Middle Years Programme

Report on the trial
MYP eAssessments
held in October and December 2013

International Baccalaureate®
Baccalauréat International
Bachillerato Internacional

Table of Contents

Table of Contents (continued)

In October and December 2013, Middle Years Programme (MYP) students from nearly 80 IB World Schools participated in trial eAssessments. The trial material, a partial examination or subject task, was delivered to schools over the internet and was downloaded to the students' computers, so they took the assessments offline. The student responses were returned to the IB over the internet, reflecting the way that the actual tests will be delivered. This means that there is no threat from internet connectivity issues during the tests themselves. The trials were held in English in October and French in December.

The trials involved five 40-minute tasks: language A, biology, history, mathematics and an interdisciplinary test.

In total, 71 IB Coordinators, 134 IB teachers and 2,367 students, almost all of whom had recent MYP experience, completed detailed questionnaires at the end of the trial tests. A total of 296 student responses were submitted and marked by 11 examiners appointed by the IB. Of these, 113 were marked by both examiners and teachers. Seven observers visited schools during the trial.

This report provides information to the MYP community about the MYP eAssessment trials—their successes and lessons learned, the views of the examiners on the responses they saw, data about student performance and, most valuably, a selection of comments from the teachers, coordinators and students who took part.

The eAssessment tasks

The five tasks in the trial of the MYP on-screen eAssessment differed in their formats, which were appropriate to the academic disciplines, such as the inclusion of graphs and diagrams in the biology and mathematics task and the use of text-based sources in the history task. The format of each task was not determined by the technology. In this trial, the tasks themselves were not subjected to scrutiny by teams of question authors (acting as critical friends on each other's papers) and external advisers, which is normal practice in live examinations.

The differences in design complexity between the tasks, and resulting technological demands, inform the conclusions that can be drawn from the student and teacher survey results. These differences occasionally explain the variance in student responses to the interface and their experience with the on-screen assessment. A sample of completed on-screen eAssessments was subject to marking by examiners and teachers. The results of this marking exercise inform and support the findings and conclusions based on the student and teacher surveys.

The students

In total, 2,367 students filled in the student survey. Almost all students had MYP experience (96%); 79% of the responses are provided by current MYP students and 17% come from Diploma Programme (DP) students with MYP experience (see Figure 1). Only 4% of the respondents had no experience in the MYP.

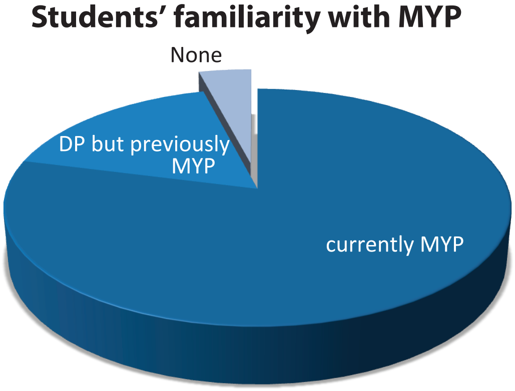


Figure 1: Share of students with MYP experience in respondents

The three represented groups not only differ in their experience of the MYP, but also in age, with current MYP students being closest to the target population for eAssessment. Where age leads to different outcomes, results are presented for the whole sample as well as for current MYP students and others separately.

It was clear that the great majority of MYP students use computers regularly in their schoolwork and under 2% had not needed or used their computer or mobile device for schoolwork in the past week.

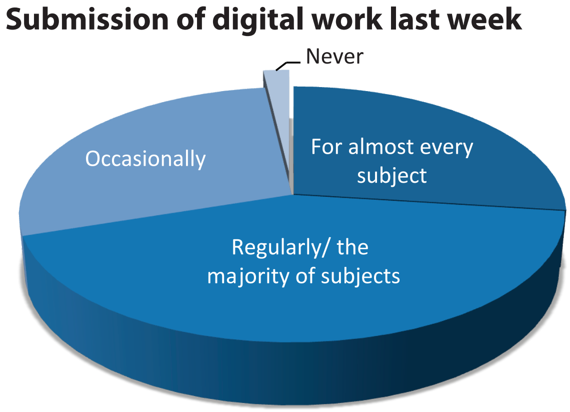


Figure 2: Share of students who submitted digital work for subjects in the last week

Twenty per cent (20%) of the students in the survey had been given access requirements or extra time for an examination at some point in the past, but they were able to complete the trial tasks without additional support. The percentage of students with experience with access requirements was about the same for each of the five tasks and more common for current MYP students. In the live eAssessments, access provision will be made.

The teachers

In total, 134 teachers, with 110 currently teaching four of the five trialled subjects, completed the teacher survey, giving feedback on all five tasks trialled. The large majority of the teachers (72%) had no experience as a DP examiner/moderator or MYP moderator; about one fifth, 18%, had experience as a DP examiner or moderator; and 11% had experience as an MYP moderator. As Figure 3 shows, there was a range of very experienced and less experienced teachers in the sample.

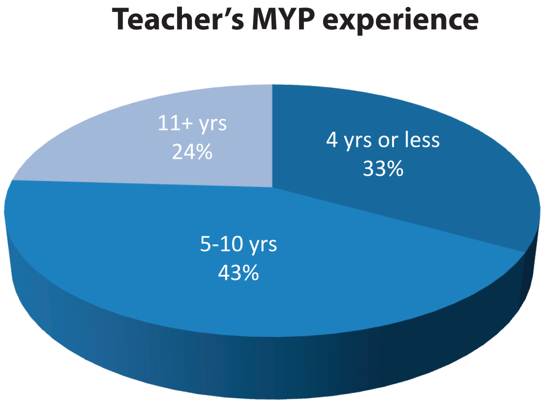


Figure 3: Teachers' teaching experience

A total of 74% of teachers required digital work from students at least a few times in a unit. This pattern was common in all five subjects evaluated in this report and corroborates the finding that almost all the students had submitted class or homework in digital format. Figure 4 shows there are large differences between subjects, with three-quarters of the history teachers requiring digitally submitted work in the majority of lessons, while digitally submitted student work is much less common in biology and mathematics.

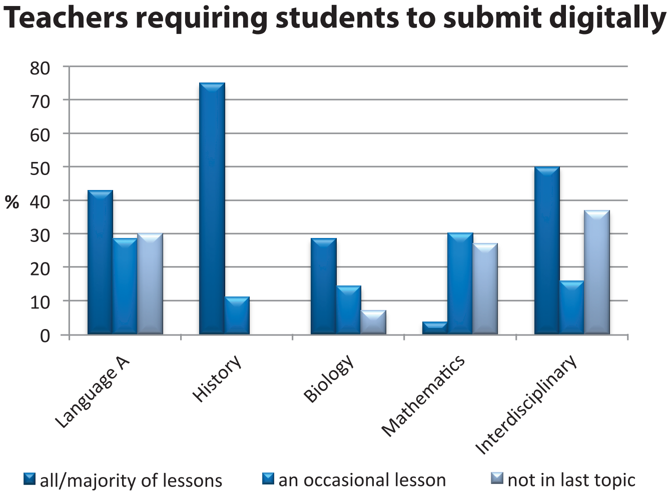


Figure 4: Frequency of required digital submission of student work by subject

The examiners and marking

In total, 11 examiners marked a sample of English language student responses to the trial tasks. Tasks were double- or triple-marked by examiners appointed by the IB in each of the trial tasks. In addition, a small number of teachers generously submitted their marking of their own students' tasks. The marking by teachers and examiners provided the opportunity to examine the students' performance on each of the questions and separate items. These performances are compared with the views of the teachers and students about the tasks described in the survey outcomes.

The small sample sizes and the lack of examiner standardization normally carried out for all IB examiners has led to results—presented in the task-specific sections of this report—that must be interpreted with caution. For example, readers of the report should not infer that differences in marking standards shown by the different examiners in the same subject will be the case in the live assessments. Rather, they can more likely be ascribed, at least in part, to the inability to standardize marking standards between examiners in the trial.

Feedback on on-screen assessment

Students' experience of technology

The large majority of students (89%) had no trouble logging into the on-screen assessment; however, 23% of the students experienced technical problems when completing the tasks. Of these students experiencing problems, most (70%) were able to solve the problem, either by themselves or with help, and proceeded with the task. Just 3% of all students could only complete parts of the task, and 58 respondents (2.5%) were not able to access the actual on-screen task at all because of technical problems (see Figure 5).

All tasks contained source materials (ie, texts, images, graphs, diagrams or videos) that could have posed technological problems. Almost 60% of the students experienced no problems in accessing the source materials during the task. Seventeen per cent (17%) of students referred specifically to problems with the videos, indicating that they failed to play or that the sound failed. The need to scroll through the sources a lot was mentioned by 12% of students, while another 9% mentioned difficulties because the source (for example, image or diagram) would not enlarge.

The most commonly mentioned problem in the student comments was the video not starting or the sound not working, and changing the resolution of the screen and its subsequent failure to revert back to the original resolution occasionally resulted in problems. A third type of problem mentioned by students had to do with the text boxes—that is, the loss of responses when going back or wanting to change responses, the need to retype responses and the inability to delete "typed in" responses. A more detailed summary of the technical problems mentioned was submitted to the software development team.

The complexity of tasks influenced the occurrence of technological problems students experienced during the task. The students accessing the language A and history tasks experienced significantly fewer problems, which is not surprising because the design of these two tasks was relatively straightforward and did not call on advanced technological capabilities. Students participating in the most technologically demanding task, biology, experienced significantly more problems during the task (32%), followed by those completing the interdisciplinary task (31%), the mathematics task (22%), history (21%), and finally the least technologically demanding task, language A, with 16%. Furthermore, the complexity of the task influenced the students' ability to solve the problem themselves or with help, with significantly more students not able to finish the mathematics or biology task due to problems.

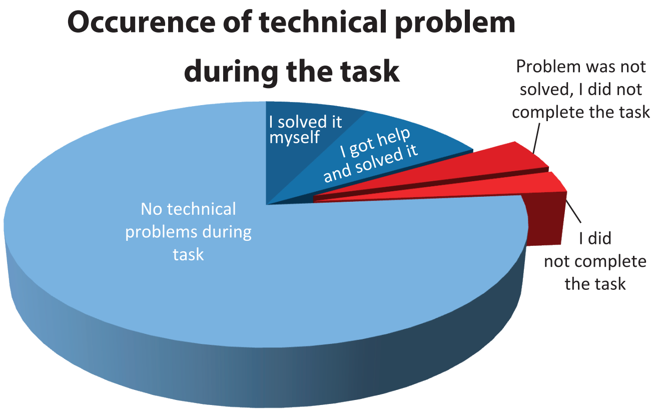


Figure 5: Did a technical problem occur during the task and, if so, how was that solved?

Students' use of the technology

On starting the test, students were presented with an overview screen, which gave background information on the whole task and listed the questions. Students could navigate through the tasks using a status menu to the right of the screen, which also showed them how much of the overall task they had completed in a progress bar. Students could bookmark a response they wished to return to later, and they could call for assistance if needed.

Students most frequently used the overview screen to read some or all of the questions and then start with the first question, followed by using the screen to choose the next question to work on—which may overlap somewhat with the third common strategy, using this screen to go back or skip a question (see Figure 6). Younger students or students who have experience with special arrangements did not differ from other students in their strategy for using the overview screen. Neither did students who completed homework on their digital devices for most of the subjects differ from the other students.

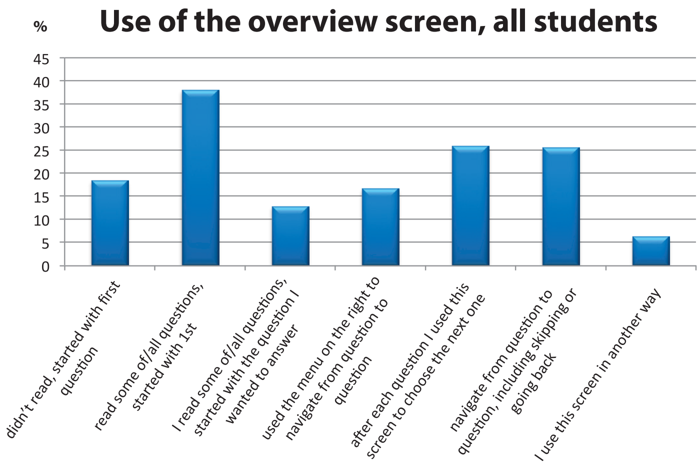


Figure 6: Use of the overview screen to navigate—all students

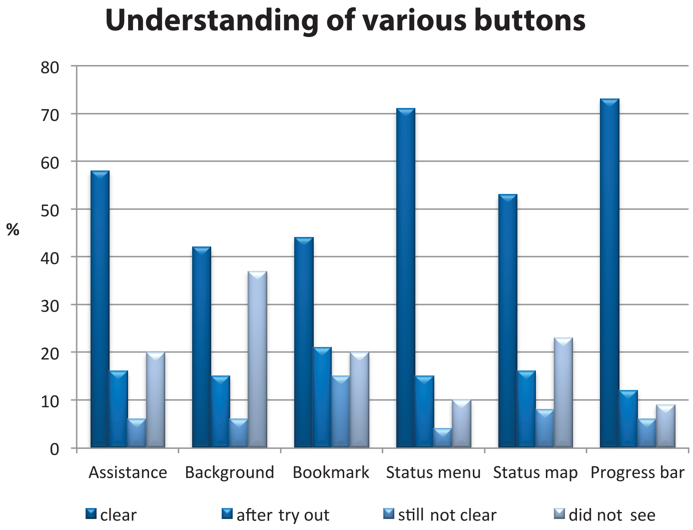


Figure 7: Identifying the purpose of buttons and icons—all students

Figure 7 shows that the blue bookmark and the background button were the most unfamiliar buttons, followed by the status map and the assistance button. While a background resource section was included on the overview screen in four of the five tasks, the background button itself (placed above the status menu on the right) was only present in the biology and interdisciplinary tasks, which explains the relatively high proportion of students who had not seen it.

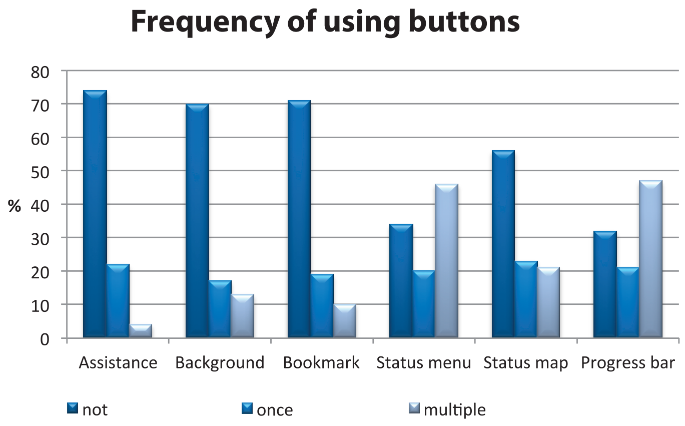


Figure 8: Frequency of using buttons and icons—all students

A total of 83% of students used the three icons meant to assist monitoring completion of items, that is, the blue tick, orange pencil and red zero percent indicator. Students with experience in special-access arrangements for examinations did not differ from the other students with regard to identifying and understanding the buttons, nor did they use the various buttons differently than the other students. In particular, they did not use the assistance button more frequently. The extent to which students were used to doing homework on their computer did not translate into a more immediate understanding of various buttons or icons in the interface, nor did it influence the frequency of use.

Issues with typing responses

In total, 73% of students were comfortable with typing their answers. Of this number, 43% felt it was an advantage and another 30% felt that typing was OK and did not slow them down. Students with experience of special-access arrangements were slightly more positive about typing, with 78% feeling that typing was OK and did not slow them down. Not surprisingly, there is a positive and statistically significant relationship between the frequency of using computers for schoolwork and the favourable response to typing. However, a positive reaction to typing was clearly related to the response types required. Of the students completing the tasks with predominantly text-based responses—language A, history and interdisciplinary—only about one in 10 (12% to 15%) felt disadvantaged by typing. However, the number of students who felt that typing was a disadvantage was significantly higher for the biology and mathematics tasks that included other response types, with numbers increasing to one in five (20%) in biology and about half (52%) in mathematics.

A number of comments referred to the disadvantages of typing responses in mathematics, indicating a preference for handwriting and drawing by hand to solve calculation or mathematical problems.

> I think E-assessments are good for subjects such as English and Humanities which are writing-based, but for subjects like Math and Science, it will be hard because it takes time in input fractions, subscripts, symbols etc. on computer.

Some comments

However, a small number of students completing other tasks with text-based answers commented that they would prefer writing by hand as they felt their answers would be better thought through or more detailed, or generally that hand writing gave them more opportunity to think. A main advantage of typing was the opportunity to edit their answers, although the current lack of a word count or spell checker was seen as a disadvantage of typing. A few students who liked typing had problems due to an unfamiliar keyboard or small technological problems, while other students mentioned that working from a screen and the noise of typing was uncomfortable in an examination situation. In the French trial, a number of teachers echoed the students' concern that using an unfamiliar keyboard disadvantaged some students.

I really liked the online examination for this particular exam because it involved a lot of writing and editing. This was easier to type because it was neater and faster to complete the test.

Typing out answers on assessments like this, for some odd reason, does not make me half as anxious as I am when I write them. I feel more confident when I type my answers, because I feel that I can elaborate more. Writing is something I do when I am emotional, and typing is when I am doing something like essays. Typing makes my focus academics and academics only.

I think that onscreen examinations are much better because there is no chance of cheating someone else's work; it is challenging because we have to do more work in comparatively less time. The examination was a really good experience as it really depended on our own intellectual and analytical skills. The students can also edit their work repeatedly if done something wrong, and there is no chance of incorrect or unfair marking through the on-screen exams.

Some students' comments

A couple of the students commented on the fact that the eAssessments made cheating in the examination harder—although that was only if the room was laid out as advised in the guidance documentation. Otherwise, as one student put it, "It is quite easy to cheat, you can just look at the screen in front of you."

Time issues

Students and teachers were divided about the allocated time. Although a large number of students seem to have been able to finish on time, frequent mention was made that perhaps 40 minutes was too short. It is unclear in this analysis if this issue is related to the native language of the students not being the same as the test language (English or French).

The 10-minute familiarization period was generally felt to be too long. Points raised inlcuded:

- that this was particularly relevant for students participating in a second task
- teachers and invigilators had to stop students from starting too early
- the length of the familiarization period would be unnecessary if students and schools could access practice assessments.

Marks per item

Marks per item were used by half of the students to determine how much effort to put into answering, with the slightly older students (former MYP students) using this indicator more often to guide their efforts. A large majority, 75%, of the teachers thought that the marks awarded were appropriate.

Almost all the teachers who participated in marking felt they were able to apply the markscheme provided (92%) with no differences between subjects. A number of teachers commented that marking this way was very different from the current assessment practice in MYP.

Use of scrap paper

Scrap paper was not provided for all tasks or in all schools. Half of the students (1,230) answered a question on the use of scrap paper. Of the 597 who indicated they had been provided with scrap paper, 62% used it. Another 151 students, 12%, indicated that they would have liked to use scrap paper, but it was not provided.

Feedback on the tasks

Difficulty level and student performance

About half of the teachers (49%) felt that the level of difficulty of the on-screen task matched the level of difficulty expected in assessment at the MYP level. Conversely, 38% thought the level of difficulty was lower, while 13% thought it was higher than expected. A total of 58% of the teachers thought that all or the majority of the examination responses they saw reflected a level of ability matching their expectations for these students.

The majority of students felt the tasks were of a level expected for the MYP. Overall, only 16% felt that the on-screen assessment was more difficult than they were used to; however, a larger proportion of students who completed the biology and mathematics tasks felt that it was more difficult than expected (29% for biology and 21% for mathematics). Looking at whether students felt that the level of difficulty in the assessment was appropriate and at a level that matched expected levels for the MYP, only 37%–40% of the mathematics and biology participants felt that the level was as expected in the MYP, compared with 60%–65% for the other three tasks. Some student comments are included in the task-specific sections.

Understanding expectations and language

The great majority of students felt that they understood what was expected of them when they took the tests. However, 29% of students felt that there was a part of a task where they did not fully understand what was expected of them, with biology (30%) and mathematics (21%) tasks mentioned more often as difficult because of this. Approximately two out of five teachers (44%) agreed with the students, with biology, mathematics and language A being mentioned significantly more often.

Almost all of the teachers, 92%, thought the language used in the tasks was familiar to MYP students. Equally, most students felt that they understood all the words and terms used in the tasks (79%). However, the mathematics task most clearly included unfamiliar terms, with 52% of the students not understanding all the words and terms. Students with experience of special-access arrangements more often felt the tasks included language or terms they believed were unfamiliar, making them unsure about what was expected of them.

Tables with the terms and words that MYP students mentioned were unfamiliar or made them unsure about expectations are included in the Appendix (Tables A.1 and A.2). It is possible that this was more apparent in non-native speakers, as some of the students mentioned that specifically as a reason they did not understand the question. After technology, language was the second most significant cause of students' insecurity about the expectations of the tasks. Other sources of insecurity were the lack of cues on how large a response was expected (big boxes for short, numerical answers, no word count tool or maximum length indicated, and so on) or the type of response expected (no clues on assessment objectives, rubrics, and so on in the task description).

Assessing conceptual understanding

A total of 70% of the teachers thought that the task assessed a level of conceptual understanding that was consistent with the MYP level. The language A task was considered to provide an opportunity to focus on organization, content and grammar within a creative response, but it also restricted students' ability to show conceptual understanding or literary analysis skills. Some of the history and interdisciplinary task items were thought not to require conceptual understanding and were interpreted as reading comprehension items. The biology task was seen to be very content specific.

Assessing skills

Most of the students, 84%, felt that the skills assessed were familiar, but again there were differences by subject, with students in biology (29%) and mathematics (27%) feeling these tasks required more unfamiliar skills. There is no statistically significant difference between younger and older students, students with or without experience of special arrangements, or students who do or do not frequently use computers for school in their evaluation of the skills and level of the on-screen assessment of these tasks.

Again, about four out of five teachers (84%) did not think a new skill was assessed. Those who felt new skills were involved mentioned:

- spontaneous and/or synergetic thinking
- linking concepts and content to current issues
- linking knowledge and skills from various disciplines
- coping with a lot of source material in the time limit
- manipulative computer skills (manipulating objects on-screen)
- content not (yet) covered, for example, continuous graphs, (quadratic) functions, statistics and logic problems
- required response types, for example, writing in response to a visual prompt.

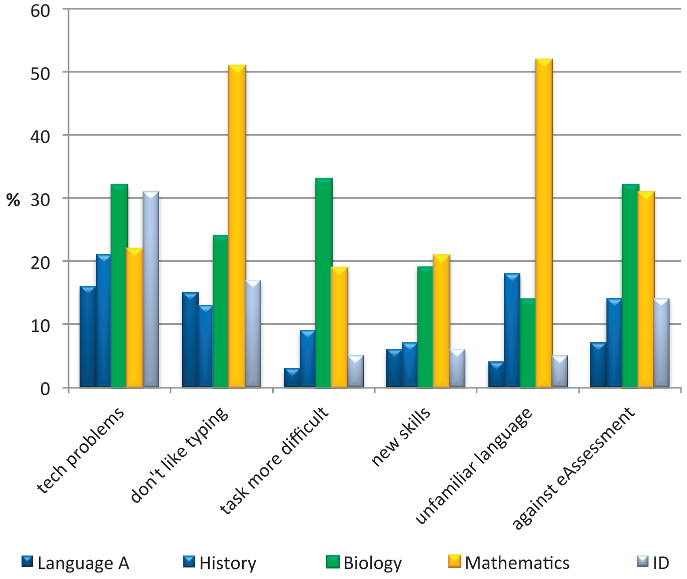


Figure 9: Students' experience per subject—all students

The language A task

For this task, students were asked to respond to a single question worth 16 marks. A total of 527 students completed a survey giving feedback on this task, and their experience is summarized in Figure 9.

Compared with the other tasks, the language A task was relatively free of technological problems, and it was perceived as easy by students. The task, not a full assessment, focused on creative writing, which in the MYP can be perceived as an easier element of assessment. Compared with the other tasks, students completing the language A task were relatively positive about on-screen assessment, with 61% probably or definitely recommending it for this subject.

The teacher comments on the language A task focused on the limiting character of the prompt or stimulus, encouraging their students to attempt to write from a naive or childlike perspective and therefore limiting them in expressing conceptual understanding or literary analytical skills. This task was seen as different from the usual MYP task, making students and teachers unsure of what to expect. While the question was intended as an option requiring only one type of response, teachers thought that, because of the format, many students would respond to all three prompts. Following that reasoning, teachers noticed that students mixed up their responses in the options. Students interpreting the options as scaffolding for the response may have followed a different logic than the question-setter, describing what they saw (option C) first, followed by their narrative of what followed (option B) in the text box. In addition, the wording of the question and the size of the text box would cue students towards short narratives. Formulations requiring explanation included "narrative", "descriptive" and "sophisticated internal monologue".

The students confirm the teachers' comments that the options were confusing, as well as occasionally highlighting the formulations mentioned above as unfamiliar. However, the most frequent comment referred to the lack of indication on the required or expected length of the response.

It would have been better if you included the fact that we only had to write one essay in the requirement. It was explained by the teacher, but writing it on the task sheet and making the questions clearer would have been a good idea.

For the prompts to be chosen, didn't know whether we had write a story, number of paragraphs, word count, rubric, requirements, everything was vague.

Nous ne savions pas vraiment quel type de texte il fallait écrire, ce qui compliquait quelque peu la tâche.

J'aurais aimé savoir sur quelles critères j'étais évaluée et quelle était la grille de correction. J'ai écrit un texte, mais je ne savais pas dans quel.

As a student who loves the subject English, I found the questions very interesting despite the fact that they were quite similar. I feel like the photo chosen and the questions asked allowed the students to imagine as much as they want and write as much as they want as well, because this photo can have different kinds of stories. Overall i really enjoyed answering the question and found myself typing on and on, to the point where I did not want to go back to class when it was over.

Habituellement, nous avons des indications plus précises pour un texte à écrire.

Monologue is something I am not familiar with.

In the explanations, some words were a bit complicated, but once put in context it made sense. Maybe the English should be just one level easier!?

Some students' comments on the language A task

The work of 60 students from 16 different schools was used in the marking analysis. All 60 students were marked by two examiners, and 31 were marked by both examiners and teachers at trial schools.

Examiner agreement

The box plot (Figure 10) and Table 1 below show a plot of the distribution and summary statistics of marks from each examiner. The mark distributions that the examiners awarded were very similar, with the senior examiner having a slightly less tightly spread set of marks than the second examiner, and both mean marks were just above 50% of the maximum mark. Both box plots are broadly symmetrical, suggesting that there is no skew in the distribution, although it is hard to draw conclusions from this due to knowing very little about the students sitting the exam.

The correlation between the marks each examiner gave was reasonable (r = 0.84), although the two examiners only gave the same mark for 25% of the students and were more than one mark apart for exactly half the students.

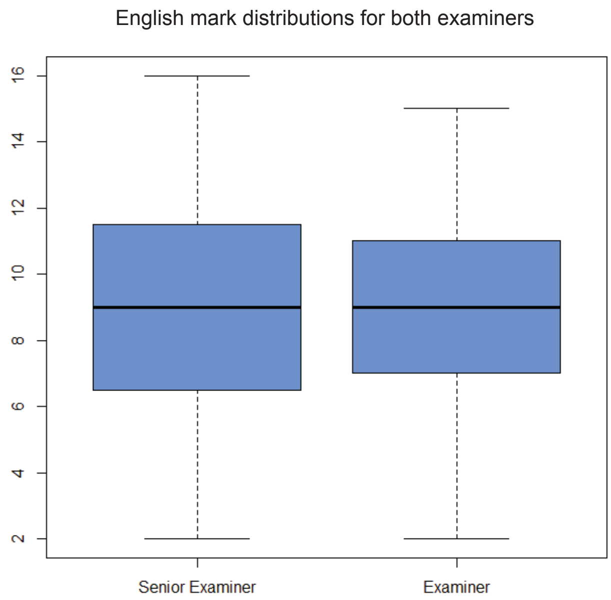


Figure 10: Examiner marks for English A

	Mean	Std. dev.	Min	Max	N
Examiner 1	8.99	2.84	2	15	60
Examiner 2	9.25	3.46	2	16	60

Table 1: Mean examiner marks for English A

Examiner/teacher agreement

The box plot and summary statistics (Figure 11 and Table 2) below show the distribution of marks given to the 31 students marked by both examiners and teachers. There were relatively few students and they were spread over a large number of teachers, so any analysis is of limited value; however, the teachers as a whole were a little more generous than both examiners. Agreement between the senior examiner and the teachers was quite low, with only 6 out of the 31 students being awarded the same mark and 12 being 4 or more marks different, which is considerable on a 16-mark item.

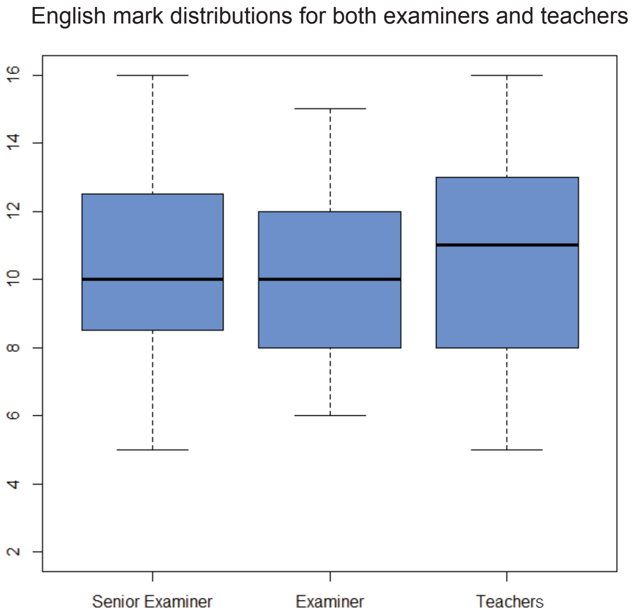


Figure 11: Examiner and teacher marks for English A

	Mean	Std. dev.	Min	Max	N
Senior examiner	10.36	2.99	5	16	31
Examiner	10.00	2.49	6	15	31
Teachers	10.71	3.10	5	16	31

Table 2: Mean examiner and teacher marks for English A

How do the outcomes match teacher and student comments on the task?

In their responses to the surveys, the teachers felt that students could have been disadvantaged by not understanding what the question wanted from them; however, the marks awarded by the examiners do not seem to indicate a general failure to respond to the question appropriately and the students themselves did not report that they struggled with this task in their survey responses. Notably, the sample of marked responses was very small.

The history task

The history task comprised three questions, with a maximum mark of 36. A total of 326 students completed a survey giving feedback on this task. Students who completed the history task had a very similar experience to those who completed the language A task (see Figure 9). The task was relatively free of technological problems, typing was an issue for a minority of students, the task was not seen as more difficult and very few felt new skills were being assessed. More students felt that unfamiliar language was used in the task; in particular, they were unsure about the thinking routine from the new MYP history curriculum that was introduced to scaffold the task: Origin, Purpose, Value, Limitation (OPVL). The students were only slightly less positive than the language A students about on-screen assessment of history, with 60% probably or definitely recommending it for this subject. Students commented that the language was relatively complex, especially for non-native speakers. In particular, the language used in the second and sometimes the third question made it more difficult to understand expectations.

When it said Origins, I didn't know what was expected. It said compare and contrast at the top, but at the subquestions it only said Value, Purpose, Limitations, Origins.

Because I'm not a Native English Speaker, it was sometimes really hard for me to understand what they want from me. I think it should be better that after a word a good and easy explanation. It will sometimes also good to have a example.

La deuxième et la troisième questions était vraiment pas clairs.

In question two, we were asked to explain the origin of the quote. I assumed they were talking about the reason for the quote. Apparently, we were expected to write about the reference of the quote. I also found the word value too broad (I would have found it easier if they would have used a more specific word).

Il y avait une question où ce qui était demandé n'était pas clair. Je ne savais pas si on voulais que je nomme les valeurs véhiculées par le texte ou la validité des sources.

Question about sources. I thought you were supposed to talk about the origin, value etc of war, not about the sources. When I asked my teacher it was clear.

La partie des origines, valeures et limites n'était pas expliquer du tout.

Some students' comments on the history task

For the history task, teachers felt that the first question was too easy, while the structure of the second question and the markscheme appeared not fully aligned, making teachers unsure of what the expectations were. It was seen as too "Diploma Programme" oriented, requiring analysis levels higher than realistic at the MYP level. However, the students and teachers participating had mostly not experienced the piloted MYP history curriculum that the task was tailored to assess.

For the marking trial, 60 students were marked by two examiners and 23 students were marked by both examiners and teachers at trial schools.

Examiner agreement

The box plots and summary statistics of the marks that the two examiners awarded are given below in Figure 12 and Table 3. We can see that the senior examiner is more generous by about two marks and didn't award many low marks.

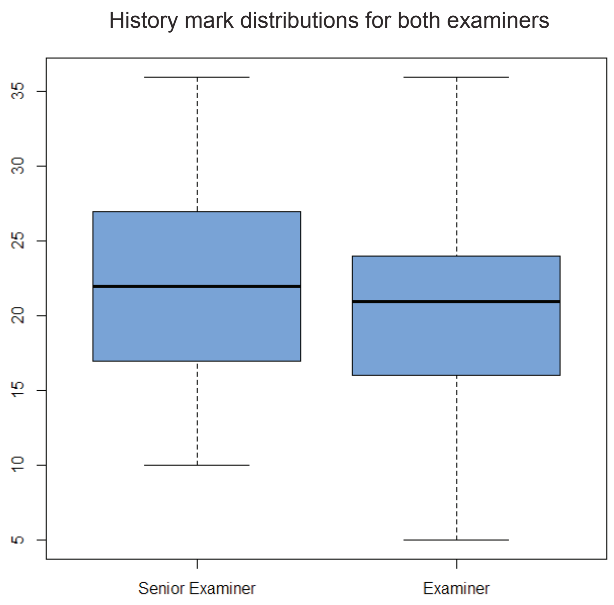


Figure 12: Examiner marks for history

	Mean	Std. dev.	Min	Max	N
Senior examiner	25.55	6.52	10	36	60
Examiner	20.38	6.58	5	36	60

Table 3: Mean examiner marks for history

The two examiners only awarded the same total mark to 8 out of the 60 students and only awarded the same marks on every question for 3 students out of the 60. There were 11 students where the difference between the examiners' total marks was 5 or more marks. The correlation between the examiners' marks is 0.93, suggesting that, while the senior examiner was the more generous of the two, the differences followed a fairly consistent pattern.

Comparison with teacher marks

There are only 23 students who were marked by both examiners and teacher, making conclusions difficult to draw, but it appears that the teachers were generally slightly less generous than both examiners, particularly the senior examiner (see Figure 13 and Table 4 below).

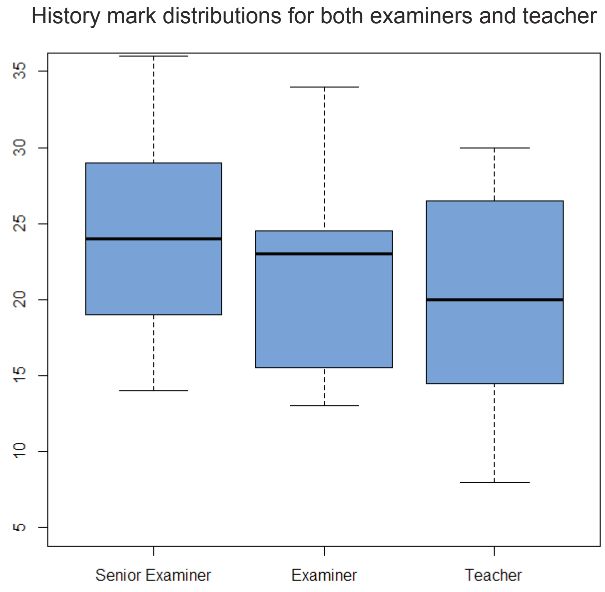


Figure 13: Examiner and teacher marks for history

	Mean	Std. dev.	Min	Max	N
Senior examiner	24.17	6.42	14	36	23
Examiner	21.39	6.16	13	34	23
Teacher	20.30	7.19	8	30	23

Table 4: Mean examiner and teacher marks for history

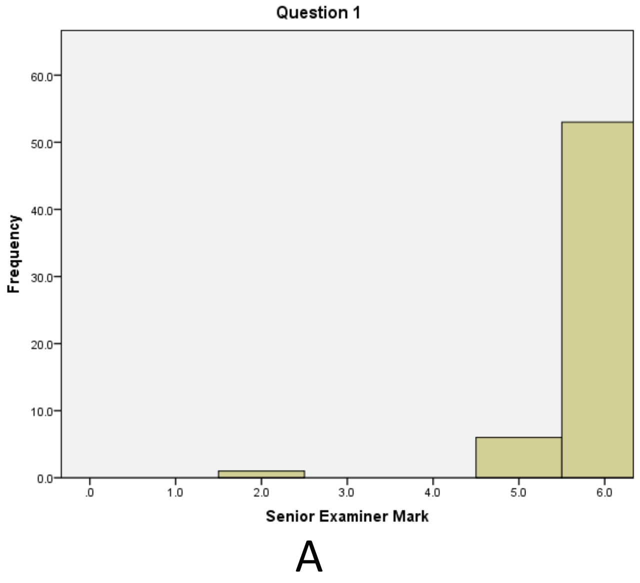
Question-level performance

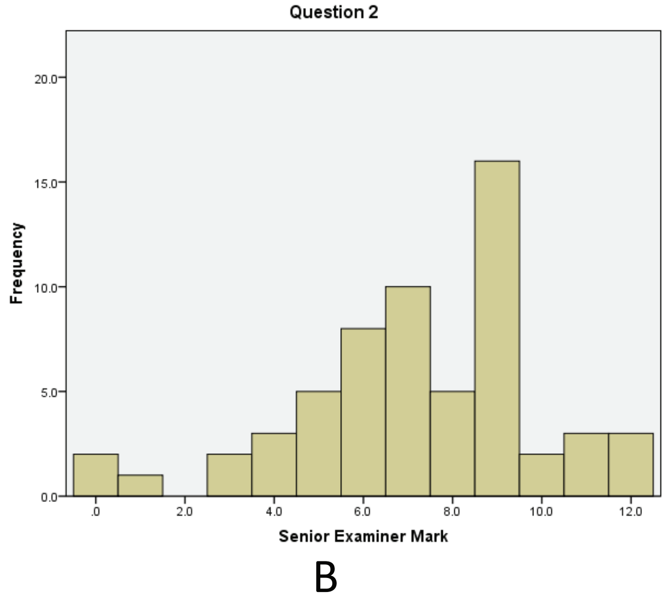
Table 5 below shows the mean mark that the senior examiner awarded each question, along with what that mean corresponds to as a proportion of the total marks available.

	Question 1	Question 2	Question 3	Total
Senior examiner	5.83 (0.97)	7.25 (0.60)	9.37 (0.52)	22.55 (0.63)
Maximum mark	6	12	18	36

Table 5: Mean senior examiner marks per question for history

The very high mean mark in question 1 suggests that students found this question very easy, which corresponds to the response from the teacher survey. However, the other two questions seemed to work much better, with reasonable mean marks and correlations, which doesn't reflect the teachers' concerns with question 2. The histograms below (Figure 14 A, B and C) show the distributions of the marks awarded on each question by the senior examiner.





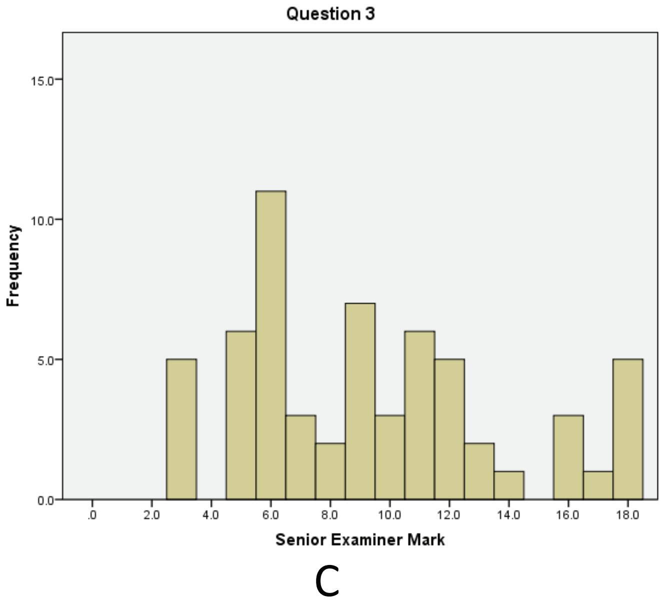


Figure 14: Mark distribution per question for history

How do the outcomes match teacher and student comments on the task?

The teachers felt that question 1 was too easy and that the markscheme for question 2 would not allow students to score well. However, this question had a higher mean mark than question 3, suggesting that it was not such a problem, and the students in the survey did not feel that they struggled with the demands of the task. The fact that students performed very well in question 1, with a mean mark of 5.83 out of 6, supports the teacher feedback that this was an easy question.

The biology task

The task contained five questions with sub-questions and a maximum mark of 30. A total of 330 students completed a survey giving feedback on this task. The student experience of on-screen assessment of biology was different from the other tasks in a number of aspects (see Figure 9). In particular, the task was seen as more difficult by the students, requiring biological content knowledge and new skills. Also, many students commented on problems with the videos, sometimes losing time due to videos not playing or freezing.

Furthermore, students commented on the difficulty in using the drawing tool. Together with the higher number of students feeling disadvantaged by typing, this leads to a significantly more negative attitude towards on-screen biology assessment for the future, with 29% unable to recommend it for this subject (45% would recommend it).

J'ai perdu beaucoup de temps. je ne me souvien plus combien, mais J'en ai perdu beaucoup (minimum 10 minutes). J'essayais de démarer les vidéos mais il était impossible de les lires. puisqu'il y avait beaucoup de questions qui était avec des vidéo, il y a beaucoup de questions que je n'ai pas pu répondre.

I lost 20 minutes of the test trying to figure out how to draw a diagram. Eventually, I did one whole diagram, but at one point the pen did not work properly. Also, I found it frustrating that every time I drew a line, I had to re-click the pen. Even when I did, the line would have an option that would move it automatically, therefore when I tried to connect lines, it kept resizing the original line, which was also really frustrating.

It is clear enough but I need the rubric so I know what is expected in the grading system.

... there were things like calculations that are simple enough but I couldn't do it because I did not expect that I will have to calculate anything for biology and that there is a calculator in the program.

Il était difficile de savoir ce qu'on attendait de nous pour le schéma au premier numéro. De plus, il était parfois difficile de savoir à quel point on voulait qu'on développe dans les différentes questions.

Some of the questions I am not fully familiar with since we haven't covered the such topics in school before. I just had to use my knowledge that I got from exploring biology.

Some students' comments on the biology task

Teacher comments on the biology task focused on the language used, leading to students taking the prompts too literally and producing responses of a lower quality than they might have done if addressed in more "MYP familiar" language. Providing students with cues towards the assessment objective of a task or question might have mitigated this, as MYP students are used to working with these in assessment situations. In general, the questions were thought to be at the right level for the MYP. The measuring task could have been explained a little better, and the lack of a calculator was confusing, while the item requiring a diagram showing interactions to be drawn was vague. Teachers and students were unsure about what kind of diagram and what type or level of interaction was expected.

For the marking trial, 55 students were marked by two examiners, and 18 students were marked by both examiners and teachers at trial schools.

Examiner agreement

There is a significant difference in standard between the two examiners, which can be seen from the summary statistics and the box plots (Table 6 and Figure 15). The second examiner is more generous than the senior examiner. It is also noteworthy that the overall mean mark awarded by the senior examiner was very low, at just over one-third of the maximum mark.

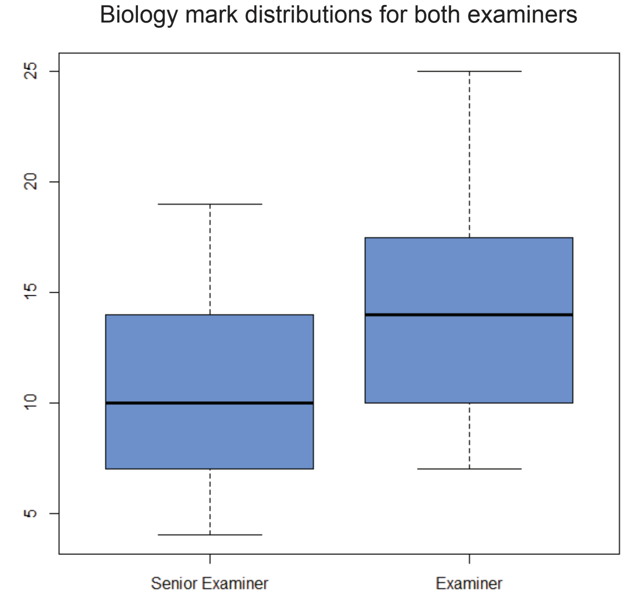


Figure 15: Examiner marks for biology

	Mean	Std. dev.	Min	Max	N
Senior examiner	10.71	4.01	4	19	55
Examiner	14.36	4.75	7	25	55

Table 6: Mean examiner marks for biology

The histograms below show the mark distributions awarded for each question (Figures 16–20). The pattern of the second examiner, being more generous, is spread across all questions, although it is most noticeable for questions 1 (Figure 16), 2 (Figure 17) and 5 (Figure 20).

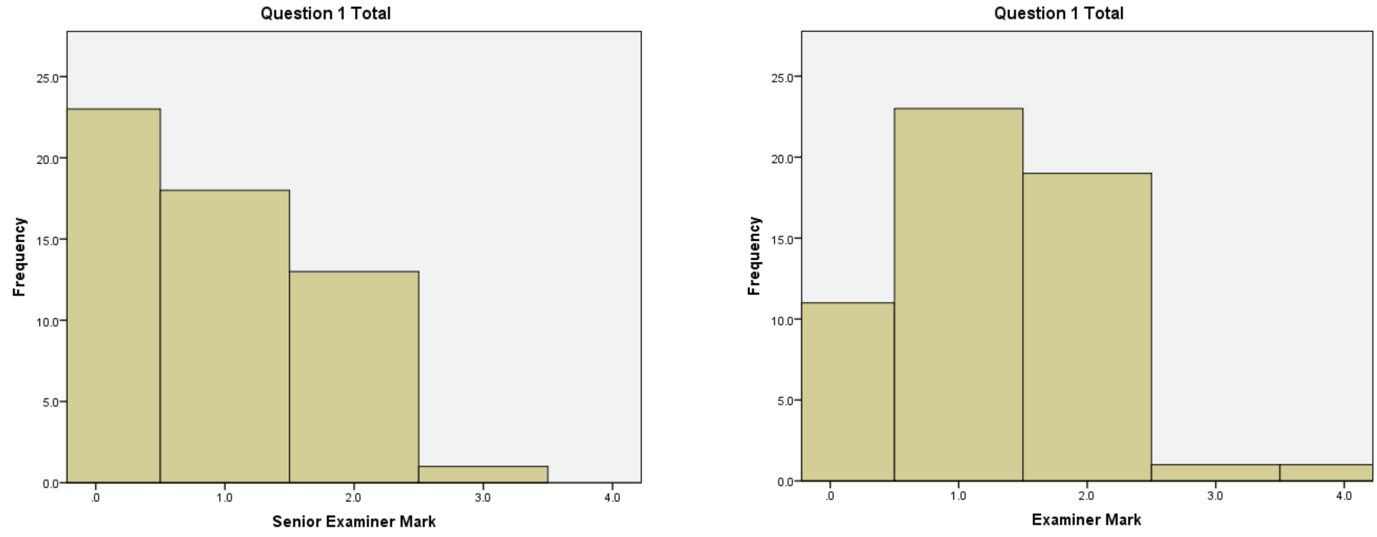


Figure 16: Mark distribution compared for biology question 1

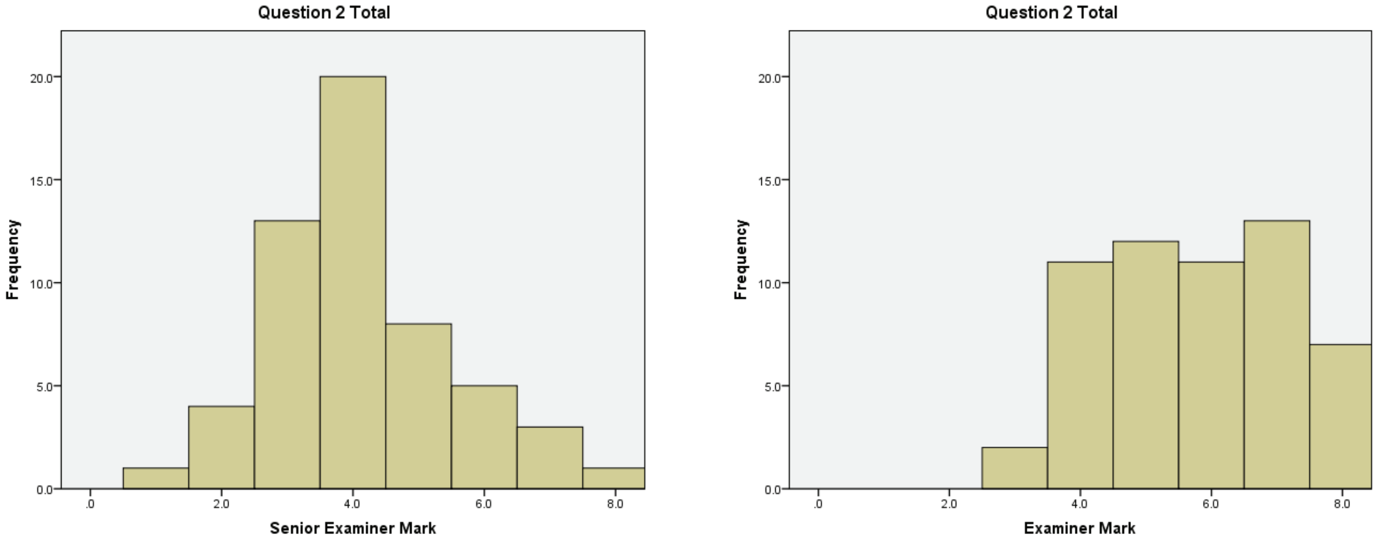


Figure 17: Mark distribution compared for biology question 2

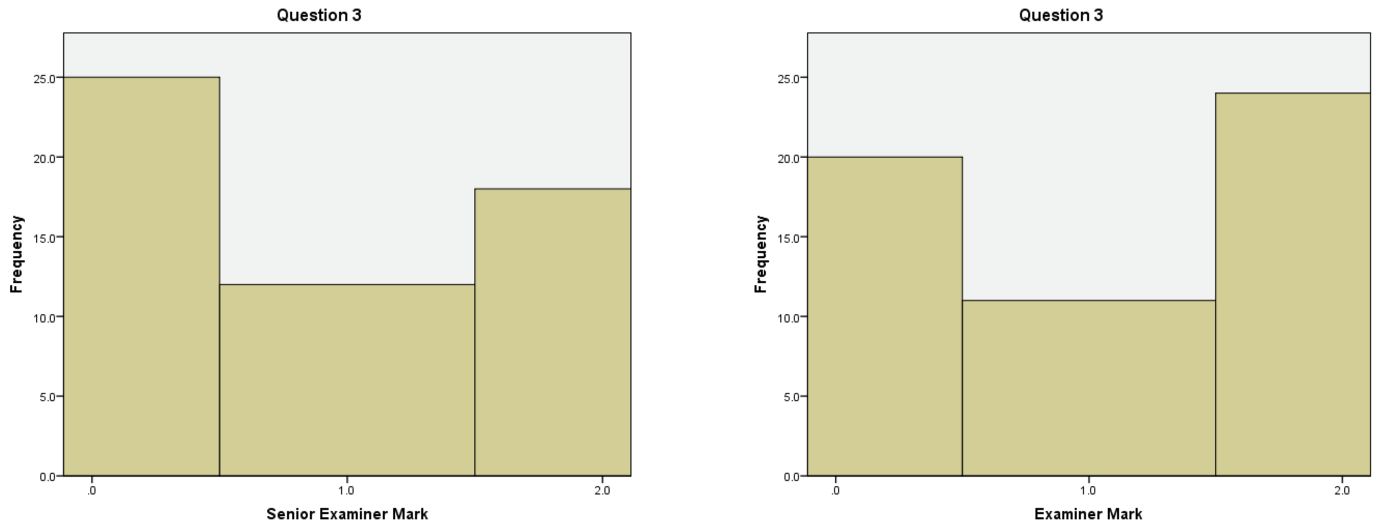


Figure 18: Mark distribution compared for biology question 3

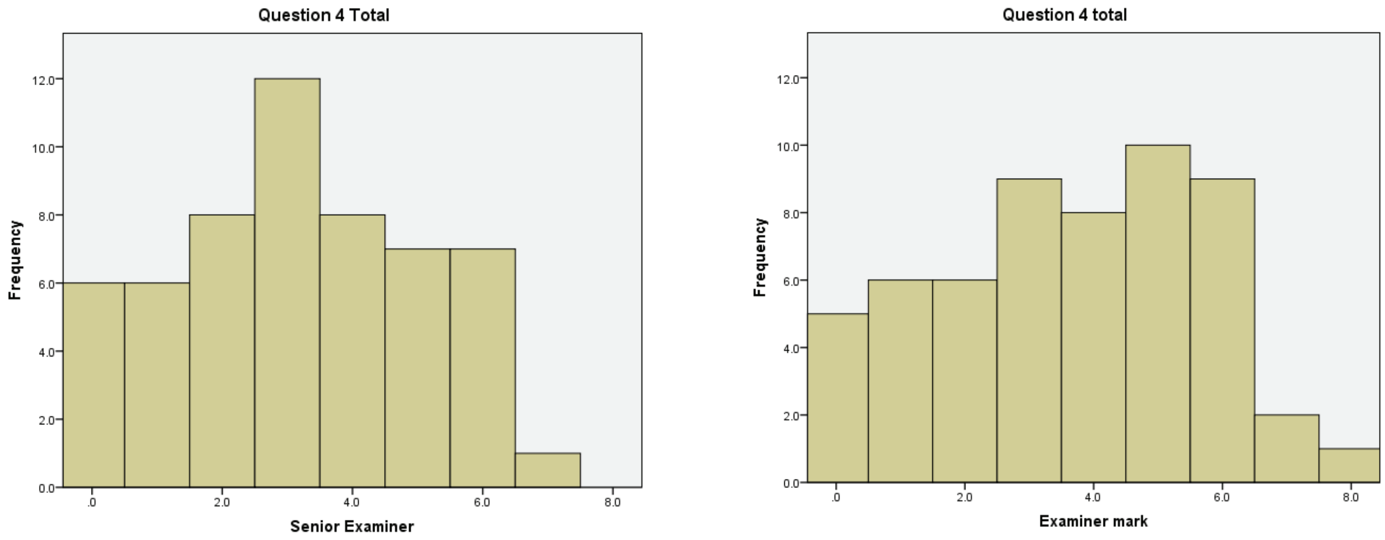


Figure 19: Mark distribution compared for biology question 4

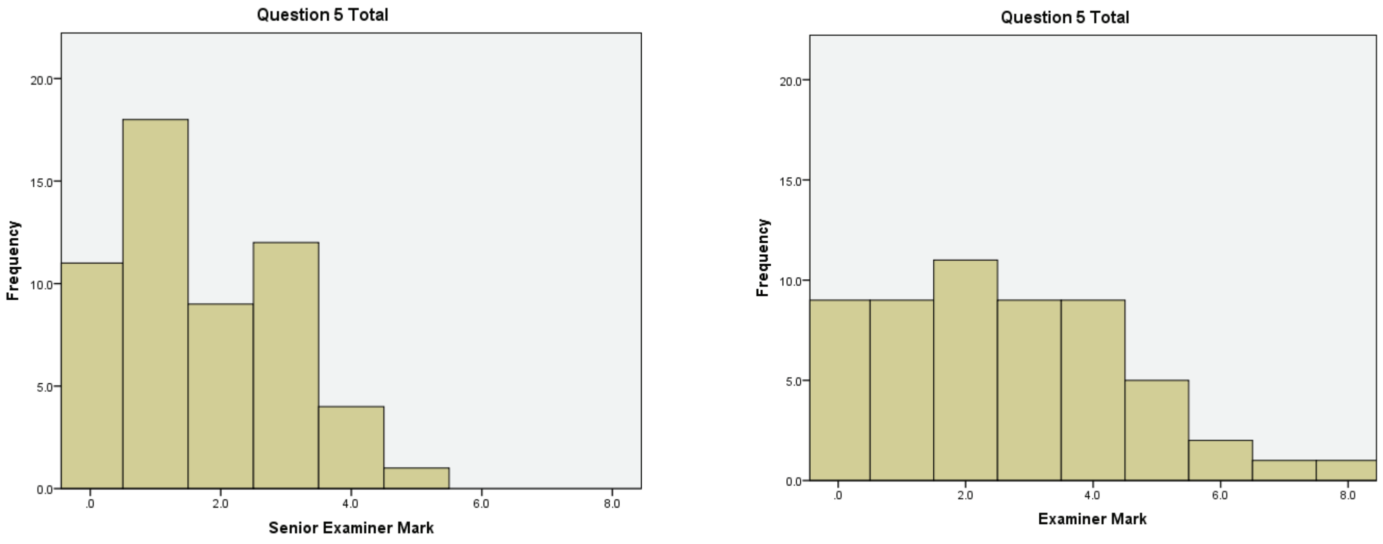


Figure 20: Mark distribution compared for biology question 5

The two examiners only awarded the same total mark for 4 of the 55 students, and they only awarded the same marks on every question for 1 student out of the 55. The correlation of overall marks for the two examiners was reasonable (r = 0.84), suggesting a strong pattern between the two examiners' marks in terms of the order in which they ranked the students.

Comparison with teacher marks

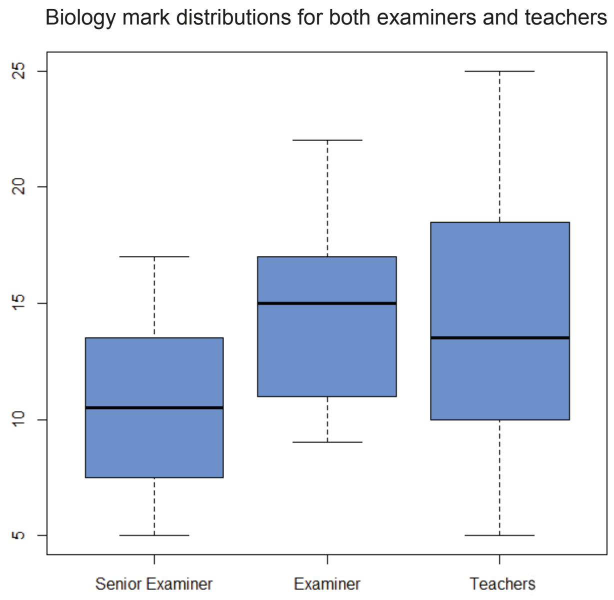


Figure 21: Examiner and teacher marks for biology

	Mean	Std. dev.	Min	Max	N
Senior examiner	10.67	3.48	5	17	18
Examiner	14.56	4.05	9	22	18
Teachers	14.28	5.26	5	25	18

Table 7: Mean examiner and teacher marks for biology

A sample size of 18 is too small to make meaningful comparisons. However, from the summary statistics (Table 7) and where the quartiles lie (in Figure 21), you can see that the teacher distribution of marks is more similar to the second examiner than the senior examiner, although much more spread out.

Question-level performance

Table 8 shows the mean mark that the senior examiner awarded each question, along with what that mean corresponds to as a proportion of the maximum mark available.

	Senior examiner	Maximum mark
Question 1	0.86 (0.22)	4
Question 2	4.13 (0.52)	8
Question 3	0.87 (0.44)	2
Question 4	3.16 (0.40)	8
Question 5	1.69 (0.21)	8
Total	10.71 (0.36)	30

Table 8: Mean senior examiner marks per question for biology

The lowest mean marks were awarded for questions 1 and 5 and the highest for questions 2 and 4 (the histograms in the "Examiner agreement" section show than neither examiner awarded a mark of 0 for question 2).

How do the outcomes match teacher and student comments on the task?

The teachers felt that the task was about right for MYP students, although there were some concerns about whether students would understand what was required of them and the measuring tool was seen to be confusing. The students also felt that this task was slightly harder than they were used to, which is supported by a relatively low overall mean of 10.71 out of 30. However, the senior examiner clearly marked more harshly than the second examiner or the teachers, where the mean mark was closer to 50% of the total mark. Questions 1 and 5 were the most difficult; students performed relatively well in question 4, which required them to use the measuring tool.

The mathematics task

The mathematics task contained four questions, each comprising smaller sub-questions. The total mark was 33. A total of 647 students took part in the mathematics task and, as can be seen in Figure 9, these students had a very different experience than all the other students participating in the trial. The mathematics task included an innovative HINT button, which allowed students who could not solve a quadratic equation to receive help in the solution, so that they could use the equation to answer subsequent questions. The students had to accept a mark penalty in exchange for the correct solution to the quadratic function if they used the HINT button.

In all, 65% of the students completing the mathematics task used the HINT button, 7% did not notice the button, 10% did not need to use it, and about 15% did not use it because they did not want to lose marks. Figure 22 indicates that the majority of students found the HINT button helpful. The high percentage of students requiring help in solving the equation indicates that the question was too hard for this level. Specifically, the entering and use of the equation was problematic for 40% of the students.

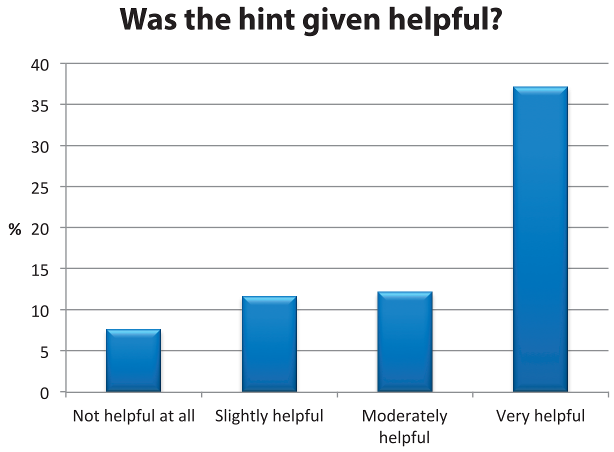


Figure 22: How helpful was the hint given—students completing the mathematics task

The unfamiliar language and the necessity to type led to strong negative reactions, leading to about one-third of the students rejecting on-screen mathematics assessment for the future, with 40% unable to recommend it for this subject (41% would recommend it).

The use of specific mathematical symbols for the equation was mentioned the most in the teacher comments on the mathematics task. Issues raised were centred around the inability to use the correct symbols (for example, not on the keyboard, problems with the equation tool) as well as students not being familiar with the content required. In particular, the use of quadratic functions is dealt with at different times in different schools, and this disadvantaged some students in this trial. The last question was perhaps too difficult for many students. Teachers echoed the biology feedback and asked for the inclusion of assessment objectives in the task as a way to cue the students better to the expected level of response.

Students mentioned many of the issues raised by the teachers in their own comments, such as the unfamiliar content (quadratic functions, modal class, and so on). Many comments focused on the problems in trying to type in equations using either their own keyboard or the function tool provided. Question 4 was found to be too difficult by many students, some specifically referring to the last sub-question.

> In the very beginning, I like how the test would ask me to count the numbers of cars passing through the junction, I feel that it is very realistic or sort of involved in a real life example where as we usually learnt about fixed theories which most of the time is not applicable to this fast paced world.
>
> For math only, I would rather hand writing because I can be more flexible with the formulas and I get to change the sentences.

Pas le meilleur moyen pour mathématique

... Bref, ce n'était vraiment pas facile de saisir les équations. Il aurait aussi été préférable d'ajouter une petite calculatrice ou des symboles comme la multiplication, la division, la soustraction et l'addition.

... we could not type x square by using our keyboard.

Il est très difficile décrire à l'ordinateur des règles avec des variables et de l'algèbre. Impossible d'illustrer certaines opérations.

I didn't know how to make a quadratic function so I couldn't answer most of question 4.

I had no way to work out the answer using the equations because I did not have a calculator.

It's kind of hard to right down the equation using computer and takes a lot of time. I prefer writing it down, rather than typing it in computer.

À la dernière partie, première question. Je ne comprenais pas le vocabulaire.

Some students' comments on the mathematics task

For the marking trial, 60 students were marked by the three examiners and 15 students were marked by the three examiners and teachers at trial schools.

Examiner agreement

The box plots and descriptive statistics (Figure 23 and Table 9) show that two examiners were fairly similar and the third was slightly harsher.

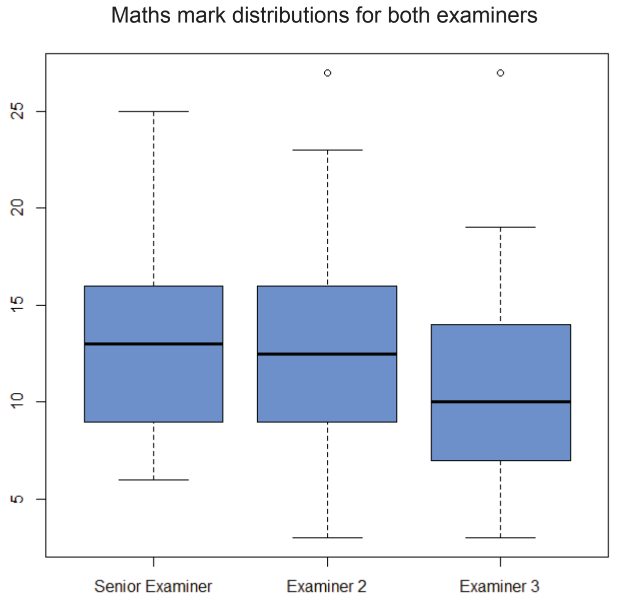


Figure 23: Examiner marks for mathematics

	Mean	Std. dev.	Min	Max	N
Senior examiner	12.85	4.14	6	25	60
Examiner 2	12.73	4.81	3	27	60
Examiner 3	10.83	4.57	3	27	60

Table 9: Mean examiner marks for mathematics

Overall, 65% of the marks were within 2 marks of the senior examiner. The correlations between the examiners' total marks were high. Between the senior examiner and the second examiner the correlation was 0.90 and between the senior examiner and the third examiner the correlation was 0.88, suggesting that the differences between the senior examiner and the other examiners followed a strong linear pattern and that the rank orders were similar.

Comparison with teacher marks

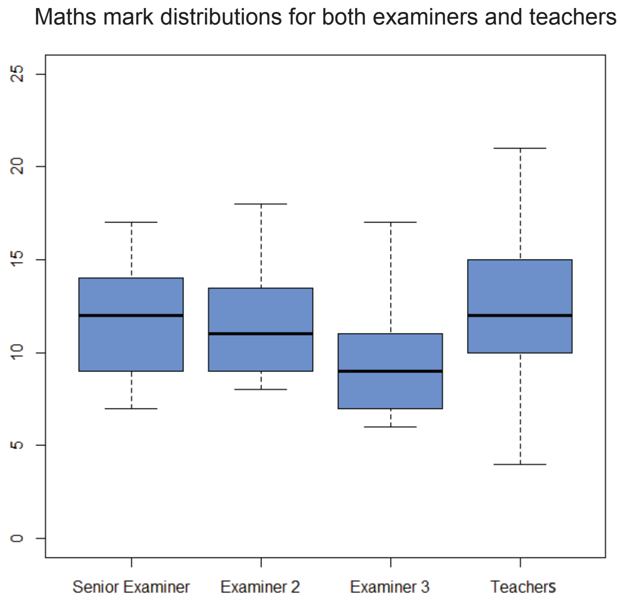


Figure 24: Examiner and teacher marks for mathematics

	Mean	Std. dev.	Min	Max	N
Senior examiner	11.93	3.10	7	17	15
Examiner 2	11.53	3.07	8	18	15
Examiner 3	9.20	3.28	6	17	15
Teachers	12.20	4.57	4	21	15

Table 10: Mean examiner and teacher marks for mathematics

A sample size of 15 is too small to make meaningful comparisons. For these 15 students, the teachers marked slightly more generously and gave a wider range of marks than all of the examiners (Table 10).

Question-level performance

The mean marks are somewhat low at 39% of the maximum, and no examiner awarded a mark above 27 out of 33.

For ease of interpretation, the questions as a whole are looked at first and, because question 4 was worth a high proportion of the total marks and because one part included the use of the HINT button, the sub-questions for question 4 will also be reviewed.

Whole questions

Table 11 shows the mean mark the senior examiner awarded each question (as a whole), along with what that mean corresponds to as a proportion of the maximum mark available.

	Senior examiner	Maximum mark
Question 1	1.97 (0.66)	3
Question 2	2.17 (0.72)	3
Question 3	5.17 (0.86)	6
Question 4	3.55 (0.17)	21
Total	12.85 (0.39)	33

Table 11: Mean examiner marks for mathematics

From the mean marks in the table, it appears that question 3 was the most accessible question for the students. Students found question 4 difficult to answer, with a mean mark of only 17% of the maximum mark for the question. As this question was worth over 50% of the total marks for the assessment, this had a major effect on the total mark for the task and explains the low correlations between questions. The IB's routine scrutiny procedures for live examinations will ensure that such a situation would not arise in the future.

Question 4

Table 12 shows the mean marks that the senior examiner awarded each part of question 4 and that mean's proportion of the maximum mark available.

	Senior examiner	Maximum mark
Question 4a	0.12 (0.12)	1
Question 4b	0.18 (0.04)	5 (+hint)
Question 4c	1.05 (0.35)	3

Question 4d	0.93 (0.31)	3
Question 4e	0.32 (0.32)	1
Question 4f	0.95 (0.12)	8
Question 4 Total	3.55 8 (0.17)	21

Table 12: Mean senior examiner marks per sub question for question 4 of the mathematics task

Students obviously found this question difficult, with an overall mean score of about 3.5 and only one student achieving more than 10 out of 21. Sub-question parts A, B (which featured the HINT button) and F were particularly difficult, with mean scores between 0.04 and 0.12 of the maximum mark for the questions. The histograms in Figure 26 below show the distribution of marks awarded by the senior examiner for each sub-question. It is noteworthy that a high percentage of students scored 0 for both parts A and B and either 0 or 1 on part F.

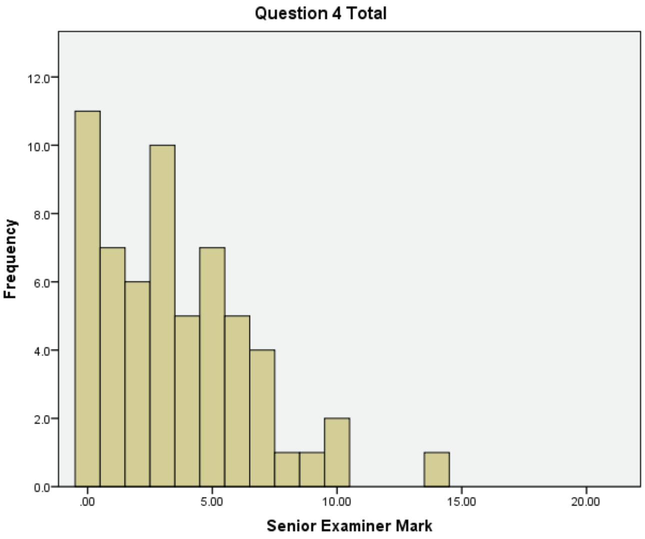
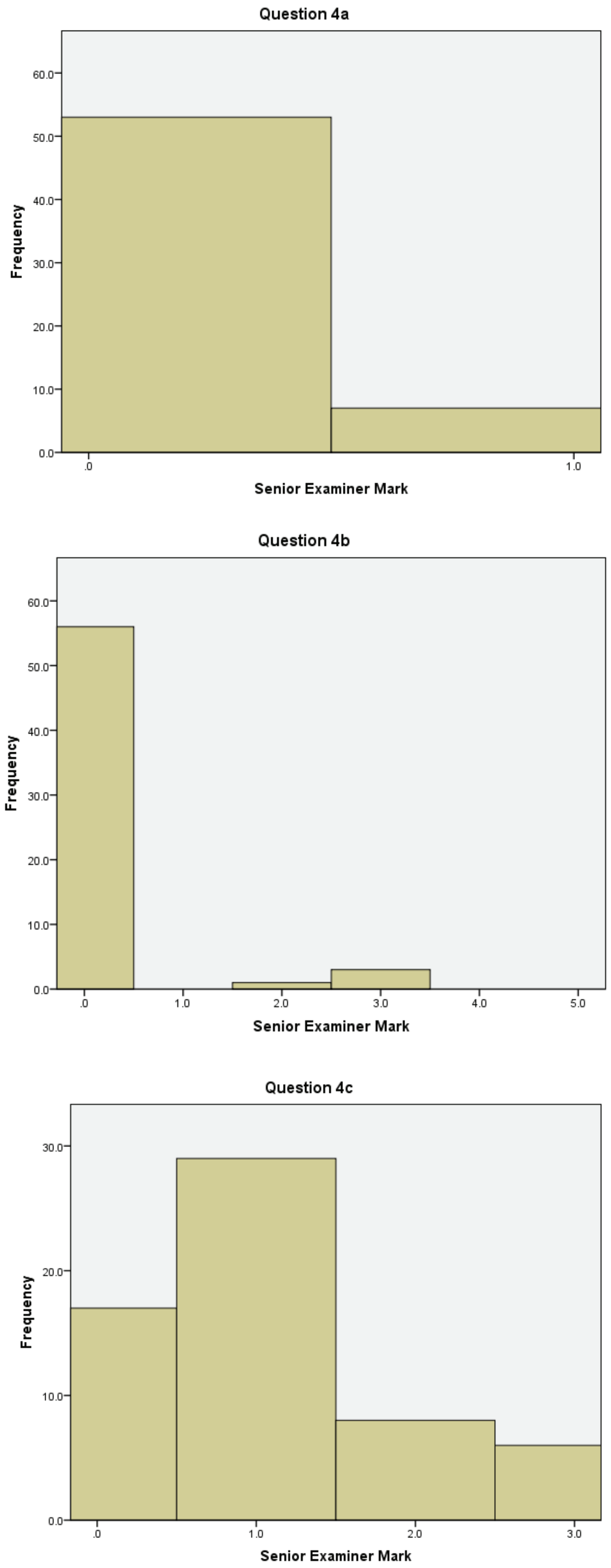


Figure 25: Mark distribution for mathematics question 4



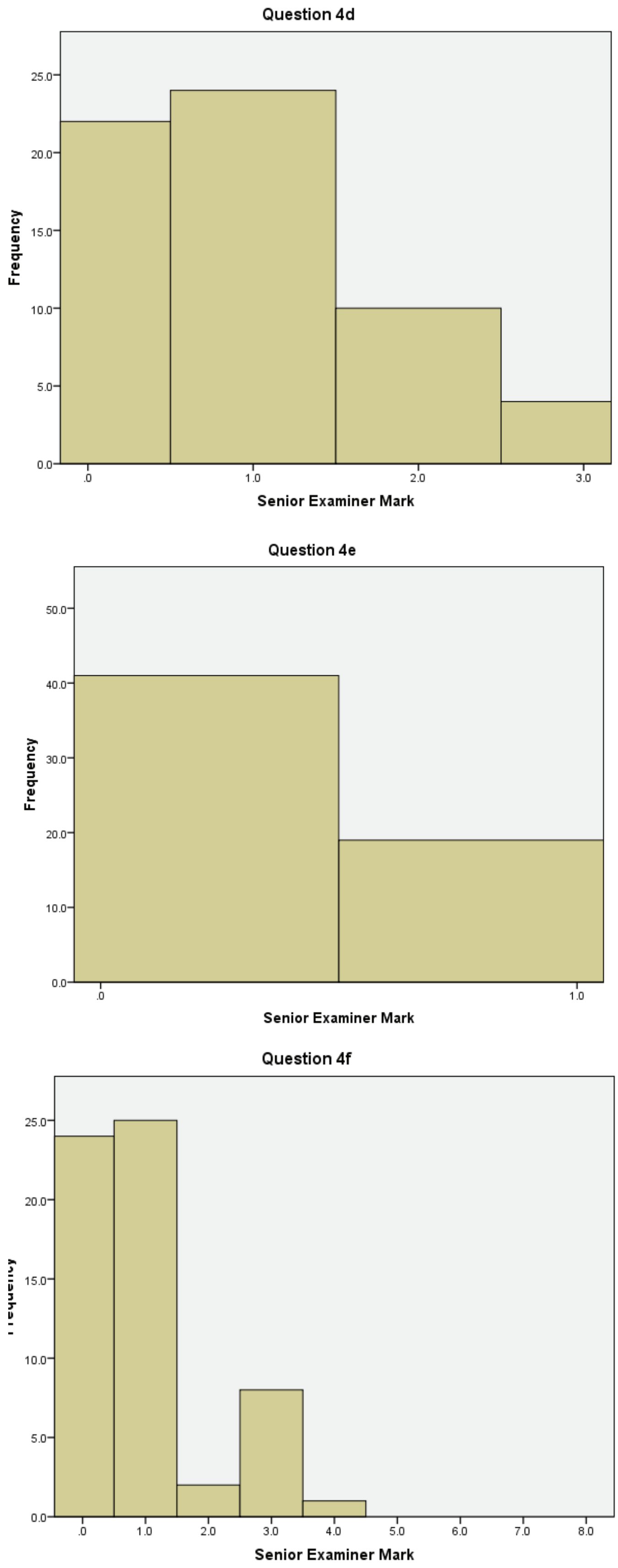


Figure 26: Mark distribution for the sub-questions of mathematics question 4

How do the outcomes match teacher and student comments on the task?

Teachers expressed concern about the demands of question 4, explaining that students struggled with using formulas and quadratic functions. The very low mean mark for this question entirely endorses this concern. Almost one-third of the students surveyed felt that the mathematics questions were harder than they expected. The overall mean was 12.8 out of 33, but question 4 mainly accounts for this, considering that the means in all other questions were well above half marks.

The interdisciplinary task

The interdisciplinary task had five questions and contained 20 separate source material items, including a graph and a diagram. The maximum mark for the task was 20. A total of 537 students completed a survey giving feedback on this task. The response type was text based, so only a minority felt that typing was a problem. Furthermore, the task was not felt to be too difficult but matching the level that MYP students could expect. Only a small minority felt new skills were assessed or that there was unfamiliar language used, making them unsure about expectations. However, the occurrence of technological problems during the task was relatively high, which may have contributed to a slightly less positive attitude towards future on-screen assessment, although it is similar to the attitude of the students completing the history task, with 54% in favour of on-screen assessment for the future and 20% unable to recommend it.

There were only a few comments from teachers on this task. Issues mentioned were that words used were unclear and could be interpreted in different ways and that the allocated time was too short for some students in the French trial. The number and size of sources may have contributed to various computer problems and loss of time. This observation was confirmed in the students' comments, which listed a variety of computer problems such as freezing and loss of responses. Students commented on the repetitive nature of the questions, as they seem to have had trouble understanding the prompts. The latter is evident in both the trials, so it may not be due to non-native speakers misunderstanding the language.

J'ai réussi à regarder le vidéo, mais je n'ai pas réussi à le finir. Le vidéo a pris au moins 5 minutes à télécharger et j'ai du quitter car l'écran a gelé.

The program was frozen and I had to change my computer in order to continue the task.

After i was finished, the timer counted down to zero but then it froze. I waited but it didn't unfreeze itself so i had to restart my computer (loosing a bunch of other things I had open) and it didn't save the task.

Au début le logiciel a geler et nous n'avons pas pu ouvrir le logiciel de nouveau. Nous avons du recommencer l'ordinateur pour que cela fonctionne. Ensuite j'ai eu un problème puisque mes réponses du numéro 1 ont été effacer lorsque je suis passée au numéro 2.

I sometimes was not very sure of how much I was supposed to write. For example I didn't know if I needed to answer in one sentence or a paragraph.

Les questions semblaient un peu trop répetitives.

The way of structuring questions was somehow confusing. I recalled that I answer a question with a statement, then provided with supporting materials from the text after it because it is a two point answer. When I continue on the task, the next question asked me for supporting evidence for my answer in the previous question.

Some students' comments on the interdisciplinary task

For the marking trial, the work of 59 students was marked by two examiners, and 26 were marked by both examiners and teachers at trial schools.

Examiner agreement

The box plots and descriptive statistics for each examiner show that, overall, the distributions of the marks awarded by the two examiners are very similar (Figure 27 and Table 13).

Looking at the distributions of marks the examiners awarded shows a very different story, particularly with respect to question 1 (Figure 28). The distribution of the senior examiner's marks shows a large number of students being awarded 4 marks, with very few below and not many awarded 7 or 8 marks, whereas the other examiner awarded the maximum to more than half the students for this question. The second examiner is then harsher on the remaining questions (see "Question-level performance"), which is what causes the distribution of the overall scores to be similar.

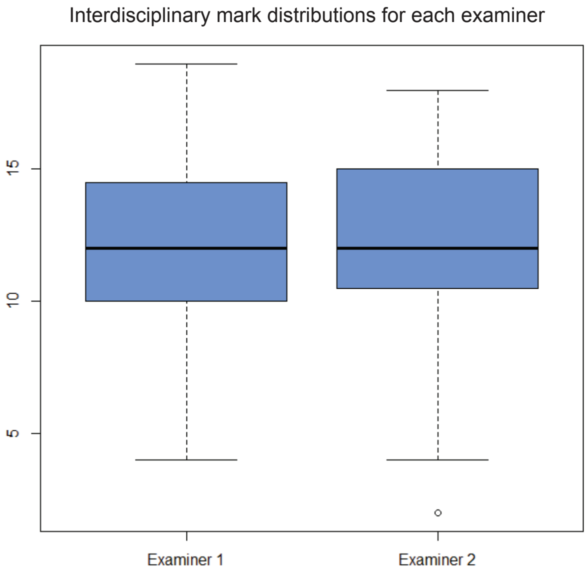


Figure 27: Examiner marks for interdisciplinary

	Mean	Std. dev.	Min	Max	N
Senior examiner	11.95	3.42	4	19	59
Examiner	12.37	3.87	2	18	59

Table 13: Mean examiner marks for interdisciplinary

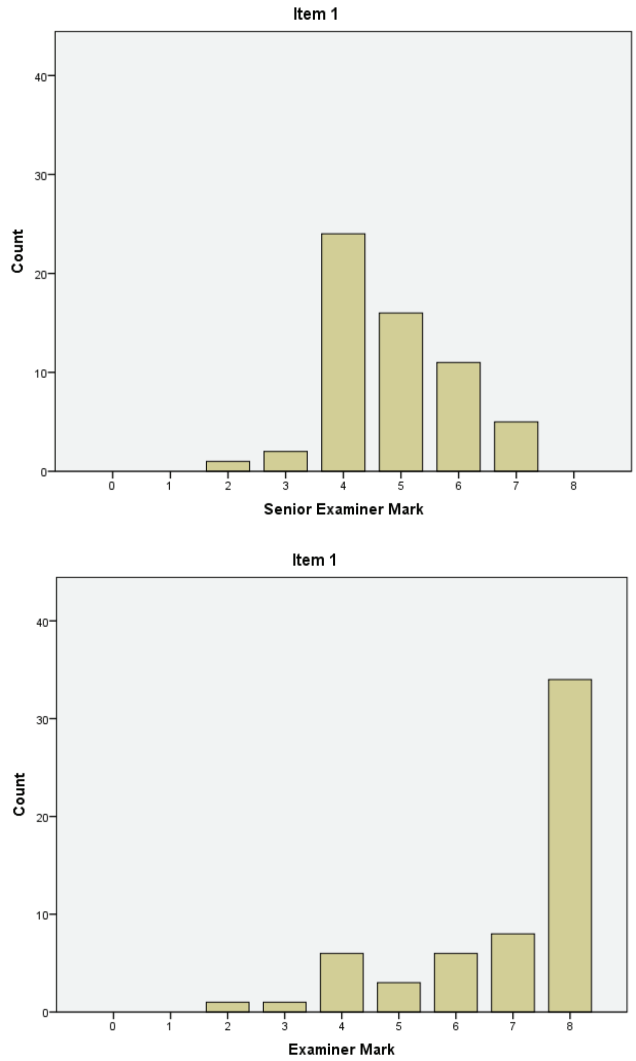


Figure 28: Examiner mark distribution for interdisciplinary item 1

The two examiners only awarded the same total mark on 10 out of the 59 students and they only awarded the same marks on every question for 1 student out of the 59. The correlation of overall marks was also relatively low (r = 0.72).

Comparison with teacher marks

Because there were only 26 students marked by both examiners and teachers, it is difficult to draw any conclusions, but it seems that the teachers awarded lower marks than both examiners and with a slightly larger spread.

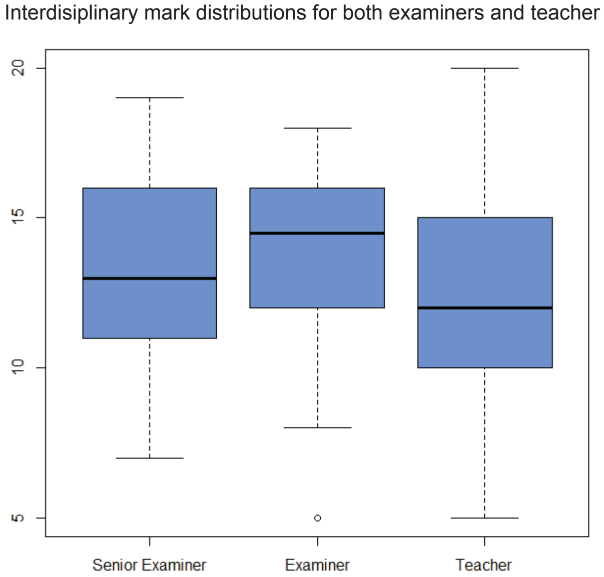


Figure 29: Examiner and teacher marks for interdisciplinary

	Mean	Std. dev.	Min	Max	N
Senior examiner	13.42	2.99	7	19	26
Examiner 2	13.73	2.49	5	18	26
Teacher	12.69	3.10	5	27	26

Table 14: Mean examiner and teacher marks for interdisciplinary

Question-level performance

Table 15 shows the mean mark awarded by the examiners for each question, along with what that mean corresponds to as a proportion of the maximum mark available.

	Senior examiner	Maximum mark	Combined	Total
Question 1	4.82 (0.60)	6.92 (0.87)	5.86 (0.73)	8
Question 2	1.25 (0.63)	0.88 (0.44)	1.07 (0.54)	2
Question 3a	2.63 (0.66)	2.58 (0.65)	2.61 (0.65)	4
Question 3b	1.28. (064)	0.80 (0.40)	1.04 (0.52)	2
Question 4	2.03 (0.51)	1.20 (0.30)	1.62 (0.41)	4
Total	12.02 (0.60)	12.37 (0.62)	12.09 (0.60)	20

Table 15: Mean examiner marks per question for interdisciplinary

Given the clear difference between the marks awarded by each examiner on each item (particularly in question 1) and not having much information about the cohort that took the test, it is difficult to draw any conclusions about the items relative to one another, but it appears that question 4 was found to be slightly harder than the other questions.

Coordinator feedback on the logistics of on-screen assessment

A total of 71 coordinators from 23 countries completed the feedback survey for the October and December trials. Their responses follow. In all of the questions, the majority of coordinators agreed with the positive statements about the administration of the eAssessments—and 70% of coordinators would recommend the use of on-screen examinations to other coordinators. However, there was a significant minority of coordinators who found the administration difficult and would not wish to repeat the experience. The detailed feedback given by the coordinators will be very helpful in improving the guidance provided by the IB and was very gratefully received.

Setting up the technology

The statement "Setting up technology required for the on-screen tasks was straightforward" elicited the most negative response from coordinators, although 53% agreed to some extent with it. There was a significant number of coordinators who found the set-up challenging, and this statement elicited the highest number of "strongly disagree" responses.

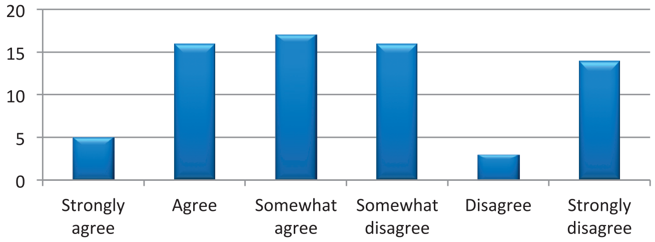


Figure 30: Coordinator agreement with "Setting up the technology required for the on-screen tasks was straightforward"

Administration of the on-screen tasks

A total of 65% of coordinators agreed with the statement "The whole administration of the on-screen tasks from the perspective of a coordinator was a generally positive experience", although the most popular response was "somewhat agree".

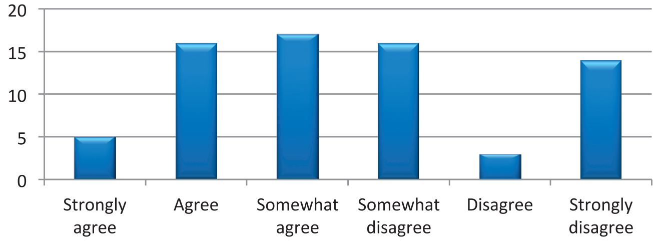


Figure 31: Coordinator agreement with "The whole administration of the on-screen tasks from the perspective of a coordinator was a generally positive experience"

School readiness to administer on-screen examinations

A total of 64% of coordinators agreed with the statement "My school is ready to administer on-screen examinations in formal timed conditions", although, the most popular response was "agree".

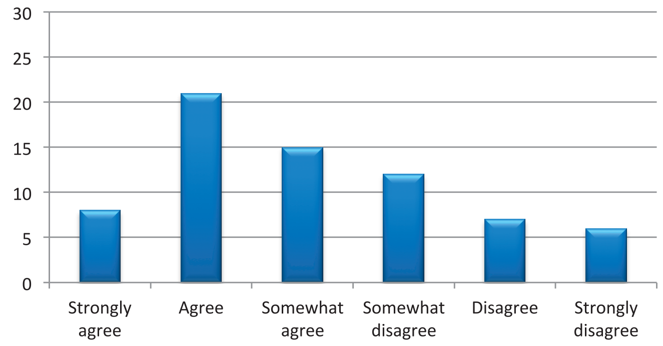


Figure 32: Coordinator agreement with "My school is ready to administer on-screen examinations in formal timed conditions"

Recommending on-screen assessment

It was pleasing to see that 70% of coordinators agreed with the statement "From the perspective of a coordinator who must organize and conduct examinations, I would recommend on-screen examinations to other coordinators". It was also this statement that elicited the highest number of "strongly agree" responses, although the most popular response was again "somewhat agree".

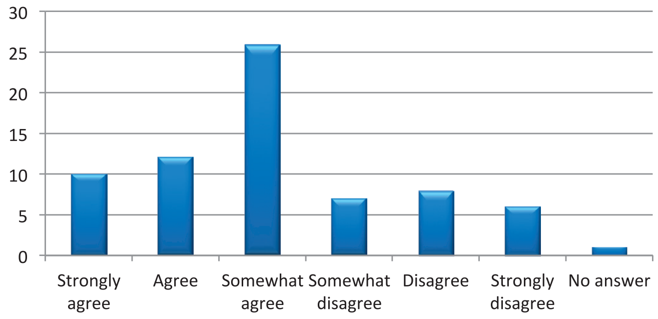


Figure 33: Coordinator agreement with "From the perspective of a coordinator who must organize and conduct examinations, I would recommend on-screen examinations to other coordinators"

The following quotations are taken from coordinators' written comments on the trial eAssessments.

> These are exciting exams, but in some ways they are quite different to what students are used to. It is vital that there are practice exams in the run-up to the real thing.
>
> I'm happy we did it once, though it is quite a commitment, for both teachers and students.

> The on-screen aspect is great, when the kinks get worked out. It's clearly the way to go.
>
> Je suis heureuse d'avoir pu me familiariser avec ce mode d'évaluation. Puisqu'il semble qu'il sera imposé, je souhaite vivre chacune des étapes à mesure que celles-ci évolueront.
>
> Participating in the trial was extremely valuable and we very much appreciated the chance to try them out. They actually worked very well and it went quite smoothly.
>
> The students found the test more enjoyable than the routine kind of assessments that they are used to.
>
> We do not have enough computers or resources to administer eAssessments in multiple subjects to our entire MYP cohort. This is unlikely to change any time soon.
>
> Nous disposions de 32 postes. L'installation fut faite par le technicien de l'école. Il est très important que le technicien soit disponible tout au long de l'évaluation.

Some coordinators' comments

Conclusions

The majority of the students (58%) would like to do more of their assessments on-screen/in this way, with 14% being undecided. Almost one-third of the students (28%) preferred a more conventional assessment system. The difference in preference for typing exams between students with or without experience of special-access arrangements did not translate into a difference in preference for conventional or eAssessment. However, there was a statistically significant relationship ($p < 0.01$) between typing fluency and their recommendations. Of the students who would like to do on-screen assessment more often, the overwhelming majority (89%) felt that typing was an advantage or at least did not slow them down. Computer literacy and familiarity is most probably an important indicator in favour of a preference for on-screen assessment, and there is a positive correlation between students who used computers a lot in school and those who favoured on-screen examinations. A selection of student comments in favour of on-screen assessment is given below.

J'ai bien apprécié l'expérience et je crois que c'est un meilleur moyen de faire un examen car nous, les jeunes d'aujourd'hui, sommes beaucoup connecté aux ordinateurs. Je me suis senti mieux qu'assise devant une tâche papier, je me sentais plus dans mon "environnement". Je crois que cela serait aussi le cas pour plusieurs autre

I didn't expect so much use of multimedia like videos. Although I did enjoy it, I didn't expect it.

Typing is ok but it felt weird because I'm used to writing my answers.

Ce fut amusant de participer a cette mise à l'essai

...I felt it was quite good and effective way of testing our critical thinking skills. I also believe that this exam which is on screen is more effective than writing with pen or pencil because it requires long explanations.

If only final exams could be like this :'(

It is advantageous to type for written answers however with mathematical equations and calculating it was more difficult to have to type my answers.

I think the onscreen examinations are better suited to humanities and English subjects which require more writing ..., not for maths where it is slightly difficult to input all the operations.

Some students' positive comments on on-screen assessment

Not unsurprisingly, the results show that students were significantly less positive about more on-screen assessment when they:

- experienced problems when logging in
- experienced problems during the task
- thought typing was a disadvantage
- felt there were parts of the task where they were insecure about what was expected
- encountered language or concepts that were unfamiliar
- felt the level assessed was more difficult than expected
- felt unfamiliar skills were being assessed.

Only the first three of these reasons for tending to reject the eAssessment tasks are directly related to on-screen assessment. The remaining four would reduce enthusiasm for a more conventional assessment as well. In Figure 34, the students' and teachers' recommendations for on-screen assessment are given per subject.

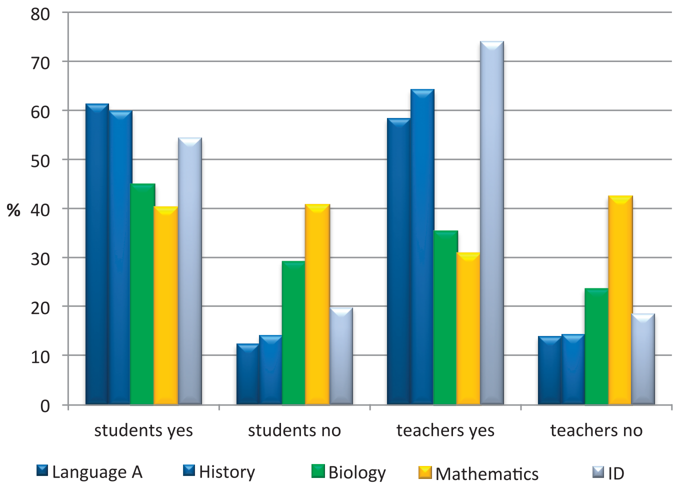


Figure 34: Students' and teachers' recommendation for on-screen assessment per subject

The examiners and teachers were provided with a markscheme, but given the nature of the trial and the time constraints, they were not trained on the application of the markscheme as is normally the case, nor was their marking monitored through "seeding" as it will be in live assessments. This difference will have contributed significantly to the disagreements between the examiners. **However, and reassuringly, the rank order of students and questions proved to be quite consistent in all of the tasks between both the examiners and the teachers.**

The trial tasks in English and French had three main purposes. These were:

1. to confirm that concept-based assessments using global contexts meet the needs of MYP students and are valid, robust and reliable
2. to demonstrate that on-screen assessments are possible and practical in IB World Schools
3. to learn where improvements to the assessment materials and the technology need to be made so that the live assessments are as good as they can possibly be when they are introduced in 2015 (live pilot) and 2016 (first sessions leading to MYP certification).

In these respects, the trials were a success. There were a number of issues raised where the IB assessment team needs to improve aspects of the assessments, and "lessons learned" data have been developed and are being analysed. The trial material was not quality assured in the same way as live materials in the IB. The assessments are so innovative that everyone concerned with their production—the task authors, the examiners and the staff—were required to develop completely new procedures, and we must expect that mistakes are made in these first attempts.

In addition to the many technical lessons we learned from the trial, we also drew the following important assessment design conclusions.

1. Typing text on screen posed few problems for the majority of students.
2. A means must be found of making it possible for students to use numbers and symbols in mathematics and sciences more intuitively.
3. Any manipulative/technical skills required by the tests must be familiar to the students, who should be given opportunities to practise before their first live test.
4. Expectations about word count, mark allocation, and so on, should be made clear to students so they can plan their responses appropriately.
5. If there are any separate requirements, such as calculators, they should be communicated in advance of the tests.
6. Organizational and logistical requirements for on-screen assessment need to be communicated clearly to coordinators (and school IT support).

Considering MYP-specific requirements

7. The content and language used in the on-screen assessments needs to be familiar to MYP students.
8. The language used needs to be as simple as possible to reflect the age of MYP students and to accommodate students using their second language.
9. No "recall of content" questions should be asked unless the information required is included in the guides.
10. Reference should be made to the assessment objectives in each task.
11. Questions should be scaffolded where appropriate.
12. Where hints are provided, it must be made clear to students that marks will be sacrificed.

Assessment and marking

13. Examiners must be properly trained and standardized before each session.

14. Markschemes must be aligned with the questions in terms of reference to assessment objectives.

This report on the trials should not conclude without a final expression of gratitude to everyone concerned: to the teachers who put their students forward for the tasks and who completed the survey and especially those who submitted marked responses from their students; to the coordinators who set up the trial tasks and who worked with IT staff in schools to ensure that their students could access the task materials; to the senior examiners and examiners who authored the trial material and marked the student responses; and, most importantly, to the 2,500 students who worked their way through the tasks and provided such invaluable and considerate feedback on their design and content. We are extremely grateful for all your support. It is fitting to end the report with some final comments from our students.

> I definitely think onscreen tests are practical and effective for assessments and would prefer is all my assessments were done this way, it keeps it fair for everyone and the additional features like the status bar, navigation tabs and clock make individual organisation and time management more effective and easier.
>
> I found this task very interesting and interactive. Especially for this exam, all the background sources were very helpful. This exam really makes the student think, it's different from the expressing your knowledge, it's about applying it. Wonderful experience.
>
> I generally didn't like these examinations. Maybe it's because I'm more acquainted with the traditional style of testing, but overall I think the pen paper style is much more appropriate for examinations.
>
> I think the onscreen examination was very good. I've done only written exams and I think it is quite hard as some people write slowly. Doing the exams on a computer also gives more access to different resources which helps to the understanding of the task. The background information is very useful but I think that we didn't have enough time for completing the exam as there were many questions.

Some students' comments

Appendix

Table A.1: Reasons students did not understand expectations

Reason
Didn't know the topic (prompts on OPVL history)/ subject content (for example, DNA/modal class/ model/method (biology)/functions/quadratic/reading graph (maths))
Drawing the diagram (Q1), what type? Drawing the graph
Didn't understand the question, didn't understand the phrasing; e.g. origins instead of sources...
Understanding the measuring ruler, need to calculate was unclear, calculation in biology was unexpected, no calculator or average function
Prompts were unclear, what sort of story was expected; one option or more in one box?
Did not know what was meant by some words, e.g. evaluate, explain (using the model), compare and contrast, discuss limitations, origin of quote/sources
Did not know the length of response that was expected, answer boxes too big/small

Function to difficult without calculator/ needed hint to understand the question

I do not speak English / too difficult language for a not native speaker

No criteria or rubric to help me determine what was expected

Mathematics was a lot of writing, less calculation

Nearly missed questions because not visible, need to scroll down/navigation

Didn't understand the use of working out boxes

I was unsure about be able to go back to answers

Questions would pop up out of order (computer problem?)

Source materials looked irrelevant

Table A.2: Words unfamiliar to students

Bacterium

Biological terms

Courses of principled action

Difficult English words

Functions

Junction

Modal class

Monologue

Narrative, what kind meant

Origin of quote

Pit latrines

Purpose, limitation, value

Quadratic (function/graph)

Rampant

Statistics

Validity

Table A.3: New skills assessed mentioned by students

Analysis and investigation

Bacteria DNA replication

Calculating frequency in a time period

Calculating graphs, long equations etc. without calculator

Compare and contrast

Comparing graphs and reflection

Computer literacy, speed of typing

Descriptive writing

Developing a function from a graph

Diagram with interactions

Dialogue writing

Finding the mean

Graphs, reading and using to

Measuring the diameters (without a ruler)

Observation (of videos)

Relation to real life problems

Measuring tasks on bacteria

Timed creative assessment/creative thinking on the fly (no preparation)

Typing picture analysis

Use of quadratic functions

Using bar graphs to write equations

Using multimedia as sources

Very small unit sizes