

ELPAC Final Validation Report – Executive Summary

ELPAC Final Validation Report – Executive Summary	1
1. Introduction	2
1.1. Limitations of the study caused by the delivery date	2
1.2. Main issues.....	2
2. Findings	3
2.1. Test development process.....	3
2.1.1. Test design and item writing.....	3
2.1.2. Quality control and test analyses.....	4
2.1.3. Dissemination and the ELPAC website (www.elpac.info).....	5
2.1.4. ELPAC Test Guidelines.....	6
2.1.5. Paper 1 - Mode of Delivery.....	7
2.1.6. Paper 1 – statistics and reliability	8
a. Reliability	8
b. Bias analysis.....	9
2.1.7. Paper 2 – Inter-rater reliability	11
2.1.8. Training of administrators, markers and examiners.....	13
2.1.9. Standard setting	13
2.2. Sustainability	17
2.2.1. Maintaining human resources	17
2.2.2. Developing and trialling new test versions.....	17
2.2.3. Training new administrators, markers and examiners	18
2.2.4. Monitoring test administration, interlocutors and assessors	18
2.2.5. Maintaining funding	18
2.2.6. Quality control of the testing system.....	19
2.3. Review of competition	19
3. Conclusions	24
3.1. Test development.....	24
3.2. Test design.....	24
3.3. Dissemination	24
3.4. Guidelines.....	24
3.5. Paper 1 – mode of delivery.....	24
3.6. Paper 1 – reliability and bias analysis.....	24
3.7. Paper 2 – inter-rater reliability during the last major trial.....	24
3.8. Training workshops	25
3.9. Standard setting	25
3.10. Sustainability.....	25
3.11. Competition with other tests.....	25

1. Introduction

This executive summary was prepared by the ELPAC project team (Adrian Enright and Magdalena Vecerova) based on the original report prepared by the Lancaster Language Testing Research Group from the Department of Linguistics and English Language (LAEL). The research Group was co-ordinated by Professor J. Charles Alderson.

The report was received by EUROCONTROL on 13 July 2007. All quotations from the original validation report are in blue italics.

1.1. Limitations of the study caused by the delivery date

It was agreed that the report would be submitted in mid-July 2007 as the ELPAC test release is planned for 31 July 2007. For this reason, Lancaster could not evaluate all the data as they were not available by the time the validation study was conducted (e.g. final trialling and standard setting were conducted in late May and first half of June which was in line with the ELPAC project plan). Data for the validation analysis were submitted in April and May 2007. Wherever possible the ELPAC project team submitted plans on how to address the issues which Lancaster could not evaluate directly (e.g. Lancaster evaluated only trialling versions but the project team informed them about how the final versions would be constructed based on the trialling and standard setting outcomes).

Validation is, nevertheless, not only about analysing what happened but about approaches taken and to be taken. The ELPAC project team intends to continue using best practices in language testing in order to retain the validity of the test. This goes in line with a comment from Lancaster:

EUROCONTROL has commissioned this report in order to validate the first versions of the ELPAC test. It is somewhat early to comment on the test review process since the test has not yet gone live. However, the procedures followed to date are suitable and it is not too early to put in place a timetable for regular review of each test. In particular it will be important to ensure that the test keeps pace with changes in the job description of and demands on air traffic controllers.

1.2. Main issues

This executive summary was prepared by the ELPAC project team in Luxembourg. It highlights the main issues as presented by Lancaster and also addresses areas for improvement. The executive summary contains direct quotations from the report and further comments from the ELPAC project team in Luxembourg. Should you require more information on a specific issue not addressed in this executive summary, please contact Adrian Enright, the ELPAC project leader: adrian.enright@eurocontrol.int

2. Findings

2.1. Test development process

Lancaster commented positively on the approaches employed during the test development. The main issues and recommendations are listed below.

2.1.1. Test design and item writing

The item writers have received item writing training during the course of the development of ELPAC; this has been achieved through training and test development meetings. There are systematic procedures for the review, revision and editing of items. All items are piloted and items that do not function well are either discarded or re-written and re-piloted. Given the organic way in which the test has been developed so far, with co-operation from ELEs and ATCOs in the different EUROCONTROL member states, training and feedback for item writers has emerged from the numerous meetings.

The test paper preparation process was documented by Jayanti Banerjee (JB) during a visit to EUROCONTROL in April 2007 and is also described in the document 'ELPAC item writing.doc'. Paper 1: Recordings are sent to item writers in advance of item writing meetings so that they can text-map the recordings and design as many items as possible. At this stage, item writers are encouraged to exploit the recordings in as many ways as they can see. At the meetings, text-maps are compared and only the items that emerge from consensus in the text-mapping are retained. These items are then discussed and either adopted, re-written or discarded based on the comments that they receive. All test versions contain a balance of Tower / Enroute / Approach scenarios, a balance of routine and non-routine situations, a balance of accents and a balance of gender. In the case of ATC-ATC interactions, the speakers are always one female and one male so that they can be easily distinguished.

Paper 2: The test management team works closely with the item writers to develop scenarios that can be used in the Speaking test. These are then trialled on every category of controller (Tower / Enroute / Approach). The Task 1a test versions are made up of a combination of prepared pilot messages (PPM) and unusual situations (US). All the 'US' messages relate to a single thread that is followed up in Task 1b. For the linked Tasks 1a and 1b candidates are given one of a choice of charts (depending on their specialism). The photographs used in Task 2 are trialled to ensure that they can be discussed by any category of controller.

This process of paper design and setting appears very thorough. We are pleased that the procedures are explicitly documented and are therefore available to new personnel in the test development team.

However, item writing in the future might increasingly be out-sourced to item-writers who are not as closely involved in the ongoing design and administration of the test. Consequently, we would advise the test management team to write clear guidelines for item writers and also consider how they might provide systematic feedback on the items produced.

Comment: As stated above, item writing in the future will be extensively outsourced. Clear Guidelines for both Paper 1 and Paper 2 item writers and recommended qualifications have been produced at the beginning of July 2007. The initial training of future item writers is scheduled for autumn 2007.

2.1.2. Quality control and test analyses

Each item is thoroughly discussed and re-written (if necessary) at item development meetings. Each item then goes through an extensive trialling process. Some items are trialled and improved two or three times before they are deemed ready for use. The pilot sample sizes for Paper 1 (the listening test) have typically been around 150 – 200 test-takers while for Paper 2 (the speaking test) the trials have typically involved about 30 test-takers. The numbers for Paper 2 are slightly lower than would be ideal but still adequate, given the many rounds of trialling that have taken place.

Apart from piloting the tests, the team always gathers self-assessment data from the test-takers as well as feedback about the test and the administration processes from test-takers, administrators and examiners. All the test data is subjected to statistical analyses (item analyses, correlations, and Rasch analyses). The feedback is quantified where possible and all open-ended statements are carefully transcribed and collated. Detailed reports are written on every trial and made available to all team members. Changes to the test are agreed at post-trial team meetings.

We would argue that the test development process has been of high quality and demonstrated a commendable attention to detail. Nevertheless, we are concerned about three aspects of test administration and feedback. The first is that our analyses have shown that different versions of the test are not equivalent in terms of the scores that a test-taker will receive. Therefore, we would strongly caution the ELPAC team against applying fixed cut-scores and pass marks. Secondly, there are no written guidelines that explain how assessors and interlocutors will be accredited and the time-span for accreditation. These plans are essential because their absence will undermine the quality and reputation of the test. Thirdly, we know that each national regulatory body has a procedure by which test-takers might make complaints or seek reassessments. We would recommend that a statement to this effect is published on the ELPAC website before the test becomes operational.

As we have not reviewed the final test versions, it will not be possible for us to state with any certainty whether they will be parallel in content. We can, however, say that the care with which the team has proceeded so far gives us confidence that this will be the case. Nevertheless, we would recommend that a report supporting claims of test equivalence be published on the ELPAC website as soon as possible after the launch of the test.

It should be noted that the final Paper 2 test versions will comprise tasks selected from sets which have been trialled. As we have not reviewed the final test versions, it will not be possible for us to state with any certainty whether they will be parallel in content or whether they have the potential to elicit similar functions from all candidates. However, we can say that the care with which the team has proceeded so far gives us confidence that this will be the case. Nevertheless, we would recommend that a language analysis of interviews taken from live administrations and a range of test-taker abilities be conducted and published on the ELPAC website as soon as possible after the launch of the test.

Comments:

- 1) Due to the reasons explained in the introduction, Lancaster could not review the final test versions. The trialling versions differed in the levels of difficulty which is nothing surprising. The levels of difficulty are confirmed by the trialling outcomes. The final test versions, which were constructed in July 2007 take into consideration results from trialling analyses and standard setting exercises. Only items / tasks performing satisfactorily were included in the final test versions. Every effort was made to ensure that the final test versions are parallel in terms of item / task difficulty, content, length of individual snippets (Paper 1) and specific language functions assigned to test

ELPAC

items (Paper 1). A report on how final versions were constructed will be published on www.elpac.info in late summer 2007.

- 2) EUROCONTROL is organising accreditation workshops for test administrators, Paper 1 markers and Paper 2 examiners in autumn 2007 and will continue with this training in 2008. Successful participants will be recommended for accreditation by their national regulatory bodies. Each test administrator will have to submit a report (using our template) on testing sessions – at least every three months. This should include information on examining too. We will also ask the administrators to provide us with a selection of recordings of all levels awarded (fail, 4 and 5) for evaluation – coded and with the candidate name deleted in order to monitor the quality of examining. All examiners will be invited to an annual refresher course.
- 3) The appeal procedure is different in different countries and we will not interfere with it. We will only recommend the minimum time before the test is taken again by candidates who failed it.

2.1.3. Dissemination and the ELPAC website (www.elpac.info)

The ELPAC website was created in September 2006 and used for dissemination information about the ELPAC project. Sample tests for both test papers were published on www.elpacsample.info which is accessible from the main ELPAC page. Lancaster commented positively on the amount of information published and on how the page is kept up-to date.

The ELPAC website (<http://www.elpac.info>) and the associated Sample Tests website (<http://www.elpacsample.info>) are excellent channels for disseminating information about the test. In addition, the document entitled 'ELPAC presentations.doc' provides a detailed list of the different presentations made to different audiences. It is clear that major efforts have been made to reach a wide range of stakeholders including the heads of Air Traffic Management Training institutions, Training managers from EUROCONTROL member states Pilots, Controllers, English Language Teachers, Language Testers, Regulators and Unions. A range of dissemination methods have been used: from presentations at regular briefing meetings, to conference presentations to magazine articles. This is impressive and we hope that the team will continue to disseminate the test in order to build up its profile and test-taking population.

We have reviewed the ELPAC website (<http://elpac.info/index.php>) to ensure that it is easy to navigate and that it is possible to alter the text size or to enlarge the site on-screen. While this might not be relevant for air traffic controllers who navigate the site, it could be relevant to other stakeholders such as teachers of Aviation English, Air Navigation Service Providers, individuals from National Supervisory Authorities and individuals from ATC Licensing Authorities.

The website can be navigated easily as the main and sub menus are all presented on the left-hand side of the screen. It is also very attractive and uncluttered while still providing much essential and supplementary information. The overall look is professional and efficient. Visitors who prefer to read text in a larger font are able to change the font of the text web-pages (though not of the menu bars). It is also possible to enlarge the site on-screen by using the zoom function. All the pages can be viewed in PDF format and/or printed by clicking on an icon on the webpage. This is a useful facility.

ELPAC

The information for test-takers is clearly marked, as is the section containing all the sample tests. The search facility (on the right-hand side of the web page) works well and allows users to navigate directly to pages of interest to them. The test-taker poll (also on the right-hand side of the web page) is interesting and can help test-takers to feel involved in the test design process, making them feel that their views are contributing to the design and validation of the test.

We looked particularly at the usefulness of the information on the website for test-takers and generally found clear explanations of the test for prospective candidates. Nevertheless, we feel that a number of points could be addressed:

- The opening page for the sample listening test provides a brief outline of the test structure (http://elpacsample.info/index.php?option=com_content&task=view&id=42&Itemid=39). In addition we would have liked a short downloadable document that provides screenshots from the listening test and presents (minimally) the instructions and example items from each section. This will help to orientate test-takers better to the demands of each section.*
- For the Speaking test, a downloadable copy should be provided of the tasksheet used for Tasks 1a and 1b and the image used in Task 2 (the picture discussion) so that test-takers can follow the test better as it progresses.*
- A printed copy should be provided of the instructions that will be issued during the Speaking test. These are standardised and a test-taker's ability to understand them should not be part of the assessment process.*

Comments: The sample task sheets for Paper 2 will be uploaded to the website prior to the test release. Instructions for Paper 2 are explained in the "Information for test takers", as is the format of Paper 1. For the moment this, together with the published sample tests, is considered sufficient by the ELPAC development team.

2.1.4. ELPAC Test Guidelines

Lancaster reviewed Guidelines from all trials. The final guidelines for live testing were produced during the last project meeting (18-22 June 2007) and were thus not available for a review. However, they do not differ substantially from the material reviewed by Lancaster.

Clear guidelines have been put in place for administrators and invigilators. Administrators are given a Powerpoint presentation that shows them (with screenshots from the Admin tool) how to register the test-takers for the test.

The most recent guideline document is 'Guidelines for administering the ELPAC Test_formatted 110507.doc' which was discussed and updated at an ELPAC meeting in April 2007. These guidelines are very detailed and we would expect that administrators and invigilators should have no trouble following the procedures laid out. Indeed, feedback on the guidelines has been excellent (see, in particular, 'ELPAC March07trial_admin rep_final_updated120407') and the test management team is to be complimented on having met the needs of administrators and invigilators so successfully. We note, however, that notes still remain in the text e.g. the contact telephone number for ENOVATE A.S. still needs to be checked (p.9 and 11) as well as the contact e-mail address (p.9). We would recommend that these details be updated before the test goes live.

The markers for the Listening test are given clear instructions on how to complete the marking task including a Powerpoint presentation which (with screenshots from the marking tool)

ELPAC

shows them exactly what they have to do for each section. We think that these guidelines are exemplary.

The guidelines for assessors and interlocutors in Paper 2 are comprehensive and well-organized. The description of these guidelines has covered all the important aspects involved in administering the test such as the roles of assessors and interlocutors, general and specific guidelines for assessors and interlocutors, and the role of the third assessor.

The comprehensiveness of these guidelines has been confirmed by examiners (93.3% of examiners in May 2007 trial have agreed that the guidelines provided enough information on how to use the rating scale and the assessor sheet).

Comment: The Guidelines for live testing are part of the self-study package for accreditation workshops for test administrators, markers and examiners. All phone numbers and contact details were updated and new up-to-date PowerPoint presentations created.

2.1.5. Paper 1 - Mode of Delivery

Paper 1 is web based and this caused some concern among stakeholders. Lancaster was aware of this issue and thus evaluated the mode of delivery thoroughly.

Paper 1 is web-delivered and since this mode of delivery is relatively new (though growing in popularity and use) it was important to review the delivery system, its functionality, security levels, and capacity.

The delivery system has been designed and is supported by ENOVATE A.S. (<http://enovate.no/en/>). The software it uses is called BASE (the BITE Assessment System); a system that is also used for the Norwegian national competency tests where more than 250 000 pupils have been tested in their English reading skills since 2004. This widespread use of the software by a national testing system gives us considerable confidence in the robustness of the tool.

In addition, our review of the ENOVATE A.S. website revealed that BASE is a flexible tool that provides item and test authoring tools for a variety of item types including drag and drop, true/false and multiple choice. This fits the needs of the ELPAC test very well. Apart from web-authoring, BASE also provides a range of administration tools including both automatic and human marking as well as the facility to generate test statistics.

The bandwidth usage for the ELPAC test is great because of the large sound files that are used in the test. However, the servers that support the ELPAC test can handle at least 1000 concurrent users. The ELPAC server is also clusterable which means that if traffic on the server pushes it to its capacity, ENOVATE A.S. will be able to add one or more hardware servers that will divide the traffic and workload in order to maintain the responsiveness of the system.

The Paper 1 interface is uncluttered and easy to process. It is possible to view entire items on the screen and this facilitates the test-taking process.

In summary, we are satisfied with the proposed delivery of Paper 1 and are confident that ENOVATE A.S. will be able to support the system. We are particularly pleased to note that dedicated support is provided for the ELPAC test and that all test administrators will be provided with a telephone number and e-mail address so that they can contact ENOVATE A.S. directly should they experience technical difficulties. However, we would like to suggest that traffic on the server is monitored carefully and that new hardware servers are added well before the maximum operating capacity of the current server is reached.

Comment: ENOVATE will monitor the traffic on the server on a regular basis, as proposed by Lancaster. They will also be responsible for the system maintenance to ensure that the good quality of the tool is maintained.

2.1.6. Paper 1 – statistics and reliability

Lancaster reviewed data from three trials (October 2006, March 2007 and May 2007). The previous trials (November 2005 and May 2006) were reviewed in Lancaster’s interim report (June 2006) and recommendations applied to the later trials.

a. Reliability

We have checked the test reliability of each version, inspected the versions for item-level bias, and also reviewed the difficulty of the test parts. When doing so we bore in mind the fact that the later test versions had evolved from earlier versions that had been trialled at an earlier stage in the trialling programme.

We re-analysed test reliability using Rasch analysis. Reliability on both test forms G and H is excellent (Cronbach alphas above .9). A few items did not perform well and were identified for editing. All the evidence (in terms of the reliability of subsequent trial versions) suggests that this took place. Test reliability for versions L and M was also very good (Cronbach alphas above .9). Additionally, reducing the number of items from 64 to 60 appears to have successfully saved test time without compromising reliability. Reliability for versions P and Q was slightly lower (Cronbach alphas .89 and .88 respectively), but this is likely to be partly due to the smaller number of candidates (n=52). Our analysis shows that some items were not performing satisfactorily but the test development team have previously demonstrated that they are highly skilled at locating and editing such items, and we are confident that these will be dealt with appropriately.

Table 2 shows the separation statistics and Rasch reliability estimations of various elements in the test for versions L and M. Separation refers to the number of statistically different performance strata that the test can identify in the sample (i.e. a separation value of 2 indicates there would be two levels of performance consistently identified by the test). Reliability refers to the reliability in difference of ability of candidates or difficulty of items (where values closer to 1 are better except in the case where separation is zero).

TEST	Examinee		Item		Controller Type	
	Separation	Reliability	Separation	Reliability	Separation	Reliability
L	3.16	.91	6.9	.98	.00	.00
M	3.12	.91	7.0	.98	.00	.00

Table 2: Rasch reliability and separation for test versions L and M

Table 2 reveals that both tests are reliably able to distinguish examinees into approximately three levels, which we would take to correspond to ICAO levels 3, 4 and 5. It further reveals that the 60 items may be separated into seven distinct levels of difficulty (the item separation value). This suggests that if the items were grouped together in terms of difficulty, we would see that there are approximately 7 gradations of difficulty. This is a good result because it means that there will always be several items which are well-targeted at the proficiency level of any particular test taker.

We should point out, however, that our analysis of the difficulty of test parts (see 6.5.3, below) indicates that these 7 gradations do not correspond directly to the parts of the test

ELPAC

(which are intended to be ranked in order of difficulty). This means that although the test contains items with multiple levels of difficulty, the difficulty values of individual items are not necessarily related to the part of the test in which they appear; easy items might appear at the end of the test (in Part 6) and difficult items might appear nearer the beginning of the test (such as in Part 3). See 6.5.3 for more discussion of this issue.

The separation and reliability values have not been given for test versions P and Q because of the much smaller sample size in the May 2007 trialling. However, Rasch analysis was conducted on this data and the reliability for these test versions appears to be very comparable with that reported for tests L and M. There are several items that performed badly in the tests P and Q, but the file named item_analysis_JKLMNPQ demonstrates that these have already been identified. As there are no more test trials planned for modifying these items, it is recommended these are not included in the final test versions.

We are therefore satisfied that there is a sufficiently large pool of very reliable items available for the test developers to choose from when constructing the final test versions. Although we are not able to guarantee the exact reliability of the final test versions, the high quality of the test versions we have seen gives us every confidence that the final test versions will be as reliable and will display the same patterns of item and controller separation as the ones we have reviewed. Nevertheless, we would recommend that the reliability statistics for the live versions should be published on the ELPAC website as soon as possible after the release of the test.

Comment: Items performing badly were not included in the final test versions. Future test trials will be organised during live testing (candidates will be encouraged to take a few more items at the end of the test). The procedure for future test trialling is described in the Guidelines for Administering the ELPAC test and candidates are informed about it in the updated “Information for test takers” which is available on www.elpac.info.

b. Bias analysis

In the analysis for the October 2006 data, items on the six test parts were grouped together and ANOVA was performed to test whether different ATCs performed significantly differently to each other on each of the six different test parts. There were found to be no significant differences, i.e. different ATC types did not on average score differently to other ATC types on any of the six test parts. The limitation of this analysis is that it combines up to 15 items together within each test part, and so if any single item was disadvantaging certain ATCs this would go undetected.

We conducted the very same ANOVA analysis on the March 2007 data in order to establish if the same was true. We found several significant differences between ENR and TWR/APP in several test parts. On test version L (df=194) the significant difference was on test part 1 ($p < .05$). On test version M (df=193) the significant differences were on test parts 1 and 6 ($p < .05$).

However, in all cases the size of the effect was not large; the most substantial difference was a mean score of 10.6 for ENR as opposed to 8.3 for TWR/APP on test part 6 of version M. Further, we contend that this analysis may be flawed because countries with traditionally high proficiency in English contributed a proportionally high number of ENR, while countries with traditionally lower proficiency in English contributed a proportionally higher number of TWR/APP. Therefore the test scores on the test parts might confound performance as related to nationality with performance as related to ATC function.

It has already been determined that the tests overall are unlikely to be biased against different controller types. Based on ANOVA, it would also appear that the test parts did not

ELPAC

substantially advantage or disadvantage any controller types. However, this does not guarantee that individual items within the tests did not favour one type of controller over another. Therefore, in order to understand whether individual items within the test were biased, we have conducted our own bias analysis of versions L and M. We would have liked to perform the same analyses for versions P and Q but, due to a smaller sample of test-takers and uneven spread of controller-types in the May 2007 trials, it was considered that this data would not be appropriate for a bias analysis.

This investigation was conducted by means of an item-level bias analysis in a Rasch model. The analysis considered only the performances from controllers who performed in either one or two roles. Controllers who performed all three roles, i.e. TWR/ENR/APP were removed from the data set. In order to estimate bias, the Rasch model estimates the level of ability for each group of ATCs, and difficulty values for all items for each group of ATCs.

Tables 3 and 4 show the instances when the difference in difficulty values for items was significant between ATC types. It has been noted that there were different numbers of individuals in each ATC group. For example, there were only 5 TWR/ENR. However, the Rasch estimations are robust to sample size (in fact, it is harder for the model to find significant differences when the sample size is small). Further, when the data were checked, it was found that these 5 TWR/ENR did not score especially high or low on the test overall, suggesting they were of average ability.

When we consider the effects of item bias we look for z-score values above +2 or below -2, which are considered significant. We may also look for large bias measures (above 1.0). The analysis revealed that certain individual interactions at the item level were significant; these are shown in Tables 3 and 4, below.

Tables 3 and 4 show that there are approximately 10 significant bias interactions which are potentially problematic in each test. Of these, none of the interactions have especially large z-scores. The items that are shaded have large bias measures (above 1.0), and we have flagged these as requiring follow up to look for the cause of this bias.

Bias Model			TypeCon	Items	measr	
Measure	S.E.	Z-Score				
-.89	.44	-2.02	TWR	L_P2_6	.57	easier for TWR
-.84	.31	-2.67	ENR	L_P4_2	.24	easier for ENR
-.79	.38	-2.10	TWR/APP	L_P4_3	-.79	easier for TWR/APP
.56	.27	2.07	ENR	L_P3_2b	-.16	harder for ENR
.69	.34	2.04	TWR/APP	L_P5_1e	.92	harder for TWR/APP
.74	.31	2.41	TWR/APP	L_P4_2	.24	harder for TWR/APP
.77	.31	2.50	TWR/APP	L_P6_2e	.22	harder for TWR/APP
.97	.41	2.37	TWR	L_P3_3b	-.57	harder for TWR
1.49	.68	2.20	APP/ENR	L_P6_2a	-.04	harder for APP/ENR
1.51	.76	2.00	APP/ENR	L_P3_3c	.94	harder for APP/ENR
1.62	.73	2.20	APP	L_P1_3a	-2.13	harder for APP

Table 3: Item-level bias for Test L

Bias Model			TypeCon	Items	measr	
Measure	S.E.	Z-Score				
-1.82	.80	-2.28	APP	M_P6_2c	5.37	easier for APP
-.97	.33	-2.97	TWR/APP	M_P2_4	2.50	easier for TWR/APP
-.66	.31	-2.14	TWR/APP	M_P5_3d	1.59	easier for TWR/APP
.71	.30	2.34	TWR/APP	M_P1_5b	-.50	harder for TWR/APP
1.45	.61	2.36	APP	M_P6_3e	-.35	harder for APP
1.80	.84	2.13	APP/ENR	M_P5_3a	1.51	harder for APP/ENR

	1.95	.94	2.07		TWR/ENR	M_P1_5a	-.35		harder for TWR/ENR
	2.09	.94	2.23		TWR/ENR	M_P1_5b	-.50		harder for TWR/ENR
	2.14	1.06	2.01		APP	M_P5_3e	2.24		harder for APP
	2.68	1.15	2.33		TWR/ENR	M_P6_1a	-3.00		harder for TWR/ENR

Table 4: Item-level bias for Test M

In presenting this data we acknowledge that items are not independent and are often confounded with other items because they are grouped according to passage. We also acknowledge that there may be factors other than controller-type that could contribute to this bias, for example because of the proportion of controller-types that came from countries with traditionally stronger or weaker proficiency in English language. Additionally, we note that the tables are difficult to interpret because any item that disadvantages two ATC types, e.g. TWR/ENR, should also disadvantage both ENR and TWR separately, but this does not appear to be the case. Nevertheless, despite the limitations of this analysis, we feel any possibility of bias requires follow-up. Therefore, we would strongly recommend that before the highlighted items are included in the final test versions, they be reviewed again by different ATCs who are also item-writers to check why they could be easier or harder for certain controller types.

In summary, test reliability is good and the overall chi-square shows no significant effect overall. Consequently, we would support the statement that the overall test neither favours nor discriminates against any controller type.

Comment: As explained above, it seems to be highly unlikely that the test would disadvantage a certain group of controllers. The construction of the final versions of Paper 1 considered the results of bias analysis on item level and problematic items were excluded.

2.1.7. Paper 2 – Inter-rater reliability

No statistical analysis was carried out on the Paper 2 data of October 2006 trialling because, it was considered that examiners required more training: interlocutors did not always follow guidelines and assessors did not understand the concept of progressive assessment. In addition, there was not thought to be enough guidance on how to assess phraseology. Review of ratings was therefore deferred until the March 2007 trials.

Despite over 170 candidates sitting the test in March 2007, complete data is not available for a full reliability analysis. ELE and ATCO assessors marked different criteria when observing the same candidate, and since raters also worked in isolated pairs in different countries it was impossible to achieve a data set connected by common candidates or mixed rater pairings. Nevertheless, Table 6 reveals an inter-rater reliability coefficient based on final scores awarded by ELEs and ATCOs. All coefficients are acceptably high (above .87), which means that, providing assessors really did come to final levels independently before agreeing on a level, they are assessing candidates and interpreting rating scales in similar ways.

		ATCO:	ELE: final level	agreed level
Spearman's rho	ATCO: final level	Correlation Coefficient	.873**	.906**
		Sig. (1-tailed)	.000	.000
		N	174	172
ELE: final level		Correlation Coefficient	.873**	.987**
		Sig. (1-tailed)	.000	.000
		N	174	171
Agreed level		Correlation Coefficient	.906**	.987**
		Sig. (1-tailed)	.000	.000
		N	172	171

** . Correlation is significant at the 0.01 level (1-tailed).

Table 6: Correlation coefficients for inter-rater agreement

A further analysis was conducted by calculating the number of occasions on which examiners agreed exactly, agreed to within one level, and agreed to within two levels. It was found that when ELEs and ATCOs worked independently they agreed exactly on which level to assign in 90% of all tests, and disagreed by just one level in the remaining 10% of tests. It appears that in the 10% of tests where there was disagreement, assessors were able to arrive at an agreed level after discussion.

When the May 2007 trial data was analysed, it was noted that there appears to be missing data, where assessors are arriving at final scores without marking all the subscores on the tasks. It is recommended that this be followed up to see why. Possible reasons may be lack of time, lack of a suitable sample of language, or (as appears in one or two cases) that assessors were hesitant to give a mark of level 3 or below. It was further noted that this data set contains a higher proportion of failing or borderline candidates, and it is useful to see to what extent assessors agree how to score these candidates. The correlation coefficient for the two sets of final scores awarded by the ELEs and ATCOs was .78 (from 55 candidates). However, this relatively low coefficient is likely because the scores are bunched around levels 3, 4, and 5, and candidates do not exhibit a wider range of proficiency. This time, in 20% of cases (i.e. 11 out of 55 candidates) the ELEs and ATCOs disagreed by one point when arriving at their final scores independently. In 12% of cases (i.e. 6 out of 55 candidates), this disagreement occurred over the pass-fail threshold, i.e. one assessor awarded a 3 and another assessor awarded a 4. Nevertheless, in every case, assessors seem to have reached a joint agreement on whether to pass or fail the candidate. In summary, though the May 2007 exhibits slightly lower inter-rater reliability than the March 2007 data, the overall picture is that assessors are generally in agreement.

While these results are encouraging, more data is required to establish whether candidates are being assessed under equal conditions in different countries and rated similarly by different examiners. In particular, it is necessary to establish that assessor pairs from one country would grade a candidate at the same level as an assessor pair from any other country. Research shows (Reed and Halleck, 1996; Brown, 2003) that assessors create an impression of a candidate's proficiency during an interview and react accordingly in their manner of questioning, which in turn influences how their partner views the candidate. In light of this, we would ideally like to see evidence that rater pairs from different countries are able to assess even borderline candidates in a uniform way, not only from videos but from live tests conducted by different assessors. This suggests that the test designers should ideally demonstrate more than the fact that raters can all rate the same video in the same way; they also need to demonstrate that examiners would interview and then rate the same candidate in the same way.

Clearly such an investigation is not feasible. Instead, we would recommend two actions:

1. The consistent and strict implementation of examiner training and accreditation procedures. The procedures that have been devised have been reviewed in 7.6

ELPAC

(above) and should be monitored carefully in order to ensure that standards are maintained.

- 2. A small study of rater behaviour using trained raters and benchmarked performances. The use of benchmarked performances would enable raters to demonstrate that they would all assign the agreed-upon score to a candidate when they watch a candidate's video-recorded performance.*

Comment: As has been already stated, for live testing only successful participants to the ELPAC accreditation workshops will be accredited for the role of Paper 2 examiners. More information on the proposed training can be found in the following chapter (chapter 7.6 in the original report from Lancaster).

2.1.8. Training of administrators, markers and examiners

One-week accreditation workshops have been organized for test administrators, examiners and markers. The requirements for registration at these workshops as well as the timetables have been well publicized by putting these on the ELPAC website for the past several months.

We were impressed with the consistency and comprehensiveness of record-keeping in relation to the training workshops. The guidelines for trainees are very comprehensive and well-organized. They clearly state the requirements of participation and the responsibilities of trainees before, during and after the workshops.

The plans for the actual content of the full one-week training workshops are sound and comprehensive. The amount of time allocated to scoring actual samples of speaking performances is very reasonable.

We also found the provision of a self-study package for trainees a definite positive feature of the training procedures. The self-study packages are very comprehensive. We were particularly pleased with the inclusion of benchmarked performances in the self-study package. Moreover, requiring trainees to submit their self-study reports at least one week before the workshop is also a good feature in the procedures to ensure that all trainees are prepared and also to give ELPAC members sufficient time to process this data.

We were also impressed that only examiners who pass the training week will be accredited.

Comment: The self-study packages were finalised in mid-July and on 13 July 2007 posted to the contact persons in all countries which applied for the autumn 2007 workshops. This would give all participants enough time to complete the self-study preparation and thus the deadline for submitting the materials was set at two weeks before each workshop. Further information can be found on the CD ROM with self-study packages.

2.1.9. Standard setting

In response to recommendations from the Lancaster Interim Report in June 2006, four ELPAC standard setting sessions were organised in the first half of 2007. The sessions were prepared and conducted by Dr. Rita Green, consultant to the ELPAC project. Lancaster received data from the first three sessions as the fourth session was organised only in June (18-23/06) when Lancaster had already been working on the validation report.

ELPAC

According to Kaftandjieva (2004: 20), the selection and training of judges is part of procedural evidence of validity in standard setting. As Alderson (2005: 66) points out, expertise is an essential characteristic for judges. For ELPAC this implies a panel with expertise in both ELT and Air Traffic Control in order to judge items and performance in relation to the performance standards, i.e. the ICAO scale.

The selection of judges appears reasonable. There is a mixture of background with English Language Experts (ELEs) and Air Traffic Controllers (ATCOs) participating. This is positive, as it reflects the mixture of ELEs and ATCOs in the ELPAC team and therefore addresses the need for expertise in both areas.

The relevant standard setting literature recommends that judges should have good knowledge of the test taking population (Hambleton, 2001; Raymond and Reid, 2001). The ELPAC judges have participated in test trialling and have taken Paper 1 and watched samples of Paper 2, which contribute to an acceptable understanding of the test-taking population, given that the test has not been administered yet and that the only population so far has been the test takers in the trials.

The number of judges was small (Meetings 1 and 2 with 6 judges and Meeting 3 with 7 judges) and this should have been acknowledged in the report, along with an explanation as to why more people could not have been invited. We therefore recommend that future standard setting meetings involve more judges, minimally 10 but more is better.

We note that a different group of judges was chosen for Meeting 3 and understand that this is because the same judges were not available for this meeting as had been available for the previous two meetings. This is a pity because it would have made sense to use the same panel, so as not to spend a lot of time on familiarisation. We recommend that this constraint is more clearly explained in the reports.

We understand that one of the purposes of the standard-setting meetings was to set levels for individual items. This is why it would have been preferable to use the same panel of judges. However, if there was overlap in the selection of items for the standard setting meeting (where some of the items had been standard set by the previous group of judges) then the data gathered in Meeting 3 could provide evidence for the equivalence of the test versions as well as inter-judge consistency. We would like to point out here that we are speculating because there is little information in the reports on the standard setting meetings to explain the composition of the test papers, their relationship to other versions of the test and the possible advantages of using a different panel of judges.

An essential component of any standard setting procedure is the judges' familiarity with the performance standards (Council of Europe, 2003: Ch. 3). RG's reports document procedures to familiarise judges with the ICAO scale (Meeting 1, p. 8; Meeting 2, p. 6; Meeting 3, p. 7), as this is the standard against which classification decisions will be made. In order to provide accurate classification of examinees into the ICAO levels, judges should be able to distinguish between the levels and have a good understanding of the performance expected at each level. The familiarisation task in the ELPAC meetings involved giving the ELPAC scale descriptors to the judges without an indication of their level and asking them to assign a level to each descriptor.

The reports acknowledge that the familiarisation results are not very encouraging. In Meeting 1, 39 descriptors were used and incorrect level placements ranged from 7 to 14 descriptors. In Meeting 3, incorrect placements varied from 1 to 16 descriptors. However, there are a number of positive issues to consider. Better results were obtained when the task was repeated on the following day, which suggests that the discussion following the first round of judgements helped the judges understand the ICAO scale better. For example in Meeting 1 (p.12) and Meeting 3 (p. 9) the range of misplaced descriptors was reduced (0-10 and 1-10 descriptors respectively). Also, the Meeting 2 results showed higher levels of familiarity with the ICAO

ELPAC

scale (the task was carried out once in Meeting 2). There are two possible explanations for the improved familiarity. First, 5 out of 6 participants had taken part in Meeting 1 and had probably achieved better understanding of the scale. Second, 18 instead of 39 descriptors were used in Meeting 2. As a result, the descriptors were longer and probably it was easier for the judges to assign correct levels. The report correctly points out that breaking down the descriptors into their constituent units might make the task more difficult (Meeting 1, p. 13); however, the small unit option is to be preferred because judges should be familiar with all parts of a larger descriptor and we therefore recommend that the descriptors used in Meetings 1 and 3 be preferred in future standard setting meetings.

In their paper at the Language Testing Research Colloquium in Barcelona, 8 – 10 June, 2007, Moder and Halleck (2007) criticised the content of the ICAO scales (the performance standard). The reports of the three Meetings also raise a number of issues regarding this scale. Listening, as documented in the reports of Meetings 1 and 3, is not very well represented in the ICAO scale. Moreover, the judges pointed out that they had problems with the terminology used in the scales, for example lack of definition of words such as 'complex', 'short', 'basic', etc (Meeting 1, p. 12). They attributed their failure to identify the correct level of the descriptors during the familiarisation tasks to wording issues. In fact, research (e.g. Alderson et al., 2006; Fulcher, 1993, 2003; North, 2000) has shown that interpreting words such as 'complex' and 'basic' is not easy for scale users. This appears also to be the case in the standard setting meetings. The Lancaster Interim Report in July 2006 (p. 22) comments positively on the attempt to quantify such terms by using percentages in order to help Paper 2 raters (who use a version of the ICAO scale to rate performance) and we continue to commend this action. However, we would recommend that the judges' evaluations of the ICAO rating scale be compiled and sent to the ICAO with recommendations for improvements to the rating scale. Additionally, we would argue that raters' difficulties with the rating scale should be considered in any investigations of inter-rater reliability in Paper 2.

Overall, the panel was just barely acceptable for making judgements and for setting cut-off scores for ELPAC. Even though familiarity was reported to be low in the first round of judgements, group discussion probably enhanced the judges' understanding of the performance standards, despite the terminological issues in the ICAO scale.

Although not explicitly mentioned in the reports, the ELPAC meetings have adopted a Yes/No Angoff variant (Council of Europe, 2003: 91), which has also been used in the DIALANG project (Alderson, 2005: Ch. 6; Kaftandjieva, Verhelst, and Takala, 1999) and has therefore proved to work for item-based tests. The method appeared to work for performance tests as well, as documented in the Trinity CEFR project (Papageorgiou, 2007). We would encourage adding a short rationale about method selection in future standard setting meetings.

The ELPAC reports show that the coordinator followed typical steps expected in a standard setting study: selection of participants with knowledge of the domain, familiarisation with performance standards, estimates of item difficulty and judgements about cut-scores for the three performance standards (ICAO levels 3, 4 and 5). Multiple rounds of judgements were carried out for the setting of cut-scores, with group discussion taking place between the rounds, a process met frequently in standard setting studies (see Hambleton and Plake, 1995). There was comparison of judgements with empirical data for Paper 1 and ratings of performance for Paper 2, which is also a common procedure in standard setting to help panellists achieve a better understanding of the items. No problems are reported with the rating forms used during the meetings. The reports do not mention problems that judges had with their task, which suggests that the task was clear and that the PowerPoint presentation at the beginning of the meeting clarified what the panellists had to do. This should increase our confidence in the suggested cut-off scores.

However, the standard setting literature suggests that cut-scores differ across standard setting methods. In order to increase confidence in the suggested cut-score we would encourage combination of more than one method in future standard setting meetings. Zieky and Perie (2006) is a useful resource of how test-centred and examinee-centred methods can

ELPAC

be applied. We would also recommend that more emphasis is given to empirical data when discussing item difficulty with the judges during the meetings and that the cut-scores are estimated based on the empirical data rather than solely on the judges' perceived level of item difficulty.

The following points are suggested for future standard setting meetings, even though we would acknowledge that some are difficult to achieve due to practicality and logistics:

- 1. Recruit a larger number of judges.*
- 2. Then use the same judges or justify the use of different groups of judges.*
- 3. Document the rationale behind the selected standard setting method.*
- 4. Compare item statistics to judges' estimates. When the judges cannot predict difficulty accurately, empirical data should be preferred. Intra-judge consistency in this case can be checked by how well judges' estimates of item difficulty correlate with the empirical data (Kaftandjieva, 2004: 25).*
- 5. Investigate the usefulness of empirical data in helping judges make decisions.*
- 6. Use a combination of different standard setting methods. Compare proposed cut-scores derived by different methods.*
- 7. Document cut-score validity evidence, in particular generalisability evidence (Kaftandjieva, 2004: 21) such as decision accuracy.*
- 8. Obtain written feedback from the judges about the clarity of the judgement task, the reasons for low confidence (if this is the case), the organisation of the meeting, etc. Hambleton (2001) can be consulted on the design of a feedback questionnaire.*

The ELPAC standard setting reports provide satisfactory evidence of procedural validity. However, we would recommend more exploration of generalisability evidence. We would also encourage further exploration of how recommendations from the standard setting meetings will be interpreted in terms of reported test scores. If a candidate at Level 4 does not answer the number of Level 3 and Level 4 items to be qualified at Level 4, but answers some Level 5 items correctly, should this count in favour of the candidate? And how will scores be reported for such a candidate?

Finally, when data becomes available from live test administrations, it would be appropriate to organise a larger standard setting study, building on the experience from the meetings conducted so far. The study should provide sufficient generalisability evidence for the cut-score to be adopted in future. If it can be demonstrated that subsequent live test forms are equivalent, then the recommended cut-score from this study can be used in the future with only slight adjustments, without necessarily repeating a large-scale standard setting meeting.

Comments:

1. The ELPAC development team is aware of the limitations of the standard setting sessions. Therefore, when constructing the final test versions, the statistical (empirical) findings were compared with the results of the standard setting exercises and whenever there were huge discrepancies, the items were reviewed again in light of all comments received (from test takers, administrators, examiners, item writers and item reviewers). Problematic items were not included in the final test versions and will be modified and re-trialled in autumn 2007.
2. The ELPAC development team will use the above listed recommendations for all future standard setting sessions.

2.2. Sustainability

2.2.1. Maintaining human resources

ELPAC developers have done an excellent job in identifying (and exemplifying) the precise human resources required for running the test in a typical situation, together with the necessary qualifications for staff members. This has covered the need for test administrators, examiners and markers as well as back-up staff. In addition to providing this information, we recommend reporting the approximate costs of the provision of these human resources.

In light of the reported difficulties the current ELPAC development team have faced in maintaining a balance between their contribution to ELPAC and their other responsibilities, we recommend recruiting and training new item writers (ideally full-time) for the development of new test versions as well as recruiting/appointing a full-time project manager and a language testing expert.

Nevertheless, as stated in the Lancaster Interim Validation Report, there is a risk of compromising standards if new members come on board unaware of the rationale for decisions. Thus far, the documentation of the project and dissemination of information has been impressive. We recommend sustaining these record-keeping and dissemination procedures to ensure that every relevant decision is recorded and passed over to new test development members.

Comment: Thorough record-keeping will be continued to avoid any misunderstandings from new team members.

2.2.2. Developing and trialling new test versions

Up to this point, item writers have been selected from team members in the contributing Member States. After the test is released at the end of July 2007, new items will be developed by independent item writers who are recruited on a part-time basis and paid for each item accepted for trialling. We are not sure what measures have been put in place to train item writers or to give them feedback on their items. We would recommend that this be considered in order to ensure that a pool of good item writers is built up over a period of time as this will alleviate a lot of the pressure on the core team at EUROCONTROL.

Given the logistics of the situation, as explained by the ELPAC development team, it sounds unfeasible to trial new test versions as complete sets. We find the alternative plan of trialling new Paper 1 items during live testing adequate as long as careful measures are taken to ensure that the items are tested on sufficient numbers of candidates and the algorithm to work this out is thoroughly tested beforehand.

As for the Paper 2 trialling plans, measures should be taken to ensure that examiners and interlocutors involved in trialling are trained and fully familiar with carrying out Paper 2.

Once the trialling system is in place it will also be crucial to have staff with sufficient expertise to carry out statistical analyses of items and to compare item difficulty levels with those in the live test.

Comment: Some suitable and experienced item writers have already been identified and details of future co-operation should be discussed and agreed upon during summer 2007. Detailed guidelines for new Paper 1 and Paper 2 item writers were produced at the beginning of July 2007. The ELPAC language testing expert will be in charge of providing training to the item writers, collecting

ELPAC

and collating items / tasks and also providing feedback. Item writers will be paid only for items / tasks which will be accepted for trialling.

2.2.3. Training new administrators, markers and examiners

We are impressed with the amount of work which has gone into planning accreditation workshops as well as preparing the training packages and other relevant documentation (e.g. confidentiality letters, administration procedures, etc).

We recommend collecting feedback from participants in these workshops in order to help improve the training procedures.

Comment: Feedback from participants to all workshops organised at IANS Luxembourg is collected. The ELPAC project team will use the feedback from ELPAC accreditation workshops to review and improve the training on a regular basis.

2.2.4. Monitoring test administration, interlocutors and assessors

The current proposal of requiring test administrators in different Member States to write a comprehensive report on each test administration is sensible given the logistical constraints. However, it is not clear who will be in charge of analyzing this potentially substantial amount of data.

We recommend designing a standard template for collecting this information. This will ensure the comprehensiveness and comparability of the data.

Sample recordings from Paper 2 examiners are collected periodically and analysed thoroughly. Proper feedback is then sent to Member States. This is an adequate procedure for monitoring interlocutors and assessors.

Comment: A template for the regular administrator report was designed at the beginning of July 2007. Sample recordings (coded, without candidate names) will be collected and analysed and feedback will be provided to all ELPAC examiners. ELPAC examiners will also be encouraged to attend an annual refresher course.

2.2.5. Maintaining funding

It is evident that considerable effort has been made by ELPAC developers to stress the need for adequate and on-going resources for sustaining ELPAC. As the budget discussion document shows, the test development team have assessed the various possible scenarios for running the system along with the financial implications for each scenario. This is a vital procedure as it ensures that the financial implications for each option are considered before any action is taken.

As already stressed by the ELPAC development team, it is crucial that adequate funding plans are in place as soon as possible to avoid compromising test validity. Failure to maintain proper resources for sustaining the test after July 2007, will not only compromise test validity and reliability but will also mean that all the resources which have been put into developing the test would be wasted.

ELPAC

Therefore, we recommend that different Member States take urgent steps towards planning ELPAC sustainability in each state. These plans should cover all the aforementioned aspects.

Comment: At the HRT meeting 27 (March 2007) the EUROCONTROL Member States confirmed their support for the sustainability of the ELPAC project for the next 5 years and agreed a regular review of progress. Detailed information on the resources necessary for ELPAC test implementation was also presented at the HRT 27 and is available on www.elpac.info.

2.2.6. Quality control of the testing system

In addition to maintaining funding for ELPAC, we would recommend that the team put in place a system for monitoring and maintaining the quality of the ELPAC testing system. External (and impartial) experts should be contracted to carry out specific tasks that will ensure on-going quality control. These tasks should include:

- 1. Standard setting for all new versions of the tests (including the setting of cut-scores).*
- 2. Scrutinising annual reports and making recommendations for test renewal.*
- 3. Monitoring the performance of the test, as well as test administrators, Paper 1 markers and Paper 2 assessors/interlocutors.*
- 4. Advising on, and where necessary assisting with the conduct of concurrent and predictive validity studies*

Comment: All the above listed recommendations will be implemented.

2.3. Review of competition

We have reviewed a number of tests that appear to be competitors for ELPAC. We focus our review on the following:

- who the test-takers are*
- the format of the tests*
- claims and evidence about validity and reliability*
- limitations that we have found in the competition*

We reviewed the following tests:

- RELTA (<http://www.relta.org>): Provided by RMIT English Worldwide, 'a global English language training and testing business, within RMIT University Australia'*
- TEA (<http://www.maycoll.co.uk/aviation-english/index.htm>): Test of English for Aviation, provided by Mayflower College.*
- TELLCAP (<http://www.tellcap.ru>): Test of English Language Level for Controllers and Pilots. The website was only available in Russian (which we have translated). This is probably administered by the Automated Air Traffic Controlling Center (AATCC) in Moscow.*
- EPTA (<http://air.gtelp.co.kr/eng/index.asp> and http://gtelp.co.kr/e_gtelp/gst/web.pdf): English Proficiency Test for Aviation, administered by G-TELP KOREA.*
- TELPA (<http://www.telpa.org> and <http://www.aviationenglish.org>): Test of English Language Proficiency for Aviation, provided by IAES (International Aviation English Service), a global Aviation English Training company.*
- VAET (http://harcourtassessment.com/NR/rdonlyres/121BCB12-7231-41C6-B051-D3BF98624D7A/0/WhitePaper_CaseForAutomation.pdf and <http://www.icao.int/trainair/meetings/gtc10/PIVP5.pdf>): Versant Aviation English Test, provided by Harcourt Assessment*

ELPAC

- IATA-Berlitz Proficiency test (http://www.icao.int/icao/en/ro/nacc/meetings/2005/NACC_DCA2/nacc02ip11.pdf): Test developed by IATA and Berlitz, a language services provider.

Test takers

Table 8 summarises information on test takers of the other tests. In the third column we have added some notes. Even though all tests are similar to each other and to ELPAC in that they aim to test two groups of test takers or more, they are dissimilar in that some of them have adopted more than one form depending on the test takers' background, whereas others have not.

Test	Test-takers	Notes
1. RELTA	Licensed pilots and air traffic controllers or those who have completed their transition training (pilots) or traineeship (air traffic controllers)	There are two forms of the RELTA: one for pilots and one for air traffic controllers.
2. TEA	Pilots and air traffic controllers	No mention of separate forms
3. TELLCAP	Pilots, cabin crews and air traffic controllers	No mention of separate forms Started as a test for pilots and was later revised to include cabin crew and ATCs
4. EPTA	Pilots, air traffic controllers and aeronautical operators	No mention of separate forms
5. TELPA	Pilots and air traffic controllers	There are two forms of the TELPA: one for pilots and one for air traffic controllers
6. VAET	Pilots and air traffic controllers	Various test versions are produced by using a random combination of items
7. IATA	Pilots and air traffic controllers	Limited information available

Table 8: Test takers for other aviation tests

Format

Overall, we observed that the tests have a common emphasis on the skills of speaking and listening. However, the format of the tests varies considerably, including how components are weighted and which types of tasks are used. We were unable to locate sample papers and videos of speaking performances.

Test validity and reliability: claims and evidence

In order to explore claims about the validity and reliability of the competitor tests we conducted a thorough review of the information available on the online resources for each test.

The detailed claims for reliability and validity are presented in Appendix II. We would like to note that these aviation tests are extremely high stakes, not just for the test-takers but for every potential airline passenger. They should, therefore, be held to the highest standards. Consequently, we were extremely surprised and concerned to find a complete lack of evidence for validity and reliability in any of the competitor tests. This is illustrated in Appendix II by the empty cells in column 4. In fact, the sources we reviewed contained claims but no evidence to support validity and reliability arguments. The exception is **VAET** for which we found evidence such as correlations between speech recognisers and human raters and a careful treatment of issues of validity and reliability. Additionally, **RELTA** seems to address issues of test development, but again, no evidence is provided, such as a reliability index.

The remainder of the tests give considerable cause for concern:

1. **TEA** displays a notable lack of even simple references to validity and reliability, apart from the claim 'research indicates that TEA is a valid, standardised, accurate and

ELPAC

secure testing solution'. The lack of a listening component clearly suggests construct under-representation. Listening is tested only by explaining to the examiner what the recording is about. There is only one rater, which raises the issue of reliability.

2. The **TELLCAP** website was the only one among those we reviewed with audio samples of the interview and this is a positive feature. However, the website makes extensive reference to training but no evidence is provided, such as agreement indices when rating performances during training. Performances in live administrations are rated by only one rater and this calls into question the reliability of the rating process. We were also surprised to find that the only language option of the website is in Russian which makes it difficult for stakeholders to access the information provided.
3. **EPTA** and **TELPA** appear surprisingly similar in their claims, even to the extent of using identical wording for their test descriptions. They use three and two raters respectively, which is likely to have a positive influence on reliability. However, neither of the test providers present any validity and reliability evidence.
4. We were unable to find any relevant claims or evidence for the **IATA-Berlitz** Proficiency Test.

Overall, much to our surprise, despite the claims we located, we were unable to find any convincing validity and reliability evidence in the online resources, with the exception of **VAET**. Given the important consequences of an aviation test, such lack of evidence is extremely worrying.

Limitations

After exploring the relevant electronic resources, we summarise a number of limitations of the ELPAC competitors in Table 9. As we have already mentioned in 13.3, we are concerned about a number of aspects of validity, reliability, test development and administration. The most important issue is that the available resources do not seem to provide enough evidence of test development and validation.

Test	Limitations
1. RELTA	<ul style="list-style-type: none"> • No information on the equivalence of test versions. • No reliability indices were given. • Some listening tasks are not repeated and some are played twice. No rationale is given for this. • The responses are recorded and rated "by an accredited rater after the test is completed". Rater reliability could be an issue here.
2. TEA	<ul style="list-style-type: none"> • Section 2 Part A: the boundaries between assessing listening and speaking skills are not clear. Will the test-takers' "talk" about the recorded clips be assessed purely in terms of listening comprehension or will the speaking criteria be applied? • The test seems to be weighted in favor of speaking with less focus on listening. • None of the rating scale components (pronunciation, structure, vocabulary, fluency, comprehension and interaction) apply to assessing listening abilities apart from "comprehension". • According to the website, the test and aviation courses focus on plain English only. Phraseology is not assessed.
3. TELLCAP	<ul style="list-style-type: none"> • Plain English does not appear to be tested. • No information on number of versions and equivalence. • No information on rater selection and training procedures. • One rater only. • In a linked PPT document there is mention of developing test versions and inviting external experts for test evaluation but no further details are available.
4. EPTA	<ul style="list-style-type: none"> • The test seems to be weighted in favor of speaking with less focus on listening. • "Part 1 is audio-mediated and/or computer-based". It is not clear which is the case.

	<ul style="list-style-type: none"> No information on trialling, number of versions and reliability indices.
5. TELPA	<ul style="list-style-type: none"> No information on test trialling. No information on reliability indices. No information on the provision of a practice/sample test.
6. VAET	<ul style="list-style-type: none"> No information on actual task types. No information on whether there are separate versions for pilots and ATCs or just one version. No information on the provision of a practice/ sample test.
7. IATA	No information

Table 9: Limitations of other aviation tests

Annex 2: Details of reliability and validity claims by competitor tests

Test	Aspects	Claims	Evidence
RELTA	Reliability	<ul style="list-style-type: none"> An extremely reliable test 	None provided on test website
	Validity	<ul style="list-style-type: none"> A valid and effective measure of language proficiency An appropriate and valid test for pilots and air traffic controllers An appropriate test instrument for assessing pilot/air traffic controllers in relation to the 6-band ICAO rating scale 	None provided on test website
TEA	Reliability	<ul style="list-style-type: none"> Courses for raters to enable experienced Aviation English teachers to understand and apply the ICAO language proficiency scale. Research indicates that TEA is a valid, standardised, accurate and secure testing solution 	None provided on test website
	Validity	<ul style="list-style-type: none"> Research indicates that TEA is a valid, standardised, accurate and secure testing solution 	None provided on test website
TELLCAP	Reliability	<ul style="list-style-type: none"> Test administered by a “specially trained examiner” Evaluation of recorded performance by a “certified” rater “the best language teachers and translators from the Moscow office of the Automated Air Traffic Controlling Center (AATCC) were trained to conduct interviews” The test was piloted with ATCs from the Moscow office of the AATCC 	None provided on test website
	Validity	<ul style="list-style-type: none"> Items cover routine and non-routine topics Listening test: “real flight crew –ATC communication” “inviting external experts for test evaluation” Informing pilots and ATCs about test format and content-familiarisation 	None provided on test website
EPTA	Reliability	<ul style="list-style-type: none"> Three raters conduct the rating independently. Then, they convene to compare and discuss the rating result “Reliable: Based on tried and tested theories and procedures” 	None provided on test website
	Validity	<ul style="list-style-type: none"> “Effective: closely aligned with ICAO requirements for an English proficiency test in aviation context” “Valid: executes the goal of testing English proficiency in an aviation context” 	None provided on test website

ELPAC

Test	Aspects	Claims	Evidence
TELPA	Reliability	<ul style="list-style-type: none"> • “Reliable: Based on tried and tested theories and procedures” • Two raters score the taped responses independently • All raters are required to undergo training on how to rate English language proficiency 	None provided on test website
	Validity	<ul style="list-style-type: none"> • Test specifications are written and reviewed by members from applied linguistics and aviation experts • Questions developed and reviewed by national and international experts • Effective: Closely aligned with the ICAO requirements for an English proficiency test in an aviation context • Valid: Executes the goal of testing English proficiency in an aviation context • Appropriate: Tasks are aviation-specific 	None provided on test website
VAET	Reliability	<ul style="list-style-type: none"> • Software is “trained” on a pool of non-native speech samples • A number of arguments in favour of machine scoring, using evidence from an empirical study which correlated ratings between a speech recogniser and human raters 	<p>$r=0.97$ (Balogh, 2006: 2)</p> <p>Balogh, J. (2006). A Case for Automation in Aviation English Language Assessment. Retrieved 28/06/2007, from http://harcourtassessment.com/NR/rdonlyres/121BCB12-7231-41C6-B051-D3BF98624D7A/0/WhitePaper_CaseForAutomation.pdf</p>
	Validity	<ul style="list-style-type: none"> • Expected as well as unexpected events are tested • Machine can measure ability on the highly debatable interactions scale • Scaling is based on the ICAO rubrics • Telephone administration entails high authenticity in the aviation context (close simulation to radiotelephony communication) • Items are recorded in “a range of accents and speaking styles of aviation professionals” • Various test versions are produced by using a random combination of items 	see Balogh (2006: 3)

Conclusions

After reviewing the competition, we conclude that ELPAC should be able to compete strongly with all other tests in terms of overall quality. There has been considerable research into the validation of ELPAC and we recommend that this is clearly stated on the website and other resources to indicate the difference in quality with the majority of the aviation tests on the market. In particular, we recommend that all evidence for the validity and reliability of the test should be published on the ELPAC website.

3. Conclusions

The Lancaster validation report makes a number of conclusions:

3.1. Test development

The procedures followed to date are suitable.

3.2. Test design

This process of paper design and setting appears very thorough.

3.3. Dissemination

The ELPAC website (<http://www.elpac.info>) and the associated Sample Tests website (<http://www.elpacsample.info>) are excellent channels for disseminating information about the test. It is clear that major efforts have been made to reach a wide range of stakeholders including the heads of Air Traffic Management Training institutions, Training managers from EUROCONTROL member states Pilots, Controllers, English Language Teachers, Language Testers, Regulators and Unions. A range of dissemination methods have been used: from presentations at regular briefing meetings, to conference presentations to magazine articles.

3.4. Guidelines

Clear guidelines have been put in place for administrators and invigilators. The markers for the Listening test are given clear instructions on how to complete the marking task including a Powerpoint presentation which (with screenshots from the marking tool) shows them exactly what they have to do for each section. We think that these guidelines are exemplary. The guidelines for assessors and interlocutors in Paper 2 are comprehensive and well-organized. The description of these guidelines has covered all the important aspects involved in administering the test such as the roles of assessors and interlocutors, general and specific guidelines for assessors and interlocutors, and the role of the third assessor.

3.5. Paper 1 – mode of delivery

We are satisfied with the proposed delivery of Paper 1 and are confident that ENOVATE A.S. will be able to support the system. We are particularly pleased to note that dedicated support is provided for the ELPAC test.

3.6. Paper 1 – reliability and bias analysis

We are satisfied that there is a sufficiently large pool of very reliable items available for the test developers to choose from when constructing the final test versions. Although we are not able to guarantee the exact reliability of the final test versions, the high quality of the test versions we have seen gives us every confidence that the final test versions will be as reliable and will display the same patterns of item and controller separation as the ones we have reviewed. Test reliability is good and the overall chi-square shows no significant effect overall. Consequently, we would support the statement that the overall test neither favours nor discriminates against any controller type.

3.7. Paper 2 – inter-rater reliability during the last major trial

All coefficients are acceptably high (above .87), which means that, providing assessors really did come to final levels independently before agreeing on a level, they are assessing candidates and interpreting rating scales in similar ways.

3.8. Training workshops

We are impressed with the amount of work which has gone into planning accreditation workshops as well as preparing the training packages and other relevant documentation (e.g. confidentiality letters, administration procedures, etc).

The plans for the actual content of the full one-week training workshops are sound and comprehensive. The amount of time allocated to scoring actual samples of speaking performances is very reasonable.

We also found the provision of a self-study package for trainees a definite positive feature of the training procedures. The self-study packages are very comprehensive. We were particularly pleased with the inclusion of benchmarked performances in the self-study package. Moreover, requiring trainees to submit their self-study reports at least one week before the workshop is also a good feature in the procedures to ensure that all trainees are prepared and also to give ELPAC members sufficient time to process this data.

We were also impressed that only examiners who pass the training week will be accredited.

3.9. Standard setting

The ELPAC standard setting reports provide satisfactory evidence of procedural validity.

3.10. Sustainability

ELPAC developers have done an excellent job in identifying (and exemplifying) the precise human resources required for running the test in a typical situation, together with the necessary qualifications for staff members. This has covered the need for test administrators, examiners and markers as well as back-up staff.

It is evident that considerable effort has been made by ELPAC developers to stress the need for adequate and on-going resources for sustaining ELPAC. As the budget discussion document shows, the test development team have assessed the various possible scenarios for running the system along with the financial implications for each scenario. This is a vital procedure as it ensures that the financial implications for each option are considered before any action is taken.

3.11. Competition with other tests

After reviewing the competition, we conclude that ELPAC should be able to compete strongly with all other tests in terms of overall quality. There has been considerable research into the validation of ELPAC and we recommend that this is clearly stated on the website and other resources to indicate the difference in quality with the majority of the aviation tests on the market.