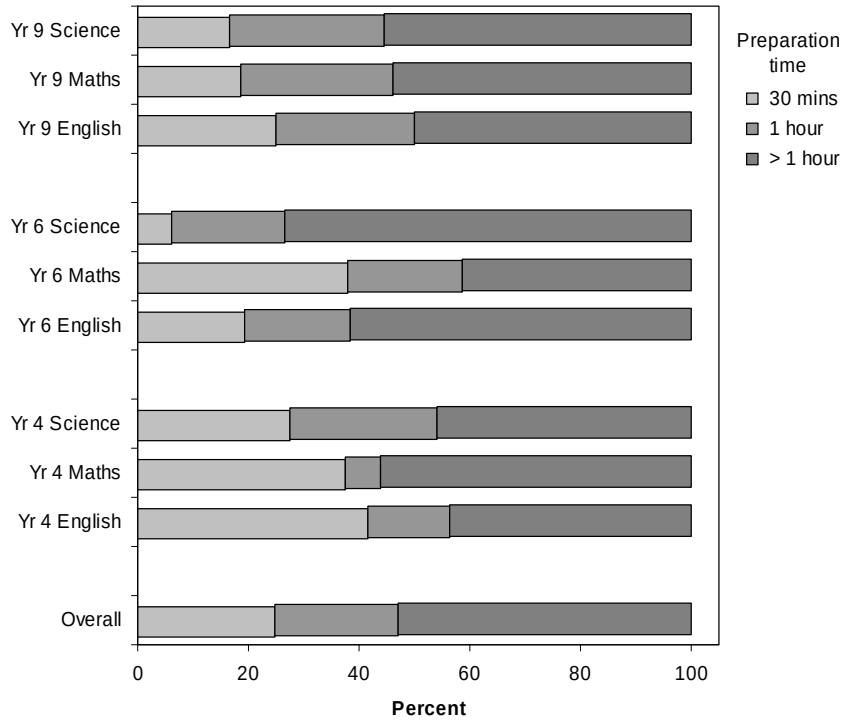


Figure 13: Time spent preparing students for the QCAT (by year level and KLA)

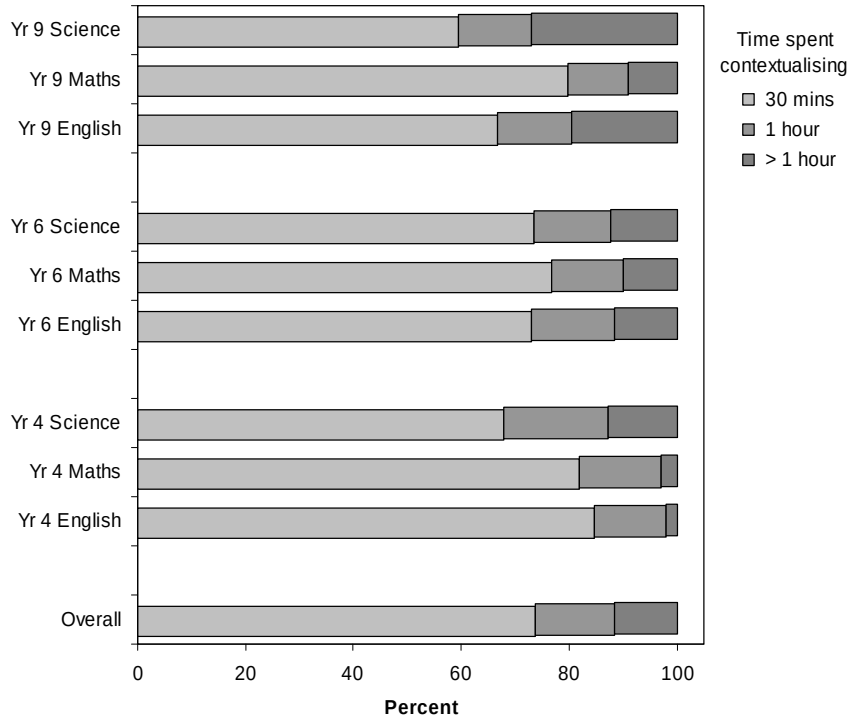


Figure 14: Time spent contextualising the QCAT with students (by year level and KLA)

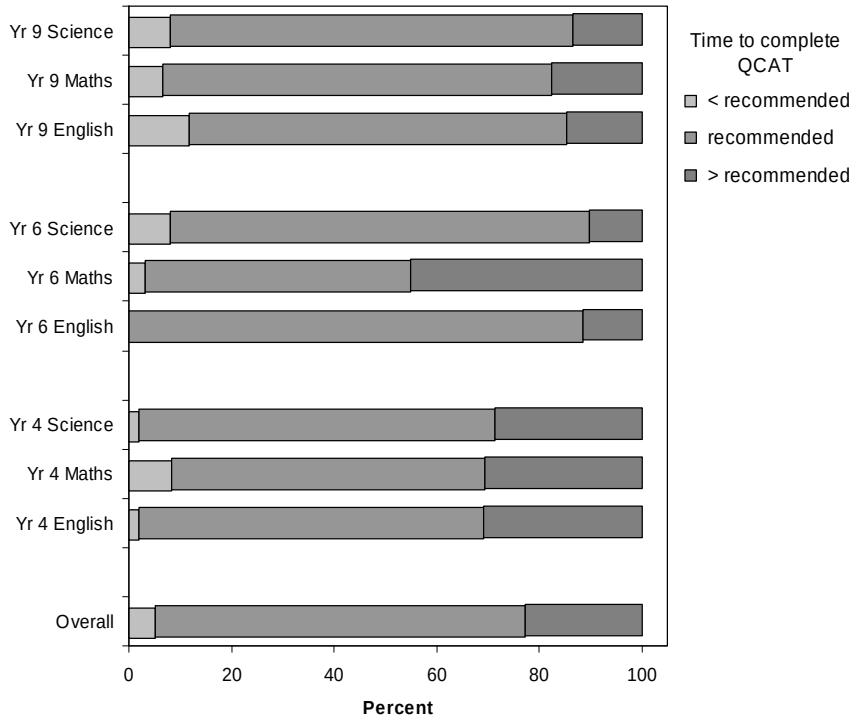


Figure 15: Time taken by students to complete the QCAT (by year level and KLA)

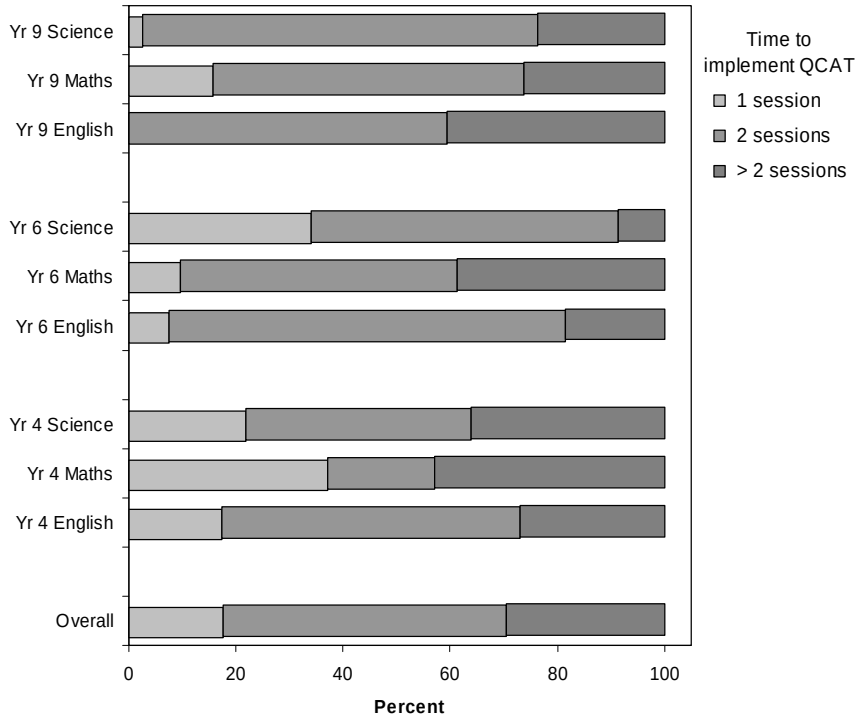


Figure 16: Number of sessions taken to implement the QCAT (by year level and KLA)

proportion of teachers ticked "30 minutes", but an even larger proportion ticked the "30 minute" category if the teachers were responding with respect to Year 4 English and Mathematics, and a smaller proportion ticked the "30 minute" category if the teachers were responding with respect to Year 9 Science. That is, the tendency was for Year 4 English and Mathematics teachers to spend less time contextualising the QCAT with students, and for teachers of Year 9 Science to spend more time contextualising the QCAT. It is noted though that these effects are not large and indeed they may be a result of no more random variation.

Similar tests were performed on the response distributions for the questions concerning "Time students took to complete the QCAT" and the "Number of sessions taken to implement the QCAT." In both cases there were statistically significant effects for KLA and year level ("Time students took to complete" – in a Kruskal-Wallis test, $\chi^2(8) = 25.9$, $p = 0.001$; "Number of sessions" - in a Kruskal-Wallis test, $\chi^2(8) = 23.6$, $p = 0.003$). Figures 15 and 16 show the distributions across the three response categories for each KLA and year level. From Figure 15 it can be seen that most teachers claimed that students took about the recommended time to complete the QCAT, but if the teachers were responding with respect to Year 6 Mathematics, fewer teachers claimed that the students took the recommended time and more teachers claimed that the students took more than the recommended time. With respect to the "Number of sessions to implement QCAT" (Figure 16), it appears that the Year 6 Science QCAT, on the whole, took less time to implement, and the Year 9 English QCAT, on the whole, took more time to implement.

A little more than half the teachers (52%) reported that there were students who did not undertake the QCAT because of absence from school, 8% because of "special considerations", and 1% because of some other reason.

There was a series of questions asking teachers' about the QCAT documents: *Teacher guidelines*, *Student booklet*, *Guide to making judgements*, and *Sample responses*. Figures 17, 18, 19, and 20 give the mean ratings for each statement about each document in turn. Each figure shows the ratings separated by KLA and 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, and that the suggested level of support to students was appropriate. The conclusion holds across the year levels and across the KLAs, with the exception that teachers of Year 6 and Year 9 Mathematics expressed somewhat less agreement for the last statement (*Suggested level of support to students was appropriate*). This conclusion is evident in Figure 17, but the conclusion is not based solely on the evidence in the Figure. A series of MANOVAs (Multivariate ANalysis Of Variance) were run – one for each question. The results of the MANOVAs are presented in summary tables contained in Appendix 3. The conclusions above, and the conclusions regarding the other documents, are based on the evidence derived from the MANOVAs, and the figures provide illustrations.

With respect to the *Student booklet* (Figure 18), there are two notable features. First there is a general decline in teachers' agreement with the statement concerning "*Students understood what they were expected to do*", particularly for Years 6 and 9 Mathematics and Science teachers and Year 4 Mathematics teachers. Second, Year 9 Mathematics and Science teachers were more critical of QCATs' capacity to engage students.

It seems that, on the whole, teachers were more critical of the *Guide to making judgements* (Figure 19) than the other documents; more so for teachers of Year 9 Science.

Year 9 Science teachers were critical of the *Sample answers* (Figure 20); in particular, Year 9 Science teachers did not agree with the first statement: *The sample responses provided clear examples of the quality expected in student work.*

Also, there are statistically significant effects for each statement in the question about the use to which data gathered from the QCAT implementation will be put, but, as can be seen in Figure 21, the effects are small. If there is any one trend that stands out, it is that Year 9 Science teachers are the least sure that the usefulness of the data generated by the QCAT implementation.

Even though statistical significance was achieved for most of the statements, the size of the effects were generally small (less than 10% - see η^2 in Appendix 3). There were just three exceptions:

- *The suggested level of support to students was appropriate*;

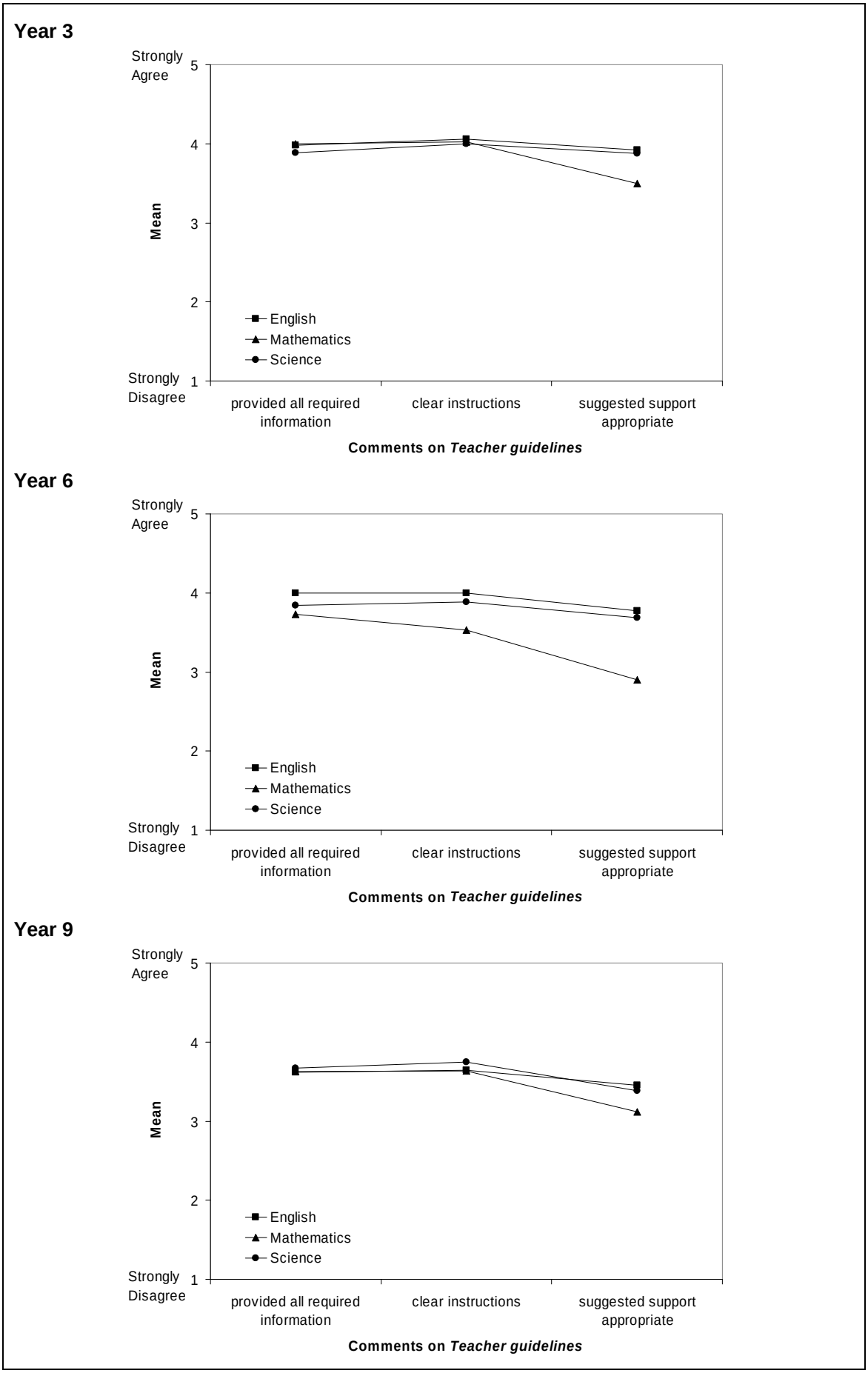


Figure 17: Mean ratings for three items dealing with teachers' perceptions of the Teacher Guidelines

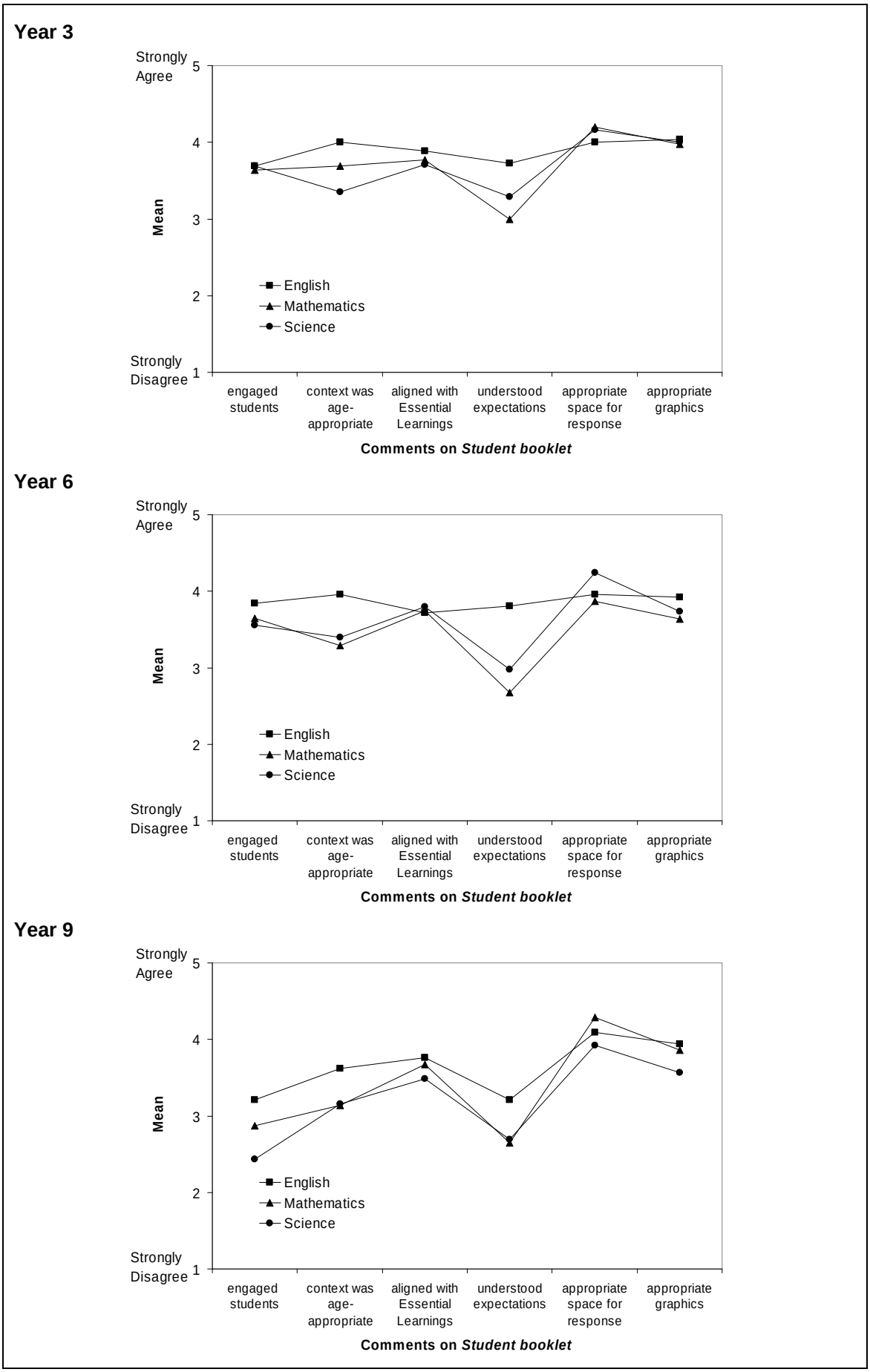


Figure 18: Mean ratings for six items dealing with teachers' perceptions of the Student booklet

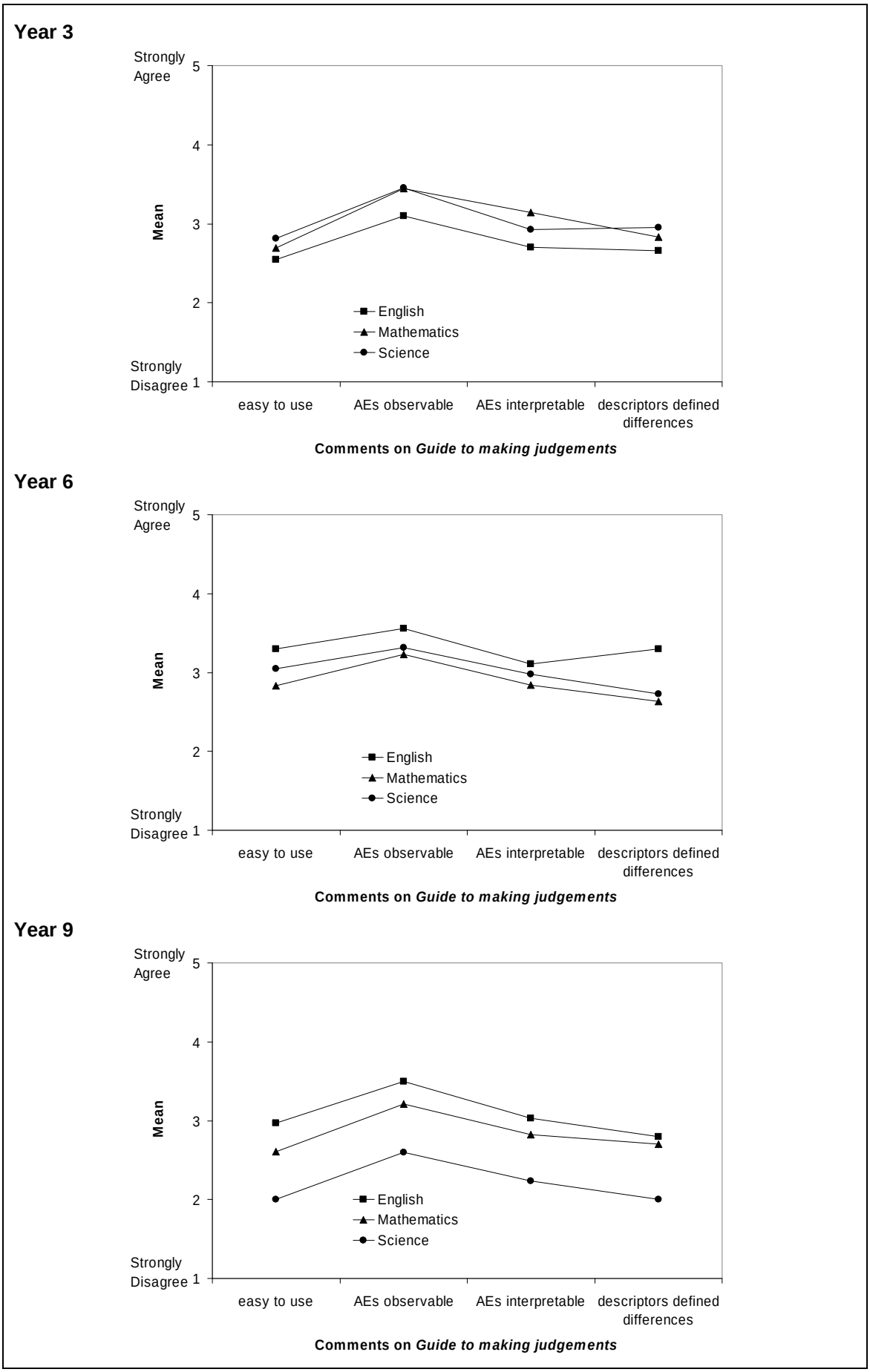


Figure 19: Mean ratings for four items dealing with teachers' perceptions of the Guide to making judgements

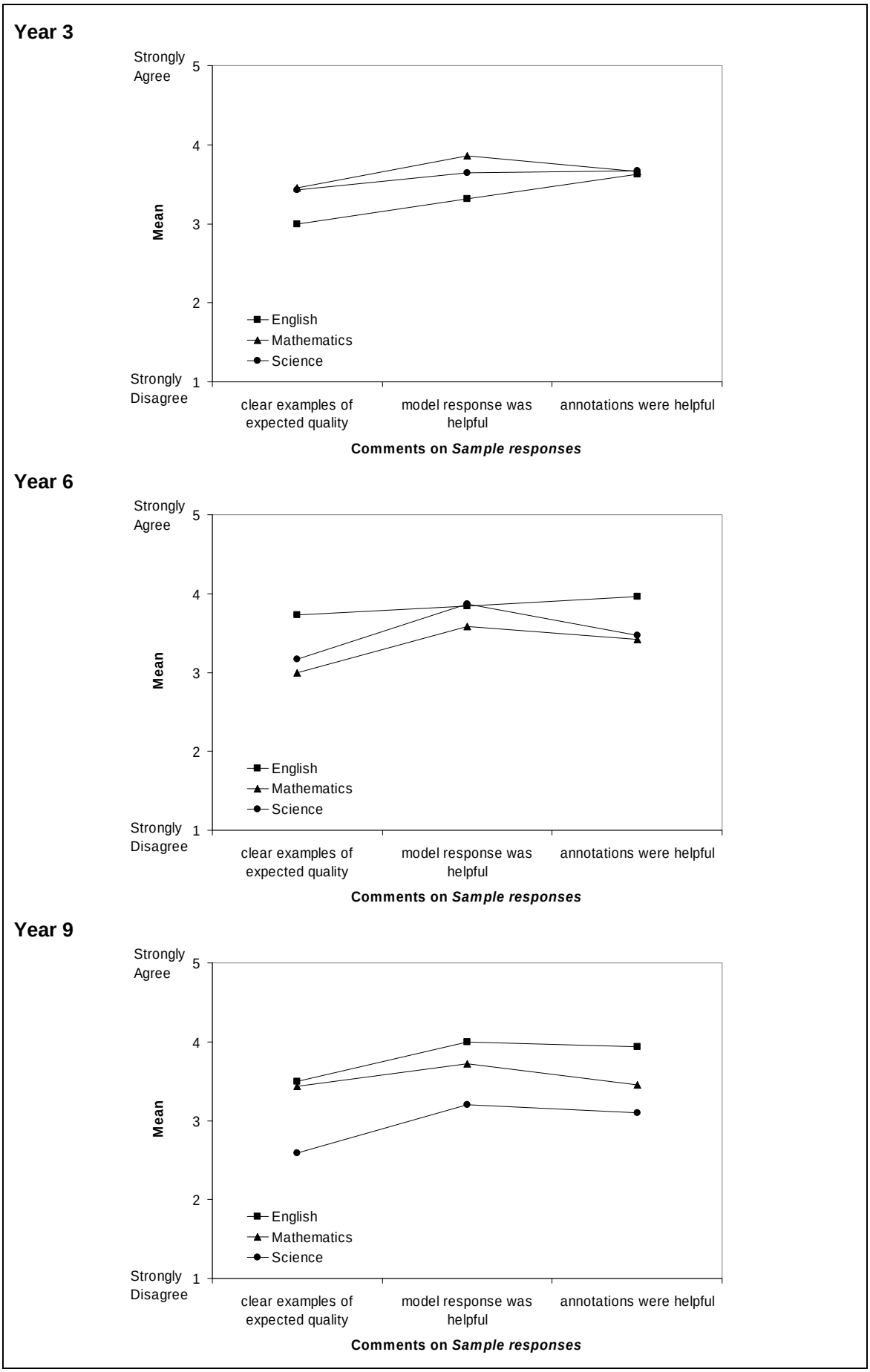
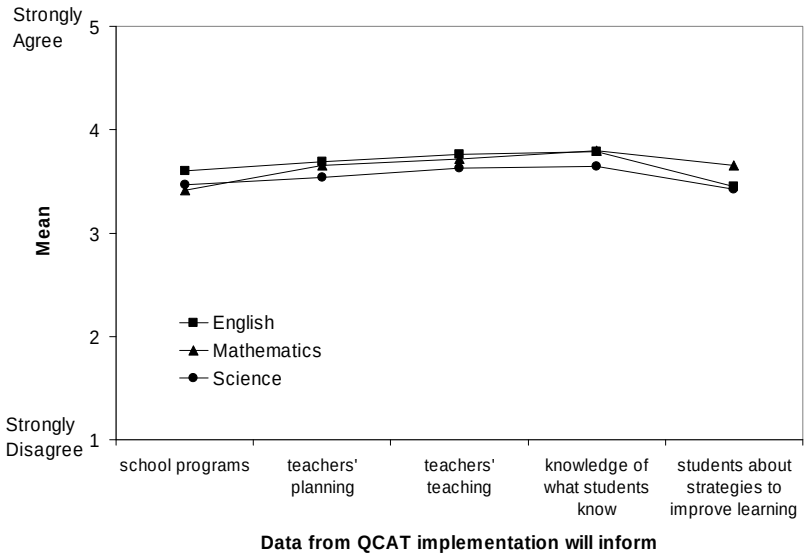
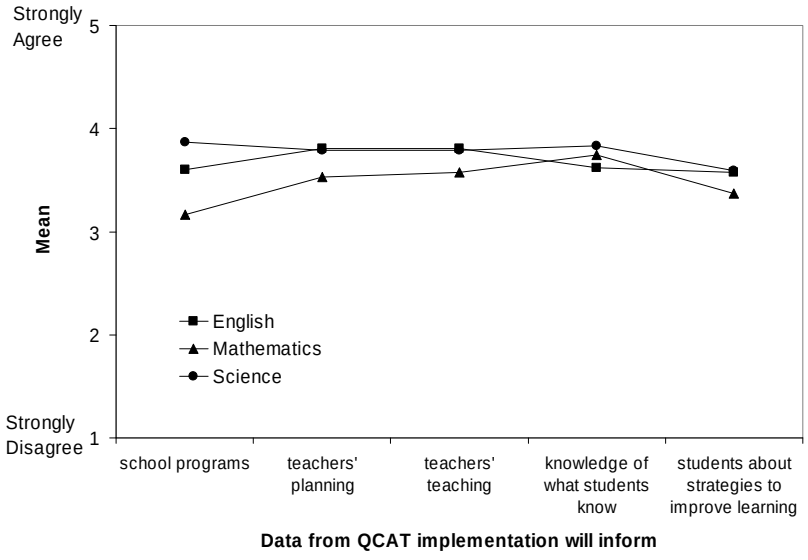


Figure 20: Mean ratings for three items dealing with teachers' perceptions of the Sample responses

Year 3



Year 6



Year 9

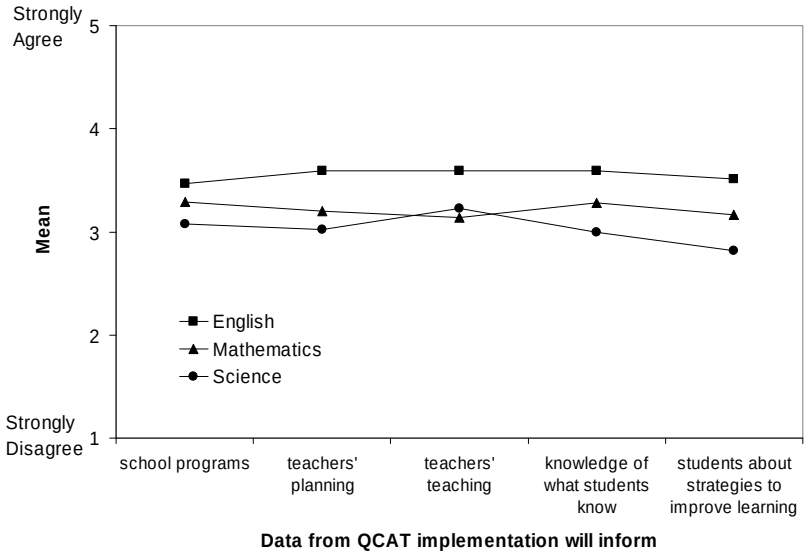


Figure 21: Mean ratings for five items dealing with teachers' beliefs about the way in which the QCAT data with inform their teaching, planning and programming

- *The QCAT engaged the student*; and
- *Students understood what they were expected to do*.

The consistent effect for all three was that Year 9 Science and Mathematics teachers, on the whole, disagreed with the statement.

Each of the questions about the documents contained space for teachers to write a comment about the documents. Following is a summary of the main concerns raised by the teachers:

- *Teacher Guidelines* were not explicit about which students should complete the QCATs, what level of support should be provided to those students who do complete the QCATs, and how much scaffolding to provide.
- All *Student booklets* were criticised to some extent for various reasons: cultural relevance of the contextual material; context was not age-appropriate; the material in general was not relevant; the questions were too difficult. Teachers claimed that students, particularly of Mathematics, were not sure how much detail to provide when asked to "Explain how you get your answer". Mathematics and Science teachers were concerned about the levels of literacy required of students to understand what was being asked and to produce a response. Some teachers complained that terminology was either inappropriate (e.g., the American term "socket") or was not generally known by students (e.g., "turbines", "graphic artist") or needed further explication ("school community").
- Much criticisms was levelled at the *Guide to making judgements*: descriptors for Assessable Elements should align with a letter grade; Assessable Elements should not draw upon information across a number of questions; there was not sufficient differentiation between descriptors and thus letter grades; not enough guidance on how to weight Assessable Elements when assigning an Overall Letter Grade; how to assign a Letter Grade for an Assessable Element that in one part was one letter grade but in another parts was a different letter grade; minor errors amounted to a large penalty; some descriptors appear to be inappropriately placed (e.g., to what extent should "spelling" be assessed as part of "interpreting" and "reflecting").
- Teachers wanted a larger range of responses in the *Sample responses*; including responses where the letter grades for the Assessable Elements do not always aligned with the overall letter grade; sample responses that were more realistic and indicative of the sorts of responses that they encountered; the model

response was not a realistic response (i.e., teachers questioned whether the model responses were written by students); for at least one QCAT, teachers claimed that the A and the B responses should have been reversed; for at least one QCAT, teachers claimed that the annotations contradicted or were not consistent with the *Guide to making judgements*.

- Other comments included: that it was a time-consuming process (making judgements); some Year 9 teachers questioned the timing of QCATS (in the same year as NAPLAN); and those that could attend PD found it useful.

The majority of teachers used "conference/consensus" methods to establish consistency of judgement (57% use "conference/consensus" alone, and another 21% use "conference/consensus" with another method. A total of 35% used "calibration" methods (but only 14% used "calibration" on its own) and 5% used "expert" methods. Half the teachers worked with teachers from other schools to help develop consistency.

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 Assessable Elements that levels of agreement dropped, and both the markers and the teachers expressed difficulties with grading the Assessable Elements. Also, if the amount missing data can be taken as an expression of difficulty, then the teachers had minimal difficulty when awarding Overall Letter Grades, but experienced some degree of difficulty when grading the Assessable Elements. When confronted with these difficulties, some teachers either did not attempt to grade the Assessable Element at all

or they awarded multiple grade presumably aligning with the multiple questions that contributed to an Assessable Element.

Two groups of teachers that stands out as expressing difficulty one way or another were Year 9 Science and Year 9 Mathematics teachers. They were awarding E grades at much higher levels than teachers grading other QCAT, their rates for missing data for Assessable Elements were higher than the rates for other teachers, and they expressed higher levels of disagreement with a number of statements concerning the implementation of the QCATs in their schools.

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

Queensland Government

Queensland Studies Authority
Partnership and Innovation

QCATs extended trial 2008

This feedback form should be completed by the teacher who implemented the QCATs.
We welcome multiple responses, if more than one teacher implemented the QCAT in the school.

School name:	**Contact person:**
Contact details:	

1. Which QCAT did you implement?

☐ 4 English ☐ 4 Mathematics ☐ 4 Science
☐ 6 English ☐ 6 Mathematics ☐ 6 Science
☐ 9 English ☐ 9 Mathematics ☐ 9 Science

2. To which education authority does your school belong?

☐ State (EQ) ☐ Catholic (QCEC) ☐ Independent (ISQ) ☐ Other _____

3. What type of school?

☐ Primary ☐ Secondary ☐ P–10/P–12 ☐ Special ☐ Other _____

4. After receiving the QCATs, what did you do to prepare for implementation of the QCAT?

5. After receiving the QCATs, how did you prepare students for the QCAT?

6. How much time did you spend preparing students for the QCAT?

☐ 30 minutes ☐ 1 hour ☐ More than 1 hour

7. How much time did you spend contextualising the QCAT with students (setting the scene)?

☐ 30 minutes ☐ 1 hour ☐ More than 1 hour

8. How long did the students take to complete the QCAT?

☐ About the recommended amount of time ☐ More than the recommended amount of time ☐ Less than the recommended amount of time

9. How many sessions did you take to implement the QCAT?

☐ 1 session ☐ 2 sessions ☐ More than 2 sessions

10. If any students did not undertake the QCAT, give the reason/s

☐ Absent ☐ Special consideration ☐ Other _____

Suggested improvements:

11. Comment on the *Teacher guidelines*:

	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
The *Teacher guidelines* provided all the information I required	☐	☐	☐	☐	☐
The instructions were clear	☐	☐	☐	☐	☐
The suggested level of support to students was appropriate	☐	☐	☐	☐	☐

Suggested improvements:

12. Comment on the *Student booklet*:

	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
The QCAT engaged the students	☐	☐	☐	☐	☐
The context was age-appropriate	☐	☐	☐	☐	☐
The QCAT was aligned with the targeted *Essential Learnings*	☐	☐	☐	☐	☐
Students understood what they were expected to do	☐	☐	☐	☐	☐
There was an appropriate amount of space for students to respond	☐	☐	☐	☐	☐
The graphics were appropriate and not distracting	☐	☐	☐	☐	☐

Suggested improvements:

13. Comment on the *Guide to making judgments* (GTMJ):

	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
The *GTMJ* was easy to use to make judgments about the overall quality of student responses	☐	☐	☐	☐	☐
The task-specific Assessable elements were observable in student responses	☐	☐	☐	☐	☐
The task-specific Assessable elements were easily interpretable by teachers and students	☐	☐	☐	☐	☐
The task-specific descriptors clearly defined the qualitative differences in student responses	☐	☐	☐	☐	☐

Suggested improvements:

14. Comment on the *Sample responses*:

	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
The *Sample responses* provided clear examples of the quality expected in student work	☐	☐	☐	☐	☐
The model response was helpful	☐	☐	☐	☐	☐
The annotations were helpful to explain how the evidence in the sample matched the descriptors	☐	☐	☐	☐	☐

Suggested improvements:

15. The data gathered from the QCAT implementation will help to inform:

	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
• Our school programs	☐	☐	☐	☐	☐
• My planning	☐	☐	☐	☐	☐
• My teaching	☐	☐	☐	☐	☐
• My knowledge of what students know and can do	☐	☐	☐	☐	☐
• My students about strategies to improve their learning	☐	☐	☐	☐	☐

Suggested improvements:

16. 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 _____

17. Did teachers from your school work with teachers from other schools to help develop consistency of teacher judgments?

☐ Yes ☐ No

18. General comments:

Appendix 3: Summary of MANOVAs

Teacher Guidelines Wilks' $\Lambda = 0.86$, MV $F(24, 1265) = 2.73$, $p < 0.001$, $\eta^2 = 0.048$

Variable	F (8, 446)	p	η^2
Provided all info I required	1.72	.092	
Instructions were clear	2.76	.005	.048
Tukey's HSD:	No statistically significant differences		
Suggested level of support appropriate	6.89	<.001	.112
Tukey's HSD:	Yr 6 M & Yr 9 M < Yr 4 E, Yr 4 S & Yr 6 S Yr 6 M < Yr 6 E & Yr 6 S		

Student Booklet Wilks' $\Lambda = 0.62$, MV $F(48, 2130) = 4.45.$, $p < 0.001$, $\eta^2 = 0.075$

Variable	F (8, 445)	p	η^2
Engaged students	12.28	<.001	.184
Tukey's HSD:	Yr 9 M & S < Y4 E, M & S; and Yr 6 E, M & S		
Context age-appropriate	4.70	<.001	.081
Tukey's HSD:	Yr 4 S, Yr 6 M & Yr 9 M & S < Yr 4 E Yr 9 M < Yr 6 E		
Aligned with *Essential Learnings*	0.54	.50	
Students understood what was expected	8.54	<.001	.147
Tukey's HSD:	Yr 4 M and Yr 6 M & S and Yr 9 M & S < Yr 4 E Yr 6 M and Yr 9 M & S < Yr 4 S Yr 6 M & S and Yr 9 M & S < Yr 6 E		
Appropriate amount of space for response	0.96	.003	.051
Tukey's HSD:	Yr 6 M < Yr 9 M		
Graphics were appropriate	1.45	.023	.040
Tukey's HSD:	No statistically significant differences		

Guide to making judgements Wilks' $\Lambda = 0.88$, MV $F(32, 1628) = 2.01.$, $p = 0.001$, $\eta^2 = 0.035$

Variable	F (8, 452)	p	η^2
Easy to use to make overall judgement	4.31	<.001	.072
Tukey's HSD: Yr 9 S < Yr 4 S and Yr 6 E, M & S and Yr 9 E			
Assessable elements were observable	3.60	<.001	.061
Tukey's HSD: Yr 9 S < Yr 4 M & S and Yr 6 E & S and Yr 9 E & M			
Assessable elements easily interpretable	2.76	.006	.047
Tukey's HSD: Yr 9 S < Yr 4 M & S and Yr 6 E & S and Yr 9 E & M			
Descriptors defined qualitative differences	4.14	<.001	.069
Tukey's HSD: Yr 9 S < Yr 4 M & S and Yr 6 E & S and Yr 9 E & M			

Sample response Wilks' $\Lambda = 0.86$, MV $F(24, 1380) = 2.74.$, $p < 0.001$, $\eta^2 = 0.047$

Variable	F (8, 451)	p	η^2
Clear example of expected quality	4.25	<.001	.071
Tukey's HSD: Yr 9 S < Yr 4 M & S and Yr 6 E and Yr 9 M & E			
Model response was helpful	3.21	.001	.055
Tukey's HSD: Yr 9 S < Yr 6 S and Yr 9 E			
Annotations were helpful	2.79	.005	.048
Tukey's HSD: Yr 9S < Yr 4 S and Yr 6 E and Yr 9 E			

Data will inform Wilks' $\Lambda = 0.84$, MV $F(40, 1868) = 1.88.$, $p = 0.001$, $\eta^2 = 0.034$

Variable		F (8, 440)	p	η^2
School program		2.39	.016	.042
Tukey's HSD:	Yr 9 M & S < Yr 6 S			
Planning		3.20	.002	.056
Tukey's HSD:	Yr 9 S < Yr 6 E & S			
Teaching		3.30	.001	.058
Tukey's HSD:	Yr 9 M < Yr 4 E and Yr 6 E & S			
Knowledge of what students know		3.58	<.001	.062
Tukey's HSD:	Yr 9 S < Yr 4 E, M & S and Yr 6 M & S			
Students about strategies to improve learning		2.64	.008	.047
Tukey's HSD:	Yr 9 S < Yr 4 M and Yr 6 E & S			