BRITISH COUNCIL

Aptis

Forward thinking
English testing

Aptis formal trials feedback report

Aptis technical report (ATR-2)

Barry O'Sullivan, British Council

August 2012

Contents

Executive summary

Four questionnaires were devised to collect feedback on the Aptis 500 trial. The questionnaires were aimed at:

1. Candidates
2. Examiners
3. Administrators
4. Invigilators

The findings from an analysis of the responses to the items contained in these questionnaires are presented in this report.

The candidate questionnaire

	Main findings	Commentary
Item 1	Just under half the respondents had taken an English language test on a computer.	We should monitor this situation carefully in the future. It may be necessary to routinely ask this question as part of the registration process in order to conduct statistical bias analysis of the results. It is important to ensure language ability and computer test familiarity (and/or computer literacy) are not confounded in the test performance.
Item 2	Seventy-five per cent of the respondents felt that the test was attractive and appealing, five per cent disagreed.	Reassuring to a large extent, though it might be useful to understand why the five per cent disagreed. Since it is impossible to please all candidates we would always expect some level of disagreement, and later comments on the font size in the reading test suggest that there may be a problem with the 'look' of the test.
Item 3	Just three per cent felt that the instructions were not clear and easy to follow.	This is a very positive outcome as we were dealing with a population with a broad spread of ability, some of whom may have been expected to struggle with the English instructions.
Item 4	Nine per cent did not feel that the test offered them an opportunity to show their true level of English.	Given this was a trial, and there were a number of technical issues with the test delivery, this is not a bad result. However, it may be useful for us to explore the area more in the future, possibly encouraging research on the topic in the Aptis Research Awards.
Item 5	Ten per cent had difficulty using the computer during the test.	This appears to have been due to a range of factors, some related to the test and the lack of experience of the candidate with the technology, others with local technical issues (poor equipment, outages etc). It is something, however, we need to explore further (see the comment for Item 1).
Item 6	Fifty per cent would recommend the test, 30 per cent would not.	This item would have benefited from an additional open ended response element asking for a reason for the decision (this should be included in a future iteration of the questionnaire). It may be that we were somewhat naïve in asking candidates to recommend a test (any test) and that this question would be better asked of policy and decision makers.
Item 7	General satisfaction with the test and the platform.	A very positive set of responses all round. Respondents seemed to feel the test offered a good estimate of their ability and that it was easy to take, with clear instructions. In terms of individual papers, most comments focused on the vocabulary paper.

Item 8	Twenty per cent expressed satisfaction even when asked to point out weaknesses with the test.	This was again good news for the test. However, some issues with technology (including the problem of having multiple keycodes) and a wariness of taking the test with limited computer literacy were highlighted. Among the other criticisms were time (some felt it was too long, others too short) and perceived level (some felt it was too high, others that it was too low). Specific paper related problems were also highlighted. These included problems reading the input texts in the reading paper as the font was too small (and could not be manipulated by the candidates), a lack of time and impeding background noise in the speaking paper, and the lack of an automatic word counter in the writing.
Item 9	Fifty per cent positive, 50 per cent negative.	Many respondents were very happy with the test. Some were unhappy and others simply wished to offer advice. The main issues that arose related again to level, some people feeling it was too high, others felt it was too low, and there were also comments on the technical problems and on the issue of multiple keycodes (an area that has been addressed).

The examiner questionnaire

	Main findings	Commentary
Item 1	All examiners felt their training to prepare them was sufficient for the examining role.	This is a very positive result as it is important administrators feel confident in their training and in their ability to perform the role competently.
Item 2	The provision of more practice, discussion and feedback would improve training for the exam.	The Aptis team had already planned to develop online training and refresher materials for examiners. It would appear from this feedback that we should go ahead with this plan at the earliest possible opportunity. It may also be useful to include a discussion forum in any examiner-focused website.
Item 3	All examiners felt SecureMarker was easy to use.	Again this is a very important finding as it is necessary that examiners feel secure and confident in their ability to interact with the system.
Item 4	More than 60 per cent of examiners felt there was a time when they were unable to mark a script or interview.	This is potentially quite problematic as it is important that examiners are at all times able to award marks. Analysis of the responses to Item 5 may offer more information about this.
Item 5	The main reason why people had problems marking appear to be related to technical problems with the system.	The most common problem recorded by the examiners was related to the sound quality of speaking test. Another problem was related to the fact the screen would not always fit onto the computer screen that the individual was using to mark. There were other issues with the marking scheme not downloading and problems with loading time.
Item 6	A number of technical issues, generally relatively minor, were pointed out by examiners.	The issues that were highlighted centred around things like sound quality and the physical presentation of the work on screen. All of these issues will be will be fixed by the Aptis team.
Item 7	Ninety per cent of the examiners felt that it was easy to mark the items that were given, within 48 hours.	Those who indicated there had been problems also indicated the problems were associated with technical issues. This was because at times it was impossible to mark because the system was inaccessible or because the system had prevented the individual from marking further items.
Item 8	Sixty per cent of the examiners felt that they had received too few items.	Many of the comments on how to improve the system suggested that examiners would like to receive more work and that this work should be more systematically spread where possible. Unfortunately, this is not always possible within the Aptis system due to the way the test is administered.

The administrator questionnaire

	Main findings	Commentary
Item 3	Just over half of the administrators had some experience of administering a computer test.	The lack of experience does not appear to have affected the performance of the group. With this high of a percentage lacking in experience, we might have expected to encounter more problems. However, the later items do not suggest that there were many serious problems.
Item 4	Just over half of the administrators had tried out the test prior to the trial test.	This is potentially problematic; as it is only in actually taking the test that the administrator can get a real 'feel' for the different components. It should be highlighted in the manual that all administrators should attempt at least part of the test prior to administration.
Item 5	The majority of administrators felt the information in the Exams Manager User Manual was clear and easy to use.	The comment we received from the only administrator who disagreed was related to a draft manual. It is important, however, to continue to monitor the effectiveness of all manuals.
Item 6	Half of the respondents felt there was something missing from the manual.	Most of the comments related to this indicated that administrators felt the need for more information on error messages and how to deal with these errors. There was also some concern expressed with a lack of detail in the manual, which interfered with the ease of administration.
Item 7	Just one respondent felt there were problems with uploading candidates into Surpass.	This issue was related to uploading more than two speaking candidates at a time This was probably due to the size of the files, an issue that has been addressed by BTL Group.
Item 8	Just one administrator indicated that there was a problem scheduling test packages.	Unfortunately, the administrator did not indicate what the source of the problem was, so we cannot comment on the issue here.
Item 9	A quarter of the respondents indicated that they'd experienced problems registering candidates.	One of the administrators indicated there had been some problems with keycodes and pins that did not work properly. Another complained it was not possible to edit candidate information in cases where errors were discovered. A third administrator indicated that registration had been done using support from BTL Group throughout CSV spreadsheet. These technical issues should be dealt with and feedback given to the developers as it is important that all problems like this are recorded and addressed.
Item 10	One third of administrators indicated that they could not access and print results easily.	Most of these issues where related to the fact results were not actually available. One administrator admitted they had not yet tried to perform this action.
Item 11	All besides one administrator felt confident about administering future Aptis test sessions.	The comments made by this administrator indicate that any reluctance appears to be related to the difficulty with uploading the speaking component. As mentioned above this has been addressed in a recent BTL Group fix.
Item 12	Most suggestions related to technical issues.	Half of the comments related to specific technical issues around the administration of the test. These are issues that should be dealt with through discussions with BTL Group. Specific comments on the administrator manual are relevant and should be addressed. These comments referred to troubleshooting and error messages, and also to specific guidelines related to test administration such as ID check, system outage etc.

The invigilator questionnaire		
	Main findings	**Commentary**
Item 3	Over half of the invigilators had not invigilated a computer-based test before.	This finding is very similar to all the other questionnaires, and is important especially in light of some of the issues around the invigilator manual described below.
Item 4	No invigilator felt the instructions in the User Manual were unclear.	This is a very positive finding, as it is important that invigilators fully understand what their job entails, especially for a computer-based test, which can be quite technical in nature.
Item 5	Five of the individual teachers indicated they had to deal with problems during the test sessions.	The problems the invigilators referred to were related mostly to confusion around multiple keycodes and two other technical issues with the test. Some of the responses to Item 7 will help us to understand more of what needs to be done with in this regard.
Item 6	Only one invigilator felt unable to deal with all of the issues that came up during the test session.	Unfortunately, since there was no follow-up question with this item we are unable to get to the heart of what the issue or issues might have been. It is important to continue to monitor this aspect of invigilating.
Item 7	Three of the invigilators offer significant feedback with regards to the manual.	The invigilators highlighted a a number of issues that were related to the manual, the test itself, and to the invigilator screen.

Recommendations

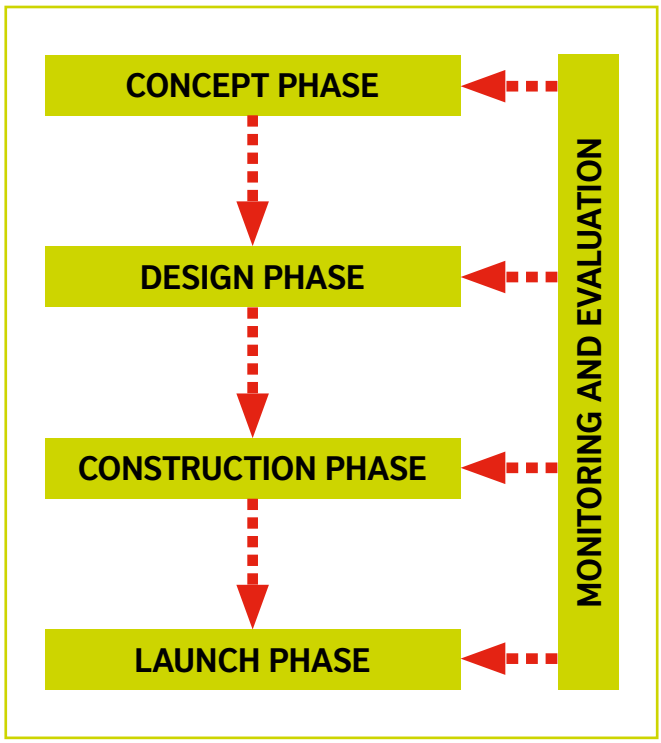
1. Monitor candidates with regard to computer test experience and computer literacy.
2. Ensure that it is possible to change the font size in the reading texts.
3. Allow candidates to change the font size on the test screen.
4. Add an automatic word count function to each of the writing output text boxes.
5. Make online testing and refresher materials available for examiners.
6. Allow the marking screen to fit the computer screen the marker is using.
7. Ensure the marking scheme downloads; this seems to be a problem with the speaking test.
8. Encourage all invigilators to try out the test prior to administration.
9. Add detailed solutions to the invigilator manual to facilitate troubleshooting.
10. Fix specific test-related problems:
    a. Avoid pausing the test so that candidates do not lose testing time.
    b. Make it easier to pin-point all candidates in the event of an emergency to pause the exam.
    c. Allow the invigilator or candidate to lower the volume of the headphones.
    d. Reduce the number of keycodes required.
11. Fix invigilator screen problems:
    a. Allow the invigilator to filter only by more than one criterion at a time.
    b. Allow the invigilator to filter for tested candidates either by date or test component.
    c. Allow the invigilator to unlock all candidates (for one component) in bulk.
    d. Allow the invigilator to see the time remaining for each test.

1. Background

This report forms part of the development cycle of the Aptis tests. Over a period of two years, from 2010 to 2012, the British Council has been committed to developing and bringing to market a major new English language testing system, based on the joint concepts of flexibility and accessibility. The tests that have emerged from the development process became operational on 20 August 2012.

The development cycle of any high quality test includes a number of clearly described stages (see Figure 1 for the development model used in Aptis). This report tells the story of the final part of the construction phase and the formal trials of the new test, which took place in May and June 2012.

Figure 1: The Aptis development model



1.1. Quality assurance

One very important element of test construction is that of quality assurance. During the development process, researchers typically create task and item versions for localised (low level) trialling during the design phase. These are presented to stakeholders, in this case to experienced teachers across the British Council teaching centre network and trialled with small groups of test takers who represent the target population. It is through this process that the test developer slowly builds up a picture of the items and tasks that will go into the final test version.

When we reach the construction phase, we have a clear idea of how the individual papers will look. We will also have some evidence that the individual papers (such as writing or reading) offer a cohesive, broad-ranging and accurate reflection of a candidate's ability. When we begin the process of constructing the test we are faced with the practical issues of replicability (is it feasible to create multiple versions of a task, each of which can be shown to be equivalent), resource availability and cost (are the tasks realistic in terms of development time and cost, delivery and scoring).

Since Aptis is designed to be delivered using a range of formats (computer, telephone and pen and paper) we must be certain that these all work well and do not impact negatively on a candidate's test score. The current report focuses on the delivery of the computer version of Aptis, as this was regarded as the primary concern immediately prior to launch.

1.2. The computer delivery system

Aptis uses the Surpass assessment platform, developed by our partner, BTL Group, in the UK. This platform is already used by a number of UK examination boards, though the system has had to be considerably strengthened for use with Aptis. The main areas of development have centred on the delivery and scoring of performance-based tests (writing and speaking) and around data management and analysis within and outside the system.

2. The formal trial

The formal trial of the Aptis computer delivery system was held in May and June 2012. The focus of the trial was mainly on how well the test elements worked together, and also on how stakeholders responded to the delivery platform, the marking system and the administration guidelines. The main stakeholder groups were:

1. Candidates
2. Examiners
3. Administrators
4. Invigilators

Separate questionnaires were developed to access quantitative and qualitative data from each group, in order to identify specific issues around the delivery of the test. The questionnaires are presented and the associated responses analysed in the following sections.

2.1. Candidates

Since past experience has demonstrated the lack of value (and resultant validity) in asking test candidates to respond to long and complex questionnaires immediately following a test, we decided to keep the instrument short and to the point. For this reason just one yes/no item, five Likert scale items (agree/disagree) and three short open response items were included. These are discussed, together with the candidate responses below.

2.1.1. Item 1

Have you ever done an English language test on a computer before?
This item was designed to establish evidence of the experiential characteristics of the test population. Since the delivery platform is an essential part of the test, we felt it important to know how experienced the candidates were likely to be with regard to taking computer-based or delivered tests.

The responses to this question (Table 1) showed that a slight majority of the respondents had no experience with the delivery format. Later in this report we will return to this to explore how it might have affected their attitude to the test, the delivery of the test and to their perception of how well it reflected their language level.

Table 1: Candidate questionnaire Item 1

	Response %	Response count
Yes	47.9	138
No	52.1	150

2.1.2. Item 2

The look of the test was attractive and appealing.
This item was designed to gather feedback on the physical appearance of the test. This is sometimes referred to as 'face validity', a term that is not particularly well regarded by assessment professionals as it reflects a concern with affect rather than language ability. It is, nonetheless, an important aspect of the process, as poor presentation can negatively impact on candidate perception of the test and result in inconsistent performances.

The results indicate that three quarters of the respondents agreed that the test appeared appealing, one fifth had no opinion either way (it is likely that they were unaffected by the presentation, an equally positive outcome for Aptis), while just five per cent were in disagreement. All-in-all this should be seen as a very positive set of responses.

Table 2: Candidate questionnaire Item 2

	Response %	Response count
Strongly agree	14.6	42
Agree	60.4	174
Neither agree nor disagree	20.1	58
Disagree	4.2	12
Strongly disagree	0.7	2

2.1.3. Item 3

The instructions were clear and easy to follow.
We were interested to know how candidates felt about the instructions for taking the test as this was felt to be a key element, and more important in light of the responses to Item 1, which showed that less than half of the respondents were familiar with computer-based tests.

The results (Table 3) show that almost 90 per cent of the respondents agreed the instructions were clear and easy to follow, with about 10 per cent not holding an opinion (again positive in that they were not negatively disposed or affected by the instructions). Just three per cent felt negatively about the instructions, and the later items regarding negative observations will be reviewed for further information on this.

Table 3: Candidate questionnaire Item 3

	Response %	Response count
Strongly agree	38.9	112
Agree	48.3	139
Neither agree nor disagree	9.7	28
Disagree	3.1	9
Strongly disagree	0.0	0

2.1.4. Item 4

The test gave me an opportunity to show my true level of English.
Two thirds of the respondents replied positively to this item (Table 4), with 25 per cent indicating no preference. Approximately ten per cent did not feel the test offered them this opportunity. This result will again be followed up in our later analysis of the qualitative responses in order to gain a better understanding of why these respondents were not satisfied with the test in such a way.

Table 4: Candidate questionnaire Item 4

	Response %	Response count
Strongly agree	17.7	51
Agree	48.3	139
Neither agree nor disagree	25.0	72
Disagree	6.9	20
Strongly disagree	2.1	6

2.1.5. Item 5

I had no difficulty in using the computer during the test.
Related to Item 1, this item was designed to highlight any significant issues encountered by candidates. This is important as computer-skills related problems that impact on a candidate's performance may mean that we are confounding skills, thus reflecting negatively on the validity of Aptis. Table 5 indicates that four out of five respondents agreed with the statement and another ten per cent not expressing any opinion (we should again see this as a positive outcome). The concern here is with the ten per cent who indicated that they experienced problems in using the computer. The results from the open-response questions will be analysed to see if additional evidence is available.

Table 5: Candidate questionnaire Item 5

	Response %	Response count
Strongly agree	35.4	102
Agree	43.4	125
Neither agree nor disagree	11.1	32
Disagree	8.3	24
Strongly disagree	1.7	5

2.1.6. Item 6

I would recommend this test to other people.
In this item we had hoped to gather an estimate of the overall satisfaction with the test. However, results suggest something else may have been happening in the responses that we had expected. While the degree of satisfaction with the test can be seen as very high, based on the responses to the other items, there is a very different picture being told here. Less than 30 per cent of the respondents indicated they would recommend a test to other people, while almost 50 per cent indicated they would not. It is possible this question was directed at the wrong stakeholder group. It may be naive to a certain extent to believe that a candidate, who may be happy with the test from his or her own perspective, would actually recommend any test to other people.

Table 6: Candidate questionnaire Item 6

	Response %	Response count
Strongly agree	17.4	50
Agree	11.1	32
Neither agree nor disagree	24.0	69
Disagree	41.3	119
Strongly disagree	6.3	18

2.1.7. Item 7

The main advantage of this test is...

For the three open response items (Items 7, 8 and 9), all responses were coded into categories that emerged in the analysis process. In each case there were responses that were not categorised. These were either because they were blank, contained non-word responses or were written in a non-Roman script (Korean in all cases).

Table 7: Candidate questionnaire Item 7

Category	%	Number
General		
Satisfaction	18	54
Washback	4	14
Candidate ability	17	51
Affect	1	5
International standards	0	1
Platform	16	48
Quick results	2	6
Approach		
Test structure	7	23
Familiarisation test	1	4
Instructions	5	16
Time	5	15
Test papers		
Grammar	2	7
Vocabulary	3	9
Reading +	0	2
Reading -	0	1
Listening	4	14
Speaking	2	6
Writing	3	9

Table 7 shows the summary of the comments made by respondents to Item 7. The most commonly referred to categories were positive responses to the test in general, to the test as a measure of ability and to the platform. Between them these categories counted for over 50 per cent of all comments.

Some examples of these comments include:

relatively short and simple to take

I think that is a good test, in fact you can see your level at the end. Besides the test does not take too much time, so it is great.

that it is easy to do. Easy to follow the instructions and to answer the questions

The test is very interesting, it is simple to do it

It is easy to fix my answer, because I just use my keyboard, so I can fix my answer

quite comprehensive and includes all the important skills required to enhance a particular language

To check your proficiency in the English Language, both written and spoken. Also understanding of various accents

it's really helpful for those who want to know their real english level, and it's easy to answer

The test structure received some praise:

Assess your 4 skills in English without further assistance

The different type of question in each part of the exam

that you can know your true level of English (there are various tasks in test)

Of the test papers, the listening received most praise, with comments such as:

Each student has his/her own headphone which can eliminate any noise in the classroom especially during the listening test.

It has various accent in Listening part

Different with other tests, i could listen the listening again whenever i want

2.1.8. Item 8

The main disadvantage of this test is...

Forty-six (15 per cent) of the responses were not categorised. Of the categorised comments, 58 (20 per cent) were actually positive and included comments such as:

I do not think there was any disadvantage of this test. In fact I would like to be part of such tests more often

I did not find any disadvantage

I think there are no disadvantage in the test

I think It doesn't have any dissadvantage if you compare it with a normal exam

The main areas of concern related to the technology. There were a number of comments related to local technology issues:

Some technical problems specially in using Mic. in speaking part.

that sometimes the answers are not registered probably because the candidate didn't press the key strongly

During the writing test, there was a system error, so i could not finish the test

THA MAIN DISADVANTAGE WAS THE LISTENING TEST DUNRIG WHICH I HAD SOME DIFFICULTIES WITH MY SPEAKER. ON THE OTHER HAND, THE LISTENING WAS VERY CLEAR.

It depends 100 per cent on the internet connection, so if the connection is not optimum (as in here) then you will have problems

Other problems highlighted focused on the need to use multiple keycodes when starting each test paper:

The main disadvantage of this test is so many keycode

One disadvantage was constantly having to enter the key code. It was just a bit annoying.

the Keycode and password typing may confuse applicants

A number of respondents complained about the time allowed for the test, some believed the test was too pressured in terms of time:

very time consuming

It is quite long, so it does get tedious sitting for such a long time. Doing it in parts would be simpler

TOO MUCH TIME SPENT ON A CHAIR

Table 8: Candidate questionnaire Item 8

Category	%	Number
General +	20	58
General -	1	5
Setup and technology		
Technical problems	8	28
Use of computer	6	22
Platform load time	2	6
Approach		
Time	11	32
Level	3	13
Inaccurate	1	3
Not attractive	1	3
Possibility to cheat	0	2
Approach too narrow	0	2
Crowded room	0	2
Unclear target population	0	2
Slow Results	0	2
Unclear task	0	2
Unfamiliar Item types	0	1
Instructions	0	1
Test papers		
Grammar and vocabulary		
Grammar	0	1
Vocabulary	3	9
Reading in general	7	19
No text highlighting	0	1
Reading too short	0	1
Font size in reading	3	9
Listening in general	2	8
Speaking in general	4	13
Speaking noise	1	5
Writing	2	7
Automatic word count writing	1	5

Others felt the test was too long:

There isn't enough time to do some question

Is a little bit long, specially the writing test.

Duration of test seems to be short on reading

I think reading test needs more time

Similarly, while the majority of respondents who referred to timing felt the test was too difficult:

This test is little higher or tuff for the begginers, the standard is quite high for the young learners

The level of this test is very difficult for me

The standard is quite high and would be difficult for people whose level of English is not that good

Other respondents felt the opposite:

The questions are quite easy.

Does not seem to test the more advanced level of knowledge

Looking at the other comments some interesting points were made:

People would not take it seriously, compared to IELTS

PEOPLE CAN CHEAT BECAUSE THIS IS ONLINE TEST

I don't think the test covers enough questions to evaluate one's English level properly

These comments reflect some interesting perceptions of both Aptis and online tests in general. This attitude will have to be tackled at some point in the near future. The messages that need to be attached to this type of test is that it is as secure as traditional pen and paper tests and also that it is potentially as sound a measure of language ability.

Looking to the test papers, we found comments on the vocabulary in general:

here were a lot of difficult vocabularies which are not common

some vocabulary is difficult for me and some qusetion I don't understand

This suggests that some, probably low level, candidates are finding the tasks difficult to handle due to the complexity of the vocabulary. This is not surprising given that the tasks within each paper gradually increase in difficulty, from both the language and cognitive processing perspectives.

Other respondents felt that the vocabulary paper itself was difficult:

Vocabulary is difficult.

the vocabulary teset is very difficult for me because I don't know a lot of words. so main disadvantage of this test has got vocabulary section.

Again, it should be recognised that the vocabulary Items tests are in sets of five, and that each set is less frequent than the previous set, and more complex for the lower level candidate.

The various comments on the reading paper focused on the content of the reading texts, with a number of candidates finding them less than inspiring:

The reads weren't very interesting

Many more felt the reading paper was too difficult for them:

Reading very hard for me

Reading it very difficult

This is again not surprising, given that many candidates were at a low level of ability and would have found the later tasks in the paper to be quite complex and beyond their level of ability.

The most commonly expressed concern with the reading was related to the size of the font used in the texts, mainly the final long text:

the main disadvantage is time for to answer some questions like the speaking and the size of the letter in the reading is very small.

The size of fonts is a little small

In the last part of reading section, there is a long text. However, it is little bit hard to see whole at once

Reading. Letters is so small and it's difficult to read them

This is an important issue that we should address as we may be disadvantaging candidates with poor eyesight. In the setup, it should be possible for candidates to manipulate the look of the test to suit their physical requirements. Another respondent commented that the menu for the reading test was difficult to read, another instance of not formatting the screen before the test proper:

I coundn't read the menu on Reading test. The text in the menu bar was too small to read. It would be better to make it bigger to read properly.

Among the other comments was the call for a highlighting tool for the reading paper:

in the reading test i needed to highlight certain lines in the paragraphs and I couldn't

When commenting on the speaking test, respondents tended to focus on two issues, that of a lack of time for the speaking tasks, and the format itself:

while speaking you have very little time to think about what you are going to say

speaking needs more time

The speaking, because you have to take into account that time is short and the microfone was´t the best

I prefer a personal interview

You don't talk to a real person

The speaking part felt a little bit uncomfortable because you are speaking with a computer and not with a person, it confuses you

In addition, respondents pointed out the problem of trying to take the test when other candidates were also taking the test in the same room:

When you are doing the speaking test, you listen to other classmates at the same time than you are doing it, and it's difficult to get concentrate

noises at the background of the room

Since a group is seating and giving the speaking test there is a lot of noise and low concentration when one has to give their own speaking test

One respondent even offered a solution to this problem:

Speaking test should be in proper rooms – one person should be in glass box, as in lingophonic cabinets. Today others disturbed me to make speaking test

When it came to the writing paper, many comments centred around the limitations of building the paper on a single topic:

the writing test presents just a topic to write about. It must have differents topics to write about.

It is hard to answer the writing test because in Korea the club activity is not working well

In writing, question is academic

However, one comment highlighted a technical issue that may need more thought:

windows for the writing part too small (esp for the longest piece) - i.e. candidates can see very limited part (4 rows?) of their work and need to scroll the text

If this really is the case we will ask BTL Group to increase the size of the response box to allow the writer to see all of the response. Another technical issue that we should consider implementing is the provision of an automatic word counter for the writing responses. The fact that this was missing from the test was mentioned by a number of respondents:

no automatic word count for the writing task.

Hard to count the numbers of letters on writing part

maybe you could add a word count tool to the writing component

This is particularly problematic when combined with the small window size as it would make it quite difficult for candidates to count their word totals during the test.

2.1.9. Item 9

Please write any comments you have about the test here. Fifty-eight (20 per cent) of the responses were not categorised. Of the categorised comments, 143 (49 per cent) expressed positive views about the test:

It was an interesting assessment. Enjoyed doing it

It was quite interesting and kept me glued to it. It is a very fine tool to test one's proficiency in English language

I enjoyed attempting this test and I feel it should be a part of our curriculum so that students are also able to develop on their skills.

The content was very good. Wide variety of questions were asked

this is very good expirience for me. online test is very good

For me, the test was great cause we could evaluate all the skills and the teacher explained how we can use the technology

It was really good. I think is a great opportunity to know our level and how it`s going

It was a good exam, I would take it again

Though we were reminded that we still had some work to do:

I didn't like the test

This exam can't test all aspects of language ability it is just suitable for testing grammar and vocabulary... thanks

I think it can be another great test compared to IELTS or IBT after changing some level of problems

The references to other tests, even though they are not our competitors, are very interesting as it points to the quality of the test presentation and delivery. This is something we must get right if we are to succeed. Candidates make judgements on a test using their own criteria and these kinds of comparisons cannot be avoided.

Table 9: Candidate questionnaire Item 9

Category	%	Number
General		
Positive general	49	143
Hesitant general	0	1
Negative general	4	13
Technical		
Technical problems	1	5
Keycodes	0	2
Approach		
Low level	2	8
High level	4	12
Preparation	1	4
Uneven level	1	4
More time needed	1	3
Too long	0	2
Negative instructions	0	2
Hesitant time	0	1
Suggest report	0	1
Test papers		
Positive grammar	0	1
Negative grammar	0	2
Positive vocabulary	0	1
Hesitant vocabulary	0	1
Negative vocabulary	2	7
Negative reading	3	9
Font size	0	1
Negative listen	2	7
Negative speaking	5	15
Negative writing	3	9
Positive writing	0	1

Many of the other comments in this item simply re-stated issues that had been pointed out in the earlier items. Many comments referred to the perceived level of the test (too high or too low):

The seemed too easy.

I found it not challenging enough to determine the level of English of the candidates.

I am not sure on which level of English was the tested focused on. From my point of view it seemed easy. Maybe it needs better distinction of levels.

I think that the test was very difficult for me besacuse the level was very high

It very difficult

i think it dificult for me some test

Thank you very much for your test. It was very interesting and instructive, but very difficult

There were a number of comments related to test preparation, which indicate a possible weakness, though it should be pointed out that not all centres carried out the expected familiarization process:

this test need parporation course to do it and more exersices and books to study

it was FUN to be honest. and i was lot comfortable doing a test this way than the normal test. oh and there was a small problem in the familiarization part in writting test the time ended so quickly that i didt get to finish it. since that is like the 1st part of this TEST ,they should give more time to familiarize with the test.

Since it is a new approach, our users/ candidates/ students should be given provided with the detailed professional demonstration probably twice before the test day to ensure quick/ acurate understanding and also, to minimise anxiety or last moment stress/ nervousness among the new users.

Comments on the unevenness of the test papers suggest that the candidates had not been well prepared for the test as they show that they did not understand the way the test papers were constructed (each with a range of tasks and items that gradually became more and more difficult):

I love this test. It was fun. but the level was quite unclear. something was easy and sth was hard

Probably the one thing which I could mention about the test above - is the strong gap between used level or exercises. There was no middle diffficulty, the questions were really hard or very easy. No middle level, in my own view.
Thank you.

Most criticism was levelled at the speaking paper, focusing on the issue of surrounding noise:

I think it's better, if everyone takes the speaking test in other classroom, or do it by part differents no everybody at the same time.

It's hard to focus when Speaking for all candidates in the same time.

I enjoyed it and I believe on the speaking part it will need a privacy area

The lack of a human interlocutor:

Speaking must be a real test with real people

I would separate the speaking part from the other parts because I think it's easier to talk personally to someone instead of talking to a computer, it is a weird feeling.

The lack of time:

The speaking was very fast, no time to answer the question.

I liked it, but there was not enough time for speaking :)

The comments on the reading paper highlighted the difficulty reading the small font size:

It was uncomfortable to read the paragraphs in the reading section. It would be nicer if the size of fonts is bigger.

The lay-out of the reading needs a bit of improvement...
We could hardly read and can not scroll up and down easily.

One respondent referred to a hitherto unknown issue; inadvertently skipping a familiarisation test:

I accidently skiped familiatisation for reading and there were no way to re-do the step which would have been very stressing in real situation.

This may not be a serious issue as only one person records having done it, though it does point to the fact (yet again) that the person may not have been properly prepared for the test. This is a more important problem for us in the longer term and we need to ensure that test users understand the importance of adequate test preparation.

2.1.10. Summary of candidate questionnaire findings

In this section we briefly summarise the main findings of the candidate questionnaire responses. While the outcome of the questionnaire was overwhelmingly positive, with most candidates happy with the test from various perspectives, there were a number of issues pointed out that will need to be addressed in the coming months. These are discussed below.

	Main findings	Commentary
Item 1	Just under half the respondents had taken an English language test on a computer.	We should monitor this situation carefully in the future. It may be necessary to routinely ask this question as part of the registration process in order to conduct statistical bias analysis of the results. It is important to ensure language ability and computer test familiarity (and/or computer literacy) are not confounded in the test performance.
Item 2	Seventy-five per cent of the respondents felt the test was attractive and appealing, 5 per cent disagreed.	Reassuring to a large extent, though it might be useful to understand why the 5 per cent disagreed. Since it is impossible to please all candidates we would always expect some level of disagreement, and later comments on the font size in the reading test suggest that there may be a problem with the 'look' of the test.
Item 3	Just 3 per cent felt the instructions were not clear and easy to follow.	This is a very positive outcome as we were dealing with a population with a broad spread of ability, some of whom may have been expected to struggle with the English instructions.
Item 4	Nine per cent did not feel that the test offered them an opportunity to show their true level of English.	Given that this was a trial, and that there were a number of technical issues with the test delivery, this is not a bad result. However, it may be useful for us to explore the area more in the future, possibly encouraging research on the topic in the Aptis Research Awards.
Item 5	Ten per cent had difficulty using the computer during the test.	This appears to have been due to a range of factors, some related to the test and the lack of experience the candidate had with the technology, others with local technical issues (poor equipment, outages etc.). It is something, however, that we need to further explore (see the comment for Item 1).
Item 6	Fifty per cent would recommend the test, 30 per cent would not.	This item would have benefited from an additional open ended response element asking for a reason for the decision (this should be included in a future iteration of the questionnaire). It may be that we were somewhat naïve in asking candidates to recommend a test (any test) and that this question would be better asked of policy and decision makers.
Item 7	General satisfaction with the test and the platform.	A very positive set of responses all round. Respondents seemed to feel that the test offered a good estimate of their ability and that it was easy to take, with clear instructions. In terms of individual papers, most comments focused on the vocabulary paper.
Item 8	Twenty per cent expressed satisfaction even when asked to point out weaknesses with the test.	This was again good news for the test. However, some issues with technology (including the problem of having multiple keycodes) and a wariness of taking the test with limited computer literacy were highlighted. Among the other criticisms were time (some felt it was too long, other too short) and perceived level (some felt that it was too high, others that it was too low). Specific paper related problems were also highlighted. These included problems reading the input texts in the reading paper as the font was too small (and could not be manipulated by the candidates), a lack of time and impeding background noise in the speaking paper and the lack of an automatic word counter in the writing. These paper-related items should be addressed in upcoming test reviews.
Item 9	50 per cent positive, 50 per cent negative.	Many respondents were very happy with the test. Some were unhappy and others simply wished to offer advice. The main issues that arose related again to level, some people feeling it was too high while others felt it was to low. There were also comments on the technical problems and on the issue of multiple keycodes (an area that has been addressed). A number of paper related problems were restated here.

2.2. The examiners

A total of 22 examiners responded to the questionnaire designed to gain feedback on their experience. The responses to the items are presented, as with the candidate questionnaire, by item.

2.2.1. Item 1

Did the training sufficiently prepare you for your examining role?

All respondents indicated the training had been sufficient in this regard. This is a satisfying outcome as it indicates our training approach is working well. This is particularly important as it supports the anecdotal evidence from training events where examiners have generally passed through the accreditation test without difficulty and where their attitude to the whole experience has been very positive.

2.2.2. Item 2

How could the training be improved to prepare you for your examining role?

One examiner expressed some concern that more practice on the computer would be helpful prior to the training (or possibly at the beginning of a training session). There were a variety of comments related to the actual training event, though no clear pattern emerged, except to suggest the sessions be lengthened to allow for additional time on each aspect of the training event. A typical comment was:

> *perhaps more practice and discussion but ok anyway*

Most comments were about post training. Here, the most commonly referred to aspect was the provision of additional practice texts, the provision of a forum for additional discussion and more feedback:

> *By providing more practice marking opportunities*
>
> *more possibility to practise on dummy tests online with correct scores given at the end of the exercise*
>
> *Perhaps further group meetings after we had done some test marking so questions could be raised and answered altogether which would avoid the duplication of answers to email queries and give examiners the possibility of hearing each others' experiences.*
>
> *Ideally we - new Examiners - would have a forum or a general meeting to discuss our marking experiences.*
>
> *more practice & regular feedback on our work.*
>
> *Inputs on where we are right and wrong with reasons could be given*

These are clearly some areas where we should aim to improve our system. We have already discussed the creation of an online training and refresher system and it appears from this feedback we should work to deliver these as soon as it is feasible.

Table 10: Examiner questionnaire Item 2

Category	%	Number
Pre training		
Computer practice	5	1
During training		
More discussion	5	1
More examples	5	1
Demonstration of marking	5	1
More demonstrations	5	1
Handouts in speaking	5	1
Post training		
Additional practice	38	7
More discussion	16	3
Online training	5	1
Feedback	16	3

2.2.3. Item 3

Did you find SecureMarker easy to use?

As with Item 1, all examiners indicated that they felt it was easy to use. Again, this is a very positive outcome.

2.2.4. Item 4

Did you ever feel that you could not mark a script and/or interview?

About two thirds of the examiners felt that this had never been a problem. However, the remaining group (eight examiners) had felt that they could not mark a script or interview at some point in their work. The following item (Item 5) explores this further).

Table 11: Examiner questionnaire Item 3

	Response %	Response count
Yes	38.1	8
No	61.9	13

2.2.5. Item 5

Please provide examples of problems you faced with SecureMarker.

Despite indicating that there had never been a problem, 20 examiners responded to this item (when we would have expected responses from the eight who responded positively to Item 4). The responses are summarised in Table 11.

The most commonly cited issues were related to technical problems encountered. These referred to the visual presentation of the work:

Viewing the pictures for the Speaking Tests but it was fine when I moved from my laptop to a different computer. It was sometimes difficult to hear the candidates.

For the writing task, I could not see the bottom of the page with its answer because I use a small laptop. On the large laptop or PC, the entire page is visible. This forced me to hunt for a large screen laptop to be able to do my work. I tried 2 small taptops, but the result was the same.

The repetition of items (actually, these were Control Items – this issue has since been resolved):

1. repetition of the same items, 2. suddenly blocking me out with a comment 'you are not allowed to mark'

Problems understanding what work was to be done:

Sometimes the screen said that I had completed marking the quota while the quota section still showed unassessed tasks.

Access and upload:

No upload of quota at times

Sometimes I could not access the tasks I was to assess

Initially, I was not able to locate the items that required marking

Marking scheme not downloading:

The marking scheme did not download for Speaking tasks

marking scheme missing,

Loading is slow and this consumes time. Sometimes the assigned items become inaccessible. Though it displays the reason we don't know how to troubleshoot and continue.

The most commonly referred to problem with the test itself was the quality of the sound on some of the files:

Not with the marker, but with some of the recordings. Seems like insufficient attention is paid either to the equipment used or that test takers are not properly instructed.

Several very poor quality Speaking items

inaudible sound files on the speaking task

Table 12: Examiner questionnaire Item 5

Category	%	Number
Technical issues		
Technical problems	35	7
Loading time	15	3
Confusing system feedback	10	2
Text not seen on screen	5	1
Unable to access task	10	2
Test Issues		
Limited output	5	1
Sound quality	25	5
Version?	5	1

2.2.6. Item 6

What improvements to SecureMarker would you suggest?

The responses to this item were all related to specific technical improvements to the system (see Table 13). The issue of system speed has been addressed, while the remaining comments will be investigated in the coming months to explore their feasibility.

Table 13: Examiner questionnaire Item 6

Category	%	Number
Speed up	5	1
Automatic screen size adjust	5	1
Automatic fill of details	5	1
Remove '0' default	5	1
Font size in mark scheme	5	1
Reduce logout time	5	1
CI feedback	5	1
Simplify display	5	1
Sound quality	11	2

Specific comments included:

I suggest the screen size should be such as to adjust pages to every size of laptop, without letting the lower part of the page 'drop off' the lower part of a small sized laptop.

On logging on, we should be provided an autofill option (as in gmail) so that we can save some time.

When submitting marks, I often have to delete the default zero before I put in the assessed mark. Could that space be left blank, so we can just fill in the mark without having to first delete the zero?

Provide an option to go back on a marked item

Would it be possible to further simplify how tests are displayed for access - this may just be a case of getting more used to how they are displayed.

It logs out automatically too quickly if there is a break

The font size of the marking scheme should be increased.

2.2.7. Item 7

Were you easily able to mark your items within 48 hours?

As can be seen in Table 14, all but two examiners responded that they could manage this requirement. There were four comments added to the item.

Table 14: Examiner questionnaire Item 7

	Response %	Response count
Yes	90.0	18
No	10.0	2

Two of the examiner comments focused on the technical problems encountered:

As mentioned above, at times the items were rendered inaccessible.

I experienced error problems which stopped me from marking any further items

The other two comments indicated that these examiners felt that the management of the system was working well:

Sufficient time allocation especially when we have pre-allocated the time.

It was great that there was constant support with fast e-mail replies if we did have queries.

2.2.8. Item 8

Did you receive the expected number of items for marking?

It seems from Table 15 that there is an issue here that should be dealt with as a large percentage of examiners are not receiving the number of test performances to mark they had expected.

The comments of the examiners indicate many would like to receive more:

Can handle more work & hope to get more items in the future.

A few more items can be given. Suggested 45 hours per examiner

I don't know how the allocation system works. 48 hours to mark is good but I would personally like more items please.

Others would like more systematic structure (not something that can happen all the time in this type of test, where the volume of tests has tended to date, be less frequent than when the test is fully operational.

Even distribution that allows me to work for two hours everyday.

A fairly equal distribution of available tests amongst the assessors.

SEND NEW ITEMS AS SOON AS FIRST ONES HAVE BEEN MARKED

If it's possible to give people an indication of how many hours a week we would be working then we'd feel a bit more organised/it would be a bit more systematic

One examiner seemed confused about how much marking was expected:

I am not sure how many items I was suppose to be receiving

Table 15: Examiner questionnaire Item 8

	Response %	Response count
Yes	40.0	8
No, too many items	0.0	0
No, too few items	60.0	12

2.2.9. Summary of examiner questionnaire findings

The main findings from the examiner questionnaire are presented below.

	Main findings	Commentary
Item 1	All examiners felt that their training to prepare them sufficiently for the examining role.	This is a positive result as it is very important administrators feel confident in their training and in their ability to perform the role competently.
Item 2	The provision of more practice, discussion and feedback would improve training for the exam.	The Aptis team had already planned to develop online training and refresher materials for examiners. It would appear from this feedback that we should go ahead with this plan at the earliest possible opportunity. It may also be useful to include a discussion forum in any examiner-focused website.
Item 3	All examiners felt SecureMarker was easy to use.	Again this is a very important finding as it is necessary examiners feel secure and confident in their ability to interact with the system.
Item 4	More than 60 per cent of examiners felt there was at time when they were unable to mark a script or interview.	This is potentially quite problematic as it is important that examiners are at all times able to award marks. Analysis of the responses to Item 5 may offer more information about this.
Item 5	The main reason why people had problems marking appear to be related to technical problems with the system.	The most common problem recorded by the examiners was related to the sound quality of the speaking test. Another problem related to the fact the screen would not always fit onto the computer screen the individual was using to mark. There were other issues with the marking scheme not downloading and problems with loading time.
Item 6	A number of technical issues, generally relatively minor, were pointed out by examiners.	The issues that were highlighted centred around things like sound quality and the physical presentation of the work on screen. All of these issues will be dealt with by the Aptis team.
Item 7	Ninety per cent of the examiners felt it was easy to mark the items that were given within 48 hours.	Those people who indicated there had been problems also indicated that the problems were associated with technical issues. This was because at times it was impossible to mark because the system was inaccessible or the system had prevented the individual from marking further items.
Item 8	Sixty per cent of the examiners felt that they had received too few items.	Many of the comments on how to improve the system suggested examiners would like to receive more work and that this work should be more systematically spread where possible. Unfortunately, this is not always possible within the Aptis system due to the way the test is administered.

2.3. The administrators

A total of 12 administrators responded to the questionnaire. The responses to the items are presented, as with the other questionnaires, by item.

Items 1 and 2 referred to the examiner him/her self (i.e. name and country) so will not be reported here.

2.3.1. Item 3

Have you administered a computer based test before?
In the same way that about half of the candidates had experienced a computer delivered test prior to these trials, about the same proportion of administrators report having experience of administering such a test (see Table 16).

Table 16: Administrator questionnaire Item 3

	Response %	Response count
Yes	58.3	7
No	41.7	5

2.3.2. Item 4

Have you taken any components of this test yourself yet?
The same pattern of response applies to Item 4. It is a bit concerning that almost half of the examiners had not actually tried out the test before administering it. It is very important that all administrators regularly re-visit the Aptis test so as to retain an adequate level of familiarisation with the test, as we feel this would assist the whole administrative process.

Table 17: Administrator questionnaire Item 4

	Response %	Response count
Yes	58.3	7
No	41.7	5

2.3.3. Item 5

All the information in the Exams Managers User Manual was clear and easy to follow.
The overwhelmingly positive response to this item is reassuring, though the single disagreeing response should be explored. The respondent commented that:

> *It was not the final version, and it still shows editing notices on the right side. It was not easy to read from time to time.*

> *It would be useful to monitor the situation with the manual from time to time in order to ensure that it meets the needs of the examiners.*

Table 18: Administrator questionnaire Item 5

	Response %	Response count
Strongly agree	16.7	2
Agree	66.7	8
Neither agree nor disagree	8.3	1
Disagree	8.3	1

2.3.4. Item 6

In your opinion, was there anything missing from the manual?
Half of the respondents reported that there were elements missing from the manual (Table 19).

Table 19: Administrator questionnaire Item 6

	Response %	Response count
Yes	50	6
No	50	6

The most common problem highlighted was related to the error codes, the need for a table of such codes with an explanation was highlighted by four of the respondents:

> *Error codes and what does each one mean, like i was getting error code 801,803,866 I did not find a table or explanation in the manual.*

> *More points in the trouble shooting section. More explanation of the error numbers*

> *When we run the speaking module we found the unknown error message but we couldn't found the action we should do for the this, so we asked to London. The User Manual should be finalised before released. Confused as it is not the final version.*

> *Normally, When I receive any kind of manual to administer the test, we just need to follow the step. For example, 1st section : prior to the test, Test Administration, Post Administration etc.. However, the User Manual is not easy to follow between the section, we found some missed step and we needed to ask to London or the centre who did already. Also, more error message explanation should be in there.*

The final comment above also refers to the lack of accurate detail in the manual:

> *I think it was too early to issue the manual when the final page has not been developed yet. It was confusing because the test package site that we got didn't seem to be the final version. I had to spend a lot of time to find where to create 'centre users' or some of the front setting-up parts.*

The issue of administering multiple sessions in an array of centres was also mentioned:

> *How to administer several sessions in several centres at the same time.*

2.3.5. Item 7

I was able to upload candidates into Surpass without any problems.

Few issues were reported for this item (Table 20), though one administrator felt there had been a problem:

> *I was having troubles starting or uploading the speaking test for more than 2 candidates at a time.*

Table 20: Administrator questionnaire Item 7

	Response %	Response count
Strongly agree	33.3	4
Agree	25.0	3
Neither agree nor disagree	33.3	4
Disagree	8.3	1

2.2.6. Item 8

I was able to schedule test packages without any problems.

Even fewer issues were reported for this item (Table 21), though again one administrator felt there had been a problem:

> *I had some problems to schedule familiarization tests for the recent batch was done on 24 July 2012*

Table 21: Administrator questionnaire Item 8

	Response %	Response count
Strongly agree	58.3	7
Agree	25.0	3
Neither agree nor disagree	8.3	1
Disagree	8.3	1

2.3.7. Item 9

I was able to register candidates without any problems.

While this was generally positive (Table 22), three administrators felt there had been problems. Their comments again referred to keycodes:

> *I had some problems while doing it, some keycodes and pin did not work properly*

Problems editing candidate details:

> *Editing candidate information in case you discover errors was not possible.*

One administrator reported that registration had been:

> *done by BTL Group support throught CSV spreadsheet*

Table 22: Administrator questionnaire Item 9

	Response %	Response count
Strongly agree	41.7	5
Agree	25.0	3
Neither agree nor disagree	8.3	1
Disagree	25.0	3

2.3.8. Item 10

I was able to access and print results without any problems.

The responses to this item were quite mixed (Table 23) with four administrators feeling there had been problems. Of the four responses negative however, one indicated that he/she had yet to try doing this, while the others reported various problems:

> *We didn't print out the result as we didn't receive it yet.*

> *Some of the speaking tests did not upload automatically and we will have to force the upload.*

> *Results were not available*

Table 23: Administrator questionnaire Item 10

	Response %	Response count
Strongly agree	41.7	5
Agree	8.3	1
Neither agree nor disagree	16.7	2
Disagree	33.3	4

2.3.9. Item 11

I feel confident about administering future Aptis test sessions.

The responses to this item were generally positive (Table 24) with just one administrator feeling unconfident. The issue raised by the administrator has been recognised by Aptis and a fix has now been developed by BTL Group which should mean a dramatic improvement in the upload time.

> *I like the way Aptis test works but I feel that it needs to have some more work on the way the component speaking is being uploaded, the file gets to big and some of them does not upload itself.*

Table 24: Administrator questionnaire Item 11

	Response %	Response count
Strongly agree	33.3	4
Agree	25.0	3
Neither agree nor disagree	33.3	2
Disagree	8.3	1

2.3.10. Item 12

Please write any other comments you have about the Exams Manager manual or your 500 pilot experience.

The responses to this item related to technical issues:

> *I liked the bulk uploading systems on candidates part. I wish there were 'tick boxes' on Candidates, Scheduling, and Invigilation pages so that I can delete the unnecessary datas efficiently rather than clicking on each line. Also, it seems to be showing too many items on one page and sorting function does not word effectively. It takes too much time to find the information I want especially when it comes to 4 familiar skills + 4 skills session which we did for piloting. We were supposed to deliver Speaking Module to candidates, but due to some systematic errors (ex. Error 801), we could not help but skip the Speaking test. We successfully delivered G&V, L, R, and Writing modules, but on writing section, some of the candidates had gone through typing freezing incidents while they were on the tasks. They could type numbers and symbols keys except for the letters. It just froze sporadically (both on question number and remaining time), so I could not catch which part was the problem. Other than that, it was much easier to invigilate and conduct piloting session.*

> *Connectivety problem with server, which caused delay of end of exam and students being late.*

> *Managing technical problems with candidates and rescheduling a test for them is quite difficult and not straightforward. I would suggest a quicker way to include a candidate in a session instead of voiding his/her session and rescheduling a session individually.*

> *Few caniddates had error message (Error: 801) of disconnection during the test and it took long to reconnect. In this case we had to close and restart again. The good thing was that it started from the same point where it was disconnected.*

> *What is the Familirisation Test? We were running the f. tests for the core component (G&W) + writing. How to fillter the candidates assigned to the exact test date? We saw other uploaded candidates for installation test. We did not find any candidates results in the Results tab (empty). Why is the Familirisation tests locked by invigilator after putting in keycodes by candidates, This did not happed when they entered keycodes and PINs to login to tests. (G&W and Writing)*

To limitations in the manual:

> *Just that there should be more points in the trouble shooting section.*

> *I think we should have an Invigilator Script for the test day, more information on ID check, what to do in case of the system stop working, etc.*

Finally, three comments were very supportive:

> *Many thanks*

> *It was fabulous.*

> *The manual was detailed and helpful and the experience was good.*

2.3.11. Summary of findings from the administrator questionnaire

The main findings from the examiner questionnaire are presented below.

	Main findings	Commentary
Item 3	Just over half of the administrators had some experience of administering a computer test.	The lack of experience does not appear to have affected the performance of the group. With this high of a percentage lacking in experience, we might have expected to encounter more problems. However, the later items do not suggest there were many serious problems.
Item 4	Just over half of the administrators had tried out the test prior to the test.	This is potentially problematic; as it is only in actually taking the test that the administrator can get a real 'feel' for the different components. It should be highlighted in the manual that all administrators should attempt at least part of the test prior to administration.
Item 5	The majority felt the information in the Exams Managers User Manual was clear and easy to use.	The comment we received from the single administrator who disagreed was related to a draft manual. It is important, however, to continue to monitor the effectiveness of all manuals.
Item 6	Half of the respondents felt there was something missing from the manual.	Most of the comments related to this indicated that administrators felt the need for more information on error messages and how to deal with these errors. There was also some concern expressed with a lack of detail in the manual, which interfered with the ease of administration. These are clearly areas that need to be addressed in the immediate future.
Item 7	Just one respondent felt there were problems with uploading candidates into Surpass.	This issue was related to uploading more than two speaking candidates at a time. This was probably due to the size of the files, an issue that has been addressed by BTL Group.
Item 8	Just one administrator indicated there was a problem scheduling test packages.	Unfortunately, the administrator did not indicate what the source of the problem was, so we cannot comment on the issue here.
Item 9	A quarter of the respondents indicated they'd experienced problems registering candidates.	One of the administrators indicated there had been some problems with keycodes and pins that did not work properly. Another complained it was not possible to edit candidate information in cases where errors were discovered. A third administrator indicated that registration had been done using support from BTL Group throughout CSV spreadsheet. These technical issues should be dealt with and feedback given to the developers as it is important all problems like this are recorded and addressed.
Item 10	One third of administrators indicated they could not access and print results easily.	Most of these issues were related to the fact that results were not actually available. One administrator admitted they had not yet tried to perform this action.
Item 11	All except one administrator felt confident about administering future Aptis test sessions.	The comments made by this administrator indicate that any reluctance appeared to be related to the difficulty with uploading the speaking component. As mentioned above this has been addressed in a recent BTL Group fix.
Item 12	Most suggestions I related to technical issues.	Half of the comments related to specific technical issues around the administration of the test. These are issues that should be dealt with through discussion with BTL Group. Specific comments on the administrator manual are relevant and should be addressed. These comments referred to troubleshooting and error messages, and also to specific guidelines related to test administration such as ID check, system outage etc.

2.4. The invigilators

A total of seven invigilators responded to the questionnaire. The responses to the first two items (which again asked for name and country) are not reported here.

2.4.1. Item 3

Have you invigilated a computer based test before?
The pattern of response to this item is very similar to that of the other questionnaires. Approximately half of the invigilators had invigilated a computer test and the others had not, as can be seen in Table 25.

Table 25: Invigilator questionnaire Item 3

	Response %	Response count
Yes	42.9	4
No	57.1	3

2.4.2. Item 4

The instructions in the invigilator User Manual were clear and easy to follow.
As can be seen on Table 26, no invigilator felt there was a problem with the manual in this regard.

Table 26: Invigilator questionnaire Item 4

	Response %	Response count
Strongly agree	42.9	3
Agree	14.3	1
Neither agree nor disagree	42.9	3
Disagree	0.0	0

2.4.3. Item 5

Did you have to deal with any problems during the test session?
Five of the seven respondents indicated they had such a problem. Their problems were related to multiple keycodes:

> *Some of the keycodes didn't work*

> *Internet connectivity issues, and one candidate could not take the grammar and vocabulary component given the confussion generated for having to enter one code and one pin for each component, I think it should be just one access code.*

> *Candidates complained that they needed to log in and out for every single component and also, familiar session of each session was too long.*

Problems encountered in using the keyboard during the writing:

> *During the writing session, some of users cannot typed some of mark so they needed to stop.*

The lack of agreement between the listening questions on screen and the audio (numbers not matching):

> *Listening question on the screen and the listening voice did not match.*

And finally, the fact the computer froze during the writing tutorial:

> *computer got stuck during the tutorial test of writing*

> *Computer was stuck after familiarisation test writing.*

All of these issues were reported by the candidates and administrators and should be addressed.

2.4.4. Item 6

I was able to deal with all issues which came up during the test session.
As can be seen on Table 27, just one invigilator felt they were able to deal with all issues that arose during the test. Unfortunately, no follow-up question was included to further explore what this issue might have been.

Table 27: Invigilator questionnaire Item 6

	Response %	Response count
Strongly agree	14.3	1
Agree	42.9	3
Neither agree nor disagree	28.6	2
Disagree	14.3	1

2.4.5. Item 7

Please write any other comments you have about the invigilator manual or your invigilation experience below.

In total, six of the invigilators responded to this item. Two of these indicated that all was fine with the system and a third did not actually include a comment. The remaining three responses are presented below.

> *The invigilator manual is too general. It needs more detail on on site, at the spur of the moment troubleshooting. We paused the exam as if there had been an emergency and the pause effectively starts about two minutes later and the time keeps running. It was quite difficult to pin point all candidates in the event of an emergency to pause the exam. There was no way of lowering the volume of the headphones because the screen locks. The countdown of the questions in the listening bit is incorrect which generates a lot of uneasiness in candidates thinking that three of their questions will not be considered for their results. The invigilator screen is quite difficult to manage since it shows all candidates at the same time and you can filter only by one criterion at a time. The constant refreshing of the screen makes the admin of the invigilator information quite difficult since it keeps jumping to the beginning even though you are reviewing information at the end.*

> *The writing is lacking a word counter. The speaking I think it is very good. Deffinitely the issue of delivering 8 sheets to each candidate it makes it look a bit messy. 5 access codes to the 5 components should be written in a single sheet.*

> *Administration was easy as the system for Invigilators is user-friendly. I would like to make a few points which were not clear during examination: 1) filtr - I was not able to apply filter for tested candidates either by date nor the component. There were other candates (made by support during installation) and it was bit confusing. 2) there was not option to unlock all candidates (for 1 component) in a bulk. It was confusing especially when creen kept refreshing by itself all the time. 3) There is no check on invigilator screen on the remining time of the test. 4) What is Familiarisation Test? We run Familiarisation test for Grammar and Vocabulary and Familiarisation Test for Writing only.*

These comments are valuable in that there is a great deal of very technical information included. It is clear some improvements are needed to ensure the invigilator system is improved. Each of the issues need to be addressed as we proceed with the delivery of Aptis. The comments can be viewed under three broad headings as shown below.

The manual

1. The invigilator manual is too general. It needs more detail on site, at the spur of the moment troubleshooting.

Test-related problems

2. We paused the exam as if there had been an emergency and the pause effectively starts about two minutes later and the time keeps running.
3. It was quite difficult to pin point all candidates in the event of an emergency to pause the exam.
4. There was no way of lowering the volume of the headphones because the screen locks.
5. The countdown of the questions in the listening is incorrect, which generates a lot of uneasiness in candidates thinking three of their questions will not be considered in their results.
6. The writing is lacking a word counter.
7. The issue of delivering 8 sheets to each candidate makes it look a bit messy. 5 access codes to the 5 components should be written in a single sheet.
8. What is a Familiarisation test? We run a Familiarisation test for grammar and vocabulary and a Familiarisation test for writing only.

Invigilator screen

9. The invigilator screen is quite difficult to manage since it shows all candidates at the same time and you can filter only by one criterion at a time. The constant refreshing of the screen makes the administration of the invigilator information quite difficult since it keeps jumping to the beginning even though you are reviewing information at the end.
10. Filter - I was not able to apply a filter for tested candidates either by date or by component. There were other candidates (made by support during installation) and it was bit confusing.
11. There was no option to unlock all candidates (for 1 component) in a bulk. It was confusing especially when screen kept refreshing by itself all the time.
12. There is no check on invigilator screen on the remaining time of the test.

2.4.5. Summary of the invigilator questionnaire findings

The findings of the analysis of the responses to the invigilator questionnaire are reported below.

	Main findings	Commentary
Item 3	Over half of the invigilators had not invigilated a computer-based test before.	This finding is very similar to all of the other questionnaires, and is important especially in light of the issues around the invigilator manual described below.
Item 4	No invigilator felt the instructions in the User Manual were unclear.	This is a positive finding, as it is very important invigilators fully understand what their job entails, especially for a computer-based test, which can be quite technical in nature.
Item 5	Five of the individual teachers indicated they had to deal with problems during the test sessions.	The problem the invigilators referred to were related mostly to confusion around multiple keycodes and two other technical issues with the test. Some of the responses to Item 7 will help us to understand more of what needs to be done in this regard.
Item 6	Only one invigilator felt unable to deal with all of the issues that came up during the test session.	Unfortunately, since there was no follow-up question with this item we are unable to get to the heart of what the issue or issues might have been. It is important to continue to monitor this aspect of invigilation.
Item 7	Three of the invigilators offer significant feedback with regards to the manual.	The invigilators highlighted the whole array of technical issues that were related to the manual, the test itself, and to the invigilator screen. These issues also relate to a number of potential weaknesses with the actors platform that should be addressed.

3. Conclusions

The surveys administered as part of the formal trials of Aptis (the 500 trial) suggest the test is well considered by all of those associated with it. While the majority of candidates, examiners, administrators and invigilators were happy with the test and with the associated manuals, a number of issues arose which required action. These issues are highlighted within the actual analysis and summarised at the end of each of the sections. It should be noted here almost all of the comments related to technology and delivery problems, with very few comments related to the actual test or the content of the test.

Since the form trial ended, the issues highlighted in this report have been addressed by the Aptis team and BTL Group. The exercise demonstrates our commitments to ongoing quality assurance and improvements of the test itself, and all of the complex systems which support its delivery, a scoring and reporting.

3.1. Recommendations

The following recommendations are based on this analysis:

1. Monitor candidates with regard to computer test experience and computer literacy.
2. Ensure it is possible to change the font size in the reading texts.
3. Allow candidates to change the font size on the test screen.
4. Add an automatic word count function to each of the writing output text boxes.
5. Make online testing and refresher materials available for examiners.
6. Allow the marking screen to fit the computer screen the marker is using.
7. Ensure the marking scheme downloads; this seems to be a problem with the speaking test.
8. Encourage all invigilators to try out the test prior to administration.
9. Add detailed solutions to the invigilator manual to facilitate troubleshooting.
10. Fix specific test-related problems:
    a. Avoid pausing the test so that candidates do not lose testing time.
    b. Make it easier to pin-point all candidates in the event of an emergency to pause the exam.
    c. Allow the invigilator or candidate to lower the volume of the headphones.
    d. Reduce the number of keycodes required.
11. Fix invigilator screen problems
    a. Allow the invigilator to filter only by more than one criterion at a time.
    b. Allow the invigilator to filter for tested candidates either by date or test component.
    c. Allow the invigilator to unlock all candidates (for one component) in bulk.
    d. Allow the invigilator to see the time remaining for each test.