The 2009 Year 9 English, Mathematics and Science 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

19 February 2010

Table of Contents

Table of Tables

Table of Figures

Executive Summary

- This report is concerned with the Year 9 English, Mathematics and Science QCATs for 2009.
- Five data collections inform the analyses presented in the report:
    - Statewide data - Overall Letter Grades plus students' gender, Indigenous status and ESL status from schools across the State;
    - Data from 150-schools – The schools returned completed *Student booklets* that represented a typical response for each Overall Letter Grade;
    - Double marking for 80 schools – The *Student booklets* from 80 schools (selected from the 150-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 statewide 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; but not much separated non-ESL students from ESL students.
- When making their on-balance judgement for the Overall Letter Grade, teachers assigned relatively more importance on the 3rd and 4th Assessable Element – elements assessing "Reflection", "Communication", "Thinking and Reasoning", and "Constructing Texts; and less importance on the 1st and 2nd Assessable Elements, in particular to "Knowledge and understanding".
- The markers achieved satisfactory levels of agreement when awarding Overall Letter Grades and when awarding letter grades for the Assessable Elements.
- The levels of agreement between the markers and the teachers were a little less than the levels of agreement achieved by the pairs of markers for Mathematics and Science QCATS. However,

for English, the teachers and markers achieved considerably smaller levels of agreement than achieved by the pairs of markers.

- 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 markers claimed that the students generally answered well the questions that drew on "Knowledge and Understanding", or, in the case of Mathematics, questions that required calculations. They answered less well those questions that asked them to justify, compare, evaluate, or reflect.
- 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.
- The majority of teachers claimed that they used a form of moderation to achieve consistency: either calibration before grading, conferencing after grading or a combination of the two.
- English and Mathematics teachers' perceptions of the *Teacher Guidelines*, the *Student Booklet, Guide to making judgement* and the *Sample Responses* were overall positive. The responses from Science teachers tended to be less positive.
- Mathematics and Science teachers tended to disagree with the propositions 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 data collections of the Year 9 QCATs for English, Mathematics and Science.

Schools 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*; and
- *Sample responses* – containing annotated responses - were available on the QSA website.

Teachers are asked to "make a judgement" (award a letter grade on a 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.

Statewide data

This data collection is concerned with the Overall Letter Grades awarded by teachers across the State for three QCATs: Year 9 English, Mathematics and Science. The schools returned the Overall Letter Grade (i.e., the *Student booklets* were not returned, nor were letter grades for the Assessable Elements available). As well, data concerning students' gender, Indigenous status and ESL status were available

For 150 schools

Across approximately 150 schools, the typical or mid-range *Student booklet* for each Overall Letter Grade was selected and returned. Thus for each QCAT, approximately 700 *Student booklets* should have been returned (150 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 consequently, they would have doubled up on an Overall Letter Grade.

Double marking of QCATs from 80 schools

From the 150 schools, a subset of 80 schools was selected and the QCATs from these schools were assessed by two trained 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 did not agree on 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.

Focus group sessions

At the conclusion of the marking, the markers attend focus group sessions. 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 interview 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 teachers to complete on the QSA website. Appendix 2 contains the survey.

Major questions asked of each data collection

For statewide data

The questions asked of the statewide data focussed on the shapes of the distributions across the Overall Letter Grades:

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

For 150 schools

The questions asked of the 150-schools data again focussed on the shapes of the distributions, but unlike the statewide data collection where the distributions were expected to follow roughly a Normal distribution, the distributions for the 150-schools data collections were expected to be flat (because each school was asked to select a typical example of each Overall Letter Grade). The 150-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 150-schools data collection included:

- Are the distributions for each KLA 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 80 schools

The questions asked of the 80-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 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 contributed 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 the processes used by teachers to establish consistency. 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.

Statewide data

Table 1 shows that Overall Letter Grades were obtained for a little less than 40,000 students for each KLA. Table 1 also shows the number of students according to gender, Indigenous status, and ESL status. As expected, the trial included roughly equal numbers of male and female students, but non-Indigenous students and non-ESL students far out-number 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: Cross-tabulations showing the number of students in the statewide data collection who completed each QCAT by gender, by Indigenous status, and by ESL status

	Male	Female		Total
English	19304	18866		38170
Maths	19578	19145		38723
Science	19183	18830		38013

	Indigenous	Non-Indigenous	Unknown	Total
English	2438	35546	186	38170
Maths	2421	36189	113	38723
Science	2337	35551	125	38013

	ESL	Not ESL	Unknown	Total
English	842	37319	9	38170
Maths	910	37798	15	38723
Science	920	37059	34	38013

Figure 1 shows the shape of the distribution across Overall Letter Grades for each KLA. The figure shows the proportion of students (of the total number of students) awarded each Overall Letter Grade. For instance, considering the distribution for English, 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 (or 0.1 of students) received an A grade. The proportions tend to rise for letter grades B and C, and 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 shapes of the distributions for Mathematics and Science are roughly the same as the distribution for English, although somewhat larger proportions of students receive letter grades E and somewhat smaller proportions of students receive letter grade C.

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 English, 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 distribution is the result of larger proportions of girls than boys receiving the higher letter grades (A and B), and larger proportions of boys than girls receiving the lower letter grade (D and E). The patterns for Mathematics and Science are similar; that is, in Mathematics and Science, girls achieve at slightly higher levels than boys.

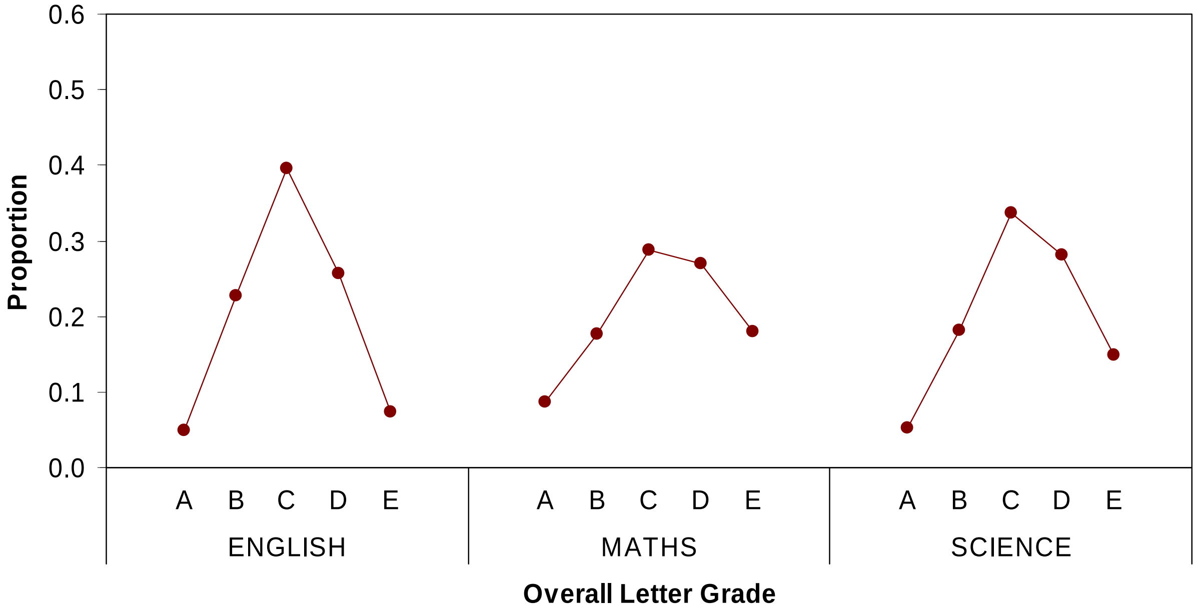


Figure 1: Distribution of responses across Overall Letter Grades for the statewide data collection for Year 9 English, Mathematics and Science

When comparing Indigenous students to non-Indigenous students (Figure 3), the separations of the English, Mathematics and Science distributions are generally large. Smaller proportions of Indigenous than non-Indigenous students receive letter grades A and B; 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.

The separation of the distributions according to ESL status (Figure 4) is not as marked as for Gender or for Indigenous status, although the overall patterns are similar. That is, for each KLA, not much separates the proportions of ESL and non-ESL students receiving each letter grade.

In summary, girls do better than boys; non-Indigenous students do better than Indigenous students; but ESL students perform at roughly comparable rates to non-ESL students for English, Mathematics and Science.

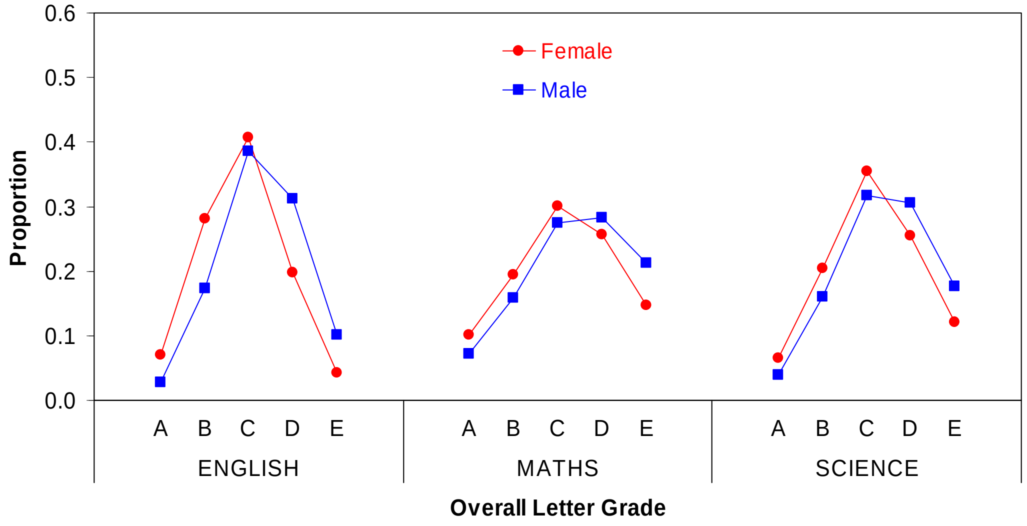


Figure 2: Distribution of responses across Overall Letter Grades for Year 9 English, Mathematics and Science separated by Gender

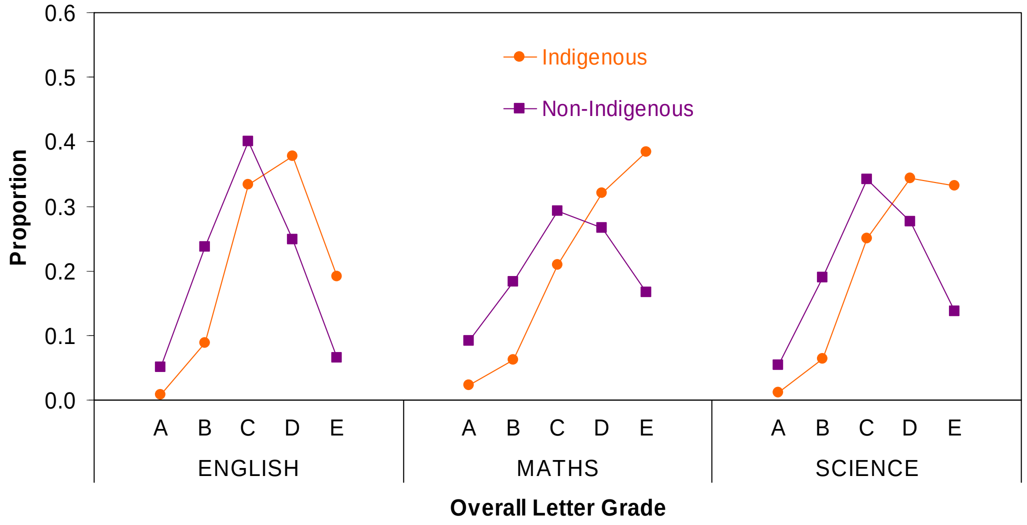


Figure 3: Distribution of responses across Overall Letter Grades for Year 9 English, Mathematics and Science separated by Indigenous status

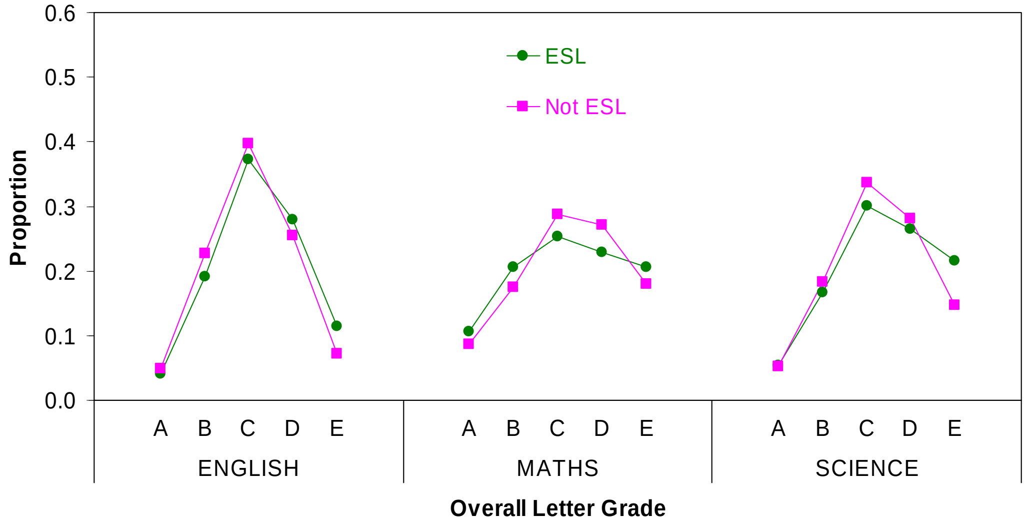


Figure 4: Distribution of responses across Overall Letter Grades for Year 9 English, Mathematics and Science separated by ESL status

150-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 English, Mathematics and Science QCATs were 152, 142 and 144 respectively. Thus, for English, it was expected that there be a total of 760 returns (5 returns X 152 schools); for Mathematics, a total of 710 returns (5 returns X 142 schools); and for Science, a total of 720 returns (5 returns X 144 schools). However, there was missing data. Table 2 shows where the missing data occurred for each KLA. The rates are not large - of the order of 2%. Most of the missing data is a consequence of schools returning fewer than five *Student booklets*.

Table 2: Patterns of missing data for Overall Letter Grade for the 150-schools data collection

English

11 booklets not returned (from 9 schools);
5 No Overall Letter Grade awarded;
Total 16 (2.1%)

Mathematics

11 booklets not returned (from 7 schools);
7 No Overall Letter Grade awarded;
Total 18 (2.5%)

Science

7 booklets not returned (from 5 schools);
6 No Overall Letter Grade awarded;
Total 13 (1.8%)

Table 3 shows the pattern of missing data for the Assessable Elements. The Table shows the number of booklets for which letter grades were missing for all Assessable Elements; the number of booklets for which Letter Grade were missing for one, two or three of the Assessable Elements; and the amount of missing data for each Assessable Element. Clearly, the rates of missing data are larger for the Mathematics and Science QCATS than for the English QCAT. Further, most of the missing data for English is a consequence of teachers failing to record letter grades for all Assessable Elements. If it can be taken that a missing letter Grade for one, two or three Assessable Elements is an indication of the degree of difficulty experienced by teachers when assigning letter grades, then Mathematics and Science teachers found it more difficult to assign letter grade for

Assessable Elements. However, no single Assessable Element was more difficult than any other Assessable Element (Table A3-1 shows the results of testing whether or not the frequencies for missing data for each Assessable Element differ significantly).

Table 3: Patterns of missing data for letter grade for Assessable Elements for the 150-schools data collection

English

26 No letter grades for any AEs;

9 instances of one, two, or three letter grades missing for AEs.

The number of times letter grades were missing by AE:

AE1	AE2	AE3	AE4
27	28	32	31

Mathematics

48 No letter grades for any AEs;

45 instances of one, two, or three letter grades missing for AEs.

The number of times letter grades were missing by AEs:

AE1	AE2	AE3	AE4
68	74	80	56

Science

44 No letter grades for any AEs;

55 instances of one, two, or three letter grades missing for AEs.

The number of times letter grades were missing by AEs:

AE1	AE2	AE3	AE4
76	81	81	68

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 each KLA will be flat, but as can be seen in Figure 5, the distributions are not perfectly flat. The three distributions follow the same shape, but it is only the distribution for English that deviates significantly from flatness (Table A3-2 in Appendix 3 shows the results of testing for statistically significant deviations from 'perfect flatness'). While it appears that the proportion of A grades is less than expected for all three KLAs, it is only with respect to English that the result is statistically so. It is also noted that the shape of the distribution for English does not deviate significantly from the shape for Mathematics nor from the shape for Science. The conclusion is that there might be a small tendency for schools not to select *Student booklets* with an Overall Letter Grade of A.

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. The doubling-up might account for the deviations from 'perfect flatness'. Given that an Overall Letter Grade of A occurs less frequently than the other letter grades in the statewide data collection (see Figure 1 above), it might have been difficult for some schools to find an A grade. In any event, whatever deviation from flatness is being exhibited in the data, it is not large.

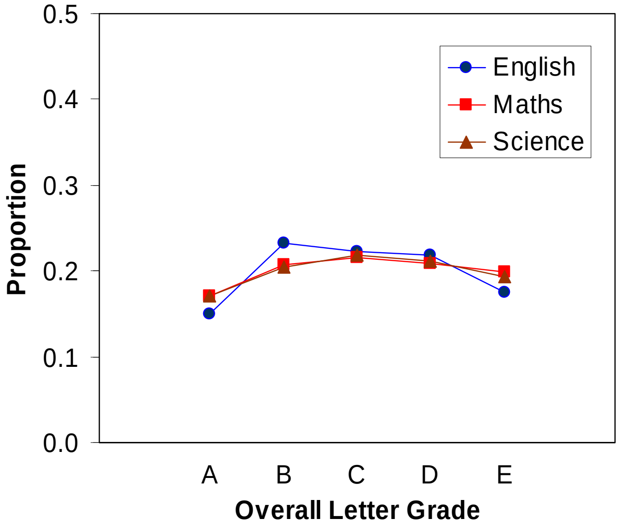


Figure 5: Distribution of responses across Overall Letter Grades for English, Mathematics and Science for the 150-schools data collection

Figure 6 shows the pattern of letter grades awarded for Assessable Elements within an Overall Letter Grade for English. 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. 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 pattern applies for Mathematics (Figure 7) and Science (Figure 8). There are one or two exceptions. Consider the pattern of letter grades awarded for the 1st Assessable Element in Mathematics. When an Overall Letter Grade of B was awarded, a substantial proportion awarded a B for the 1st Assessable Element but most teachers awarded an A. The fact that there is no descriptor aligning with a B partly explains the pattern, but it cannot be a complete explanation. The reason being that there are other Assessable Elements where descriptors do not align with letter grades (for instance the first two Assessable Elements for Science), yet the pattern is not evident for Science. Possibly, teachers experienced difficulty interpolating only for certain pairs of descriptors.

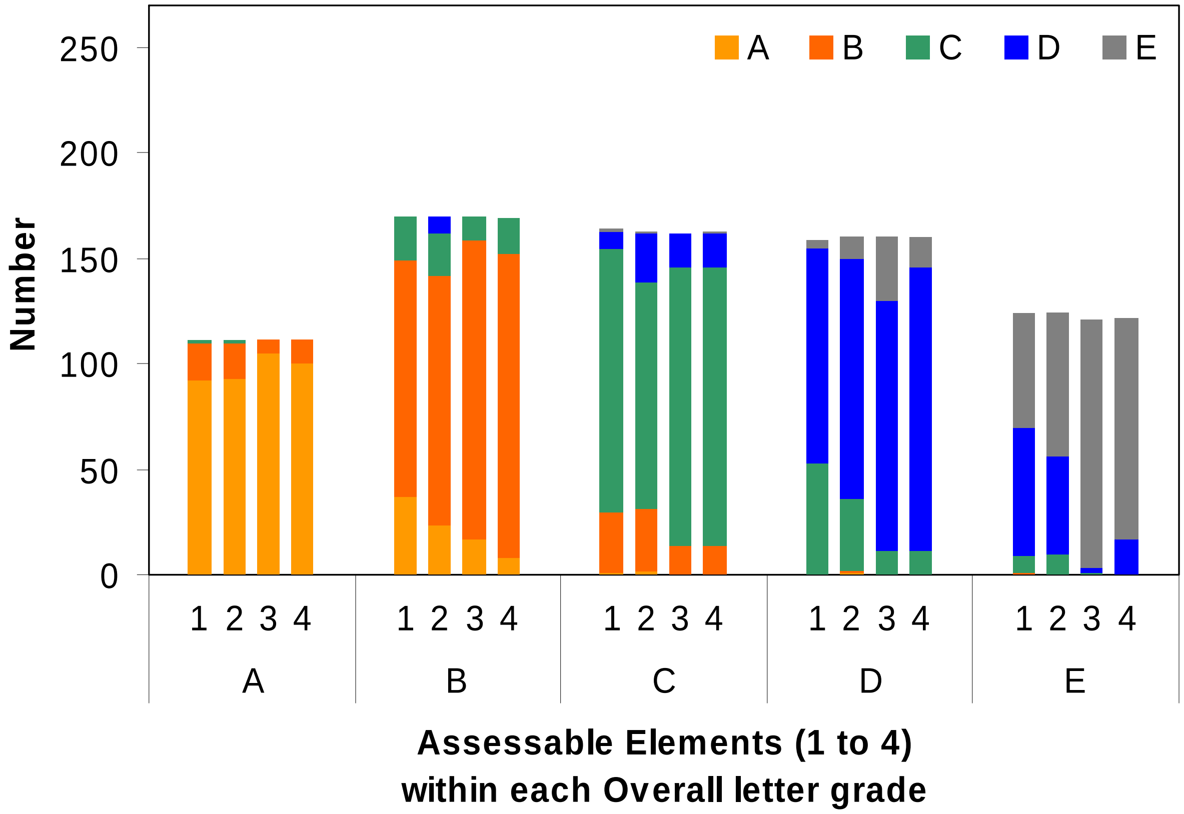


Figure 6: Pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade – English

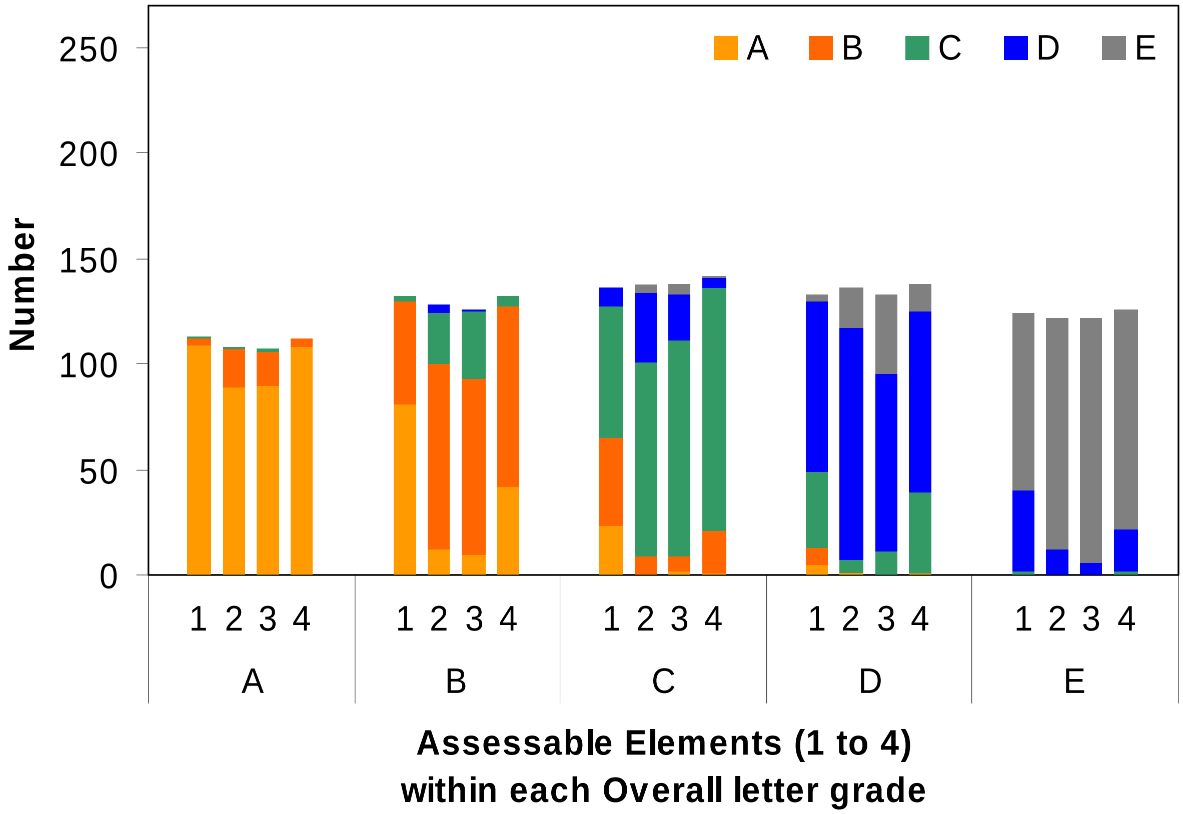


Figure 7: Pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade – Mathematics

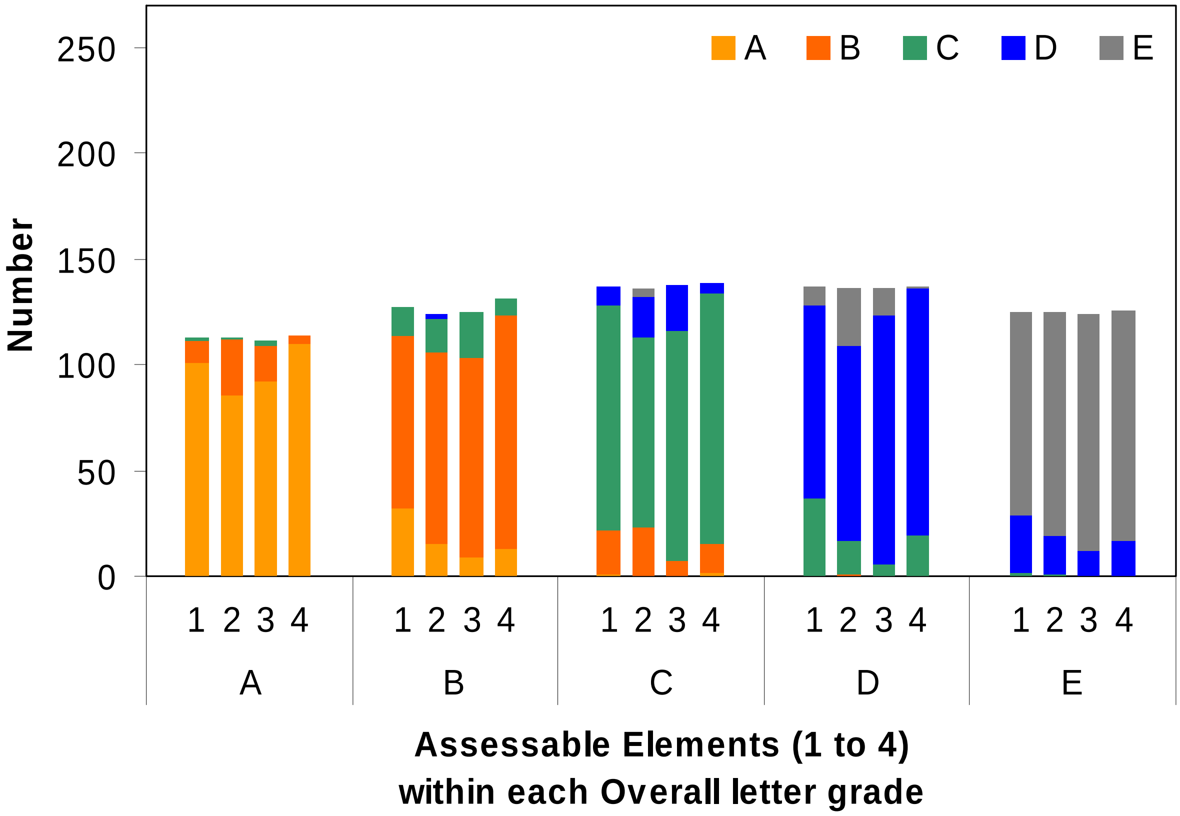


Figure 8: Pattern of letter grades awarded for Assessable Elements within each Overall Letter Grade – Science

Another element that might play a role when awarding an Overall Letter Grade is the importance teachers attach to 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 English in Table 4. An increase of one standard deviation for the 1st Assessable Element leads, on average, to an increase in the Overall Letter Grade of 0.179 standard deviations; an increase of one standard deviation for 2nd Assessable Element leads to an increase in the Overall Letter Grade of 0.142 standard deviations; an increase of one standard deviation for the 3rd Assessable Element leads to an increase of 0.353 standard deviations for the Overall Letter Grade; and an increase of one standard deviation in 4th Assessable Element leads to an increase of 0.360 standard deviations for the Overall Letter Grade. Thus, somewhat less importance is assigned to the 1st and 2nd Assessable Elements than to the 3rd and 4th Assessable Elements. There is a similar pattern for Mathematics and Science: more importance is assigned to the 3rd and 4th Assessable Element than to the 1st and 2nd Assessable Elements. A possible explanation is that the 3rd and 4th Assessable Elements are still fresh in teachers' minds when deciding on an Overall Letter Grade, and so the 3rd and 4th Assessable Elements assume greater importance.

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

Assessable Element	**Standardised coefficient**
English	
Knowledge & understanding: Appreciating texts	.179
Knowledge & understanding: Constructing texts	.142
Reflecting	.353
Constructing texts	.360
Mathematics	
Knowledge & understanding	.186
Thinking & reasoning	.258
Thinking & reasoning	.299
Communicating	.290
Science	
Investigating	.200
Knowledge & understanding	.187
Investigating	.289
Communicating	.347

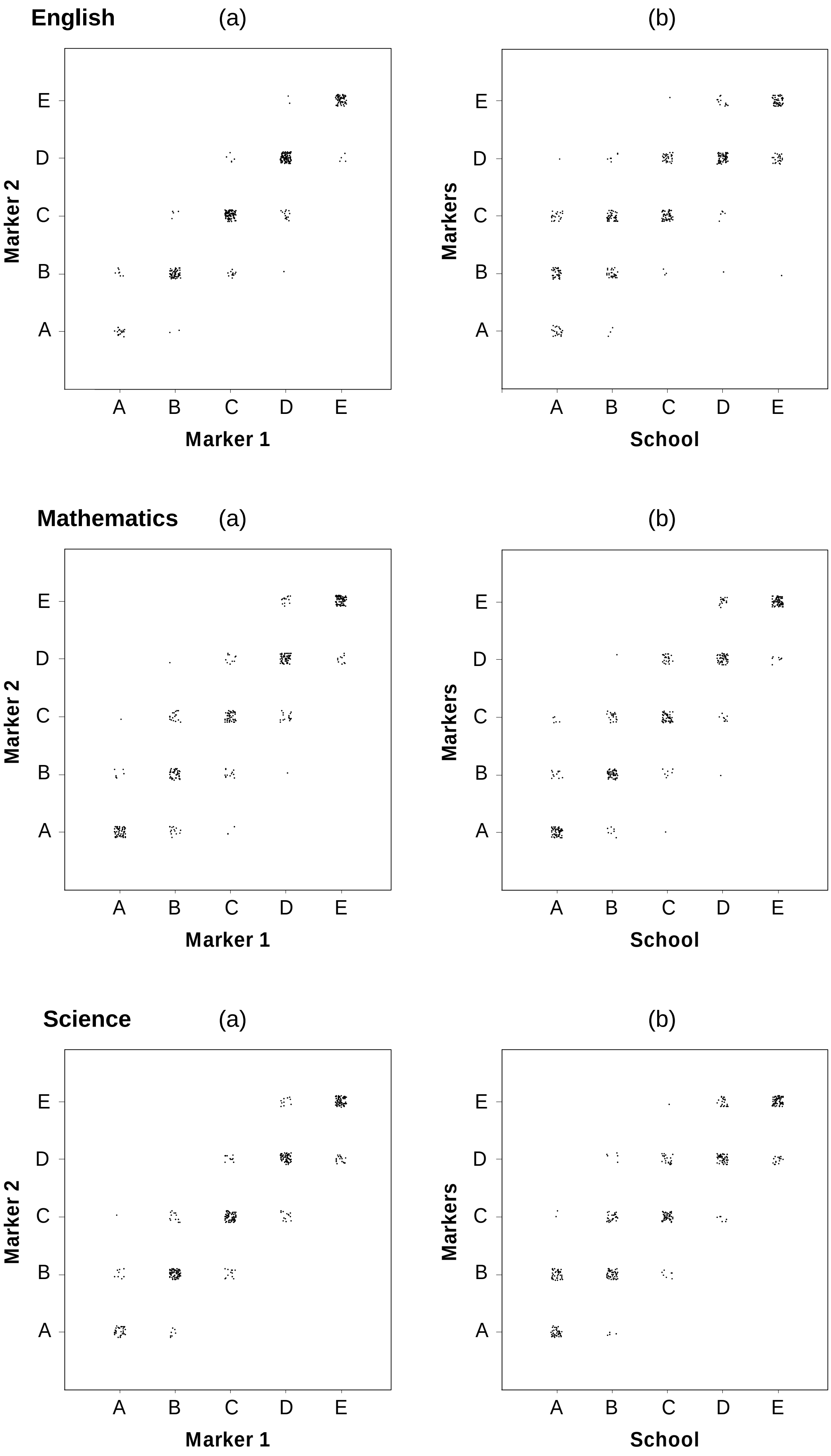
Double marking of QCATs from 80 schools

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. In addition, 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 80 schools, a sub-sample of the 150-schools data collection.

Figures 9, 10 and 11 give a visual representation of the consistency achieved by pairs of markers when awarding Overall Letter Grades. To read the top left scatterplot (English) in Figure 9, note that each point is represented by a cloud of points. Consider the point represented by the coordinates (B, B). There are 60 *Student booklets* represented by (B, B), which means that for 60 booklets, the two markers agreed when awarding the B grade. If the 60 booklets were instead represented by a single point, information would be lost – the information about there being 60 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, 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, and occasionally a point displaced two spaces off the diagonal. That is, the pairs of markers were consistent. The three scatterplots on the left in Figure 9 show that the pairs of markers were consistent, but it is noted that occasionally there are points appearing two spaces off the diagonal.

The scatterplots on the right in Figure 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 of the Figure, there are more points displaced one and two spaces off the diagonal. That is, teachers and markers did not achieve the same level of consistency as achieved by the pairs of markers.

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 80 schools that comprise this data collection are a sub-sample of the 150-school data collection. Table 5 shows where the missing data occurred for each QCAT. As was the case with the 150-schools data collection, there was more missing data for Mathematics and Science than for English, but the rates are more or less even for each Assessable Element.



English (a) (b)

Mathematics (a) (b)

Science (a) (b)

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

Table 5: Patterns of missing data for the letter grades for Assessable Elements for the 80-schools data collection

English

2 No letter grades for any AEs;

4 instances of one, two, or three letter grades missing.

The number of times letter grades were missing by AEs:

AE1	AE2	AE3	AE4
3	2	4	3

Mathematics

9 No letter grades for any AEs;

37 instances of one, two, or three letter grades missing.

The number of times letter grades were missing by AEs:

AE1	AE2	AE3	AE4
26	33	39	18

Science

30 No letter grades for any AEs;

27 instances of one, two, or three letter grades missing.

The number of times letter grades were missing by AEs:

AE1	AE2	AE3	AE4
42	44	45	35

Figures 10, 11 and 12 show consistency in the same way as shown in Figure 9, except that Figures 10, 11 and 12 show consistency when awarding letter grades for the Assessable Elements. There is some decline in levels of consistency 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 appears to be a further decline in consistency for English.

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[1], 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

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

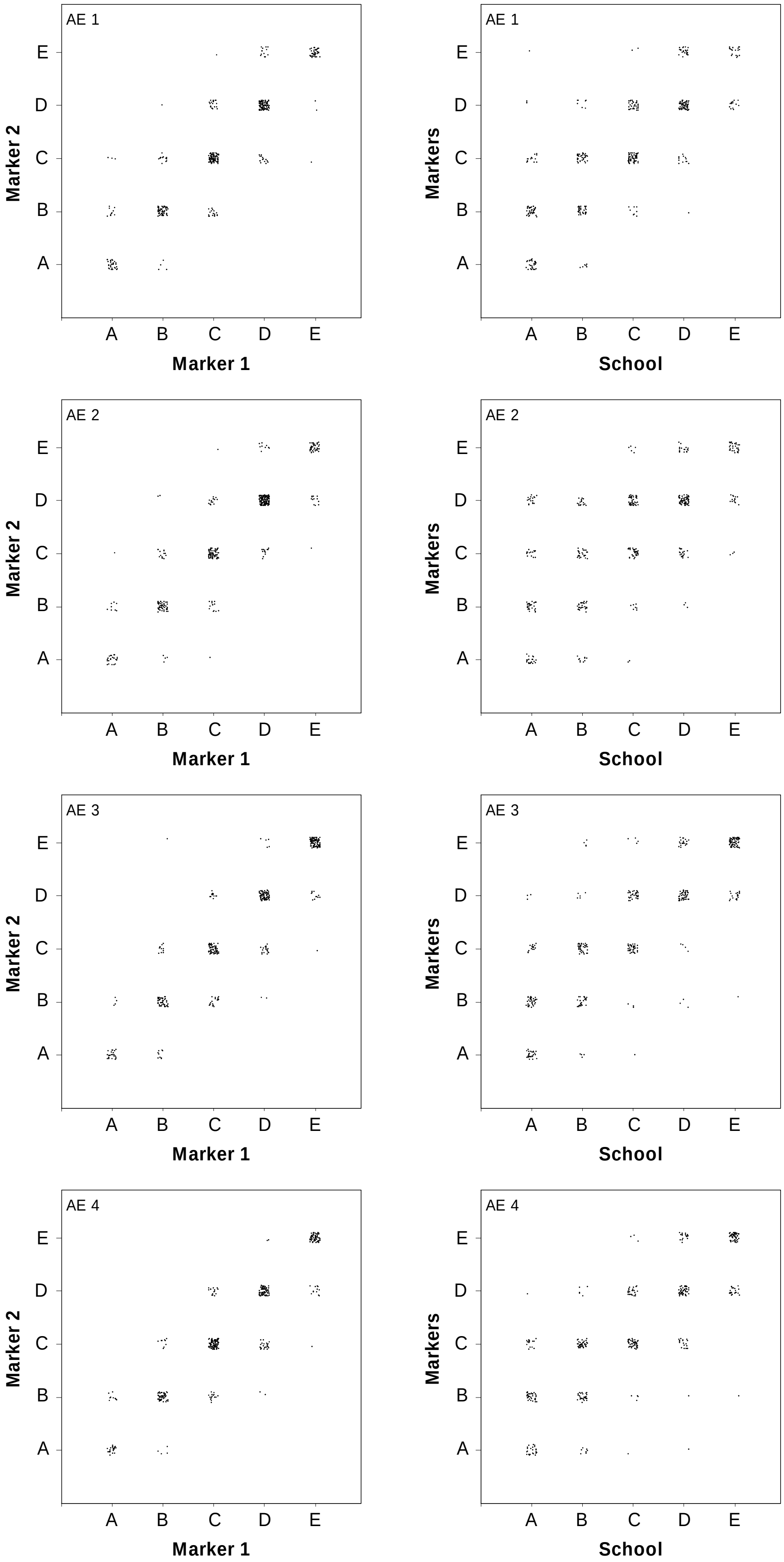


Figure 10: Agreement between pairs of markers; and between markers and the schools for each Assessable Element – English

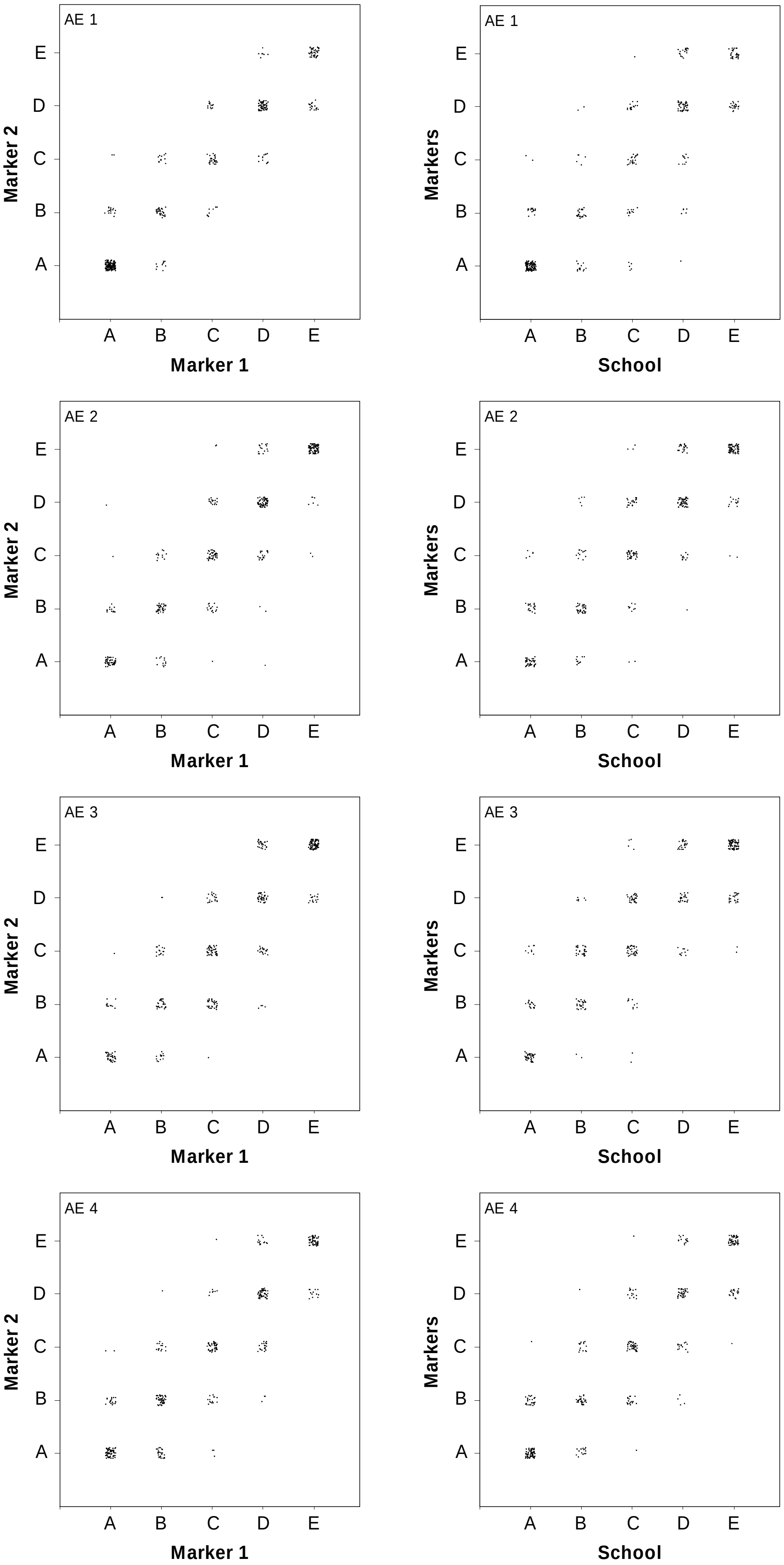


Figure 11: Agreement between pairs of markers; and between markers and the schools for each Assessable Element – Mathematics

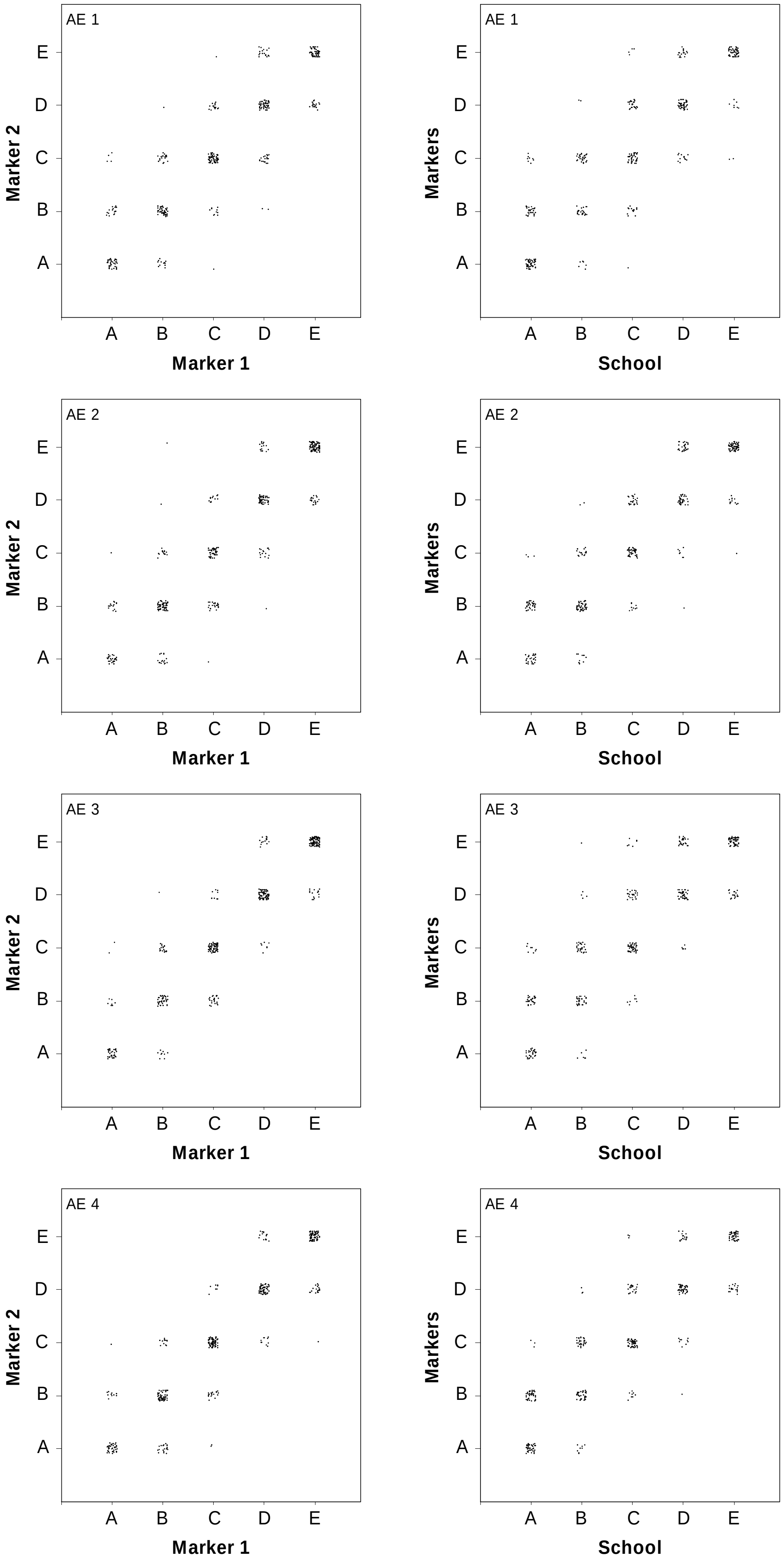


Figure 12: Agreement between pairs of markers; and between markers and the schools for each Assessable Element – Science

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 13 shows the values for Cohen's κ 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 κ values for pairs of markers are in the "Very good" range for the Overall Letter Grade, and in the "Good" to "Very good" ranges for the Assessable Elements. The values for Cohen's κ are constant across the Assessable Elements for English, but there is a drop for 3rd Assessable Element in Mathematics ("Generalisation and justification of reasoning") and for the 1st and 2nd Assessable Elements for Science ("Selection and manipulation of formulas to calculate lengths, volumes and statistical measures of central tendency", "Choice of strategies and procedures to generate solutions").

The κ values for assessing agreement between the consensus grade and the grade awarded at the schools for the Overall Letter Grades are in the "Good" range for the three KLA, and are less than the corresponding values for the pairs of markers for English and Science. That is, the markers and the teachers could not achieve the same levels of agreement as was achieved by the pairs of markers when awarding Overall Letter Grades for English and Science. For the Assessable Elements, there is a decline in consistency, particularly for English.

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 then 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 for Mathematics and Science, but less so for English. It is noted however that the rates of missing data were larger for Mathematics and Science. The Cohen's κ values for Mathematics and Science might have been closer to the values for English had more Science and Mathematics teachers marked letter grades.

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

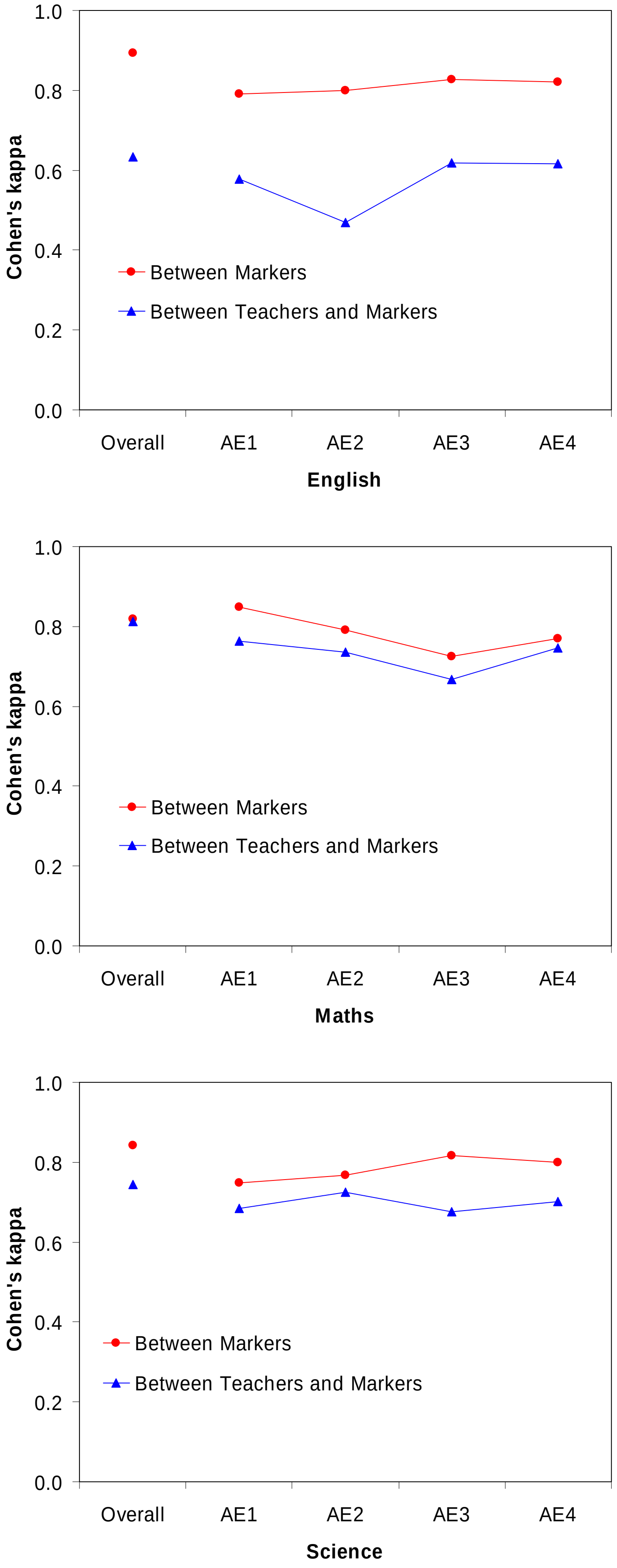


Figure 13: Coefficient of agreement (Cohen's κ) between the two makers and between teachers and markers when awarding Overall Letter Grades and the letter grade for Assessable Elements for English, Mathematics and Science

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 reasonably consistent when awarding letter grades. Overall, their assessment of their consistency aligns with the assessments of agreement presented in the previous section (see Figures 9 to 12 and, in particular, Figure 13). In addition, the Mathematics markers claimed that they were more consistent when awarding Overall Letter Grades than when awarding letter grades for the Assessable Elements. It would, at first blush, appear to be easier to maintain higher level of consistency when awarding that the letter grades for Assessable Elements because those letter grades 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). However, Figures 13 shows that their perceptions were not entirely correct. They were achieving higher levels of consistency with the 1st Assessable Element than with the Overall Letter Grade; and it is only with respect to the 3rd and 4th Assessable Elements that they achieving lower levels of consistency.

When discrepancies did occur, the markers claimed that they were not large and that mostly they did not disagree by more than one letter grade. The scatterplots in Figures 9 to 12 show that their perceptions were mostly correct. The markers claimed that disagreements were concerned mostly with borderline grades, and thus there was rarely a difficulty reaching consensus. 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;
- Consulted the *Guide to making judgements*;
- Re-read the task specific assessable element; and
- Read the relevant student responses again, pointing to evidence in the script.

Despite their claims of consistency, the markers nevertheless argued that there were types of Assessable Elements that proved more difficult to assess than others. A major difficulty occurred where they perceived the descriptors to be not sufficiently specific. For instance, the Science markers pointed to a difficulty discerning points between "consistently" and "some" with respect to the 1st Assessable Element; or the English markers pointed to a difficulty distinguishing between "well-developed arguments", "supported arguments", and "using arguments" with respect to the 3rd Assessable Element. Some Science markers argued that the problem was compounded when there was not descriptor for a letter grade. For instance, determining a point between "consistently" (an A) and "some" (a C) was important because that aligned with the B (which not aligned with a descriptor). Other claimed that interpolating in this way was not a major difficulty, but they did have trouble with extrapolating (for instance, in situations were there was a descriptor aligned with a D, but no descriptors beyond D).

Some markers claimed difficulty with awarding letter grades for Assessable Elements that drew on multiple questions. The difficulty lay with weighting the responses to the different questions when assigning the letter grade. For instance, with respect to the 2nd Assessable Element for Mathematics (which drew on evidence from two questions), some markers claimed it was difficult to determine a letter grade when one was answered well and the other was not. Similarly, some markers claimed difficulty with assigning weights to the Assessable Elements when awarding an Overall Letter Grade especially when the letter grades for Assessable Elements varied greatly. There was a related difficulty with keeping separate the components of a particular question when that question appeared in two or more Assessable Elements. For instance, Mathematics markers argued that because two "Thinking and reasoning" Assessable Elements drew on Question 10, it was difficult to keep separate the relevant aspects of Question 10. Similarly, the Science markers argued that Question 9 (a question that required a diagram to be labelled) contributed to a "Knowledge and Understanding" Assessable Element and to a "Communicating" Assessable Element. If a diagram was missing a label, they claimed they were not sure whether the student was experiencing a Knowledge and Understanding" problem or a "Communicating" problem.

The Mathematics markers claimed that students answered well those questions that depended on calculations; the Science markers claimed students answered well questions that drew on and largely required a re-presentation of information contained in the graphs; and the English markers claimed that the students answered well the questions that drew on the initial advertisement. However, large numbers of students experienced difficulty when asked to justify, compare, evaluate, or reflect, or when answers depended on students having well-developed literacy schools (with respect to both understanding the requirements of the question and to produce a response). Additionally, some markers noted that students could present well-crafted arguments, but the genre was inappropriate. For instance, in the Science QCAT, students would argue for the need for citizens to take responsibility for and understand the social consequences of their environmental actions, when the questions required scientific arguments.

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 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. Therefore, 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, ticks, ½ marks, and letter grade for individual questions. However, the markers conceded that the marks and ticks could be used more as a reminder of features in the 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, with respect to Science, scientific literacy skills, and with respect to English, justification. Generally, questions that depended on knowledge, recall and understanding were answered better than questions that depended upon justification, interpretation and reflection.

Survey

A total of 98 surveys were completed: 31 with respect to the English QCAT; 27 with respect to the Mathematics QCAT; and 40 with respect to the Science QCAT. The sample is small, and given that it is self-selected, the conclusions below need to be treated with some caution. The majority of surveys (77%) were received from State schools, with smaller numbers received from Catholic schools (11%) and Independent schools (12%). A small number of returns were received from teachers in schools located in remote areas (5%), with the remainder more or less evenly spread across rural (35%), provincial (34%) and Brisbane metropolitan (27%) 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 KLA, the education authority of the teachers' schools, and the location of teachers' schools. Figures 14, 15, 16 and 17 show the distribution of responses for each question in turn separated according to KLA (each figure also shows the overall pattern). Consider Figure 14. It shows the distribution of responses for "Time spent preparing students for the QCAT" for each KLA 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 KLA, and indeed, the results of the significance test (see Kruskal-Wallis tests in Appendix 3) indicate that, overall, there were no differences in the response patterns across the KLAs.

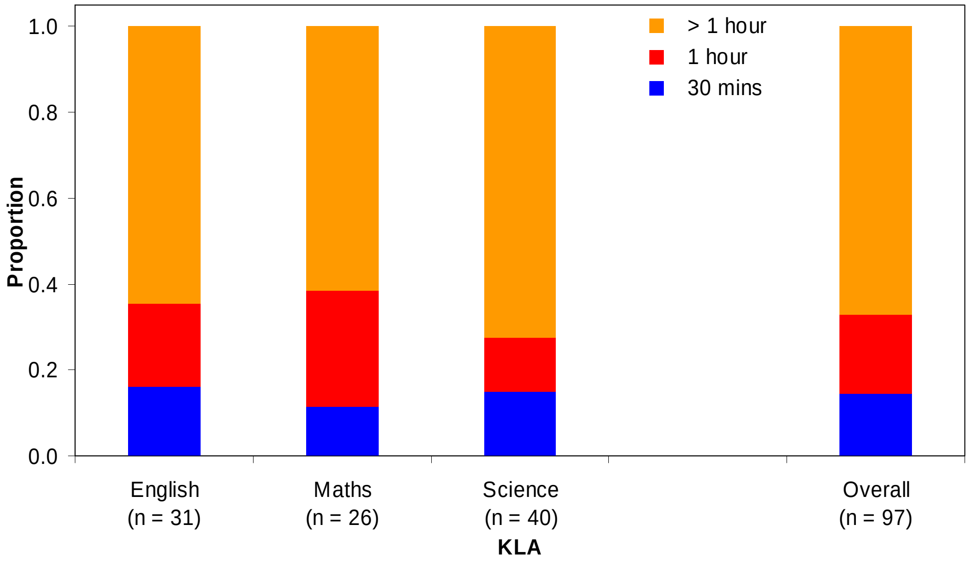


Figure 14: Time spent preparing students for the QCAT

Figure 15 shows the distribution of responses for "Time spent contextualising" The figure is structured the same way as Figure 14. Overall, a large proportion of teachers ticked "30 minutes", with smaller proportions ticking the "1 hour" and "more than 1 hour" categories. There are similar patterns for each KLA, and the minor differences are not statistically significant (see Kruskal-Wallis tests in Appendix 3). Figure 16 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 17 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. Once again, there are similar patterns evident for each KLA and any differences in the patterns are not statistically significant (see Kruskal-Wallis tests in Appendix 3).

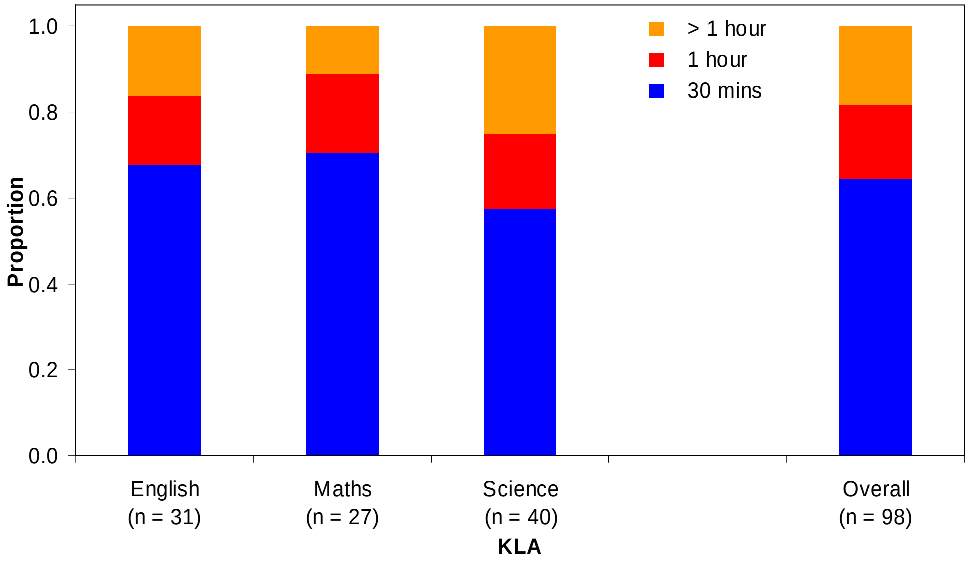


Figure 15: Time spent contextualising the QCAT with students

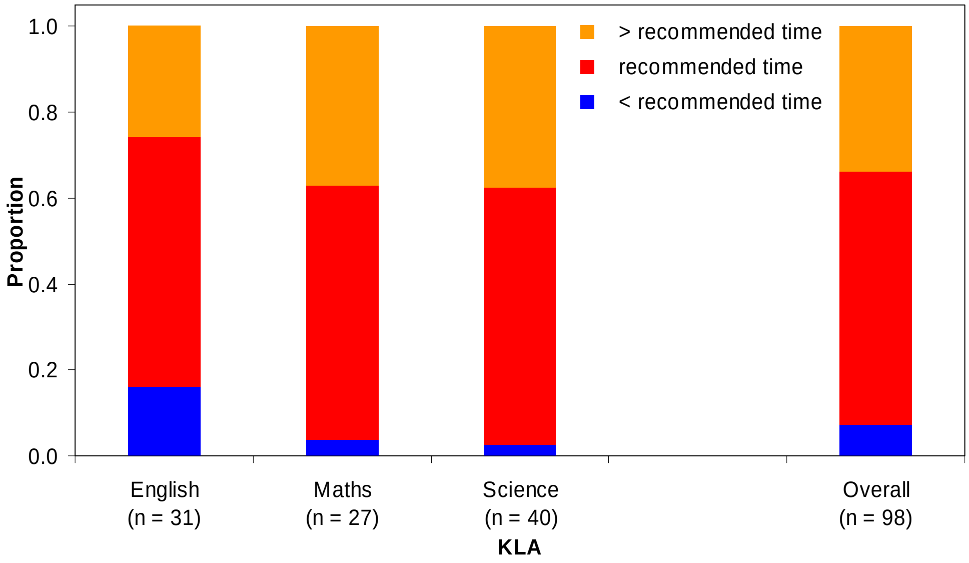


Figure 16: Time taken by students to complete the QCAT

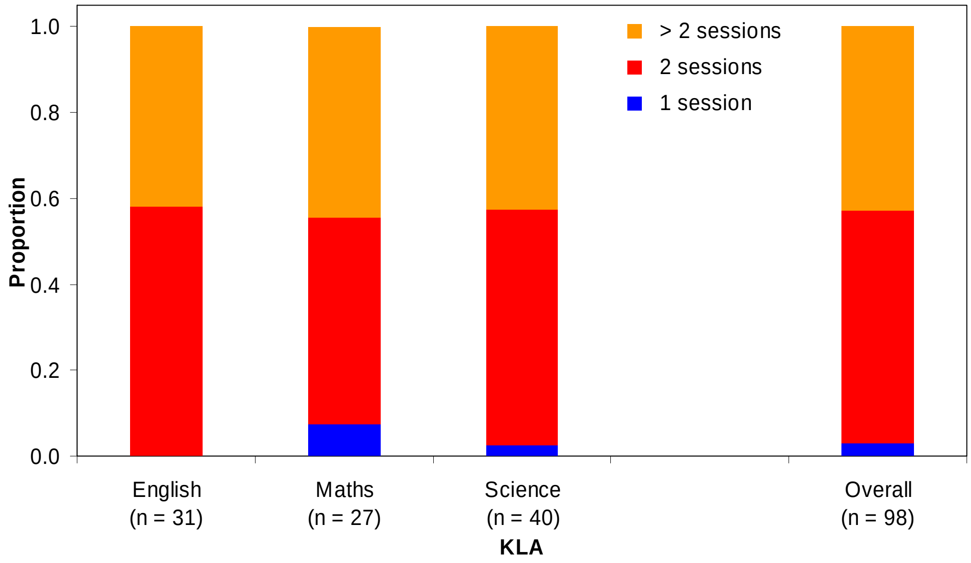


Figure 17: Number of sessions taken to implement the QCAT

Similar statistical tests show that there were no discernible differences in the patterns of response according to the education authority of the school (State, Catholic, and Independent), nor according to location of the school (remote, rural, provincial city and metropolitan Brisbane).

In summary, this sample of teachers, on the whole, took more than one hour to prepare students for the QCAT, took 30 minutes setting the scene of the QCAT with students, their students took about the recommended time to implement the QCTA (although the students of a substantial proportion of teachers took longer than the recommended time), and the QCAT was implemented during two sessions (although a substantial proportions took more than two session). There were no discernable difference in these patterns according to the KLA, education authority and location of the school.

A little more than half the teachers (53%) claimed that the QCAT was used as part of their school-based assessment. The proportion did not change significantly when the sample was split according to KLA, nor education authority of the school, nor location of the school.

Less than half the sample (40%) claimed that teachers from their school worked with teachers from other schools to help develop consistency of judgements. The proportion did not change significantly when the sample was split according to KLA. However, teachers in Catholic and independent schools were less likely to do so (though the numbers from Catholic and Independent schools are small). Surprisingly, teachers from schools located in urban areas were less likely to do so.

Table 6 shows teachers' responses to the question concerning the processes put into place to establish consistency of teacher judgements. Only 3% of the teachers claimed that their schools did not have a process in place to ensure consistency. Furthermore, most teachers claimed that their schools used a form of conferencing and collaboration either before grading or after grading (94%). Only 3% made use of an expert marker. Differences according to KLA, education authority of the school and the location of the school were not significant.

Table 6: Processes in place to ensure consistency of teacher judgements

Process	**%**
Conference/consensus (reaching agreement after grading)	35.8
Calibration (reaching agreement before grading)	22.1
Expert (one marker, no conferencing)	3.2
Combination of moderation models	35.8
None	3.2

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 18, 19, 20 and 21 give the average ratings for each statement about each document in turn. Each figure shows the ratings separated by KLA. 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 18), 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. However, Science teachers expressed less agreement with the statements, especially the statement concerning the helpfulness of the model response. Their comments focussed on the need for consistency of implementation (including instructions to students, advice to students, timing, and grading). Some teachers' comments concerning the "Model Response" contained in the *Teacher Guidelines* indicated that they might not have been aware that there were sample responses on the QCAR website.

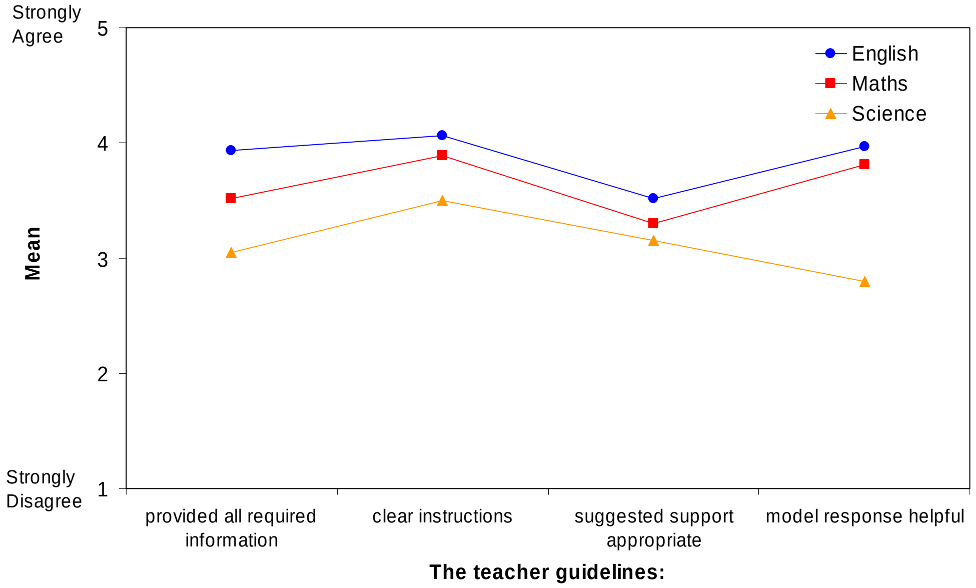


Figure 18: Mean ratings for four items dealing with teachers' perceptions of the Teacher Guidelines

With respect to the *Student booklet*, Figure 19 shows that the mean rating for four statement are above the neutral midpoint on the scale, indicating that teachers, on the whole, agreed with the propositions: that the content was age appropriate; that the QCAT aligned with Essential Learnings; and that the amount of space for students' responses and the graphics were appropriate. However, teachers, on the whole, disagreed with the statement that the QCAT engaged the students, and were undecided about students understanding of expectations. The separation of the KLAs in Figure 19 is not statistically significant. English teachers' comments focussed on the appropriateness of the advertisement; Mathematics teachers argued that the content was often more advanced than Year 9 level; and Science teachers commented that many students had difficulty comprehending the questions.

Figure 20 shows that teachers were more critical of the *Guide to making judgements* than the other documents. While the overall means remain at or close to the neutral midpoint of the scale, Science teachers on the whole tended to disagree with each of the statements concerning the *Guide*, and Mathematics teachers were on the whole undecided. The Mathematics and Science teachers commented that the *Guide* did not provide sufficient information to allow them to make reliable and valid judgements; claiming that the descriptors allow only subjective judgements, that the *Guide* was not sufficiently specific, and that being aligned across question added to the workload.

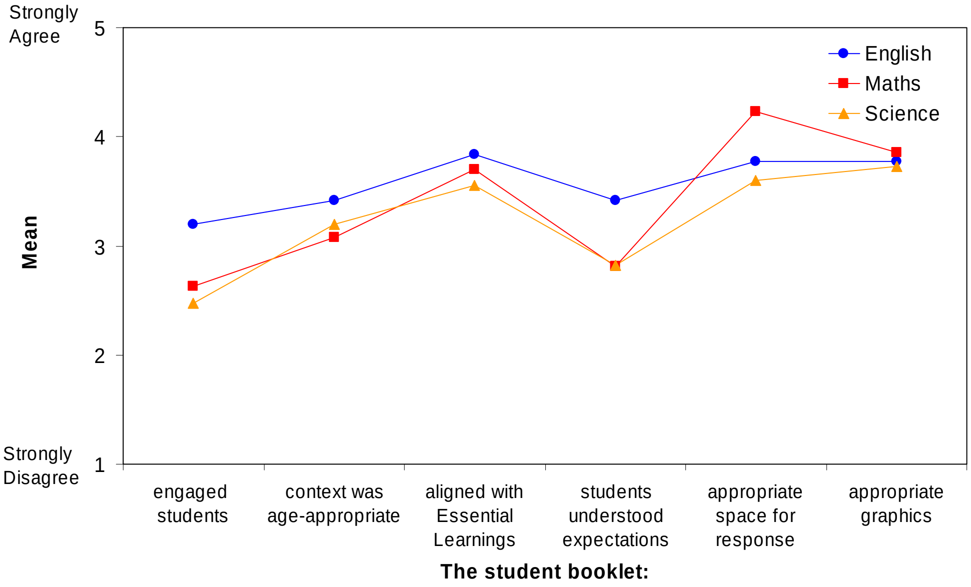


Figure 19: Mean ratings for six items dealing with teachers' perceptions of the Student booklet

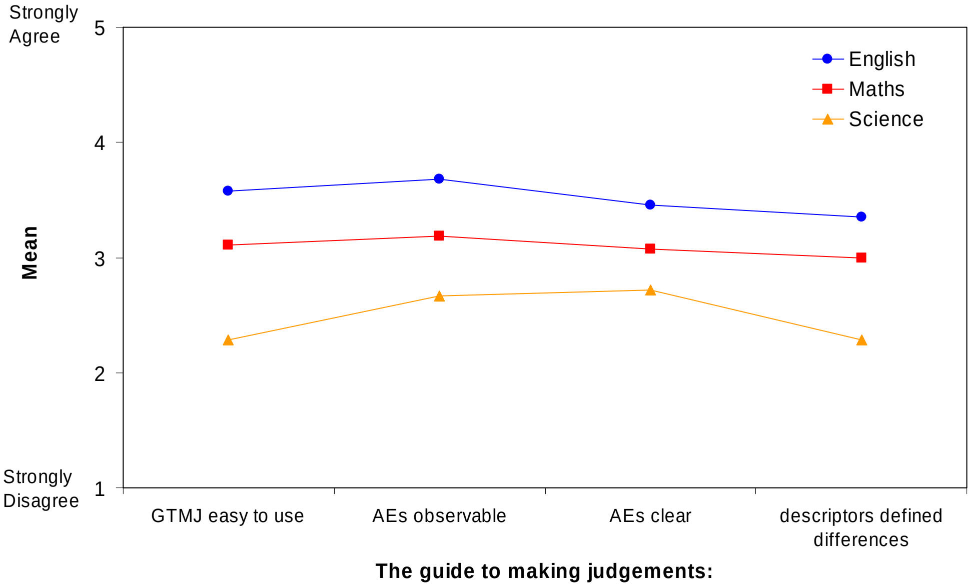


Figure 20: Mean ratings for four items dealing with teachers' perceptions of the Guide to making judgements

Finally, Figure 21 shows that the teachers were on the whole marginally in support of the propositions concerning the *Sample responses*. The separation of the KLAs in Figure 21 is not statistically significant. Teachers' comments were mixed. Some claimed that they could not assess the *Sample responses*; others claimed that they were not aware that *Sample responses* were available; while some claimed that they were generally useful. Some teachers wanted more *Sample responses* or more information within the *Sample responses* (e.g., "a range of B responses"; and they did not cover the range of responses that teachers had before them).

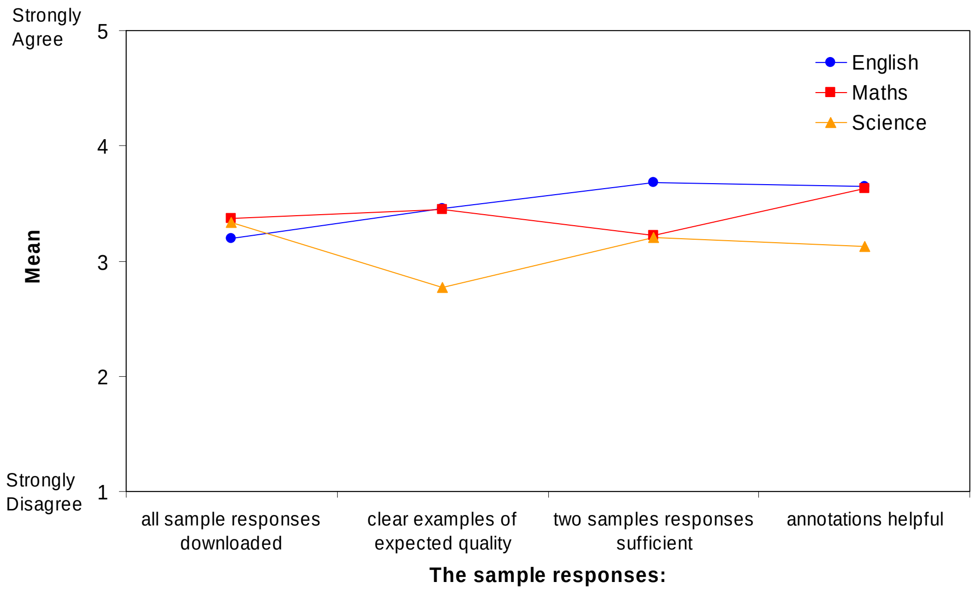


Figure 21: Mean ratings for four items dealing with teachers' perceptions of the Sample responses

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. Figure 22 shows that teachers on the whole did not agree with each proposition. Also, Mathematics and Science teachers disagreed with the first three statements about whether or not the data would inform school programs, and teachers' planning and teaching, whereas English teachers were in agreement although only marginally so.

For all sets of questions, responses patterns did not differ according to education authority (the Teacher's school was a State, Catholic or independent school) nor according to the location of the school (metropolitan, provincial city, rural).

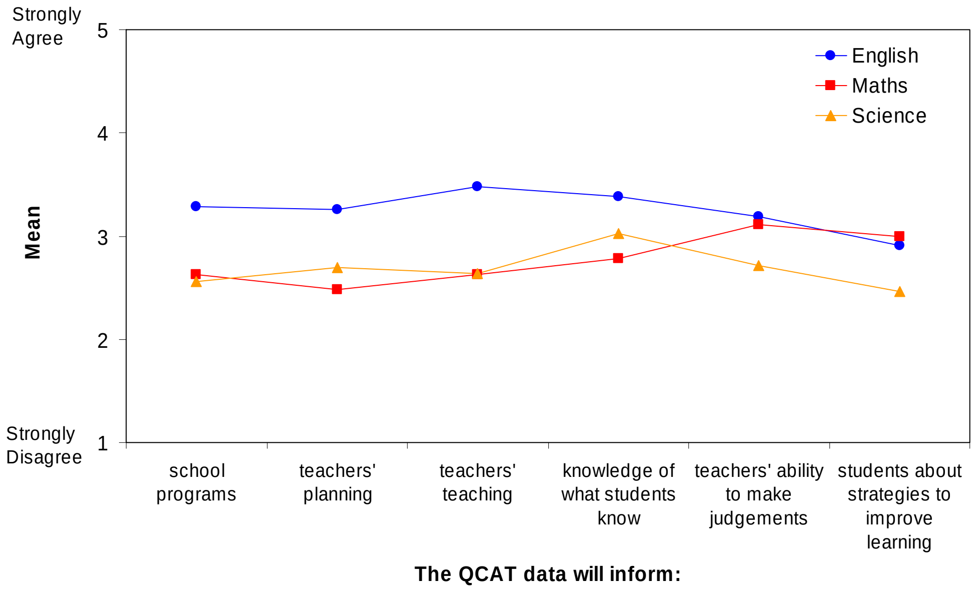


Figure 22: Mean ratings for six items dealing with teachers' beliefs about the way in which the QCAT data will inform their teaching, planning and programming

Overall, teachers argued that it would have been better for their planning purposes to know the content area of the QCAT earlier in the year or even the year before. Some argued that there was too much testing in Year 9 with QCATs and NAPLAN; other argued that the QCAT would have been better positioned in term 3. The Science and Mathematics teachers tended to be more critical of the whole process, claiming that it did not reflect the abilities of students, that students disengaged with the task, and that it was too difficult for many students.

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 Mathematics and Science teachers were able to achieve satisfactory levels of agreement with the consensus grade when awarding the letter grade. English teachers appeared to be experiencing more difficulty at achieving moderate level of agreement.

It appears that the Science and Mathematics teachers' approach to the QCATS was somewhat different from that of the English teachers. Larger proportions of Science and Mathematics students received a E grade, the rates for missing data for Assessable Elements for Mathematics and Science were larger, and Mathematics and Science teachers tended to express 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?

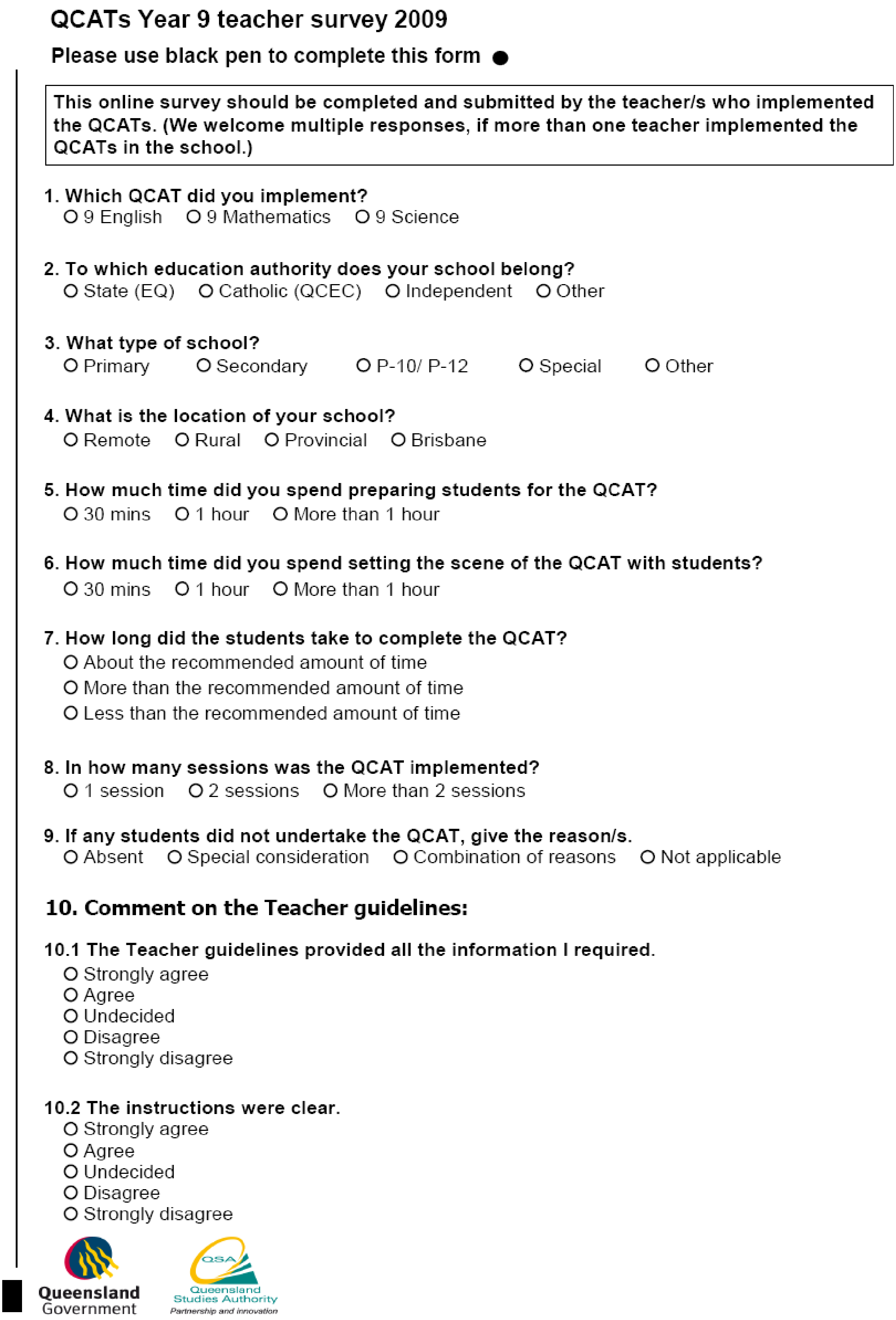
Focus area 4: Think back to any notes or marks or ticks that the teachers might have left on the QCATs.

- 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



QCATs Year 9 teacher survey 2009

Please use black pen to complete this form ●

This online survey should be completed and submitted by the teacher/s who implemented the QCATs. (We welcome multiple responses, if more than one teacher implemented the QCATs in the school.)

1. Which QCAT did you implement?
○ 9 English ○ 9 Mathematics ○ 9 Science

2. To which education authority does your school belong?
○ State (EQ) ○ Catholic (QCEC) ○ Independent ○ Other

3. What type of school?
○ Primary ○ Secondary ○ P-10/ P-12 ○ Special ○ Other

4. What is the location of your school?
○ Remote ○ Rural ○ Provincial ○ Brisbane

5. How much time did you spend preparing students for the QCAT?
○ 30 mins ○ 1 hour ○ More than 1 hour

6. How much time did you spend setting the scene of the QCAT with students?
○ 30 mins ○ 1 hour ○ More than 1 hour

7. 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

8. In how many sessions was the QCAT implemented?
○ 1 session ○ 2 sessions ○ More than 2 sessions

9. If any students did not undertake the QCAT, give the reason/s.
○ Absent ○ Special consideration ○ Combination of reasons ○ Not applicable

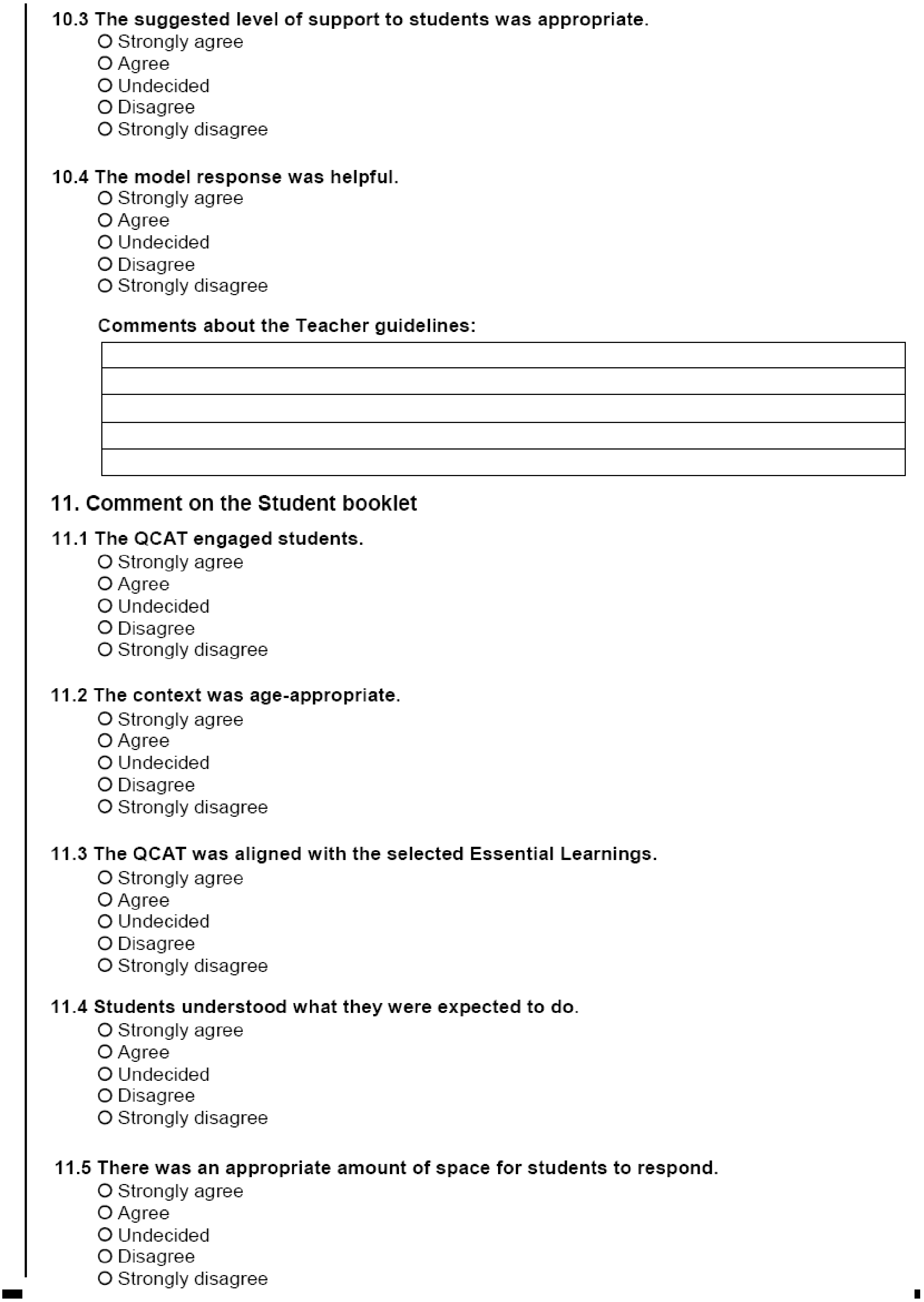
10. Comment on the Teacher guidelines:

10.1 The Teacher guidelines provided all the information I required.
○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

10.2 The instructions were clear.
○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

Queensland Government

Queensland Studies Authority
Partnership and innovation



10.3 The suggested level of support to students was appropriate.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

10.4 The model response was helpful.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

Comments about the Teacher guidelines:

11. Comment on the Student booklet

11.1 The QCAT engaged students.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

11.2 The context was age-appropriate.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

11.3 The QCAT was aligned with the selected Essential Learnings.

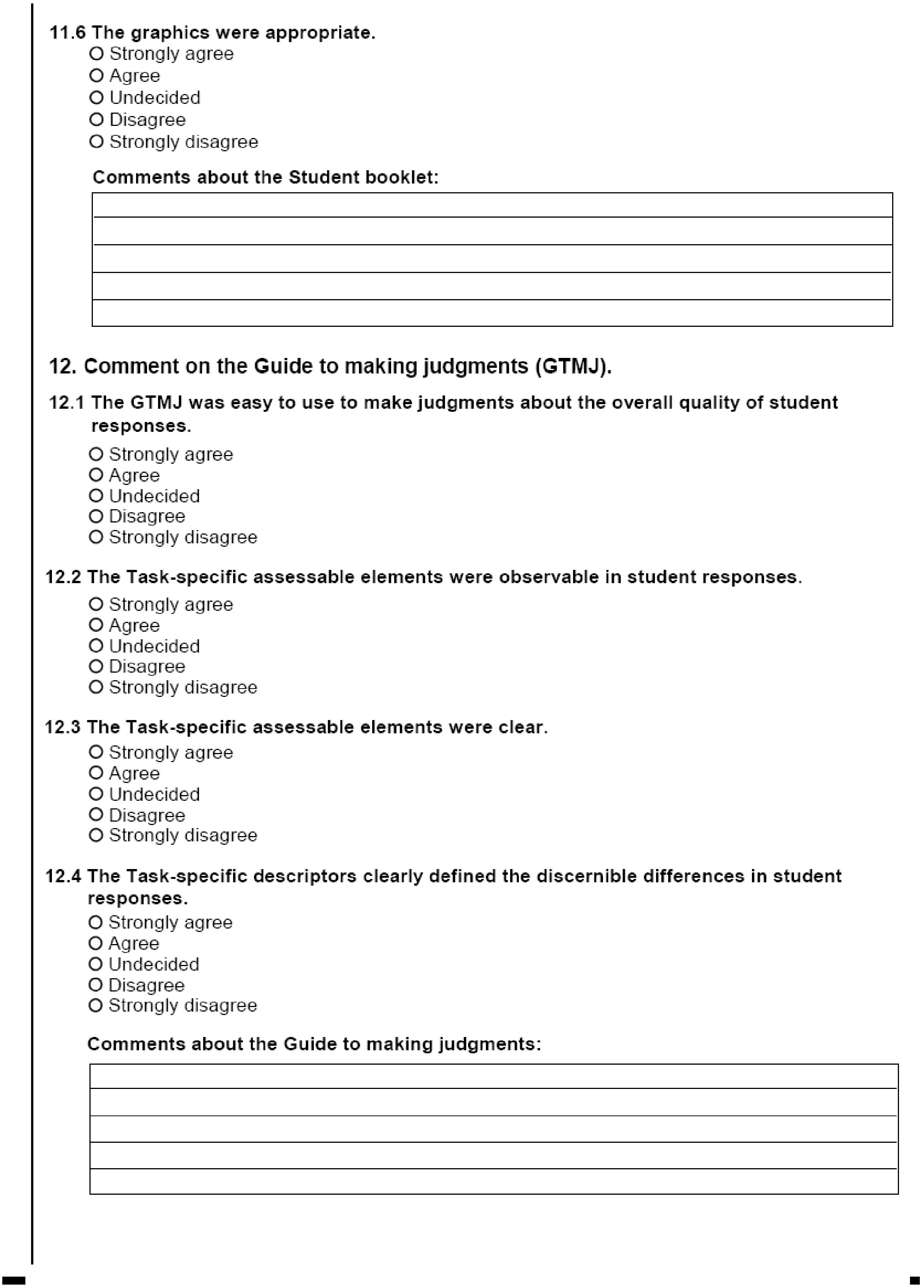
○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

11.4 Students understood what they were expected to do.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

11.5 There was an appropriate amount of space for students to respond.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree



11.6 The graphics were appropriate.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

Comments about the Student booklet:

12. Comment on the Guide to making judgments (GTMJ).

12.1 The GTMJ was easy to use to make judgments about the overall quality of student responses.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

12.2 The Task-specific assessable elements were observable in student responses.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

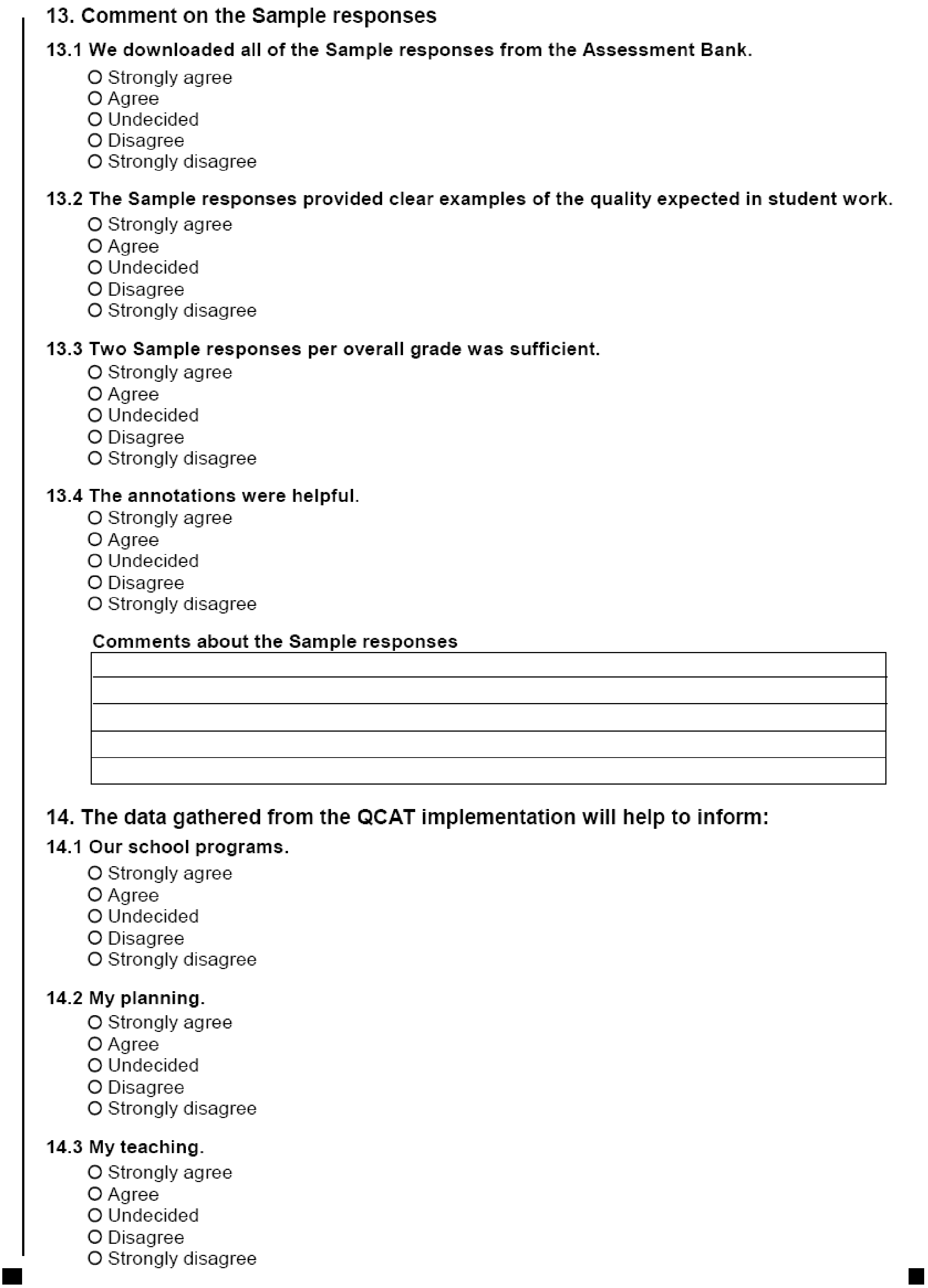
12.3 The Task-specific assessable elements were clear.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

12.4 The Task-specific descriptors clearly defined the discernible differences in student responses.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

Comments about the Guide to making judgments:



13. Comment on the Sample responses

13.1 We downloaded all of the Sample responses from the Assessment Bank.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

13.2 The Sample responses provided clear examples of the quality expected in student work.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

13.3 Two Sample responses per overall grade was sufficient.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

13.4 The annotations were helpful.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

Comments about the Sample responses

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

14.1 Our school programs.

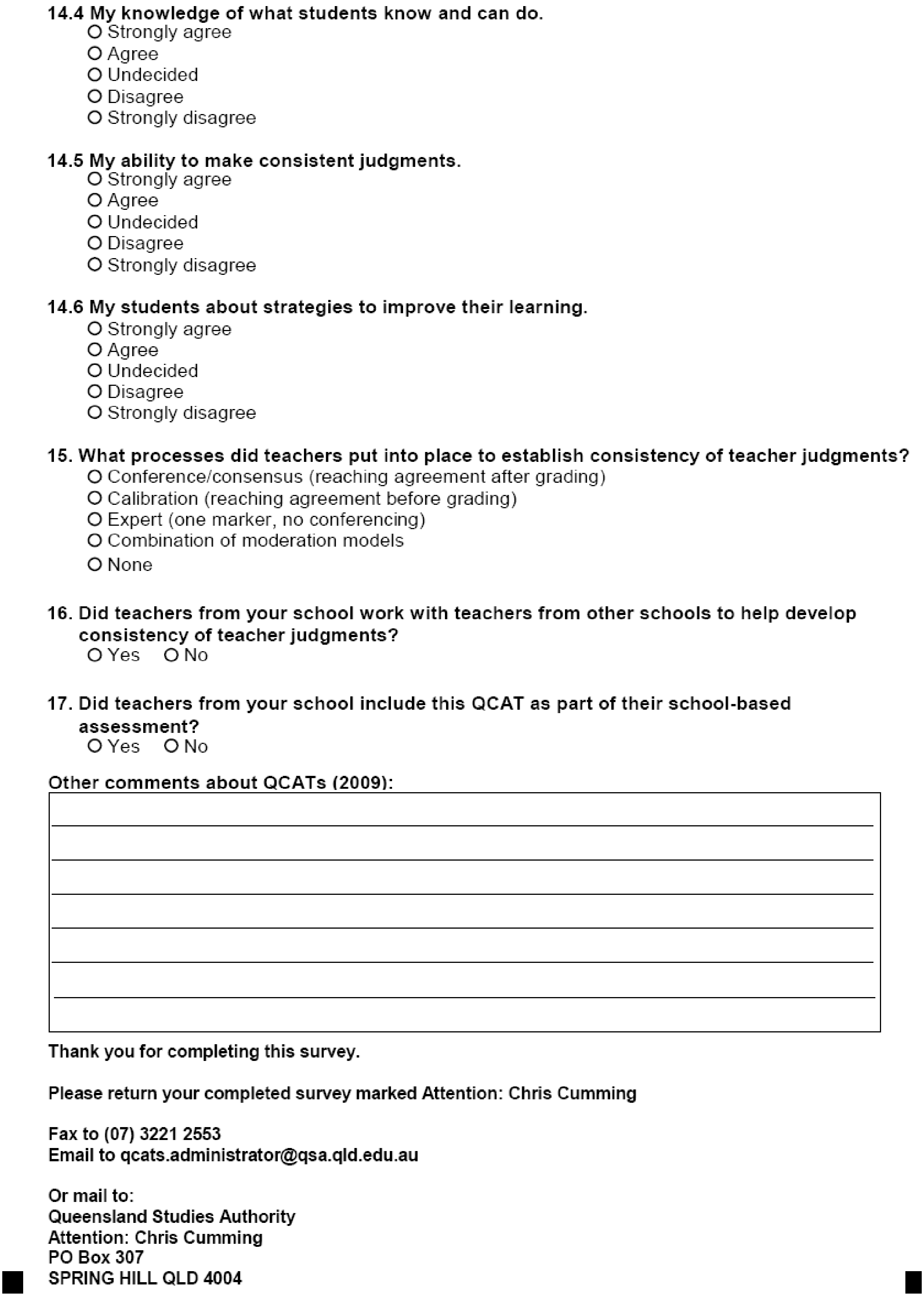
○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

14.2 My planning.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

14.3 My teaching.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree



14.4 My knowledge of what students know and can do.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

14.5 My ability to make consistent judgments.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

14.6 My students about strategies to improve their learning.

○ Strongly agree
○ Agree
○ Undecided
○ Disagree
○ Strongly disagree

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)
○ Combination of moderation models
○ None

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

○ Yes ○ No

17. Did teachers from your school include this QCAT as part of their school-based assessment?

○ Yes ○ No

Other comments about QCATs (2009):

Thank you for completing this survey.

Please return your completed survey marked Attention: Chris Cumming

Fax to (07) 3221 2553
Email to qcats.administrator@qsa.qld.edu.au

Or mail to:
Queensland Studies Authority
Attention: Chris Cumming
PO Box 307
SPRING HILL QLD 4004

Appendix 3: Summaries of Statistical Tests

Table A3-1: Chi-square tests – testing for differences among the number of missing letter grade for the 150-schools data collection

KLA	$\chi^2_{df=3}$	p
English	0.58	0.90
Maths	4.53	0.21
Science	1.48	0.69

Table A3-2: Chi-square tests - Testing for equality of number of returns across the Overall Letter Grades for the 150-schools data collection

KLA	$\chi^2_{df=4}$	p
English	18.76	0.001
Maths	4.20	0.38
Science	4.82	0.31

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 among three groups is the Kruskal-Wallis test. In the summaries below, Kruskal-Wallis analysis tests for differences for:

KLA - the three groups are English, Mathematics and Science;

Location – The three groups are Rural, Provincial, Metropolitan; and

Education Authority – The three groups are State, Catholic, and Independent.

Time spent preparing the students for the QCAT

By KLA – $\chi^2 = 0.603$, df = 2, p = 0.740

By Location - $\chi^2 = 2.355$, df = 3, p = 0.502

By Education Authority - $\chi^2 = 2.140$, df = 2, p = 0.343

Time spent setting the scene of the QCAT with students

By KLA - $\chi^2 = 1.828$, df = 2, p = 0.401

By Location - $\chi^2 = 9.753$, df = ., p = 0.021

By Education Authority - $\chi^2 = 2.371$, df = 2, p = 0.306

How long did students take to complete the QCAT

By KLA - $\chi^2 = 3.151$, df = 2, p = 0.207

By Location - $\chi^2 = 2.027$, df = 3, p = 0.567

By Education Authority - $\chi^2 = 3.060$, df = 2, p = 0.217

How many session did it take to implement the QCAT

By KLA - $\chi^2 = 0.018$, df = 2, p = 0.991

By Location - $\chi^2 = 1.619$, df = 3, p = 0.655

By Education Authority - $\chi^2 = 0.757$, df = 2, p = 0.685

MANOVAs – Survey questions 10, 11, 12, 13 & 14

In the summary below, MANOVAs test for differences for:

KLA - the three groups are English, Mathematics and Science;

Location – The three groups are Rural, Provincial, Metropolitan; and

Education Authority – The three groups are State, Catholic, and Independent.

KLA

Teacher Guidelines - Wilks' $\Lambda = 0.767$, MV F(8, 184) = 3.26, p = 0.002, $\eta^2 = 0.124$

Q10.1 F(2,95) = 5.032, p = 0.005, $\eta^2 = 0.096$

Q10.2 F(2,95) = 3.273, p = 0.042, $\eta^2 = 0.064$

Q10.3 F(2,95) = 1.172, p = 0.406

Q10.4 F(2,95) = 10.907, $p < 0.001$, $\eta^2 = 0.187$

Student Booklet - Wilks' $\Lambda = 0.776$, MV F(12, 178) = 2.002, p = 0.026

Guide to making judgements - Wilks' $\Lambda = 0.765$, MV F(8, 182) = 3.263, p = 0.002, $\eta^2 = 0.125$

Q12.1 F(2,94) = 11.323, $p < 0.001$, $\eta^2 = 0.194$

Q12.2 F(2,94) = 7.615, p = 0.001, $\eta^2 = 0.139$

Q12.3 F(2,94) = 3.281, p = 0.042, $\eta^2 = 0.155$

Q12.4 F(2,94) = 8.594, $p < 0.001$, $\eta^2 = 0.155$

Sample response - Wilks' $\Lambda = 0.853$, MV F(8, 182) = 1.884, p = 0.065

Data will inform - Wilks' $\Lambda = 0.718$, MV F(12, 178) = 2.670, p = 0.003, $\eta^2 = 0.153$

Q14.1 F(2,94) = 3.431, p = 0.036, $\eta^2 = 0.068$

Q14.2 F(2,94) = 3.094, p = 0.050, $\eta^2 = 0.087$

Q14.3 F(2,94) = 4.479, p = 0.014, $\eta^2 = 0.087$

Q14.4 F(2,94) = 1.667, p = 0.194

Q14.5 F(2,94) = 1.414, p = 0.248

Q14.6 F(2,94) = 1.979, p = 0.144

Location

Teacher Guidelines - Wilks' $\Lambda = 0.960$, MV F(8, 184) = 0.473, p = 0.870

Student Booklet - Wilks' $\Lambda = 0.813$, MV F(12, 178) = 1.621, p = 0.089

Guide to making judgements - Wilks' $\Lambda = 0.852$, MV F(8, 182) = 1.902, p = 0.062

Sample response - Wilks' $\Lambda = 0.935$, MV F(8, 182) = 0.775, p = 0.625

Data will inform - Wilks' $\Lambda = 0.887$, MV F(12, 178) = 0.913, p = 0.535

Education Authority

Teacher Guidelines - Wilks' $\Lambda = 0.947$, MV F(8, 184) = 0.639, p = 0.745

Student Booklet - Wilks' $\Lambda = 0.887$, MV F(12, 175) = 0.912, p = 0.536

Guide to making judgements - Wilks' $\Lambda = 0.967$, MV F(8, 182) = 0.387, p = 0.927

Sample response - Wilks' $\Lambda = 0.880$, MV F(8, 182) = 1.500, p = 0.160

Data will inform - Wilks' $\Lambda = 0.874$, MV F(12, 178) = 1.033, p = 0.421