Final Report on a Survey of Aviation English Tests

J Charles Alderson, Lancaster University

Executive Summary

The Lancaster Language Testing Research Group was commissioned in 2006 by the European Organisation for the Safety of Air Navigation (Eurocontrol) to conduct a validation study of the development of a test called ELPAC (English Language Proficiency for Aeronautical Communication), intended to assess the language proficiency of air traffic controllers.

As part of that study, Internet searches for evidence of other tests of air traffic control found very little evidence available to attest to the quality of the tests for licensure of either pilots or air traffic controllers.

Therefore it was decided to conduct an independent survey of tests of aviation English.

The consequences of inadequate language tests being used in licensing pilots, air traffic controllers and other aviation personnel are potentially very serious.

- A questionnaire was developed, based on the Guidelines for Good Practice of the European Association for Language Testing and Assessment (EALTA), and sent to numerous organisations whose tests were thought to be used for licensure of pilots and air traffic controllers.

- 22 responses were received, which varied considerably in quantity and quality.

- This probably reflects a variation in the quality of the tests, in the availability of evidence to support claims of quality, and in low awareness of appropriate procedures for test development, maintenance and validation.

- We also surveyed all 190 Member States of ICAO about the tests they recognise, with only 17 responses.

- The Survey of Aviation English Tests indicates that it is unclear whether national civil aviation authorities have the knowledge to judge the quality of tests.

We conclude:

1. We can have little confidence in the meaningfulness, reliability, and validity of several of the aviation language tests currently available for licensure.

2. Monitoring is required of the quality of language tests used in aviation to ensure they follow accepted professional standards for language tests and assessment procedures.

Acknowledgements

I am very grateful to all members of the Lancaster Language Testing Research Group, in particular to Jayanti Banerjee, Spiros Papageorgiou, Rima Al Zadjali, Ali van Moere and Tania Horak, for their valuable contributions to this research.

Introduction

1. By 5 March 2008, air traffic controllers and pilots were required by the International Civil Aviation Organisation (ICAO) to have a certificate attesting to their proficiency in the language(s) used for aeronautical communication. ICAO Document 9835 - "Manual on the Implementation of ICAO Language Proficiency Requirements" - lays out the principles according to which the language proficiency requirements should be met. Although several organisations made every effort to produce suitable tests by the deadline, in the event an implementation period was allowed, with a new deadline of March 2011.

2. Clearly, aviation language tests are extremely high stakes, not just for the test-takers but for every potential airline passenger, crew member and air-traffic controller, not to mention the airline companies, insurance companies, etc. Therefore, it is of the utmost importance that such tests are constructed to the highest possible standards, and it is a matter of no little concern that certificates attesting to the attainment of relevant levels of language proficiency will not be required of aviation personnel for another three years.

3. Consider, for example, the following report in the Times of London, 13 June 2008:

> "A Polish airliner came within seconds of colliding with another plane near Heathrow because its pilots had such poor English that they could not understand basic instructions from air traffic controllers. The Lot Boeing 737, carrying 95 passengers and crew, wandered the skies for almost half an hour as the pilots struggled to identify their position. A controller had to instruct another aircraft to change direction to avoid a collision.
>
> A document seen by The Times suggests that only 15 out of 800 Polish pilots flying internationally have passed the test for the required standard of English".

4. The Lancaster Language Testing Research Group was commissioned in 2006 by the European Organisation for the Safety of Air Navigation (Eurocontrol) to conduct a validation study of the development of a test called ELPAC (English Language Proficiency for Aeronautical Communication), intended to assess the language proficiency of air traffic controllers. This 18-month study produced two reports, an Interim Report which made recommendations for the improvement of the tests and the associated quality control procedures, and a Final Report, which provided a commentary on the quality of the ELPAC test and made a series of recommendations for further quality control measures. That Final Report was published on the Eurocontrol website, accompanied by a commentary on the recommendations by Eurocontrol, and is available at http://www.elpac.info/.

5. As part of the ELPAC Validation Project, the Internet was searched for evidence of other tests of air traffic control, and we were somewhat surprised to find that, although several websites were identified, very little evidence was publicly available in summer 2007 attesting to the quality of the tests. Once the ELPAC Validation Report had been submitted and

accepted by the funders, it was therefore decided to conduct an independent survey of all providers of tests intended for air traffic control. This document reports on the methodology and findings of that survey.

Methodology

6. A repeat search of the Internet in January-February 2008 confirmed the earlier findings of the ELPAC Validation Study that there was very little information available about the validity and reliability of what few tests of air traffic control language appeared to exist. At this point, we were relatively unfamiliar with the aviation language testing field, and had few contacts in the area. We felt that it would be impractical to plan to conduct interviews with informants, at least in the initial stages of the research, and so it was decided to develop a questionnaire that could be sent to any organisation that had developed tests of aviation English.

7. Since the Executive Summary of the ELPAC Validation Study had been framed by the Guidelines for Good Practice (GGP) of the European Association for Language Testing and Assessment (EALTA) (see Appendix 1 to this report), we also based our survey questionnaire on EALTA's Guidelines, which reflect professional expectations about minimum standards of practice in language testing. In addition, reference was made to the questionnaire used in an earlier survey of the practices and procedures of British ESOL examining bodies, conducted by the Lancaster Language Testing Research Group (LTRG) in 1990 (see Alderson *et al*, 1995, Appendix 2). Below, we summarise how the GGP formed the basis of a questionnaire, which was then adjusted to meet the needs of the Aviation English Tests (AET) survey. A detailed account, taken from Alderson and Banerjee (2008), can be found in Appendix 2 and a copy of the final version of the questionnaire is in Appendix 3.

8. The first drafts of our AET questionnaire were developed in several iterations of construction, critique and revision by the LTRG in the winter of 2007/8. We were conscious during this process that our survey might be very demanding on test designers' time, but we felt that the extremely high stakes of the tests justified requesting test providers to put considerable effort into providing the evidence for the quality of their instruments.

9. We were, however, conscious of the fact that many of the respondents would be aviation industry personnel, unfamiliar with some of the concepts and terminology used by language testing researchers and made every effort to ensure the comprehensibility and relevance of the questionnaire (see Appendix 2 for full details).

10. The main divergences from the EALTA's Guidelines were as follows:

1. Glossing terms: e.g. "specifications" = "test blueprint"; "construct" = "the skills and sub-skills that the test / subtest(s) are intended to measure".

2. Deleting questions:
    a. *What quality control procedures are applied?* (redundant)
    b. *How is potential test misuse addressed?* (highly unlikely)

3. Specifying
    e.g. GGP Q 1.5 Are the specifications for the various audiences differentiated?
    AET Q 1.4 Which of the following groups of test-takers are specified?
    i. *Air Traffic Controllers*
    ii. *Pilots*

iii. *Air Service Personnel*

Please add any groups your test caters for but which we have not listed.

4. Clarifying

   e.g.. GGP 4.1 What are the security arrangements?

   AET Q 5.1 How is cheating (for example, personation, bribery, copying, access to illicit materials, etc) checked or prevented?

5. Making relevant

   e.g.. GGP Q 5.3 What procedures are in place to ensure that the test keeps pace with changes in the curriculum?

   AET 6.2 What procedures are in place to ensure that the test keeps pace with changes in the ICAO requirements?

6. Adding items
   a. The GGP does not enquire about needs analyses conducted by the test developers, so two questions on this topic were added.

      AET Q2.8 *Have you conducted an investigation of test-takers' needs?*

      AET Q2.9 *Have you carried out a survey of how language is used in the context of aviation?*
   b. The GGP makes no reference to pass/fail distinctions or grade boundaries, whereas in the case of the Aviation English tests, pass/fail boundaries are clearly crucial:

      AET Q3.9 *How is the pass mark for objectively scored tests determined?*
   c. The GGP makes no reference to complaints procedures

      AET Q 5.5: *What processes are in place for test-takers to make complaints or seek scrutiny of marks, and/or re-marking of the test?*
   d. Nor does the GGP ask about the possibility for candidates to take the test again to achieve a higher score.

      AET Q 5.6: *How often can a candidate re-take the test?*

7. Modifying double barrelled questions

   e.g. GGP Q 1.10 *Are marking schemes/rating criteria described?*

   AET Q 1.8 Are the rating criteria available?

8. Changing ambiguous wording

   e.g. GGP Q 3.2 *Are the tests piloted?*

   AET Q 4.1 *Are the tests trialled?*

11. At the same time as the questionnaire was being trialled, we learnt of a parallel project being conducted by Dr Ute Knoch of the University of Melbourne into attitudes to the ICAO scales of language proficiency. We made contact with Dr Knoch and we agreed to join forces in order to avoid duplicating resources. As a result we broadened the scope of the survey to include tests for pilots and other air service personnel and named our project "Survey of Aviation English Tests".

12. It became clear in our discussions with Dr Knoch that many test providers might indeed find the survey intimidating and might be very reluctant to respond. We decided, therefore, to administer a two-stage survey, with a short initial "filter" questionnaire which asked a few

questions about respondents' perceptions of the ICAO scales, and also included a brief section on basic quality control procedures (see Appendix 4). In the filter questionnaire, we asked respondents if they would be willing to respond to a longer questionnaire about quality control procedures. Both the filter and the full survey questionnaires were made available on the Internet via Survey Monkey for ease of completion, but we also produced Word documents for those who preferred not to complete the questionnaire on the Internet. In addition, we offered to telephone or Skype those who preferred to discuss orally their responses, or who might have questions about the survey. In the event, nobody requested a telephone interview.

13. Dr Knoch will be reporting separately on the findings of the initial survey, and in this report we concentrate on presenting the results of the main questionnaire and our attempts at maximizing responses.

Informants

14. Although we had conducted the validation study of the ELPAC test and gained some understanding of the demands of communication in the aviation context, we were by no means part of the community of language educators responsible for the delivery of courses in aviation language (predominantly English). Nor were we aware of who the main and subsidiary players might be in the field of aviation language testing, despite, as mentioned above, conducting a comprehensive Internet search to see which tests might by now have been made available. This second search failed to produce any new information about tests or test providers, and what information that was publicly available in January - February 2008 (between one and two months before the ICAO deadline of March 5th, 2008) was very sparse. This finding confirmed our conclusions in the Final Report to Eurocontrol, and indeed also the conclusions presented at LTRC 2007 by Gene Hallek and Carol Moder, on the quality of aviation language testing in general.

15. However, in order to survey as wide a sample as possible of organisations involved in aviation language testing, we decided to contact language testers and those engaged in teaching aviation language, to identify what tests might be out there and which organisations might be involved in developing aviation language tests. The letter we wrote to LTEST-L and to the EALTA discussion list is presented in Appendix 5, and Dr Knoch sent a similar letter to aviation lists and newsletters.

16. The efforts that were made to identify respondents resulted in a list of 30 contacts and our correspondence with these contacts yielded the names of further individuals and organisations. We eventually sent out the initial filter survey letter and link to the questionnaire website (see Appendix 6) to 30 contacts. In the three months that the survey was available, we received 20 responses, including two requests for a Word version of the questionnaire. Of the 20 respondents, 13 agreed to respond to the main, follow-up questionnaire. Of those that said they were not willing to answer the follow-up questionnaire, none were (yet) involved in delivering tests for licensure of pilots or air traffic controllers.

17. During the period that the filter questionnaire was available, we were contacted by a number of well-informed people who kindly provided us with the names of further individuals and organisations and who reassured us that our survey was indeed of considerable importance and interest. We therefore approached those additional contacts, and slowly our contacts list grew, and is still growing.

18. To date we have attempted to contact some 71 individuals and organisations, and have received some sort of reply from 42. However, a number of those replying have said that they are not involved in aviation language testing at all, or that their tests are not for the purposes of licensing pilots, air traffic controllers or other air service personnel, but are either placement or internal achievement tests. In all, we received responses to the main, follow-up questionnaire from 22 persons or organisations who are or have been or appear to have been involved in developing tests for certification (licensure) of pilots, air traffic controllers and/or other air service personnel. Although we also received responses pertaining to the following four tests, these appear not to be used for licensure, and so have not been included in our account of results:

- HungaroControl EPT
- Aviation English Proficiency Assessment
- PELA (Proficiency in English language for (student) Controllers)
- BETA - Benchmark English Test for Aviation

However, we are very grateful to the organisations that supplied us with information on these tests.

19. We received refusals to complete our follow-up questionnaire from two organisations that do appear to be involved in producing tests for licensure. One was included in the initial filter questionnaire results, mentioned above, but we received the following response after dispatching, on request, a paper version of the main questionnaire:

“I have discussed with my partners the convenience to participate in this second part of your survey and we have arrived to the agreement that we can not do it due to the nature of the specific questions included.”

We also receive the following message from one company:

“Thank you for the invitation to participate in the Survey of English Language Testing which we are declining.”

20. Quite a few of those we have contacted have yet to respond, despite reminders. Whilst this may be indicative of the lack of quality of the tests produced by such organisations and an associated reluctance to admit this in public, we cannot take non-response to indicate lack of quality, although it may well indicate lack of public accountability of such providers.

21. Making contact with those responsible for developing aviation language tests for certification/ licensing purposes has proved, and is still proving, to be a time-consuming activity. This we find rather curious, given the very high stakes involved in testing the language proficiency of pilots, air traffic controllers and others. We were somewhat surprised to discover that the ICAO itself has not chosen to approve or disapprove of any testing procedure. However, as an organisation which is part of the United Nations, the ICAO can set out the requirements of standards, and indeed has provided guidelines on implementation of language proficiency requirements for Member States (http://www.icao.int/fsix/lp/docs/Guidelines.pdf) but apparently cannot enforce these and it is the responsibility of the national civil aviation authorities to decide which tests or assessment procedures they will accept. Whether such national aviation authorities have the competence

to judge the quality of the tests available is unclear to us at present. As one correspondent noted:

> “There is a widespread misunderstanding of ICAO′s functioning, mandate and budgetary constraints. ICAO is an international legislative body, but not its police force. Both its decisions and their implementation are entirely dependent on its sovereign member States. This very sovereignty may explain in part some reluctance to take part in such a survey conducted by an English-speaking country.”

22. Nevertheless, we were eventually able to discover on the ICAO website (http://www.icao.int/anb/fls/lp/lpcompliance1.cfm) a document claiming to detail the state of play of national compliance with the ICAO requirements. We were relieved to find in that set of publicly available documents the names and email addresses of those responsible within the various national authorities for compliance with the ICAO requirements. We therefore began a separate survey in May, 2008, requesting each of the 190 authorities on the list for the names of the tests that they have recognised or approved, and we also requested contact details of those organisations providing the tests. We hoped in this way to triangulate our survey. However, we received only 17 responses by the deadline of this survey, and we report on the results in a separate document. We will be pursuing this avenue in a further attempt to identify all tests and assessment procedures likely to be used to meet ICAO language proficiency requirements by March 2011.

Caveat

23. As already mentioned, the deadline of March 5, 2008 for compliance with the ICAO requirements has been extended by up to three years (see the communication '*A36-11*: Proficiency in the English language used for radiotelephony communications' in Appendix 7). It appears that many national authorities have been unable to arrange for the development of suitable assessment arrangements, or to develop or commission their own tests, in time for the 2008 deadline. It would also appear that some authorities have been reluctant to provide suitable testing, or to ensure that adequate language teaching is available for aviation personnel in preparation for whatever assessment procedures exist.

24. It should also be pointed out that we are still by no means certain that we have identified the key players, and we would welcome additional names and addresses.

25. Moreover, several respondents have said that they are still in the process of developing their assessment procedures and tests, and are therefore not yet in a position to report. Indeed, at least two such organisations to date have enquired about the possibility of the Lancaster Language Testing Research Group assisting them with validation studies similar to those undertaken for Eurocontrol.

Reactions

26. Before we proceed to present the results of our survey, we would like to draw the attention of readers to the fact that we received numerous messages welcoming our survey and expressing the general opinion that our research was very much needed. These (anonymised) messages and other comments on our work, are presented in Appendix 8 to this Draft Report.

Results

27. The tests used for licensure which we have identified to date and which are included in the account of responses to the main questionnaire, below, are the following:

1	a) Level 6 Proficiency Demonstration b) Formal Language Evaluation
2	ALITE
3	Altec Benchmark Evaluation
4	Aviation English Proficiency Assessment
5	CXELT (Cathay English Language Test - Pilots)
6	ELP Test V1 (obsolete); ELP Test V2 (operational); ELP Test V3 (in development)
7	ELPAC (English Language Proficiency for Aeronautical Communication)
8	Aviation Language Proficiency Test (available for both English and French)
9	English for Aviation Language Test (EALT); Expert Level 6 Speaker Assessment (ELSA)
10	English Proficiency Exam for Aviators (Chile)
11	English Proficiency Test for Airline Pilots (航空英語能力証明試験)
12	English Proficiency Test for Aviation (EPTA)
13	ICAO English Proficiency Exam for Aviators
14	IELTS
15	Ilmailuhallinnon kielitaitotesti (The language proficiency test of the (Finnish) Civil Aviation Authority)
16	LANG TECH Aviation English Oral Competence Assessment
17	RELTA
18	TEA (Test of English for Aviation)
19	TELLCAP® - Test of English Language Level for Controllers and Pilots
20	Test of English for Aviation Purposes (the name may have changed after the development period)
21	Thai DCA aviation test
22	Versant Aviation English Test (VAET)

28. The institutions that appear to be involved in developing these are as follows:

1	Aviation Services Limited, New Zealand
2	Griffith University, Australia
3	Altec Internationale, LLC
4	Berlitz Inc.
5	Cathay Pacific China
6	Aerosolutions, Belgium
7	EUROCONTROL
8	Transport Canada - Civil Aviation
9	MLS International, UK
10	Universidad Técnica Federico Santa María
11	Sophia Linguistics Institute for International Communication, Sophia University
12	G-TELP KOREA
13	Universidad Técnica Federico Santa María
14	IELTS Australia
15	Finnish Civil Aviation Authority

16	Language Technology
17	RMIT University, Australia
18	Mayflower College, UK
19	Aviation English Training Center “CompLang”, Russia
20	Colegio de Pilotos Aviadores (México)
21	Department Of Civil Aviation, Thailand
22	Ordinate Corporation, USA

29. We were very pleased also to receive test validation reports as follows:

o	ELPAC Validation Study Final Report
o	Japanese pilots test - English Proficiency Test for Airline Pilots, Sophia University (translated from Japanese)
o	Expert Opinion on external validation of TELLCAP®
o	TEA - Test of English for Aviation - Mayflower College Research notes
o	VAET - Versant Aviation English Test - Versant with Ordinate Technology

30. We are very grateful to all respondents to the main questionnaire for taking the time and trouble to give very useful and detailed responses. Since we guaranteed all respondents anonymity, responses to the various questions are presented anonymously, and where quotations are provided, these are presented in random order.

31. We have included responses from two respondents although we have doubts in both cases. In the first case, we had some doubts about tests 10 and 13 in the above lists, which appear to be so similar as possibly to refer to the same test. Since, however, some of the responses referring to the tests differed from each other, we assumed that the tests were in some sense discrete and therefore included the results below.

32. In the second case, that of IELTS, we were informed that the IELTS test is used for licensure for aviation purposes, and we have therefore included their responses, but we are aware that the IELTS test was not developed for the purposes of licensing pilots or air traffic controllers, but as evidence of proficiency in English for admission to tertiary institutions where instruction takes place in English. The IELTS test is, however, also widely (mis)used for purposes of immigration and for some forms of professional recognition, including medical councils.

33. Should clarification be forthcoming in either case, we will edit this report appropriately (contact c.alderson@lancaster.ac.uk).

34. What follows is a relatively brief summary and discussion of a set of often complex and detailed responses. The questions asked and the details of the responses are to be found in Appendix 9. The text below is organised according to the six main sections of the questionnaire. It should be noted at the outset that much of the information received was somewhat unsatisfactory, perhaps due to the unfamiliarity of respondents with language testing standards. This suggests that education of aviation authorities is necessary, as well as further research to enable clarification of any misunderstandings.

Section One: Test Purpose and Specification

35. All but one respondent said that they had produced test specifications, and although the description of the test purpose varied considerably, all but one referred specifically to the language used for communication in aviation contexts or, rather more ambiguously, to "compliance with ICAO English Proficiency Rating". However, the amount of detail presented varied enormously.

36. The amount of detail supplied in response to the question about stakeholders also varied greatly, but one response claiming that this was confidential information may have misunderstood that we were asking for names, which was not the case. Generally, the identification of stakeholding groups is not considered confidential.

37. All the test-takers for whom the tests were specified were identified with the aviation context, the most common (21 out of 22 responses) being pilots, but in only 12 out of the 22 cases were separate specifications developed for subgroups of test-takers. Given the potentially different demands on language among the various groups, this seems unfortunate.

38. Nevertheless, 21 out of 22 respondents say that they do specify the skills to be tested and the test methods used. One respondent neither specifies nor exemplifies the test methods to be used (and provides no justification for this omission), although they do specify the construct. Only 15 out of the 22 respondents say they provide examples of test-taker performance, which rather suggests that not all test-takers, or their teachers, will be adequately informed about what is expected of them.

39. Two respondents said that rating criteria were not made available, which is rather odd since it is normal practice to make such criteria public.

40. Two respondents failed to specify which ICAO levels their tests measured, but as many as five claimed that their tests measured all six levels. This seems unlikely, since a very wide range of proficiency is covered by the ICAO level descriptions. Although this is an empirical question, no evidence for the claim was forthcoming. Quite a few tests claimed to be able to measure Levels 2 to 6, and all claimed they included Level 4 – a crucial cut-off for licensure – in their measurement. Only four did not claim to be able to measure Level 6.

41. However, when asked for the evidence to support all these claims, responses varied greatly in specificity, and included irrelevant responses like "approval by civil aviation authorities" (which is clearly no guarantee of level without information on how this approval was reached), or the fact of presentation at conferences, or even by bald reference to the ICAO Proficiency Scale, with no details of how this had been related to the test results. There seems to be a degree of naïvety among some respondents as to what constitutes adequate evidence of a level.

Section Two: Test Design and Item Writing

42. When asked what professional experience was required of test developers, most mentioned both EFL teaching or research experience, and aviation experience, but one claimed no EFL experience was required, another that no professional experience at all was required and two that only EFL or applied linguistic experience were required. The relevance and suitability of these four tests to the assessment of language in an aviation context is

questionable. Moreover, why one respondent declared that this information was confidential is unclear. The response: “The test has been completed, approved by the XXX DGAC and in operation since 01/06/07” indicates a lack of understanding of the question, or of the issue.

43. A similar range of responses was provided in answer to the question about professional experience required of item writers.

44. Only just over half of the respondents (12) claimed that they provided training for their test developers and eleven explicitly stated that they provided no training for item writers either. However, 19 said they provided guidelines (nature unspecified) on test development and item writing.

45. 18 respondents said that they had systematic procedures for review and editing of items and tasks to ensure that they matched the test specifications and comply with item writer guidelines, but two said that they had no such procedures. Similarly, two respondents said that item writers do not receive feedback on their work, whilst 18 said that they did provide such feedback.

46. Fourteen respondents reported that they had carried out analyses of test-takers’ needs, whilst as many as five admitted that they had not carried out research in the aviation context. Eleven said that they had carried out a survey of how language is used in the context of aviation, without giving further details, but seven admitted that they had not conducted such research. Perhaps these respondents are unaware of the importance of doing so for a high-stakes setting such as aviation communication.

Section Three: Rating Procedures

47. Only one respondent said that performances on their speaking test were scored (rated) by computer, but three organisations did not respond to this question. Furthermore, only five respondents said that their raters were trained for each administration of their speaking test, although sixteen did say that they used exemplar performances (benchmarks) in their training. Five failed to respond to this question, so their practices remain unexplained.

48. Sixteen respondents said that they do double marking of subjectively marked tests, but as many as six either failed to respond, or said No. It is unclear why the lack of double marking is felt to be acceptable. However, as many as eighteen respondents explained what they did in the event of disagreement among raters, although four failed to respond.

49. Seventeen respondents said that they calculated inter-rater reliabilities (with three saying No, and two not responding), but when asked to give details of the resulting coefficients, only four provided details, with two asserting that the details were confidential. It seems there is no reason not to give the detailed results unless they are indeed worryingly low.

50. Thirteen respondents said that they also calculated intra-rater reliabilities and of these, nine gave details, but yet again, one (unacceptably) claimed that this was confidential information, and four said either that they did not do such calculations, or failed to respond. A worrisome six respondents either did not know whether routine monitoring of marking was carried out, or said that this did not happen, or failed to respond to this question.

51. Five either failed to respond to the question about pass-marks for objectively scored tests, or simply claimed that the question was not applicable. Of the seventeen that did respond in detail, one claimed the test was not a pass-fail test, one claimed not to understand the question, and six simply referred irrelevantly to the rating scale descriptors. Whilst two respondents acknowledged the complexity of the issue, and gave useful details of procedures, at least one failed to understand what was involved: "Using sections in the test that focus on specific areas of language. Each section requires a candidate to employ a particular language ability." Clearly, more clarification and exploration is required of this topic, or at least awareness-raising amongst the organizations responsible for deciding on a candidate's level of performance.

Section Four: Test Analyses

52. Nineteen claimed that their tests were trialled, but three failed to give a response. When asked what the normal size of the trial sample was, seven failed to respond.

53. Two respondents simply repeated that the tests were trialled but failed to provide any sort of figures, and one incomprehensibly said that "it corresponds to the regular test format". Where figures were given at all, some were as low as 6 or 10. Although a few provided acceptable figures, several were simply evasive. Again, more exploration of this important topic is clearly called for.

54. When clarification was sought of the size of the trial sample in relation to the total test population, eight either failed to respond, or claimed it was impossible to say, although no reason was given. Three not unreasonably reported that they did not yet know the size of the test population, whereas others gave estimates ranging from 2.5% to 20%. However, the large number of respondents that failed for whatever reason to give firm figures suggests that we can as yet have little confidence in the representative nature of the trialling processes. Hopefully this is something that will improve as experience with trialling and test administration develops.

55. When the question was asked whether the reactions were gathered of those involved in the tests, it appeared that test-takers' views were only collected in 17 out of 22 cases, that sixteen collected examiners' views and only fourteen gathered data on the views of invigilators. Quite why so few are interested in the views of participants is unclear.

56. When asked how the trial data was analysed, six failed to give any response at all, and several gave vague and uninformative answers like "quantitatively and qualitatively" or "statistically" without providing any details. Only three respondents gave detailed responses. When asked about how changes were agreed upon once the trial data had been analysed, four failed to respond, but most respondents referred to some form of discussion. Only one respondent referred to the re-trialling of items that had to be re-written.

57. Sixteen respondents said that different versions were produced of their tests each year and details of how equivalence between versions was established were given by fourteen respondents. Seventeen claimed that statistical analyses were carried out, but when asked what sort of analyses were conducted, one person responded: "Observational checklists", and several either gave vague answers like "various", or "being developed". One claimed "this is too comprehensive to cover here". Six, however, were able to give more or less detailed

reports of existing practice, and one referred to future intentions in some detail. It is clear that the amount and nature of acceptable analyses varies considerably as, probably, does any recognition of the need for such analyses.

58. Fifteen respondents said that their analyses are presented in a report, but only ten gave details of the sorts of analyses presented. One claimed that this was an internal process and failed to give any details at all. Only two out of the 22 respondents said that they made their reports available publicly, and both gave details of how they could be accessed. This lack of openness is a cause for considerable concern.

Section Five: Test administration

59. When asked how cheating was checked or prevented, five failed to respond, but the rest gave details which ranged from the vague "general guidelines for the conduct of tests. Regulations prohibiting cheating" and "extreme security measures are in place. Upon cheating candidates are disqualified and blacklisted", to very detailed accounts of many measures taken to prevent impersonation and other forms of cheating. It was interesting to contrast the fullness of responses from most respondents to this question with the vague responses by some to questions about test analysis, as it suggests greater importance is attached to security than to test analysis.

60. Seventeen respondents said that their test invigilators (proctors) were trained and eighteen respondents said that the administration of their tests was monitored. However, in only eleven of these cases did invigilators have to write reports.

61. Only sixteen respondents provided evidence of procedures for lodging appeals or complaints, and seventeen provided details of the number of times candidates can re-take the test, which varied considerably, from once, to no limit.

Section Six: Review

62. Sixteen respondents said that their tests were reviewed continuously or changed regularly (one claimed that this was not applicable), with only one organization claiming that tests were destroyed after they had been administered once. Details varied considerably, but the majority of respondents showed an awareness of the need for regular or continuous review. The extent to which such reviews were comprehensive or rigorous remains, however, unexplored.

63. Respondents were asked how they ensured that their tests kept pace with changes in the ICAO requirements. Most claimed, either that the ICAO requirements have not changed, or that they keep in touch with the ICAO, its website, or documents, but nobody claimed to have specific procedures in place to monitor this. Nevertheless, there appeared to be no reason for concern that test developers were not kept informed of the need for change. One respondent simply replied "Amendments to the tests as required".

64. Finally, respondents were asked whether there were any other procedures which they used to ensure the validity and reliability of their tests of Aviation English which we had not mentioned. Seven said that there were, but gave no further details, and five gave further details, three of which included validation studies, one referred to "basic research" which appeared to be some form of discourse analysis and ethnographic study, and one referred to

procedures to ensure that test-takers were as familiar as possible with the test, and had opportunities to get "official" feedback from their performance on a practice test.

65. Participants were asked if they would be willing to provide further documentation to back up their claims but, as mentioned earlier, only five provided any form of validation study or research notes, and none provided any of the specific documents requested in the list presented at the end of the questionnaire. It is, therefore, impossible to estimate the extent to which evidence is available to provide full and adequate support for many of the claims made by respondents.

Summary and (interim) conclusions

66. The responses reported in this survey varied greatly in quantity and quality, and this may well reflect a variation in the quality of the tests themselves, both in terms of the evidence available to support claims of quality and in terms of the awareness among test developers as to what constitutes appropriate procedures for test development, maintenance and validation. In only a minority of cases was there evidence of adequate concern for quality control and public accountability. Too often, confidentiality was claimed as a reason for not providing essential information. Too often, vague answers were provided which failed to answer the questions asked, and too often, certain test developers or organisations simply failed to respond to questions, or even to understand what was required, or why it is important to be able to supply relevant information.

67. Nevertheless, it was reassuring to see that at least some test developers took their responsibilities seriously, and made every effort, both to answer our questions fully, and to supply additional information in the form of validation studies. As we agreed to maintain the anonymity of respondents, we cannot identify those who represent good practice in test development, nor those who appear not to understand what good practice entails. We can point out, however, that,

- if test specifications and test content do not match the needs of test-takers,
- if test developers and item writers do not have relevant aviation and language education experience,
- if item writers and test raters are not appropriately trained,
- if tests are not adequately pre-tested on suitably representative and sizable samples of test-takers,
- if the results of such trials are not suitably analysed and actions taken to address any evident problems,
- if test-takers are not aware of how and on what they will be tested and how their performance will be rated,
- if the reliability of marking is not monitored, calculated and reported,
- if there is no statistical information available to support the claimed level of the test, the equivalence of different versions from year to year, and the comparability of the results of different tests purporting to measure the same target level of proficiency,
- and, above all, perhaps, if reports providing evidence as to the thoroughly professional quality of tests are not publicly available,

then little or no confidence can be held in the meaningfulness, reliability and validity of several of the aviation language tests currently available. Unfortunately, at present, this appears to be the case for too many of the tests that we have surveyed, and if other tests exist for the aviation context that have not responded to our survey, then it is highly likely that they, too, fail to meet minimal standards of quality. The consequences of inadequate language tests being made available to license pilots, air traffic controllers and other aviation personnel are almost too frightening to contemplate.

Further research and monitoring

68. This survey raises and, we believe, addresses an important question: how fit for purpose are currently available aviation English tests? We hope to have provided some answers, but much remains to be done. In particular, we continue to monitor relevant electronic newsletters, discussion lists, and websites. We will continue to press national civil aviation authorities for details of the tests and assessment procedures they recognise, and how they know that such tests are fit for purpose. We have already begun to interview some of our contacts for in-depth information about their tests, areas of quality and development that concern them, and insights into the processes and politics that have or have not facilitated the development of quality language tests for this very high-stakes context. We intend to extend this in-depth study to a range of participants and contexts and it is highly likely that at some suitable point in the future, we will repeat this survey with a view to seeing to what extent things have improved and to what extent test developers are willing to be accountable for the quality of their instruments and procedures. We are extremely grateful to those who participated in this study, but we are all too aware that participation was voluntary, was patchy, and that there are almost certainly many tests and assessment procedures out there that we have not yet identified or that have not come forward to give details of their processes and their quality.

69. However, there is another angle to and purpose for this report, and that relates to the use of codes of practice / of ethics/ guidelines for good practice. This project started life as an enquiry into test quality, but we quickly saw that the project represented an opportunity to see if a particular set of guidelines could be used to frame a validity study. Although the EALTA guidelines were used, in principle other professional guidelines could have been used. How we had to adjust the guidelines is reported in Appendix 2, but the fact is that we were indeed able to do so and our survey instruments seem to have proved capable of yielding useful, indeed crucially important information.

70. Whether they are wholly adequate is a lesson we have not yet learned, as we are, unexpectedly, still *in media res*. But our own hunch is that more information and research is needed in order to be able to evaluate the quality of the tests we are interested in.

71. The first and most obvious gap is the lack of documents that might support the assertions in the answers to the survey instruments. We have received to date only five validation study reports. Otherwise we have only the responses to the questionnaire to rely on, and those alone are clearly inadequate and need triangulating. The question is how? What will it cost in time and resource and effort? Who will provide the necessary resources? So far this effort has been entirely voluntary and unfunded, albeit quite time-consuming.

72. The second issue is the fact that we have not begun to address the issue of the content of the various tests: are they suitable for this aviation context? Do the tests tap the “right” constructs? Are the tasks a faithful or plausible reflection of the target language use situation? Should they be? Are Oral Proficiency Interviews suitable instruments, even if data were available on their reliability and validity (which is generally not the case)? And, of course,

who are we, language testers who know little about the aviation context, to judge? Indeed, should it be we who judge?

73. A third issue has to do with the levels/ standards of the tests themselves: is a Level 4 on one test equivalent to a Level 4 on a different test? There are suggestions in our correspondence that this is indeed not the case and that some tests are "easier" than others. That indicates that a set of comparative studies is called for, and even a series of standard setting studies applied to a range of available (and officially recognised) instruments.

74. The need for such standard setting studies then leads to a questioning of the standards themselves - the ICAO scales. Are they sufficiently explicit and relevant to guarantee that any test constructed on the basis of the ICAO scales will indeed be at the "right" level or do the scales represent shifting sands, an uncertain and unstable foundation? Even if a test would prove to be an adequate reflection of the ICAO scales, are the scales themselves adequate to identify the appropriate levels required for aviation communication?

75. Fifthly, whose responsibility is it to conduct such surveys? Who are we to hold aviation language test developers, or indeed national civil aviation authorities, accountable for what they do or fail to do, for the decisions they take as to test acceptability, or the decisions they refuse to take? What should be the role of a professional language testing organisation like the International Language Testing Association (ILTA) or the European Association for Language Testing and Assessment (EALTA - whose Guidelines formed the basis of our main questionnaire)?

76. Our answer is that if nobody is taking those decisions, or providing a degree of oversight, then it is better that we do it than that nobody does it. But is that good enough? We think not.

References

Alderson, J. C., Clapham, C. and Wall, D. (1995) *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press

Alderson, J. C. and Banerjee, J. V. (2008) 'EALTA's Guidelines for Good Practice: A test of implementation'. Paper presented at the 5th Annual Conference of the European Association for Language Testing and Assessment, Athens, Greece, May. http://www.ealta.eu.org/conference/2008/programme.htm, last accessed June 2nd, 2008

Hallek, G and Moder, C (2007) 'Discourse context, complexity and task variation in ESP assessment.' Paper presented at the 29th Annual Language Test Research Colloquium, Barcelona, Spain.

List of appendices

Appendix 1: EALTA Guidelines for Good Practice (Section C)

Appendix 2: How the EALTA Guidelines were adapted to the Aviation English testing context

Appendix 3: The final version of the AET Survey questionnaire

Appendix 1
Guidelines for Good Practice in Language Testing and Assessment, by the European Association for Language Testing and Assessment (EALTA)

For the text of the full Guidelines in English and 32 other European languages, see www.ealta.eu.org/guidelines.htm. The Guidelines are divided into three sections, addressing three different audiences: teacher trainers; classroom teachers; and institutions which develop language tests and examinations. The text below constitutes the third and final section of the EALTA Guidelines.

EALTA Guidelines for Good Practice (Section C)

Considerations for test development in national or institutional testing units or centres

EALTA members involved in test development will clarify to themselves and appropriate stakeholders (teachers, students, the general public), and provide answers to the questions listed under the headings below. Furthermore, test developers are encouraged to engage in dialogue with decision makers in their institutions and ministries to ensure that decision makers are aware of both good and bad practice, in order to enhance the quality of assessment systems and practices.

1. TEST PURPOSE AND SPECIFICATION

1. How clearly is/are test purpose(s) specified?
2. How is potential test misuse addressed?
3. Are all stakeholders specifically identified?
4. Are there test specifications?
5. Are the specifications for the various audiences differentiated?
6. Is there a description of the test taker?
7. Are the constructs intended to underlie the test/subtest(s) specified?
8. Are test methods/tasks described and exemplified?
9. Is the range of student performances described and exemplified?
10. Are marking schemes/rating criteria described?
11. Is test level specified in CEFR terms? What evidence is provided to support this claim?

2. TEST DESIGN and ITEM WRITING

1. Do test developers and item writers have relevant experience of teaching at the level the assessment is aimed at?
2. What training do test developers and item writers have?
3. Are there guidelines for test design and item writing?
4. Are there systematic procedures for review, revision and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines?
5. What feedback do item writers receive on their work?

3. QUALITY CONTROL and TEST ANALYSES

1. What quality control procedures are applied?
2. Are the tests piloted?
3. What is the normal size of the pilot sample, and how does it compare with the test population?
4. What information is collected during piloting? (teachers´opinions, students´ opinions, results,…)
5. How is pilot data analysed?
6. How are changes to the test agreed upon after the analyses of the evidence collected in the pilot?
7. If there are different versions of the test (e.g., year by year) how is the equivalence verified?
8. Are markers trained for each test administration?
9. Are benchmarked performances used in the training?
10. Is there routine double marking for subjectively marked tests? Is inter and intrarater reliability calculated?
11. Is the marking routinely monitored?
12. What statistical analyses are used?
13. What results are reported? How? To whom?
14. What processes are in place for test takers to make complaints or seek reassessments?

4. TEST ADMINISTRATION

1. What are the security arrangements?
2. Are test administrators trained?
3. Is the test administration monitored?
4. Is there an examiner's report each year or each administration?

5. REVIEW

1. How often are the tests reviewed and revised?
2. Are validation studies conducted?
3. What procedures are in place to ensure that the test keeps pace with changes in the curriculum?

6. WASHBACK

1. Is the test intended to initiate change(s) in the current practice?
2. What is the washback effect? What studies have been conducted?
3. Are there preparatory materials?
4. Are teachers trained to prepare their students for the test/exam?

7. LINKAGE TO THE COMMON EUROPEAN FRAMEWORK

1. What evidence is there of the quality of the process followed to link tests and examinations to the Common European Framework?
2. Have the procedures recommended in the Manual and the Reference Supplement been applied appropriately?
3. Is there a publicly available report on the linking process?

Appendix 2
How the EALTA Guidelines were adapted to the Aviation English testing (AET) context

In what follows, we describe how EALTA's Guidelines had to be adjusted to form the basis of a questionnaire survey.

COMPARISON WITH A PREVIOUS SURVEY

A previous survey, of British ESOL examining bodies, had been conducted by the Lancaster Language Testing Research Group in 1990, and so we compared Section C of the EALTA GGP ("*Considerations for test development in national or institutional testing units or centres*") with the original questionnaire used in that survey (see Alderson *et al*, 1995, Appendix 2). The latter questionnaire was much more detailed than the EALTA GGP, consisting of 54 Yes-No questions, often with up to as many as 12 sub-questions, divided into six sections: Syllabus, Examination Construction, Validation, Marking, Results, and Examination Revision. In contrast, the GGP has 44 questions in seven sections, headed: Test Purpose and Specification; Test Design and Item Writing; Quality Control and Test Analysis; Test Administration; Review; Washback; Linkage to the Common European Framework.

Nevertheless it was felt that most of the aspects of test design, administration and quality control covered in Appendix 2 were also addressed, albeit in different ways, by the GGP, and the draft questionnaire was thus based on the GGP. Only four additions were made to the draft questionnaire as a result of this comparison with Appendix 2 of Alderson *et al*. First, it was felt important to identify in the Aviation English Test Survey (AET) whether any needs analyses had been conducted in the course of the design of AET tests, and so two extra questions were added to the draft Section 2 on Test Design and Item Writing. Such questions had not been asked in the GGP, presumably since many of the tests and assessments being developed in Europe would be related to general purposes, and to primary, secondary or tertiary education, rather than to professional or vocational test purposes.

AET Q2.8 *Have you conducted an investigation of test-takers' needs? Please tick 'Yes' (Y) or 'No' (N).*

AET Q2.9 *Have you carried out a survey of how language is used in the context of aviation? Please tick 'Yes' (Y) or 'No' (N).*

Secondly, Appendix 2 had asked a question about procedures to decide upon pass/fail distinctions or grade boundaries, whereas the GGP makes no reference to such concepts. Since in the case of the aviation English tests, pass/fail boundaries are clearly crucial, it was decided to add the question *"How is the pass mark for objectively scored tests determined?"* to Section 3 of the AET Questionnaire (AET Q3.9).

In addition, a version of Question 49 of Appendix 2 was added to the end of the AET questionnaire which is not in the GGP, namely:

AET Q 6.3 *Are there any other procedures which you use to ensure the validity and reliability of your tests of Aviation English which we have not mentioned?Y N*

Finally, Appendix 2 made reference to machine marking, whereas the GGP seems to assume that all marking is done by humans. As we were aware that at least one AET test used computer-based testing and scoring, it was decided to make a distinction in the AET questionnaire between electronic scoring and human marking.

This comparison of the GGP and Appendix 2 also highlighted the fact that many of the GGP questions can be answered by Yes or No, with no further clarification. This is consistent with EALTA's mission to raise awareness in its constituencies, but is arguably inappropriate for a questionnaire survey. Although Appendix 2 specifically prompts for Y and N responses, this was in an attempt to make responding to the survey as easy and quick as possible, in order to maximise response rates. However, we initially felt that this was inadequate in such an important survey as the AET, and we decided to request explicitly that AET examining bodies provide the evidence to support their responses, otherwise the results would bear no credibility. Therefore, although the open-question format of the GGP was modified in some cases by prompting for Yes or No, further prompts were added to encourage submission of the evidence for the response, or of its location:

IF YES, please attach (the construct description) or give the precise URL

However, on reflection we decided, once the questionnaire had been drafted, critiqued and revised several times, that it would simply be too demanding to require such detail, and response rates would likely be poor. The final version of the AET questionnaire therefore largely consisted of Yes/ No questions, with occasional open-ended boxes to complete where we felt that it would not be too demanding for respondents to input a small amount of text. Instead, at the end of the questionnaire, we offered respondents the opportunity to send us further details by email, and provided a list of documents or information that we would find it very helpful to have access to.

ADJUSTING THE GGP TO THE AET CONTEXT

The comparison reported above with the earlier survey resulted in some changes to the GGP for the purposes of the AET survey. More important and substantial than these additions, however, were the modifications, additions and deletions made to the GGP in order to make it more suitable for the specific context of aviation English, and for a questionnaire survey rather than an awareness-raising instrument. The addition of two questions about needs analysis mentioned above was an example of the former, albeit prompted by Appendix 2, and the initial change of format from the GGP to the AET questionnaire (also mentioned above) was an example of the latter.

One important change to the GGP was the removal of the final two sections on *Washback* and *Linkage to the Common European Framework*. The latter was motivated by the assumption that this would simply not be relevant. Instead we changed two questions in earlier sections to refer to the ICAO framework. Reference to the CEFR in GGP Question 11 Section 1 on Test Purpose and Specification was changed from

> GGP Q1.11 *Is test level specified in CEFR terms? What evidence is provided to support this claim?*

to read:

AET Q 1.9a. *What ICAO level does the test measure?*

AET Q 1.9b. *What evidence is provided to support this claim?*

Moreover, in keeping with the reference to the ICAO, GGP Question 3 in Section 5, Review was changed from

> GGP Q 5.3 *What procedures are in place to ensure that the test keeps pace with changes in the curriculum?*

to read

AET 6.2 *What procedures are in place to ensure that the test keeps pace with changes in the ICAO requirements?*

The remaining five sections of the GGP became six sections in the AET questionnaire, because the LTRG group felt that the GGP Section 3 headed *Quality Control and Test Analyses* included two very different types of activity, namely Rating Procedures (which became Section 3 in the AET questionnaire) and Test Analyses (Section 4 in the AET questionnaire). In any case, we felt that the whole questionnaire (and arguably the whole of the GGP) was about quality control, and so this was a misleading and inappropriate heading for one section.

When splitting GGP Section 3 into two, we changed the order of elements, such that questions about test analysis came after questions about rating procedures in the AET questionnaire, as this seemed more logical. GGP Q 3.1 *What quality control procedures are applied?* was deleted as being redundant.

GGP SECTION 1 - TEST PURPOSE AND SPECIFICATION

This section was modified by deleting GGP Q 1.2 *How is potential test misuse addressed?* (this was felt to be highly unlikely in this context). GGP Q 1.6 *Is there a description of the test taker?* was also deleted as it was felt to be too vague, and its likely intentions were covered by the revised wording of the original GGP Q 1.5 (see below).

> GGP Q 1.8 *Is the range of student performances described and exemplified?*

was also deleted as it was felt that it overlapped with GGP Q 1.11 (see above), as it could be interpreted as referring to the descriptions of performance in the various ICAO levels, an interpretation we wished to avoid. The other questions were retained but many were reworded to make them easier to answer.

The rather general GGP Q 1.4 became the first question in the AET questionnaire, as it is assumed by GGP Qns 1.1, 1. 2 and 1.3. It was reworded from

> GGP 1.1 *Are there test specifications?*

to

AET Q 1.1 *Have you produced a set of test specifications (test blueprint)? Please circle 'yes' (Y) or 'no' (N). If YES, answer the questions 2-9b below regarding the specifications*

IF NO, go to Section 2 "TEST WRITING AND ITEM WRITING" on page 3

The open-ended GGP Q 1.1 *How clearly is/are test purpose(s) specified?*
was modified by removing the word "clear" in the new AET Q 1.2, as it is highly unlikely that any test provider would admit to their specifications not being clear.

AET Q 1.*2 According to the specifications, what is the purpose of the test?*

Furthermore, GGP Q 1.3: *Are all stakeholders specifically identified?*

was adjusted to read:

AET Q 1.3 Who are the stakeholders (those who commissioned the test or who have an interest in it) for this test? Please provide a full list.

i. ____

ii. ____

iii. ____

iv. ____

v. ____

Add more stakeholders if necessary

Also, GGP Q 1.5: *Are the specifications for the various audiences differentiated?*
was adjusted to prompt particular audiences:

AET Q 1.4 *Which of the following groups of test-takers are specified?*

Air Traffic Controllers	***Y***	***N***
Pilots	***Y***	***N***
Air Service Personnel	***Y***	***N***

Please add any groups your test caters for but which we have not listed.

____

____

Are there separate specifications (or a separate section within the specifications) for the different groups of test-takers? Please tick yes (Y) or no (N).

Minor wording changes only were made to the following questions. It was felt that respondents might not understand the word "constructs" in

GGP Q 1.7 *Are the constructs intended to underlie the test/subtest(s) specified*?

and so it was changed to

AET Q 1.*5 Do you specify the test construct (the skills and sub-skills that the test/subtest(s) are intended to measure)? Please tick 'Yes' (Y) or 'No' (N).*

Changes were made to GGP Q 1.8 *Are test methods/tasks described and exemplified?* to become

AET Q 1.6 *Do you list and exemplify the test methods/ tasks used in the test? Please circle yes (Y) or no (N).* ***Y N***

GGP Q 1.10 *Are marking schemes/rating criteria described?*

was a double-barrelled question, but we were not interested in knowing whether a marking scheme for objective questions existed and so the question was changed to read

AET Q 1.8 Are the rating criteria available? Please circle yes (Y) or no (N).

GGP SECTION 2. TEST DESIGN and ITEM WRITING

Relatively minor changes were made to this section, with the exception of the addition of the new questions on needs analysis, as described above. The double-barrelled GGP Q 2.1 *Do test developers and item writers have relevant experience of teaching at the level the assessment is aimed at?* was clarified and specified for the AET context, and the words "*in test design*" were added to GGP Q 2.2, which originally read: *What training do test developers and item writers have?*

AET Q 2.1 *What professional experience do you require of your test developers (eg. air traffic control experience and/or English as a Foreign Language teaching experience)?*

AET Q 2.2 *Do you provide training in test design to your test developers?*

AET Q 2.3 *What professional experience do you require of your item writers (e.g. air traffic control experience and/or English as a Foreign Language teaching experience)?*

AET Q 2.4 *Do you provide training in item writing to your item writers?*

The two questions GGP Q 1.3 *Are there guidelines for test design and item writing?*

and

> *GGP Q 1.4 Are there systematic procedures for review, revision and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines?*

remained intact, with the minor change of "*Are there*" to "*Do you have*" .

> GGP Q 2.5 *What feedback do item writers receive on their work?*

was reworded as follows:

AET Q 2.7 *Do item writers receive feedback on their work? Please circle yes (Y) or no (N)*

GGP SECTION 3: QUALITY CONTROL and TEST ANALYSES

As mentioned earlier, the third section of the GGP was divided into two separate sections, Section 3 Rating Procedures and Section 4 Test Analyses. The major innovation in the AET questionnaire Section 3 was to add questions about electronic scoring, as follows:

1. *Are the speaking parts of the examination scored by computer? Please circle yes (Y) or no (N).*

If your answer to Section 3, Question 1 was YES, please answer questions 2 and 3.

If your answer was 'No' please move to Question 4.

2. *Is the computer software trained for each test administration? Please circle yes (Y) or no (N).*

The GGP Qns 3.8, 3.9 and 3.10 were as follows:

> Q 3.8. *Are markers trained for each test administration?*
> Q 3.9. *Are benchmarked performances used in the training?*
> Q 3.10. *Is there routine double marking for subjectively marked tests? Is inter- and intra-rater reliability calculated?*

These were revised as follows:

AET Q 3.3 *Are human raters (of the speaking tests) trained for each test administration? Please tick 'Yes' (Y) or 'No' (N).*

AET Q 3.4 *Are exemplar performances (benchmarks) used in the training? Please tick 'Yes' (Y) or 'No' (N).*

AET Q 3.5a *Is there routine double-marking for subjectively marked tests? Please circle yes (Y) or no (N).*

If NO, please move to question 7.

AET Q 3.5b. *What action do you take in the event of disagreement between markers?*

AET Q 3.6a. *Is inter-rater reliability calculated? Please circle yes (Y) or no (N).*

If YES, please provide details of the correlations achieved.

AET Q 3.6b. Is intra-rater reliability calculated? Please circle yes (Y) or no (N)

IF YES, please provide details of the correlations achieved.

GGP Q 3.11. *Is the marking routinely monitored?* was retained unchanged, with the Yes/ No boxes added.

As mentioned already above, a new question, AET Q 3.8 to address the issue of pass marks was added at the end of this section.

Questions 2 and 3 of the GGP Section 3 were retained virtually unchanged in Section 4 on Test Analyses,

> GGP Q 3.2 *Are the tests piloted?*
> GGP Q 3.3 *What is the normal size of the pilot sample, and how does it compare with the test population?*

with the one exception of the substitution of the, in this context, rather unfortunate terms "pilot" "piloted" and "piloting" with the terms "trial", "trialling" and "trialled". Doubtless if the specific purpose tests being investigated were tests for lawyers and judges, we might wish to retain the original word!

AET Q 4.2a. *What is the normal size of the trial sample?*

AET Q 4.2b. *How big is the trial sample compared with the test population?*

> GGP Q 3.4 *What information is collected during piloting? (teachers´opinions, students´ opinions, results,…)*

was changed to the following:

AET Q 4.3 *Is the following information collected during trialling? (Please circle yes (Y) or no (N))?*

a) test-takers' views on the difficulty of the test

b) test-takers' views on the appropriacy of test tasks

c) invigilators' views on the difficulty of the test

d) invigilators' views on the appropriacy of test tasks

e) examiners' views on the difficulty of the test

f) examiners' views on the appropriacy of test tasks

GGP Qns 3.5, and 3.6 were unchanged, with the exception of the wording mentioned above

> *How is pilot data analysed?*
> *How are changes to the test agreed upon after the analyses of the evidence collected in the pilot?*

GGP Q 3.7 was changed from

> *If there are different versions of the test (e.g., year by year) how is the equivalence verified?*

to the following:

AET Q 4.6 *Are there different versions of the test (e.g. year by year)? Please circle yes (Y) or no (N).*

IF YES, how is the equivalence verified?

> GGP Q 3.12 *What statistical analyses are used?*

was changed to read

AET Q 4.7 *Are any statistical analyses carried out? Please circle yes (Y) or no (N).*

If YES, what statistical analyses are used to examine the results of the tests?

The triple-barrelled GGP Q 3.13 *What results are reported? How? To whom?*
was adjusted to be simpler to answer, as follows:

> AET Q 4. 8a *Are the analyses presented in a report? Please circle yes (Y) or no (N)*

If YES, please list all the analyses included in the report, giving details of how they are reported.

AET Q 4.8b *Are the reports available to the public? Please circle yes (Y) or no (N).*

If YES, please provide details of how we might access the reports.

GGP SECTION 4 TEST ADMINISTRATION

This section became Section 5 of the AET questionnaire. GGP 4.1 *What are the security arrangements?* was unclear and so it was replaced by the clearer question AET Q 5.1, as follows:

AET Q 5.1 How is cheating (for example, personation, bribery, copying, access to illicit materials, etc) checked or prevented?.

The next three questions (Questions 2 – 4) of the GGP were retained virtually unchanged, with the exception of the replacement of the word *administrator* and *examiner* by *invigilator*, as follows:

GGP 5.2 *Are test administrators trained?*
GGP 5.3 *Is the test administration monitored?*
GGP 5.4 *Is there an examiner's report each year or each administration?*

AET Q 5.2 *Are invigilators trained? Please circle yes (Y) or no (N).*

AET Q 5.3 *Is test administration monitored? Please circle yes (Y) or no (N).*

AET Q 5.4 *Do invigilators write a report after each administration? Please circle yes (Y) or no (N).*

In addition, two new questions were added, as follows:

AET Q 5.5: *What processes are in place for test-takers to make complaints or seek scrutiny of marks, and/or re-marking of the test?*

AET Q 5.6: *How often can a candidate re-take the test?*

GGP SECTION 5: REVIEW

The original fifth section of the GGP contained three questions, two of which were retained with only minor wording changes in the final sixth section of the AET questionnaire. However, the second GGP question (Q 5.2) was deleted as it was felt to be a very wide-ranging question and because other questions allowed this issue to be addressed, but in a more focussed manner.

GGP Q 5.1 *How often are the tests reviewed and revised?*

GGP Q 5.2 *Are validation studies conducted?*
GGP Q 5.3 *What procedures are in place to ensure that the test keeps pace with changes in the curriculum?*

AET Q 6.1 *How often are the tests reviewed and changed?*

AET Q 6.2 *What procedures are in place to ensure that the test keeps pace with changes in the ICAO requirements?*

Respondents were then invited to mention any other procedures which they use to ensure the validity and reliability of their tests, as mentioned earlier.

ADDITIONAL INFORMATION

Although we had simplified an earlier draft of the AET questionnaire by removing the request to attach relevant reports or give URLs, we nevertheless felt it would be very helpful, and in keeping with the spirit of the enquiry, to seek as much documentation or evidence as possible. We therefore added a final section to the Survey questionnaire, requesting any further information that might be available either on a website or in internal documents or reports. The list of possible documents appears after Section 6 in the questionnaire, see Appendix 3 below. This information is not requested in the original GGP.

Appendix 3
The final version of the AET Survey questionnaire

FOLLOW-UP QUESTIONNAIRE TO PROVIDERS OF AVIATION ENGLISH TESTS

SECTION 0). YOUR DETAILS

1) Name of company/ organization

2) Your name (optional)

3) Your position within the organization

4) Your telephone number/ Skype name (optional)

5) What are the names of the Aviation English test (s) you are reporting on in this questionnaire? Please list all

SECTION 1. TEST PURPOSE AND SPECIFICATION

1. Have you produced a set of test specifications (test blueprint)? Please tick ‘Yes’ (Y) or ‘No’ (N).

   Y N

   If YES, answer Questions 2-9b below regarding the specifications.
   IF NO, go to Section 2 ’Test Writing and Item Writing‘ on page 3.

2. According to the specifications, what is the purpose of the test?

3. Who are the stakeholders (those who commissioned the test or who have an interest in it) for this test? Please provide a full list.

vi. ______________________________

vii. ______________________________

viii. ______________________________

ix. ______________________________

x. ______________________________

Add more stakeholders if necessary.

4. Which of the following groups of test-takers are specified?

Air Traffic Controllers	**Y**	**N**
Pilots	**Y**	**N**
Air Service Personnel	**Y**	**N**

Please add any groups your test caters for but which we have not listed.

____

____

Are there separate specifications (or a separate section within the specifications) for the different groups of test-takers? Please tick yes (Y) or no (N).

Y **N**

5. Do you specify the test construct (the skills and sub-skills that the test/subtest(s) are intended to measure)? Please tick ‘Yes’ (Y) or ‘No' (N).

Y **N**

6. Do you list and exemplify the test methods/tasks used in the test? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

7. Do you provide examples of test-taker performances? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

8. Are the rating criteria available? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

9a. What ICAO level does the test measure?

9b. What evidence is provided to support this claim?

SECTION 2. TEST DESIGN and ITEM WRITING

1. What professional experience do you require of your test developers (e.g. air traffic control/ pilot experience and/or English as a Foreign Language teaching experience)?

2. Do you provide training in test design to your test developers? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

3. What professional experience do you require of your item writers (e.g. air traffic control/ pilot experience and/or English as a Foreign Language teaching experience)?

4. Do you provide training in item writing to your item writers? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

5. Do you have guidelines for test design and item writing? Please tick 'Yes' (Y) or 'No' (N).

Y N

6. Do you have systematic procedures for review and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines? Please tick 'Yes' (Y) or 'No' (N).

Y N

7. Do item writers receive feedback on their work? Please tick 'Yes' (Y) or 'No' (N).

Y N

8. Have you conducted an investigation of test-takers' needs? Please tick 'Yes' (Y) or 'No' (N).

Y N

9. Have you carried out a survey of how language is used in the context of aviation? Please tick 'Yes' (Y) or 'No' (N).

Y N

SECTION 3. RATING PROCEDURES

1. Are the speaking parts of the examination scored by computer? Please tick 'Yes' (Y) or 'No' (N).

Y N

If your answer to Section 3, Question 1 was YES, please answer Questions 2 and 3.
If your answer was NO please move to Question 4

2. Is the computer software trained for each test administration? Please tick 'Yes' (Y) or 'No' (N).

Y N

3. Are human raters (of the speaking tests) trained for each test administration? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y **N**

4. Are exemplar performances (benchmarks) used in the training? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y **N**

5a. Is there routine double-marking for subjectively marked tests? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y **N**

If NO, please move to Question 7.

5b. What action do you take in the event of disagreement between raters?

6a. Is inter-rater reliability calculated? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y **N**

If YES, please provide details of the correlations achieved.

6b. Is intra-rater reliability calculated? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y **N**

IF YES, please provide details of the correlations achieved.

7. Is the marking routinely monitored? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y **N**

8. How is the pass mark for objectively scored tests determined?

SECTION 4. TEST ANALYSES

1. Are the tests trialled? Please tick 'Yes' (Y) or 'No' (N).

Y **N**

2a. What is the normal size of the trial sample?

2b. How big is the trial sample compared with the test population?

3. Is the following information collected during trialling? Please tick 'Yes' (Y) or 'No' (N).

a) test-takers' views on the difficulty of the test **Y** **N**

b) test-takers' views on the appropriacy of test tasks **Y** **N**

c) invigilators' views on the difficulty of the test **Y** **N**

d) invigilators' views on the appropriacy of test tasks **Y** **N**

e) examiners' views on the difficulty of the test **Y** **N**

f) examiners' views on the appropriacy of test tasks **Y** **N**

4. How is the trial data analysed?

5. How are changes to the test agreed upon after the analyses of the evidence collected in the trial?

6. Are there different versions of the test (e.g. year by year)? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

IF YES, how is the equivalence verified?

7. Are any statistical analyses carried out? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

If YES, what statistical analyses are used to examine the results of the tests?

8a. Are the analyses presented in a report? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

If YES, please list all the analyses included in the report, giving details of how they are reported.

8b. Are the reports available to the public? Please tick ‘Yes’ (Y) or ‘No’ (N).

Y N

If YES, please provide details of how we might access the reports. .

SECTION 5. TEST ADMINISTRATION

1. How is cheating (for example, personation, bribing, copying, access to illicit materials, etc.) checked or prevented?

2. Are test invigilators trained? Please tick ‘Yes’ (Y) or ‘No’ (N).

   Y N

3. Is test administration monitored? Please tick ‘Yes’ (Y) or ‘No’ (N).

   Y N

4. Do invigilators write a report after each administration? Please tick ‘Yes’ (Y) or ‘No’ (N).

   Y N

5. What processes are in place for test-takers to make complaints or to seek scrutiny of marks, and/or re-marking of the test?

6. How often can a test-taker re-take the test?

SECTION 6. REVIEW

1. How often are the tests reviewed and changed?

2. What procedures are in place to ensure that the test keeps pace with changes in the ICAO requirements?

3. Are there any other procedures which you use to ensure the validity and reliability of your tests of Aviation English which we have not mentioned?

Y N

Many thanks for completing this questionnaire. Your responses will be very helpful to us in our research.

If you are willing to provide more details of your responses, we would be very happy to receive the details of any URL that provides further information.

Alternatively, if you could send us copies of any reports that provide more detailed evidence of your quality control procedures, we would be extremely grateful.

Specifically, we would welcome any documentation/ reports/ URLs of the detail on the following:

❑ The statement of purpose for your test(s)

❑ The construct description

❑ Lists or examples of test methods and tasks, or sample practice tests provided for test preparation

❑ Examples of test-taker performance

❑ The rating criteria

❑ The evidence to justify the ICAO level of your test (s)

❑ The different kinds of experience demanded of test designers

❑ The nature of the training the test developers have received.

❑ The different kinds of experience demanded of item writers.

❑ The nature of the training the item writers have received.

❑ Guidelines for test design and item writing.

❑ Documentation of systematic procedures for review and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines

❑ Examples of the feedback given to item writers

❑ A report of any investigation of test-takers' needs

❑ Any survey of how language is used in the AET context

❑ The software training procedures used for computer-based scoring.

□ The measures in place to monitor the performance of the computer scoring

□ The training provided to raters of subjectively marked tests.

□ Details of how exemplar performances (benchmarks) are selected.

□ Details of how double-marking is organised

□ Details of the action taken in the event of disagreement between raters

□ Details of the system adopted to monitor the quality of the rating

□ Any report on how the pass mark is determined for objectively scored tests

□ Any report of how the tests are trialled.

□ Evidence or reports of:

a) test-takers' views on the difficulty of the test

b) test-takers' views on the appropriacy of test tasks

c) invigilators' views on the difficulty of the test

d) invigilators' views on the appropriacy of test tasks

e) examiners' views on the difficulty of the test

f) examiners' views on the appropriacy of test tasks

□ A report of the analysis of trial data.

□ An account of how changes to the test are agreed upon after the analyses of the evidence collected in the trial

□ A report of how the equivalence is verified of different versions of the test.

□ Any report of what statistical analyses are used to examine the results of the tests

□ Any report on how cheating is checked or prevented.

□ Details of how test invigilators are trained

□ Details of how test administration is monitored

□ An example of an invigilator's report.

□ The complaints procedures and how requests are made to seek scrutiny of marks, and/or re-marking of the test

THANK YOU FOR YOUR COLLABORATION IN THIS IMPORTANT SURVEY.

Appendix 4
The initial filter questionnaire, devised by Ute Knoch and Charles Alderson

QUESTIONNAIRE ABOUT AVIATION ENGLISH TESTING

SECTION A) Your details

1) Name of company/ organization

2) Your name (optional)

3) Your position within the organization

4) Your telephone number/ Skype name (optional)

5) What are the names of the Aviation English test (s) you have developed, are developing or which you administer? Please list all

6) What skills do each of the above tests measure?

7) How long have each of the above tests been operational (since when)?

8) Is/ are your test(s) recognized/ used by your national civil aviation authority for licensing purposes? Y N

9) What ICAO level does/do your test(s) measure?

10) Are you familiar with the detail of the ICAO scales? Y N

If yes, please go to Section B) on the next page. If No, please go to the final section, Section C on page 3.

SECTION B)   The ICAO scale

1) Has the ICAO scale influenced the design of your test in any of the following areas:

- the skills you included in the test   Y   N

- the speakers you use for any listening materials or as interlocutors for the speaking test
Y   N

- the aspects the raters look for when rating the performance   Y   N

- how the test result is reported   Y   N

2) Do the raters of your test use the ICAO rating scale for their ratings?

Y   N

3) Do you think the ICAO rating scale is equally useful for language experts and aviation experts?

Y   N

4) Are you satisfied with the six categories on the ICAO rating scale? Please comment briefly on each:

- **Pronunciation   Y   N**
- **Fluency   Y   N**
- **Structure   Y   N**
- **Vocabulary   Y   N**
- **Interaction   Y   N**
- **Comprehension   Y   N**

5) Do you think ICAO level 4 is the appropriate operational level for pilots or ATCs?   Y N

6) Is there anything you would like to be changed about the ICAO rating scale?

SECTION C

Please make one copy of this page for each of the tests mentioned in Question A5 above.

1) What quality control procedures do you use to ensure the quality of your test? (We would be happy to receive copies of any reports, or the URL of any relevant Website)

2) Are the tests trialled before going live? Y N

3) Are the results of the test analysed? Y N
If Yes, how?

4) Do you have in-house technical experts in language testing? Y N

5) Have you ever contracted any language testing consultants to assist you with your test design, test development and/or test validation? Yes / No

If Yes, could you give us that person/ organisation's name and/or email address?

6) Would you be willing to answer a more detailed survey about the ICAO and the technical qualities of your test? Y N

That survey could be delivered in a Word document, over the web, or could be the object of a telephone interview. If your answer to 6) is Yes, which of these three options would you prefer?

If your answer is No, could you give us the name of a person we could contact to respond to the more detailed survey?

7) Would you be interested to receive a copy of the results of this survey?
Y N

If yes, please give a mailing address below

Many thanks for your time

Appendix 5
Letter to LTEST-L, EALTA discussion lists and aviation English lists

Dear Colleagues,

We are involved in conducting a survey of aviation English test designers, and we are compiling a list of people and organisations involved in developing such tests. We would be very grateful if anybody involved in writing tests of aviation English, be that for air traffic control, for pilots, or for other learners or users of aviation English, would contact us directly. Even if you are not involved in such tests, perhaps you know somebody or some organisation that is.

We are already aware of, or in contact, with the following organisations, but if any colleagues receiving this message know of others not mentioned below, we'd be very grateful if you could email us their details off this list.

Thank you very much in advance
Charles Alderson
Lancaster University
c.alderson@lancaster.ac.uk

Ute Knoch
University of Melbourne
uknoch@unimelb.edu.au

Organisations involved in developing tests of aviation English

Aviation Services Limited (ASL)
Griffith University
English Proficiency Test for Aviation (EPTA, Korea)
EUROCONTROL (ELPAC/PELA)
RMIT English Language Test for Aviation (RELTA)
Versant Aviation English Test

Appendix 6
Covering letter to the Filter questionnaire survey

Dear Colleague,

We are writing to you as part of a survey of the ICAO scales and associated tests of Aviation English. We understand that you have been involved in developing such tests and have experience of the ICAO scales.

We are conducting this survey without funding and independently of any organisation. We would be very grateful if you were willing to spare a little time to complete the attached questionnaire and supply us with relevant information. We are aware that some of this information may be confidential or commercially sensitive, and we can assure you that we will treat this in complete confidence. We will not divulge any of the information we gather to third parties and all respondents will be anonymised in any report we present or write about the results of the survey.   We will ensure that nobody can trace any information back to you or your company.

You can complete the questionnaire online by going to
http://www.surveymonkey.com/s.aspx?sm=YkkdSgbdDEtc5fNOdRzzyQ_3d_3d

Alternatively, if you prefer to answer the questionnaire in a word document, please send an email to uknoch@unimelb.edu.au and you will be sent the Microsoft Word version.

Or, if you prefer, we could telephone you and conduct the survey over Skype. Simply email us your Skype name or regular telephone number, and we will get in touch at a time and date of your choosing.

We plan to present a paper at a language testing conference on our results this spring, and therefore we would be very grateful if you could respond by April 15th if at all possible. Should this not be possible, we would be very grateful if you could let us know by when you would be able to return your responses.

If you have any questions about the survey, the instruments, or anything else relating to this letter, please do not hesitate to get in touch.

Thank you in advance for your cooperation.

Yours sincerely,

J Charles Alderson, MA (Oxon), PhD (Edin)

Professor of Linguistics and English Language
Education, Department of Linguistics and
English Language, Lancaster University,
Lancaster, LA1 4YT

Tel (+ 44) 1524 593029

Ute Knoch, PhD (Auckland)

Research Fellow
Language Testing Research Centre
University of Melbourne, Level 3,
245 Cardigan St
Carlton, Victoria, 3052
Australia
Tel: +61-3-83445206

Skype: charles.alderson

Email: c.alderson@lancaster.ac.uk

Webpage:
http://www.ling.lancs.ac.uk/profiles/284/

Fax: +61-3-83445163
Email: uknoch@unimelb.edu.au

Appendix 7
A36-11: Proficiency in the English language used for radiotelephony communications

Whereas to prevent accidents, ICAO introduced language provisions to ensure that air traffic personnel and pilots are proficient in conducting and comprehending radiotelephony communications in the English language, including requirements that the English language shall be available on request at all stations on the ground serving designated airports and routes used by international air services;

Recognizing that the language provisions reinforce the requirement to use ICAO standardized phraseology in all situations for which it has been specified;

Recognizing that Contracting States have made substantial efforts to comply with the language proficiency requirements by 5 March 2008;

Recognizing that some Contracting States encounter considerable difficulties in implementing the language proficiency requirements including the establishment of language training and testing capabilities;

Recognizing that some Contracting States will require additional time to implement the language proficiency provisions beyond the applicability date;

Whereas in accordance with Article 38 of the Convention any Contracting State which finds it impracticable to comply in all respects with any international standard or procedure is obliged to give immediate notification to ICAO;

Whereas in accordance with Article 39 b) of the Convention any person holding a license not satisfying in full the conditions laid down in the international standard relating to the class of license or certificate held, shall have endorsed on or attached to the license all the particulars in which this person does not satisfy such conditions; and

Whereas pursuant to Article 40 of the Convention no personnel having certificates or licenses so endorsed shall participate in international navigation, except with the permission of the State or States whose territory is entered;

The Assembly:

1. *Urges* the Contracting States to use ICAO standardized phraseology in all situations for which it has been specified;

2. *Directs* the Council to support Contracting States in their implementation of the language proficiency requirements by establishing globally harmonized language testing criteria;

3. *Urges* Contracting States that are not in a position to comply with the language proficiency requirement by the applicability date to post their language proficiency implementation plans including their interim measures to mitigate risk, as required, for pilots, air traffic controllers and aeronautical station operators involved in international operations on the ICAO website as outlined in accordance with the Associated Practices below and ICAO guidance material;

4. *Directs* the Council to provide guidelines to States on the development of implementation plans, including an explanation of the risk mitigation measures so as to enable Contracting States to post their plans as soon as practicable, but prior to 5 March 2008;

5. *Urges* Contracting States to waive the permission requirement under Article 40 of the Convention, in the airspace under their jurisdiction for pilots who do not yet meet the ICAO language proficiency requirements, for a period not exceeding three years after the applicability date of 5 March 2008, provided that the States which issued or rendered valid the licenses have made their implementation plans available to all other Contracting States;

6. *Urges* Contracting States not to restrict their operators, conducting commercial or general aviation operations, from entering the airspace under the jurisdiction or responsibility of other States where air traffic controllers or radio station operators do not yet meet the language proficiency requirements for a period not exceeding three years after the applicability date of 5 March 2008, provided that those States have made their implementation plans available to all other Contracting States;

7. *Urges* Contracting States to provide data concerning their level of implementation of the Language Proficiency Requirements when requested by ICAO;

8. *Requests* the Council to submit to the next ordinary session of the Assembly a report regarding the implementation of the ICAO language proficiency requirements; and

9. *Declares* that ***this resolution supersedes Resolution A32-16*.**

Associated Practices

Contracting States that are not able to meet the language proficiency requirements by 5 March 2008 should:

1. Develop implementation plans for the language proficiency requirements that include the following:

a) a timeline for adoption of the language proficiency requirements in their national regulations;

b) a timeline for establishment of language training and assessment capabilities;

c) a description of a risk based prioritization system for the interim measures to be put in place until full compliance with the language proficiency requirements is achieved;

d) a procedure for endorsing licenses to indicate the holders' language proficiency level; and

e) designation of a national focal point in relation to the English language proficiency implementation plan;

2. Make their language proficiency implementation plans available to all other Contracting States by posting their plans on the ICAO website as soon as practicable, but prior to 5 March 2008;

3. Notify ICAO of differences to the language proficiency Standards and Recommended Practices; and

4. Publish differences to the language proficiency requirements in relation to the provision of air navigation services in their Aeronautical Information Publications.

Appendix 8

Reactions to the Survey and issues raised by correspondents

In addition to responses to our questionnaires, we also received quite a number of emails reacting to our request for information, including some unsolicited emails. These were very interesting, some welcoming our survey or commenting in general on the field, and others addressing specific issues. These have been anonymised.

i) General comments

1)
How relieved I was to hear about your research into aviation English testing.
The ICAO option of leaving such an important issue as this to "market forces" leaves the clients for these tests open to any kind of commercial skullduggery. The ethics and mechanisms of testing, validity, reliability and so on is a rather esoteric world for the uninitiated. In these circumstances, I fear potential test customers may be easily blinded by "science". I am sure we all look forward to finding out what is going on out there.

2)
Absolutely fantastic to have a group of researchers with the knowledge and expertise that you offer, investigating the area of aviation English testing. Will it be possible for anyone to access the results?

3)
Glad to be of help. What you're doing is very much needed

I was very pleased that the Lancaster Language Testing Research Group is conducting a survey on Aviation English Testing. If and when possible, I would be very interested in reports your team will produce following this survey.

ICAO does not and will not produce tests. Unfortunately, ICAO does not have a register of which test the different national authorities are using/approving as such. ICAO has asked States optionally to say which test they were using.
http://www.icao.int/fsix/lp/docs/Guidelines.pdf
and http://www.icao.int/anb/fls/lp/lpcompliance1.cfm.

4)
I am extremely concerned about safety and am worried about the way tests have been used as a means to avoid meaningful change and training. I also feel that some of the tests I have seen have not been constructed by persons with a deep enough understanding of the issues pilots face.

ii) Claims without evidence

5)

In essence, non-testing specialists are making claims for test methods, practices, results that they are not qualified to make. It's such unsubstantiated claims that are in doubt, due to, in most cases, no data in support.

It's clear that many who make claims for their method or approach to testing, and
others who don't make public claims, but are continuing to promote their products
in the industry don't have training in statistical analysis for testing (I certainly don't) by which to process and substantiate their results. …….Nowhere else in the aviation industry is this allowed to happen. …….We are the guardians allowing this to continue.

In aviation English there are only two organizations that have produced and
published substantial data sets in support of their work….. If we don't support such standards, and earnestly spend part of each day either installing such rigor in all we do, or, much more pleasurable, expose those who persist in making unsubstantiated claims, and at the same time ignore irrelevant arguments, then is it any wonder the industry in many key locations remains confused?

iii) Role of ICAO and national aviation authorities

6)
My impression is that if the ICAO doesn't get off its (anatomical reference) *very soon* (next 3-4 months), this is going to develop into a monumental language-testing disaster that we'll all be talking/writing about for decades.

The ICAO has teeth only when (the most powerful of) its member states let it open the box and take them out. All states are sovereign, but some are more sovereign than others.

The point is that the LPRs haven't been given teeth -- and may never have any……
Everyone realizes that currently-working pilots' and ATCs' jobs have to be protected.
Therefore, the new LPRs *really* have to do only with personnel who will be licensed **from now on**. This is why I expect the "teeth" question to surface in the next few years (possibly sooner).

7)
It's beyond the role of XXX, the civil aviation authority of XXX, to certify English language competence of air traffic controllers and pilots in other nations. The aim is to have civil aviation authorities and air navigation service providers (and the airlines) to gain enough knowledge of the issues to be able to prepare and install permanent in-house solutions to the English proficiency issues…..In doing so, they were attempting to set the bar at an appropriate level for the industry, and to provide a counter to the stream of unsupported claims being made about language "tests" from individuals, training and publishing companies, and several national and international aviation testing, training, or regulatory organizations……
Personally, I am pleased to see the contributions that Lancaster and XXX are now making in this regard.

In this highly regulated industry, I have been aware from the outset of the ICAO initiative of a continuing stream of unsupported claims being made in regard to various forms of stature, standards or accreditation for testing.

8)
Although the XXX CAA set up the task force to design the test for aviators that I described in my responses it did not regulate that only this test must be used. Rather, following an old, and still rather common, practice in my country, they specified that any language test approved by the XXX CAA as meeting their requirements can be used to certify English proficiency required by XXX CAA (and based ultimately on ICAOs specifications). Incidentally, the person whose task it is to approve any such tests is the head and convenor of the task force I was / am a member of. I have got the feeling that now that he knows more about the issues in testing he's probably not at all happy with that task but cannot obviously ignore the regulations issued by his superiors in the XXX CAA.

Thus, the situation could be that there were several 'parallel' and official tests of English in use in XXX country, all 'approved' by the XXX CAA and administered by dozens of people all around the country.

Currently, the situation is that the test that our task group designed is used in about half the cases when English proficiency of aviators is tested in XXX country. The second test used is ELPAC, obviously with air traffic controllers. The third test is a modified version of our test (it uses the same speaking test as we do, but the other parts are somewhat different) and is designed by XXX airline. It is used in an integrated fashion in their simulator training programmes. Their test was reviewed by my language teaching / testing colleague in the task force and by the head of the task force, who also approved it officially. That test, then, was not piloted in the same way as our test was, except for the speaking test of course since it comes from our test. So far there haven't been others who have ventured to design their own tests and submitted them for official approval. Not sure if there even will be other tests -- hope not -- since XXX airline is the only really big player in the field and the test we designed probably suffices for the rest of the smaller companies etc.

Also, the total number of authorised persons to administer either our test or the XXX airlines' test is only 7 -- the people behind XXX airline's system participated in our rater training in the autumn. I gather ELPAC is responsible for their own rater training; one of our task force in fact participated in it last year. So, we try to limit the people involved in this kind of testing to those who have at least some background and training in the area, and the tests available to those that have gone through at least some real quality control.

9)
Some countries/organizations will knowingly choose crap tests. Seems to me that for your research to be complete, you should not be putting all the onus on the test providers, but shining a light on the people that select the test for their country/company, and ascertaining their motives. They, as well as the test providers, bear some responsibility. When you choose a gym to become a member of, you don't just look at the glossy marketing pamphlet, but you also do your own appraisal of whether the machines are safe before jumping on the running machine or lying under a benchpress.

10)
(My colleague) got the impression that lots of airlines/countries want to be ICAO-compliant but that their best, most experienced pilots (who are mostly in their fifties) are often only Level 2. These are the pilots whose airplanes are most safe to fly in and who no one wants to sack -- according to the industry insiders. So the focus is slowly shifting to newly graduating pilots - who have had better language training at school - and where it is easier to impose a

Level 4 qualification on them. His take on it was that everyone might adopt a more patient approach for the older, more experienced pilots.

iv) Levels and standards

11)
I can share with you my opinion that many pilots who have been certified at ICAO 4 by the X test are in my judgement closer to ICAO 2. Others here are of the same opinion, and an effort is being made to meet a higher standard, one which meets the real objective of improving safety.

12)
The issue is what any test results mean, however they're arrived at. At present for human rating in aviation English we have nothing more than the assertion from people who score speech samples that they do it well, but there's a lot of evidence that even trained graders disagree.

13)
The prevailing attitude (pressure from airlines, inter alia) is to meet minimum standards *just barely* -- if a grade of C- is passing, then I passed, and that's that. So, when officialdom and the airlines etc look at the LPR scales, what they see is "All I have to do is to figure out what the trick is to assure myself a C-"

14)
The test has been approved by XXX's aviation authority, and it is now one of a small set (of 4, I believe) of tests being used. (As far as I can see, the others don't really test a pilot's capacity to communicate, but rather, infer that capacity from discrete point item (grammar and vocabulary) tests.)

15)
I saw IELTS being flogged at the ICAO Aviation Language Symposium last year.

v) Accreditation

16)
I propose a non-profit accreditation mechanism whereby raters around the world can demonstrate a standard interpretation of the assessment criteria. Of course, the project is fraught with difficulties (issues of fairness, adequate representation, academic issues of the use of various test instruments, the fact that I'm no expert and will need support with the use of FACETS and the IT system for implementation and of course, finance) but I'm willing to give it a shot.

vi) Impact

17)
I disappointed myself (when filling out the questionnaire) by ticking the 'no' category to all too many of the questions. Still, it may be of use to you, and I'm sure will force us to understand where we need to pull our socks up.

18)

Though ICAO has no test endorsement programme at the moment, we feel that XXX test should pass a validation process. At the moment we are looking for an organization in XXX country that would be able to run necessary checks and validate XXX test. We would also like to pass validation process at a well-known organization in an English-speaking country. We do not know whether we can afford the best service in validating, but could you inform us on your requirements, both financial and organizational, to undergo validation with you?

vii) Controversy: OPI vs tests?

19)
I do not conduct testing myself (except informal or placement type) but I know of very fine experienced responsible and well trained language testers who use OPI-like testing in aviation settings to good practical results.

I wonder if the search for a perfect measuring instrument for such an intrinsically complex human attribute as language proficiency is not, to some degree, illusionary? And I rush to say I am not suggesting that we give up the effort to perfect the tools we use.

But, isn't a more workable approach to (a) consider the Rating Scale--any rating scale--a 'guide to good, experienced judgement? and (b) rather than debate ourselves into our respective `psychometric versus judging' corners, perhaps it is more useful that we continue to urge testers towards ethics and responsibility, insofar that testers should be able to document and justify their approach and methods, even if they continue to disagree on some/many of those points?

viii) Teaching and testing

20)
We believe the training provider and the certification test provider should be two separate organizations

ix) Refusals

21)
Thank you for the invitation to participate in the Survey of English Language testing which we are declining.

22)
I have discussed with my partners the convenience to participate in this second part of your survey and we have arrived to the agreement that we can not do it due to the nature of the specific questions included.

x) Need for research

23)
I haven't been involved in aviation language testing for more than a year, nor will I be involved in the foreseeable future. Therefore, I haven't recently given much thinking to the issues you deal with. However, three themes came to mind. It seems to me that the quality of final results will not be proportional to the effort serious people are investing, if these themes are not addressed…... I take the liberty of briefly expressing my views on these matters here:

(1) Although the teams that produced ICAO's scale and manual did a very good job, their results are not good enough. Their knowledge of aviation discourse was not sufficient as a basis to define their objectives, let alone the guidelines they provide. Simply, more basic research on aviation discourse was necessary, and it still is. Maybe an analogy can help. To find certain cures, simple trial and error methods have worked well; obtaining other medicines has required a solid knowledge of physiology and chemistry; some diseases are demanding deeper understanding of biophysics. We will not get the fine aviation tests that are needed if testers don't see what pilots are doing when they communicate, and they can't see that because available accounts do not show deeply enough what happens when pilots and controllers communicate.

(2) Tests should have external validity, i.e. they should have proved to properly test representative samples of the population they are aimed at. Obtaining such samples requires having good sampling frames and, in order to get these, good (administrative) records of testing in real situations are required. I don't see testing agencies producing those records – and I am not sure they would make them available– unless they are required to do so.

(3) Tests should also have internal validity, i.e. they should have proved to properly test what they are supposed to test in experimentally controlled situations. This necessarily requires using small samples, because one wants to ensure that testers are really comparable and because they get easily saturated, not to mention scores of practically insurmountable problems that would appear with large scale experiments. I don't see any thinking in this field on how to get good experiments and good statistical analyses with small samples. People seem to believe that they only have to worry about external validity, and that is not the case; one first has to care about internal validity.

There is, of course, one additional complication…... Airlines and the schools that have been contracted by airlines are major players here. Even when they want to do things well, and most of them probably do, they have to be well informed. Besides, provisions have to be taken in order to prevent the development of unfair competition, and if these provisions don't develop quickly enough, they will prefer to move forward slowly.

xi)
Discrete tests vs integral testing

The ICAO's manual strongly recommends that testers evaluate communicative capacity directly, using integral tests. Its explicit argument is that this capacity cannot be inferred indirectly, from measurements of isolated language skills or from assessments of knowledge about separate language areas. Normally, this would entail that discrete item tests ought to be avoided. But it now appears that some agencies are relying on such tests. One would naturally wonder if they have new evidence that those tests can actually predict the outcome of integral tests. One would also want to know, if either type of test can be used to evaluate communicative capacity, why have discrete item tests been preferred. These worries would lead to another question: would it be advisable to make special efforts to ensure that airlines, national authorities and other major stake-holders understand what is to be evaluated and what is the point of the manual's central recommendations?

Appendix 9
Detailed responses to the main, follow-up questionnaire

SECTION 1. TEST PURPOSE AND SPECIFICATION

1) Have you produced a set of test specifications (test blueprint)?

Yes: 21/22 No Response 1

2) According to the specifications, what is the purpose of the test?

No response 1

- To Assess Operational Aviation English Skills
- Tests plain English language proficiency in communications common to both pilots and controllers in an aviation context. Designed to effectively elicit language assessable by the ICAO band descriptors
- Determine English language proficiency.
- To assess the English language skills of those who are part of the flight deck. The test is not an assessment of aviation knowledge, but a test that focuses on linguistic skills needed to construct meaning.
- There are three sets of specs for different forms of the test: 1) XXX test for Pilots - to assess ATPL pilots for licensing purposes (in response to the ICAO LPRs) 2) XXX test for ATCOs - to assess air traffic controlles for licensing purposes (in response to the ICAO LPRs)
  3) XXX test for Light Aircraft - to assess PPL and CPL pilots for licensing purposes (in response to the ICAO LPRs and other regulatory requirements)
- To assess English language proficiency in the context of aviation in accordance with the ICAO Language Proficiency Rating Scale and its accompanying Holistic Descriptors
- To assess the communicative capability of test takers within an aeronautical environment placing a high emphasis on the operational aspect and the use of language under unexpected situations. It is also assessed the listening ability of the test takers in order to define his situational awareness and to decode relevant information from radio communications.
- Compliance with ICAO English Proficiency Rating
- Compliance with the ICAO English Proficiency Testing for Pilots and ATC
- To accurately and objectively assess the examinee's English speaking and listening ability within an Aviation context.
- This service enables organizations to assess their Pilots' and ATCOs' level of English language proficiency in an aviation context. The assessment evaluates pronunciation, structure, vocabulary, fluency, comprehension and interactions. A consolidated report with recommendations for training is then provided to the organization.
- XXX Test The purpose of this test is to confirm that pilots (aeroplane and helicopter) can demonstrate expert plain language proficiency for safe and efficient aviation radiotelephony communications in accordance with section 1.2.9 (Language Proficiency) to Annex 1 to the Convention on International Civil Aviation. Satisfactory demonstration is necessary for issue of a (XX country) aeroplane or helicopter licence from 05 March 2008.
  XXX test The purpose of this test is to formally evaluate the spoken language proficiency of pilots (aeroplane and helicopter). The test is to be able to discriminate between performances at ICAO Levels 6, 5, 4, and "3 or lower". Demonstration to at least Level 4 (Operational) is necessary for issue of an ICAO (incl.XX country) aeroplane or helicopter licence from 05 March 2008.
- Find out the level of general and aviation-related English proficiency of pilots
- Evaluate applicants for language proficiency in an aviation context. Proficiency test for speaking and listening using the ICAO rating scale and holistic descriptors.
- To determine whether the test taker complies with the minimum ICAO and European Commission

(EC) requirements for language proficiency in English in aeronautical communications between air traffic controllers and pilots, and between controllers and controllers;
To determine whether the test taker meets ICAO and EC language proficiency requirements at Level 4 or Level 5.

- Testing the Aviation English proficiency of licenced pilots and air traffic controlers
- The test measures facility in spoken aviation English and common English used in the aviation domain. That is, the ability to understand spoken English on topics related to aviation (such as movement, position, time, duration, weather, animals, etc.) and aviation phraseology and to be able to respond appropriately in intelligible English at a native-like conversational pace. In addition to covering the strict phraseology found in aviation dialog, the test will also cover common English since in emergency situations vocabulary and structure will not be predictable
- To assess English proficiency level of pilots flying international flights
- The purposes are to evaluate the testee´s capacity to communicate in English in normal and unusual flight situations, and to grade him/her on ICAO's rating scale. (Levels 2 to 5 of this scale are used. It is stated that determining whether a person's level is not 1 is the result of another test that takes place before this, and whether their level is 6 should be the result of a different test, offered to those assigned to level 5 here.)
- Assessment, in accordance with the ICAO LPR Scale, of pilots' and controllers' correspondence to holistic descriptors, as a licensing requirement in fulfilment of ICAO Annex 1 requirements.
- The purpose of the test is to measure the ability to understand spoken English (listening ability) and the ability to produce language (speaking ability) with regard to: i) phraseology used in routine and predictable contexts ii) plain English and phraseology are used in non-routine and less predictable contexts iii) plain English in aviation contexts

3) Who are the stakeholders (those who commissioned the test or who have an interest in it) for this test? Please provide a full list.

- Confidential
- The CAAs, ANSPs, the pilots and ATCs, passengers....
- There are over 6,000 recognising organisations. The test is developed through a partnership of XX:YYY, ZZZ and QQQ.
- XXX airline XXX Civil Aviation Department, YY airline, ZZ airline
- Civil Aviation Affairs (CAA), XX Civil Aviation Authority (CAA)   Civil Aviation Authority of XXX country XX Civil Aviation Department   Civil Aviation Safety Authority, XX country   Directorate General of Civil Aviation (DGCA) Directorate General of Civil Aviation and Meteorology (DGCAM), XXX country, General Civil Aviation Authority (GCAA), XXX country, XX Civil Aviation Authority, XX Air Transport Oversight Authority, XX Transport and Water Management Inspectorate, Civil Aviation Authority XX country, XX Civil Aviation Office, Civil Aviation and Meteorology Authority (CAMA) XX country, General Authority Of Civil Aviation, (GACA), XX counrty Flight Training Organisations in XX country, Airlines ANSPs   Civil Aviation Authorities
- Internal Stakeholders Board of Directors (Decision Making Unit) English for Aviation Department Examinations Unit/Exams Administration Unit Control Air (Private Aviation Consultancy Firm) Software development consultancy   External Stakeholders XX country CAA, Overseas CAAs (Licensing/Regulatory Authorities) ANSPs Airline Operators Private Pilots Aviation Training Academies (Pilot & ATC) Test Takers (ATCOs and Pilots)
- airline pilots, general aviation pilots, air force pilots and air traffic controllers
- DGAC XX country
- XX country Civil aviation authority
- Civil Aviation Authorities and organizations who want their aeronautical personnel to meet the English language requirement of the ICAO. 2. Organizations who want to evaluate the English language skills of t order to apply for a job. 5. Current pilots or air-traffic controllers who want and/or need to take the test to apply or c
- XXX organisation and IATA
- XXX Civil Aviation Authority XXX organisation Training Providers Candidates
- ICAO XX Civil Aviation Authority Task force set up for designing the test Authorised persons who can administer the test Test takers (mainly pilots but also some air traffic controllers; most of the latter take XX test in fact)

- ICAO (International) requirements. Contracting ICAO States – XXX organisation. Flight crew (pilots), Air Traffic Controllers and the safety of the traveling passengers.
- Air Navigation Service Providers (ANSPs) and Civil Aviation Authorities (CAAs) of XXX organisations.
- Pilots, Air traffic controlers, Aviation students, Aviation English teachers, trainers
- The test was developed and funded internally by XXX organisation. Other stakeholders are:

i. XX Aviation Administration– the test was developed under a cooperative research and development agreement
ii.xx organisation– tests are exclusively available through XX organisation. They deliver the tests in secure sites and accredit test-takers based on their test scores

- Airlines and airline pilots flying international flights ii. all other pilots flying international flight iii. XX governme

- The project's general coordinator and the person directing the test's application at present; ii. Myself; iii. XX researcher; iv. Colegio de Pilotos Aviadores, CPA, the organization that commissioned the test's development; iv. Consejo Nacional de Ciencia y Tecnología, a government body that sponsored the development; vi. Dirección General de Aeronáutica Civil, DGCA, XXX aviation authority, who supported the project, as part of their efforts in developing the country's standards for the field; vii. External consultant.
- Transportation Safety Oversight (XXX country) ii. Interstate Aviation Committee iii. Airline Operators iv. Air Navigation Service Providers
- Airlines, Air Navigation Service Providers ii. Pilots and Air Traffic Controllers (test takers) iii. Aviation Authorities and other Regulators (Governmental entities) iv. Station Operators (test takers) v. (XXX test provider) vi. English Language teachers vii. Raters viii. Aviation Associations

4) Which of the following groups of test-takers are specified?

Air Traffic Controllers: Yes 18 No 4
Pilots: Yes 21 No 1
Air Service Personnel: Yes 5 No 11

Please add any groups your test caters for but which we have not listed:

5, as follows

- Safety personal, nurses, doctors, engineers all associated with the aviation industry.
- Cadet Pilots
- Cadet pilots need to be differentiated from experienced pilots. Since XXX test engages background knowledge (flight operations) the test tasks need to be different for existing pilots and cadets (eg. who only fly in VFR conditions).
- There are also some small groups of people who operate on radio frequences that the requirement to take the test applies to (they are people who inform pilots / aviators about e.g. weather conditions) and who need English in their work
- Some air service personnel may be required to complete the test, however this is not covered under our current regulations.

Are there separate specifications (or a separate section within the specifications) for the different groups of test-takers?
Yes 12 No 9 No Response 1

5) Do you specify the test construct (the skills and sub-skills that the test/subtest(s) are intended to measure)?

Yes 21 No 1 as follows:

- The question cannot be answered properly in these terms. The guideline questionnaire reflects an analytic framework (product of research into pilot/controller communication). But its primary result is holistic, and the analyses support the global grade, rather than lead to it. The observer is guided to assess if the testee uses his/her language resources in combination, and different combinations could be equally successful; crucially, high performance in one area does not lead to "adding" more points in the overall score. Furthermore, the framework is a matrix of communication processes and language resources, rather than skills as such, although production and recognition skills are involved all the time. The processes are: situating utterances, taking turns, organizing messages, constructing the typical (three move, bi-strata) cycles of aviation discourse, and using the information that is communicated. The linguistic resources are grouped into four categories: phonological, lexical, syntactical and textual.

6) Do you list and exemplify the test methods/tasks used in the test?

Yes 21 No 1

One expansion, as follows:

- The project's report included the characterisations referred to in the answer to question one, as well as diagrams, photographs and scripts

7). Do you provide examples of test-taker performances?

Yes 15 No 5
Two expansions or explanations, as follows:

- Neither Yes nor No: The project's report included recordings and transcriptions of volunteer test-takers that participated during the project's development. (The recordings are played during training of new testers.) A system was set up to make recordings and observation records (of any testee) available to aviation authorities to verify testees' certificates and to check the test's procedures, and I suppose this system is being maintained; however, I believe these materials will be confidential, i.e. not available to others besides operators, eventual developers and authorities

- Yes: During the Rater Training Course, a range of test-taker performances is described and exemplified.

8) Are the rating criteria available?

Yes 18 No 2

Two expansions or explanations, as follows:

- Neither Yes nor No: The general criteria are available and have been presented publicly. I am unable to tell if specific glosses are

- Yes. Rating criteria is one of the key components in performance assessment. Although scoring of the XXX Test requires the application of the ICAO six-band language proficiency rating scale, we believe that the scale cannot be used as a stand-alone instrument. To assist those who have to apply the criteria, we provide thorough training and guidance material: i) Rater training courses (a six-day course in association with XXX ii) Manual for raters (included in our XXX Test Handbook and available online through the XXX Data Management System) iii) Refresher courses

iv) Re-certification courses v) Rated Speech Samples library (reliability scripts) with detailed rater rationales and commentaries on the candidate's performance.

9a). What ICAO level does the test measure?

No response 2

- Operational Level 4
- All
- level 4,5 & 6 - as the test is uses to gain speech samples from a range of materials

the ICAO descriptors can be used to assess the candidates language skills.
Non-operational levels are ratherly below 3.

- Below level 4, 4, 5 and 6
- XXX test: Ideally 1-5 but also capable of identifying a level 6 YYY test: Confirmation of Level 6 language profi
- From 2 to 6
- ICAO Rating Levels 2 - 3 - 4 - 5 – 6
- ICAO levels 2 – 6
- All 6 levels: Expert Level 6, Extended Level 5, Operational Level 4, Pre-Operational Level 3,

Elementary Level 2, Pre-Elementary Level 1

- Expert 6, Extended 5, Operational 4, Pre-Operational 3, Elementary 2, Pre-elementary 1
- Levels 6,5,4,below 4
- Levels 4 to 6 (focus is on 4-5)
- Expert (6) Operational (combined 4 & 5) Below-Operational (1-3)
- ICAO levels 4 and 5 (and Fail i.e. below 4)
- (3)-4-5-6
- **1** through **6**
- Levels 3 or below, level 4, level 6 (level 5 being considered for near future)
- Levels 2 to 5
- From Level 1 to Level 5 inclusive
- (2), 3, 4, 5, 6. Remark: the trial population didn't include sufficient information to make sure that

the test measures effectively at level 2. New data is now available and is being analysed
as we speak to verify if the test can measure effectively at level 2.

9b) What evidence is provided to support this claim?

No response 1

- Our Compliance Report
- Clearly-expressed rationales to support rating of candidates of all levels.
- Validation testing and research can be provided by XXX and supplied.
- The tests were piloted with a large number of pilots from different nationalities and language abilities to meas
- Test trials conducted with potential target users as well as retrospective analysis of results collected on an onç
- Test research and validation activities cross referenced with all available ICAO SARPS (Doc 9835 AUD0001 anc
- The comparison with the speech samples given by ICAO
- Approval by the XXX DGAC
- The ICAO English Proficiency Rating Scale
- XXX has been invited to several ICAO-sponsored aviation conferences to talk about the XXX test.

2. Content from the XXX test and its administration has been used by the ICAO as training and guidance material for other testing institutions.
3. The XXX test is currently being used by the XXX government as the ICAO certification test for its pilots and air-traffic controllers.

- The assessment evaluates pronunciation, structure, vocabulary, fluency, comprehension

and interactions in accordance to the ICAO Language Proficiency Scale.

- Audit by XXX person
- Judgements of the test design group - results of the first test round (trial round)

- comparison of test results with test takers' self-assessments against the ICAO scale and their background information - comparison of the results of the 3 main subtests
- the fact that the rating of speaking is done directly against the ICAO scales

- ICAO Rating Scale & Holistic Descriptors
- Trialling and standard setting of test items.
- Testing of a large number of candidates
- A complete Validation Manual is publicly available (scoring is described on page 19)
- Criteria used to assess proficiency level is based on ICAO's specifications
- The test's level specifications correspond to ICAO's levels. 2) During the development of the project, an experiment was set up to see whether there was high agreement among different volunteer testers using the questionnaire. 3) The ranking of testees resulting from the testers' assessments was the same as a previous ranking defined by the test developers.
- XXX test, in accordance with ICAO Doc 9835, is a proficiency test of speaking and listening; it is based on the ICAO Rating Scale and holistic descriptors; it tests speaking and listening proficiency in a context appropriate to aviation; and it tests language use in a broader context than in the use of ICAO phraseologies alone
- The test was designed to measure at levels 1, 2, 3, 4, 5 and 6. However, our test population didn't include enough pre-elementary test takers (ICAO Level 1) and elementary (ICAO Level 2). Therefore, we claim only to measure effectively at levels 3, 4, 5 and 6.
This claim is supported mostly by several concurrent validation studies that have been performed.

SECTION 2. TEST DESIGN and ITEM WRITING

1. What professional experience do you require of your <u>test developers</u> (e.g. air traffic control/ pilot experience and/or English as a Foreign Language teaching experience)?

- Confidential, but does not include English Teaching Experience
- EFL Teachers: experience of Aviation English Teaching, other test delivery and other test trialling experience (we are also an IELTS Test Centre). Pilots and ATCs:
no extra experience required as simply guiding test development team with authenticity of items
- No professional experience required. The test is designed for student and professional test takers. It is then referred to by admissions centers at universities, and professional organizations such as xxx councils, and is also used by immigration in (four countries) for visa purposes.
- A Masters in linguistics or Education (Focus on Language), Flight Training Manager.
- air traffic control/ pilot experience and/or English as a Foreign Language teaching experienced
- Masters in Applied Linguistics - Language testing Curriculum writing experience (proven track record)
- Language: Post Graduate applied linguistic qualification, test development experience and Aptitude/competence SME: ATC/pilot training and experience
- Our test developers are experienced flight instructors who are also specialists and researchers in the field of English language teaching
- The test has been completed, approved by the XXX DGAC and in operation since 01/06/07
- Experience teaching ESL to pilots; aeronautical engineers; general aviation service providers and the general public.
- The minimum requirements are: For all: 1. Certification in ICAO Radiotelephony Procedures and Language or equivalent. For aviation specialists: 1. Member in good standing of an officially sanctioned aviation organization. 2. 5 years experience as a pilot or air-traffic controller. 3. Must pass an English proficiency test conducted by an independent company (TOEFL, IELTS, etc.)
For linguistic specialists: 1. TESL or TEFL certification or equivalent. 2. Bachelor's degree in English, Education, or P
- The test was developed in conjunction with the University of XXX Department of Education,

XXX and XXX organisations and SME's.

- ATC and Pilot Experience in collaboration with English as a foreign language teaching and testing expertise
- See a separate account of the role of this particular test in the context of the whole testing system (there are currently 3 tests that can be used for this purpose). For this test, a task force was set up that consisted of one experienced language tester (myself), an experienced language teacher who also had long experience in item writing, and 3 aviation experts, one of whom had long experience in testing aviation English (with instruments developed by him/colleagues). The plan is that the same group continues work and designs new test versions in the coming years; at the same time, the aviation experts are trained on language testing. So new developers & item writers are introduced into the team only very slowly and carefully.
- Test Development Team – XXX - included team members with pedagogical experience in X and Y language testing experience. Team members included experienced pilot Air Traffic Controller and Flight Service Station Radio Operator.
- Test developers are provided on a part-time basis and allocated by their ANSP. We have little control over their selection although as a minimum we ask for ATC instructors and/or operational controllers with ICAO level 5/6 and for XXX test we require them to have aviation English language teaching experience (experience in test design desirable).
- PhD in Test Development and Validation MA in Applied Linguistics and Graduate Certificate in Language Assessment. Pilots Language Teachers Experienced L2 raters
- The test development team consisted of spoken language testing experts and experts in aviation language training and testing, all of who had higher degrees in a relevant field. Their qualifications are listed on pages 16 and 17 of the validation manual
- Test developers are applied linguists, specializing in language teaching, SLA and language testing
- The general coordinator of the team that developed the test and carried out the research this was based on is a pilot with ample experience. He also has an M.A. degree in anthropology, and to obtain it he wrote a thesis on aviation communication that focused on accidents that occurred due to communication problems. I, the academic coordinator, have an M.Sc. in applied linguistics and a Ph.D. in education. I have had experience in test development. One other researcher participated throughout the development. She has experience in the Teaching of English as a Foreign Language and an M.A. degree in applied linguistics. During the first stages of the research and development project, an applied linguistics student collaborated in some tasks
- University philological education (English language); ab initio course in radiotelephony; at least 10 years experience as a teacher of Aviation English or Radiotelephony; experience in developing progress tests in Aviation English and Radiotelephony; an ab initio course for raters; familiarity with research on GE and ESP testing; good knowledge of controllers and pilots jobs, ICAO documents regulating radio communication
- XXX company has been providing training to pilots, air traffic controllers, maintenance license engineers, English language teachers ... for many years. We have also developed the following courseware: i) Pilot training: ATPL, CPL, IR and PPL courseware (aerodynamics, flight performance, air law, operational procedures ....) in accordance with the existing CAA, EASA and ICAO regulations.
ii) Maintenance license training (engines, airworthiness, electrics, ...) in accordance with the existing EASA regulations (e.g. part-66 and part-147)
iii) Language training for pilots and air traffic controllers (for all levels) in accordance with the existing CAA regulations and the ICAO SARPS.Unlike many other test providers, our test development team consists of both language experts and subject matter experts (pilots, air traffic controllers, aerospaceengineers, ground instructors, flight instructors ...).
Note: We have worked together with the University of XXX to design and develop the test (e.g. development of an item selection algorithm, expert calibration of part I test items, assessing speaking and listening separately, validation plan ...).
The test development team includes: Master in Linguistics: 7; Master in Applied Linguistics: 3; Aviation English Experts; Subject Matter Experts: 14; Aerospace engineer: 1; Pilot: 5; Air Traffic Controller: 3

2. *Do you provide training in test design to your <u>test developers?</u>*

Yes 12 No 9 No response 1

3. What professional experience do you require of your <u>item writers</u> (e.g. air traffic control/ pilot experience and/or English as a Foreign Language teaching experience)?

No response 1

- EFL Teachers: experience of Aviation English Teaching, other test delivery and other test trialling experience (we are also an XXX Test Centre). Pilots and ATCs: no extra experience required as simply guiding test development team with authenticity of items
- We have a high level of experience and training through successful completion of accredited TESOL courses, master of linguistics, bachelor of languages and various other requirements including years of teaching and training experience.
- Background in assessments, and assessment development (3 years minimum). Knowledge of Aviation procedures (5 years at least).
- air traffic control/ pilot experience and/or English as a Foreign Language teaching experienced
- ESL teaching experience and quals Curriculum writing experience Language test writing expertise SMEs - both ATCOs and pilots work with the test writers
- Language: Post Graduate applied linguistic qualification, R/T familiarisation, item writing experience/aptitude and competence SME: ATC/pilot training and experience
- Idem as 1
- The test has been completed, approved by the XXX DGAC and in operation since 01/06/07
- Advanced degree in Curriculum and Assessment Vast experience with ESL Aviation English General Foreign Language teaching experience
- The minimum requirements are: For all: 1. Certification in ICAO Radiotelephony Procedures and Language or equivalent. For aviation specialists: 1. Member in good standing of an officially sanctioned aviation organization. 2. 5 years experience as a pilot or air-traffic controller. 3. Must pass an English proficiency test conducted by an independent company (TOEFL, IELTS, etc.) For linguistic sp independent company or test. 4. 1 year English teaching experience.
- Subject Matter Expert's (ie. Pilot Experience and Air Traffic Controllers) and the University of XXX Department of Education
- Formal qualifications and experience both required
- see previous answer
- Extensive experience for the aviation team members At least 5 years pedagogical experience for the linguists
- See 1. above for ATC and ELE.
- MUST have experience in Aviation industry AND applied linguistics and language testing
- There are two kinds of items: common English items and aviation English items. Common English items were developed internally by XXX company by linguists with Masters degrees and/or PhDs and language testing experience. Aviation English items were developed at XXX University by aviation language trainers with MAs and/or PhDs in Applied Linguistics or Language Testing
- Item writers are present or former airline pilots with experience in international flights
- There are no items in the test. If new simulation scenarios are developed, they should be designed by interdisciplinary teams of the same sort that produced the basic specifications and the model scenario (though not necessarily so highly qualified if the specifications are followed.) The guideline questionnaire does not require any modifications to be used in the observation of communications in new scenarios. (If it is to be improved, any changes should be based on extensive research.)
- The same as for test developers
- The XXX Test items have been written and reviewed by dedicated members of the test development team (Item Writer meetings). This team includes at least: - A language expert - An Aviation English Expert - Two SMEs - A pilot and/or an air traffic controller - An XXX Test Rater

4. Do you provide training in item writing to your <u>item writers</u>?

Yes 10 No 11 No response 1

5. Do you have guidelines for test design and item writing?

Yes 19 No response 1 Other 2, as follows:

- (See answer to questions 1.1. and 2.3.)
- We don't have guidelines for test design but we do have guidelines for item writing

6. Do you have systematic procedures for review and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines?

Yes 18 No 2 No response 1 Other, as follows:

- XXX company does not outsource item-writing to item writers. At the Item Writer meetings, items are developed, reviewed and edited.

7. Do item writers receive feedback on their work?

Yes 16 No 2 No response 1 Expansions as follows:

- Yes. At least one XXX Test Rater is present during the Item Writer meetings. The role of the Rater is to provide feedback and to assist the item writers in developing and reviewing new items.
- Yes: See page 17 of the Validation Manual for the qualifications of the item reviewers
- No: There are no items in the test

8. Have you conducted an investigation of test-takers' needs?

Yes 14 No 5 No response 1 Explanations as follows:

- The question is unclear to us. If you mean whether or not we have conducted face validity studies, the answer is yes
- We carried out extensive research on how language is used in pilot/controller communications

9. Have you carried out a survey of how language is used in the context of aviation?

Yes 11 No 7 No response 1 Expansions and explanations as follows:

- Unlike many other providers of AET, our company is specialised in aviation training. Our main customers are airlines, pilots, air navigation service providers, air traffic controllers, aviation authorities, and aerospace universities. Based on our operational experience, and based on the existing and available literature, we didn't feel the need to carry out such a survey. How language is used in the context of aviation is part of our core business.

- We carried out extensive research on how language is used in pilot/controller communications

- Yes: Test development was guided by ongoing studies on the aviation language use domain at XXX University The language and functional content of the test was sampled using the XXX corpus, consisting of radiotelephony transcripts (235,000 words), aviation standard publications (338,000 words) and general aviation publications (258,000 words).

- No. *Informally, yes, through extensive discussions with airline pilots.

SECTION 3. RATING PROCEDURES

1. Are the speaking parts of the examination scored by computer?

Yes 1 No 18 No response 3

2. Is the computer software trained for each test administration?

No response 21 Other:

- The computerized scoring models were specially trained on each item in the item pool that was developed for the XXX test. Since each test administration draws its items from the item pool, in a sense it has been trained for every test administration.

3. Are human raters (of the speaking tests) trained for each test administration?

Yes 5 No 1 No response 16

4. Are exemplar performances (benchmarks) used in the training?

Yes 7 No response 5 Expansions on Yes, as follows:

- All raters are benchmarked and assessed using authentic recordings of candidates over a range of levels.
- YES – of course
- ICAO Rating samples formed the basis for on-going English Proficiency Rating
- Yes. Trainees are required to listen to pre-rated speech samples and to successfully assign the correct level to all of them in order to pass the training course. 'Up-training' and calibration sessions are also periodically conducted to ensure consistency and validity of the ratings assigned.
- Yes, recordings
- Yes (the ICAO supplied samples)
- Yes, the ICAO benchmarks and some locally produced exemplars produced during the design & trialling stage.
- Yes. Assessment of standard set performances are use for both rater/assessor and interlocutor training.
- Yes. Raters are given benchmark audio samples which familiarize raters how performance might be levelled. As some degree of variability in scoring of different elements of ICAO LPR Scale is unavoidable, raters who rate higher or lower than the others are given more detailed training until inter-rater overall results become the same. In case underscoring or overscoring continues, the rater is dismissed

Expansion on No, as follows:

- No. In order to train our scoring models, human raters scored test-takers' performances during the data collection and development phase. The raters were not standardized or trained on a benchmark set. Rather, a select but diverse group of highly qualified raters (experienced in language assessment and aviation language training), who were familiar with the ICAO scales, were asked to rate performances using the ICAO scales (see pages 19 and

31of the Validation manual). However, recordings were extracted from the official ICAO Rated Speech Samples audios and made available to all raters. It was suggested that they listen to those files before starting the project.

In general, we took the view that they were the experts and it was not appropriate for another "expert" to tell them how to interpret the ICAO scales or test takers' performances. (It's an interesting theoretical debate: standardisation is good for reliability, but it may also lead to artificiality or even defining the construct in ways not intended by the ICAO criteria)

5a. Is there routine double-marking for subjectively marked tests?

Yes 13 No 5 No response 1 Expansions as follows:

- Yes: Each performance was independently rated at least twice
- Yes. Rating is 'blind' (without knowing whose test recording it is) and independent (without knowing how the other rater scored the sample).
- Yes. Normally, there are two independent ratings per test. If the raters disagree, a third rating can be requested (the XX system produces a discrepancy flag). If a client decides that only one rating is required (based on a decision taken by the local Aviation Authorities), we activate the sampling program. This means that the XX system automatically selects a number of tests at random for double marking.

5b. What action do you take in the event of disagreement between raters?

No responses 4

- In cases of rater disagreement, test referred to senior rater.
- Review process globally with multiple marking and blind marking of tests to check results.
- If there are two ratings per test, then a third rating is carried out. If there is only one rating per test, double markings are carried out at random (the XX system has a built-in sampling program)
- Markers discuss the grades - each marker puts forward their case for that given mark. A third party then decides which mark they take.
- We deploy a third rater (senior rater)
- Further rating by a minimum of two other trained assessors until a consensus is reached.
- Send conflict to arbitrator Have a three-way meeting to resolve differences
- Written evidence of rating presented to Arbitrator Joint meeting of raters and arbitrator
- 1) Raters will discuss the item or sample in question and try to determine the appropriate score to be assigned
2) Should the raters still fail to agree, the item or sample in question will be forwarded to the quality control group for assessment and final rating.
- Relisten to the recordings. If there is no agreement, a reassessment takes place.
- Third (senior)required to review and decide. discrepant rater recalibrated
- Double marking is encouraged, especially if the level awarded is around 4. It was not possible to require double marking because the national aviation authority responsible for commissioning the development of the test considers that it is not necessary. The reason for this is that the test flight, either, is not done by two inspectors but only one. So why should language assessment be different, they argued. At this stage, we have trained about 80 aviators to eventually function as language testers authorised to administer this (or some other authorised) test. However, the first round of assessments (i.e. all new tests of the aviators who take the test for the first time) is limited to seven persons, most of which are members of the task force that designed this test. This aims at maximising consistency of the ratings. Level 4 result is valid for 3 years, so it is expected that the number of testers / raters will increase only after the first 3 years. All speaking tests must be recorded for voluntary double rating or for later inspection or in case of a complaint / request for re-rating, and be stored for 6 years.
- Review process includes provisions for subsequent completion of a independant test and review.
- Recorded performances go to a trained third assessor for a final result. No prior information about

the candidate's performance is provided.
Ratings were subject to FACETS analysis for building scoring models. Responses that elicited unusually high disagreement were additionally given to a third rater.

- A trainer (third rater) is called on
- A third rating is required
- Senior rater adjudicates the rating. The recording is discussed at a monthly refresher class for raters

6a. Is inter-rater reliability calculated?

Yes 16 No 3 No response 2 One explanation, as follows:

- After the experiment conducted during the project's development, a statistical test for small samples was devised along the lines of Fisher's permutation test. We required that three testers agreed on assigning testees above or below the critical level 3/level 4 dividing line on at least four out of five cases. This requirement was met. The probability that this could happen by mere chance is less than 3/1000. We also required that any two (of the three) testers never differed by more than one level on their assignments of the five testees (on the 2-5 level scale). This requirement was also met. It was indicated that similar tests should be carried out when the exam was implemented regularly (or that other more commonly used ones be established when larger populations of results were available). I am unaware if this has been done so far (the test has been in use for less than a year)

  If YES, please provide details of the correlations achieved.

- Pearson Correlation Coefficient - 0.82 Average Absolute Difference - 0.2
- By giving raters a number of samples to mark and measuring the levels of deviation between grades. One level is acceptable, but deviations 2 or more would mean an marking evaluation for a rater who has deviated. A marker who has marked high or low will have training in that area to help align them to the ICAO descriptors.
- We have a database that collects and stores all data. Reports are routinely generated to allow correlations to be computed. As low as .7 but threshold for acceptance is above .8 (however this is complicated by the fact that raters whose scores are overriden are omitted from the data sheet).
- The concordance between raters is assessed by means of the Pearson Correlation Co-efficient (range -1/+1). According to most recent research (Jan 2008) Final Rating: Pearson (r) coefficient of 0.91
Individual 6 features of language ratings: Pearson (r) coefficient of 0.82
- Test-takers are rated by outside organizations. Level of correlation acceptable to the XXX DGAC
- Comparisons by the DGAC of test-takers evaluated by other outside authorized test - providers.
- The exact figures are confidential. However, we can say that the assessments have revealed a high degree of inter-rater reliability over a 1-year period.
- The standards are set in place in the training, throughout, spot checks are done and team meetings take place to ensure consistency.
- Too early yet. XXX country testing not mandated until 8 May 08
- In the first training sessions (2 days per session) in late 2007, a total of 80 aviators were trained to function as test administrators. Main focus was on acting as the interviewer and rater in the speaking test. As part of training, they rated ICAO and other sample performances. These ratings were collected but not yet analysed; the plan is to provide each individual tester with feedback on his/her rating based on these data. I'm happy to send you the results once I've calculated them but that won't be until in the autumn. The 80 aviators also took the test described here. The speaking performance were recorded and rated on the spot by 2 members of the task force, taking turns as the interviewer and rater. Any disagreements and borderline cases (about 20%) were afterwards reviewed by the whole task force and a final decision was made. Data on these rating is in principle available but has not been calculated, as the main aim was to create a common view of the levels in the task force for the future ratings.
- Inter-rater reliability - testing started only recently and there is insufficient data available at the moment.

This issue will be discussed at the XXX User Group meeting in April 2008 so that we have enough data for the first refresher workshop for examiners in Autumn 2008.

- Inter-rater reliability varies according to rater pairing and task, but as a composite estimate, inter-rater reliability is in the region of r= .75 to .8, Please note, however, that this coefficient should not be compared to inter-rater reliability in live interview tests for theoretical reasons. In live interview tests, raters usually give one rating on each subskill (i.e. six ratings in total) based on everything they hear from the test taker, and therefore it is very important for the two raters to agree. In XXX tests, there are discrete ratings on each response, and multiple responses. An accurate assessment of proficiency is constructed over responses to multiple items, for example, Pronunciation is evaluated on over 44 responses. For this reason, although it seems unconventional at first, split-half reliability is actually a more useful estimation of rating reliability and a more responsible coefficient to report (see 4.7 below)
- Confidential information
- Inter-rater reliability calculations are continuously performed. These calculations are built-in into the XX system. Senior raters have access to the inter-rater reliability calculations through the Quality module. The results are presented in the form of a Kappa Statistic, a Spearman Rank coefficient … Because of our rater standardization program, and because our raters are continuously monitored, we achieve high correlations.

6b. Is intra-rater reliability calculated?

Yes 13 No 6 No response 2 Other 1 (does not know)

IF YES, please provide details of the correlations achieved.

- Pearson Correlation Coefficient - 0.94 Average Absolute Difference - 0.13
- Assessed by means of the Pearson Correlation Co-efficient. According to most recent research Final Rating: Pearson (r) coefficient of 0.93 Individual 6 features of language ratings: Pearson (r) coefficient of 0.82
- 95% correlation achieved within ICAO rating levels
- I interpret this to mean the raters who work only with our test. Raters correlation averages greater than 90%. Greatest number of differences is found in the ICAO categories 3+ vs. 4-
- The exact figures are confidential. However, we can say that the assessments have revealed a high degree of intra-rater reliability over a 1-year period.
- Yes, rater conditions are consistent, a break is taken between each assessment to ensure their ear does not become accustomed to the candidates' accent.
- Intra-rater reliability calculations are continuously performed (not only on the trial sample). These calculations are built-in into the XX system (Quality module). The system makes sure that some tests are double marked by the same rater within a certain time period (raters are usually unaware of this). The results are available to the senior raters through the Quality Module.
- Intra-rater reliability is not calculated per se but raters are recommended to re-rate some performanaces later and see whether they awarded the same level. All differences should be discussed at a meeting with other raters in the respective country (testing centre) and reported to XXX organisation by the National Administrator (who has to submit a report on testing to XXX every three months). Note: during trialling we conducted intra-rater reliability and in most cases achieved a figure for reliability of 0.8.
- Intra-rater reliability was generally higher than inter-rater reliability. Raters were subject to an overlap of between 3% and 5% on all the responses they were presented, allowing for the monitoring of intra-rater consistency. It should also be noted that raters had to complete a training set for internal consistency/standardization prior to beginning the project
- Confidential information

7. Is the marking routinely monitored?

Yes 14 Unable to respond 1 No 3 No response 2 Expansions, as follows:

- Yes. In case there were no remarks during the first 3 months or 10 ratings/deliveries (whichever comes first), afterwards at least 10 per cent of what testers and raters do are monitored by senior raters.
- Yes Routine monitoring of test scores is facilitated by the XXX system. Monitoring takes place at three different levels:
  - The Training Center Administrator - The System Administrator - Senior raters (for standardisation purposes)

8. How is the pass mark for objectively scored tests determined?

No response 4 N/A 1 Other responses as follows:

- Profile-marked (ICAO Descriptors).
- Determined by the using organization. XXX is not a pass or fail test, it is a global indicator of proficiency that is then used by organizations to accept or reject candidates on their ability in English.
- Using sections in the test that focus on specific areas of language. Each section requires a candidate to employ a particular language ability.
- This is a huge topic. Basically comparisons are made (correlations, scatter-plot graphs, mean score comparisons - ANOVA etc) are made with external tests on lead base versions. These then become the anchor versions for subsquent developent. The Speaking and Listening XXX tests are correlated, and since the Listening test is objectively scorerd, the cut scores are determined with reference to an externally validated listening test (against XXX speaking), then this against XXX Listening. This establishes the ICAO band scores on XXX Listening.
- According to the ICAO rating scale
- Based on the holistic descriptors outlined in the ICAO English Proficiency Rating Scale.
- Not applicable to the XXX test. The XXX is a test of English speaking and listening ability, and not knowledge of aviation phraseology or procedures. As such, there are no right or wrong responses. Instead, it is the way the examinees deliver the responses that is rated.
- They need an Operational 4, raters consistently use the ICAO scale as the measurement tool. The structure of the test is consistent -- 1. verification; warmup; rated conversation; 4. wrap up. Questions are always asked according to candaidates level according to the ICAO scale.
- All tests are listened to by a rater and subjectively scored using ICAO holistic descriptors guide
- The task force members evaluated the difficulty of the test items for listening and reading tests against the ICAO levels. The results of the first test administration with the 80 aviators were considered together with these judgements to arrive at the best cut-offs for the levels. So, it is a kind of standard setting exercise but probably not as rigorous as that recommended in e.g. the Supplement to the CEFR Manual, due to lack of resources & a quite ambitious timetable imposed on the task force by the Aviation Authority.
- Sucessful completion of a minumum number of test items.
- An algorithm calculates the ICAO level (4 or 5) based on a 70% passmark. This passmark has been determined by a group of independent (standard setting) experts.
- or what is the pass mark? Question not clear.
- The pass mark for the listening comprehension test is set at 70%
- The test is objective in the sense that its results are the product of the guideline questionnaire, i.e. in the sense that raters focus on the same communication processes and use the same criteria. But it is not objective in the usual sense of the term, i.e. it does not consist of questions with closed sets of possible answers
- Part I of the test (this part is objectively scored) is computer adaptive.

SECTION 4. TEST ANALYSES

1. Are the tests trialled?

Yes 19 No response 3

2a. What is the normal size of the trial sample?

No response 7

- XX test has a 2 year production process on test development which includes extensive global pre-testing. Trial sample is in the thousands on a regular (monthly) basis.
- 10 - taken from a range of people with a known rating.
- Sometimes as low as 50 (due to availablity) but normally 80+
- Initially, whole tests were trialled, with the typical trial sample being between 20 and 40 candidates. Since the test has become operational, trial items are embedded in the live test without the candidates final score being either adversely or positively affected.
- The tests were trialed before approval was given for the final version.
- minimum 20, maximum 40.
- Yes, test trials were done when the test was developed in 2005. Participants included SME's, raters, pilots, controllers and professors from the Univ. of XXX
- There is no normal size for the trial sample, as the system has just been created and exact procedures / decisions on how to create new versions of the test in the future have not been made. There was a small scale trial with 6 people when the current test was created; the timetable and resources allocated to the project made more extensive trialling impossible. However, because of that, the first administration of the test during the training of the 80 aviators in autumn 2007 was made as pilot like as possible. The cut-off points for the comprehension tests were decided only after that. Fortunately, the small scale trialling and the experience of the task force in item writing resulted in no apparent / serious problems with the items (had that occurred, the items in question would have not been taken into account when computing the final score). The procedure for ensuring rater consistency in the speaking test during the testing of the 80 aviators was described earlier.
- 130 E 100 F
- This question is unclear to us (especially the word 'normal'). Do you want to find out how big the trial population was? Our system allows us to select tests (key tests) and ratings (key ratings) for validation purposes. These tests and ratings are then used for test analysis. Today, there are over 200 key tests
- 200 controllers - a total of more than 700 controllers. Since the release of the test more than 1100 controllers have completed the XXX test. All test takers are encouraged to participate to the trialling of new items.
- Trialling involved 478 native speakers and 628 non-native speakers, see page 18 of the Validation Manual
- New listening comprehension items are included in the listening comprehension test as trial items and are not used for assessment. (Sample size, therefore, depends on the size of the test taking sample). However, for the national test, since all the items must be opened right after the test is administered, it is virtually impossible to pre-test new items. All items are included in the assessment. Statistical checks on tested items are sometimes conducted, however.
- During its development, the test was applied to 20 pilots. During the validation experiment, it was applied to 5 pilots (see answer to question 3.6a.)
- It corresponds to the regular XXX test format

2b. How big is the trial sample compared with the test population?

No response 7

- Globally, impossible to say.
- We conduct 1 million tests a year and the trial sample is between 5 - 8% of this.
- 2.5% (Approx 400 tests per year)
- Not sure what this is asking. The current test bank is comprised of 10 base versions (unique versions) which a

then recompiled to form the test bank. Each base version contains 3 sections which must be used in their entirety. Currently there are 72 versions in the XXX test bank. So far the test has been administered to 8000+ pilots. Each base version is trialled prior to release on different groups of 50-80+ pilots. Then analysis is performed retrospectively on live versions (especially Listening).

- 5-10%
- The trial sample was 25 pilots, ATC and DGAC personnnel
- The exact figure is confidential and varies with the needs of our clients.
- 5%
- The trial sample was thus 6 + about 80. The test population is a few thousand (I must check the exact figure).
- 18 countries participated to the initial trialling phase (prior test release)with a total controller population of approximately 4700: 15% representative of population. Current trialling (see above) is about 95% for Paper 1.
- We don't know the test population yet
- A difficult question as we don't yet know how many test-takers there will be in any given year. Page 19 of the Validation Manual gives the demographic information of the trial test takers and this shows that they represent a suitably large sample
- For the airline test, all test takers provide information about the trial test items. For the national test, there are no true trial items
- The universe wasn't determined during the project's development, since the aviation authorities had not defined which tests would be approved for which sectors of pilots
- About 20%.

3. *Is the following information collected during trialling?*

a) test-takers' views on the difficulty of the test

Yes 15 No 3 No response 2 + Explanations as follows:

- Yes. This information was not collected for every single test-taker, invigilator, or examiner. Rather, a sample of stakeholders and test-takers were formally involved in review and feedback. Other test-takers' views on difficulty/appropriacy were not collected separately, but were apparent through their responses: item difficulty estimates indicated whether they found items/tasks difficult or easy; and their ability to answer the items correctly helped indicate whether they were appropriate.

- No. Only impressionistic information was gathered on these matters. (We proposed that that a questionnaire should be set up and applied to a group of testees that volunteered to take two of the tests that the authorities have approved, or to comparable samples of one population, if it can be defined properly. It is not clear if this will be possible.)

b) test-takers' views on the appropriacy of test tasks

Yes 17 No 2 No Response 3

c) invigilators' views on the difficulty of the test

Yes 14 No 5 No Response 3

d) invigilators' views on the appropriacy of test tasks

Yes 14 No 5 No Response 3

e) examiners' views on the difficulty of the test

Yes 16 No 3 No Response 3

f) examiners' views on the appropriacy of test tasks

Yes 16 No 3 No Response 3

4. How is the trial data analysed?

No Response 6

- Observational Checklists. Comparative analysis to previous candidate performances.
- Results of all pre-testing are sent back to XXX organisation for review and analysis through a department within XXX.
- The original tests used L2 speakers of English who had scores from other established English language tests. The results were compared with ability levels in other tests to measure the consistency. Presently, scores were compared with those of the original tests to see if the new tests have any weaknesses. The results are also compared to the ICAO requirements for testing and the descriptors to make sure any new test follows the format agreed with XX CAD.
- For quantitative analysis: Construct validation: statistical tools available in SPSS. Correlations, ANOVA, T-tests Feedback sessions are held and surveys are collected from interlocutors/raters. This is then interpretted by the XXX test manager before to allow decisions to be fed through to the review/re-writing of the test (versions) with the test writers. This is the process in use for each new version. The process was more vigorous and lengthy when the initial protypes of each form of XXX test was developed (eg. in 2004 for pilots). Extensive data was collected from: NNS pilots, NS pilots, test writers, raters and interlocutors. Reviews were conducted on perceptions across different nationalities and ages of prospective test users. Two theses were written by two staff members on the intial perceptions of the test by users and the data for construct validation of a prototype of the test.
- empirical and statistical analysis
- Analysis done by the XXX DGAC
- Through statistical analysis.
- Listened to; meetings; people took the tests; recorded; analysed; observed
- Evaluated by PhD qualified experienced English language test constructor
- Classical item analysis in the first instance but I'll do an IRT analysis, too, before we start designing new versions of this test later this year.
- Test-takers, markers, administrators and examiners are asked to grade statements using numbers 1 – 6 (strongly disagree, disagree, slightly disagree, partly agree, agree, strongly agree) and then add comments. We use SPSS (for graded statements) and MS Excel (for comments) to analyse all data. Results are Reported to the National Admins (reports sent by e-mail).
- Qualitatively & Quantitatively

Qualitatively: Test-takers' responses were transcribed and compared to expected responses. Also, items that many test-takers could not answer correctly were reviewed again to ensure those recorded items were of good quality.

Quantitatively: Rasch modelling - Item difficulty, discrimination, fit and analysis of item appropriateness based on native test-taker performances.
Moreover, statistical analysis involves modelling native speaker and non-native speaker answers for manner of speech (such as pace, timing, phrasing, pausing) and content of speech (such as sequence of words spoken keywords in the answer choice).

- (See answer to question 6a.)
- Versions under trial are offered to a focus group that has passed XXX test already a short time ago. Results of a trialled version are compared with previous results. In case the majority of the group shows higher or lower proficiency in at least one of the elements, the problematic fields in the test tasks are identified and corrected.
- Through statistical analysis and internal meetings

5. *How are changes to the test agreed upon after the analyses of the evidence collected in the trial?*

No response 4

- Test Development meeting - items re-written if necessary and re-trialled.
- The test production has several check and review levels which is why the test takes 2 years to produce for each test day. This is then multiplied 48 times across the 48 tests that are conducted each year.
- Through discussion with XXX CAD.
- See above
- Subsequent to discussions between internal stakeholders
- Mutual agreement between the XXX DGAC and the test designer.
- By consensus.
- Observation; trial and error
- Test Constructor and Evaluator discuss and agree changes which are then trialled again.
- They were / will be discussed in the task group that is responsible for the test.
- Review panel
- Based on data analysis, discussion with development team and guided by consultant.
- Low discriminability and advice from experienced ATCs and pilots
- Decisions were made on the basis of empirical data analysis, which in turn was guided by defined parameters. For example, the discrimination index was fixed at above.30 for each item, otherwise it would be deleted. Also, 80% of native speakers had to be able to answer an item correctly for it to be acceptable. Similarly, decisions such as the number of items required in the test, or number of items of each item-type, were dictated by data such as internal reliability and correlation and with human ratings
- Test developers at provide statistical information and information about relevance of wording of choices, etc. for items writers to change
- I am unable to answer this question, since I haven't participated in maintaining the test after it was developec
- Necessary changes are introduced after they have been discussed by linguists and SMEs of the group of developers and trialled for non-licensing purposes with aviation English students
- During the Test and Item Development Meetings (TIDM) new and existing test items are edited and reviewed

6. *Are there different versions of the test (e.g. year by year)?*

Yes 16 No 4 No response 2 Explanations as follows:

- I understand different scenarios (with different trajectories and different flight parameters have been developed). I have also learned that practical improvements to the way these are set up (physical arrangement of the rooms, computer displays and communication channels) have taken place. I don't have any specific information on these matters

IF YES, how is the equivalence verified?

No response 2

- Results of observational checklists - evidence of frequency of language functions elicited during test
- Key questions are seeded in the pre-test process. Review of results and feedback from the pre-testing and test development process cross check this data.
- Using volunteer Candidates who have been recently assessed across a range of levels.
- Concurrent validation against an anchor test as well as internal reviewing and ancedotal comparison of task types (in roleplays) by a panel of reviewers
- trialling and concurrent validity analysis
- Through trialling done by previous test-takers.
- It is intended to refresh the test after a certain number of exposures rather than year by year. New prompts will be trialled as per standard trialling procedure
- There will be a new version of the test later this year. Equivalence of the speaking test is easier to ensure as the raters will remain the same (remember there are only 7 of us at the moment), the scale remains the same, and the tasks structure will remain the same / very similar. For the comprehension test, it will be trickier but we will try to build in a trial in which same people take the two tests. The problem is to find enough people to really do that properly; doing that with hundreds of people is simply out of the question but if can get a few dozen ... We will also do item judgements (standard setting) in the task for with the new tests and compare the distribution of judgements of the 2 tests. This is something that we can do even with limited resources.
- There are plans to amend the test yearly.
- First three versions (and parts of versions 4 and 5) were trialled, analysed and standard set. Three versions are being used now. At the end of the test each test taker is asked to complete another test section for trialling. The trialling items are then analysed. We collect info on test-takers (ICAO level awarded, test version taken, ATC function, etc.) – this info is used to ensure the equivalence of test versions. At the moment we are collecting data and starting the first analyses. We hope to publish test versions 4 and 5 in summer 2008.
- Each test is randomly generated from the same item pool. Equivalence was verified with extensive data collection and overlapping items during field testing
- Trial items are checked and reliable and valid items are used in the new tests.
- In the same way as in (4) above
- We expect to carry out a logistic regression analysis (Rash or 2PL or 3PL) in the nearby future.

7. *Are any statistical analyses carried out?*

Yes 17 No 2 No Response 3

If YES, what statistical analyses are used to examine the results of the tests?

No Response 5

- Observational checklists.
- This is too comprehensive to cover here.
- text types are analysed and checked to see if the test is following the same format. Language content is checked for lexical density to make sure each test question is clear for a level 4 candidate. Responses are analysed for content length and use complex forms.
- See above
- CR agreement indices Threshold loss agreement indices (intra and inter rater reliability analyses)
- Comparison of test-takers' self evaluation and final results. Comparison of results according to level of perforı
- Various, depending on the variables to be isolated and the goal of the analysis.
- We start at level 5 and the rater listen to the candidates responses to see if the candidates meet, exceed or fail expectations at this level.

- Being developed
- Probably the mean and individual results of the old and new test versions will be compared (t-tests or analyses of variance). It's unlikely that we'll have enough data for IRT equating / comparison. In the case of judgements of item difficulty, probably the distributions of placements onto different levels by different judges in the 2 tests will be compared but as this is in the future, I haven't really thought about this yet. If this question refers to the analysis of the trial data, then yes, we've calculated some simple classical analyses such as item difficulty and reliability. More detailed analyses on discrimination and IRT will be done later (simply haven't had time for this yet) before we start developing new versions of the test.
- Frequencies, descriptive stats, reliability, item statistics, item-total stats, scale stats, bias analysis, feedback comments.
- Correlation, item analysis
- The Validation Manual gives details of validation studies involving 140 test takers whose tests were not used for training the speech recognizer or building scoring models. Included in the Manual are: a) split-half reliability for machine scores and for human scaled scores for Overall scores and for each subscore (Table 6) b) correlation matrix among Overall scores and each subscore (Table 7) c) correlation between machine scores and human scaled scores for Overall scores (Figure 3) and for each subscore (Table 8) d) correlation between machine scores and ratings provided by experts who listened to open response performance samples (page 26 and Figure 4) e) a comparison of 478 native speaker scores and 625 non-native speaker scores, showing how they are distributed among the 6 ICAO levels.
- Reliability, IRT

- A full report of the project was given to the CPA and another sent to the National Science Council. An extensive summary was given to the DGAC. An oral presentation supported with computer projected slides (ppt) was given to the DGCA and a committee of testers and airplane companies

8a. Are the analyses presented in a report?

Yes 15 No 4 No Response 3

If YES, please list all the analyses included in the report, giving details of how they are reported.

No Response 4

- The Observational Checklists were created from a list of language functions listed in Appendix B of ICAO's Document 9835 (Communicative Language Functions, Events, Domains and Tasks associated with Aviation). They were applied to test transcripts by test developers (after training and standardisation sessions) and results of frequency of function elicitation are reported by level for each part of the test.
- This is an internal process as it is with most test developers.
- It is a short report to state the changes made comparison of the results are presented. Details of the content analysis are kept on file and not made public.
- For the initial prototype this is documented in a reprot. Data and analysis for subsequent versions is recorded internally (eg on spreadsheets etc)
- Statistical information given to the XXX DGAC as per requests made furing periodical audits.
- The reports and their content are confidential, but do cover test validity and reliability, and examinee current and projected performance.
- Online System - candidates have results in 24 hours.
- Two reports: Detailed reports for project leader – all the SPSS files + Excel files + MS Word report on these f highlighted problematic areas and recomemndations on how to proceed National administrators receive a summary report - reliability + bias analysis + basic descriptive stats + explanations on what worked, what did not and what will be done about it.
- Correlation, item analysis
- See above

- No. The only group that is really interested in the characteristics of the test is the task force. The higher administration of the Aviation authority does not really know anything about testing and what makes a good test, except at a very superficial level. So we are not required to produce a report for them and all results, findings, questions, worries we have are discussed in the task force.

8b. Are the reports available to the public?

Yes 2 No 9 No Response 5 Explanatory comments, as follows

- No. Some reports have been provided to key organizations under a strict confidentiality agreement.
- Generally no. But upon request an evaluation report is made available upon request - eg a regulator may wish to review the test.
- No. Reports are available to client
- Not normally made available to the public but would be released for valid research purposes.
- No. Our Test Handbook is available to our customers, our partners and the Aviation Authorities.

At this stage, the report is not available to the public. We are looking for a University who is interested to conduct an external validation of our test (we expect to make a choice within before September 2008). The results of the external validation will be shown to the public (at least an executive summary).

- Yes. Detailed reports not available but a lot of info can be found on (URL)

If YES, please provide details of how we might access the reports.

- URL of XXX Test given (The Manual is there, as is other information. Please let us know if it is NOT easy to navigate to what you are looking for and we shall take steps to correct that. We aim to be open and informative.)

SECTION 5. TEST ADMINISTRATION

1. How is cheating (for example, personation, bribing, copying, access to illicit materials, etc.) checked or prevented?

No response 5

- Test materials kept under two-tier security system; candidates photographed; certificates include photograph with time taken & candidate's biographical data; counterfeit-proof certificates; centralised certification in XXX organisation; certificate verification website
- XXX test is used by the immigration authorities in particular, due to a strong security process. ID theft, swapping of candidates at the test, buying forged test papers and test results are all covered through a three way check system that is both paper and digitally based. A central global digital reference system is available to the immigration authorities to check online in real time a persons identity and XXX test score. If any details are doctored on their paper result that the application is rejected.

Impersonating in tests is covered through passport invigilation and digital photography of candidates on test day. These photos are then matched to their results from the day. Training is provided to invigilators and examiners on a quarterly basis across our 5,000 examiners globally. Administrators are trained by id staff through embassies and immigration on a regular basis.

- ID checks, standard legal forms signed by all employees of XXX airline, Cross-checks. Materials are not kept online or distributed outside the department controlling the test.
- Training of examiners is extensive. Manuals and policy documents are provided to each test centre.

Audits are carried out once per year. Examiners and raters have reporting processes. Security protocols and XXX test methods for delivery are provided in training and manuals.

- The test is only administered at XXX organisation or by XXX organisation staff at overseas centres. Candidate impersonation is countered by the requiremnt for photo ID and the inclusion of a photograph on the final certificate issued. The administration system assures secure and traceable handling of all examination documents, preventing copying and illicit access. The interview is recorded and the recording sent to two remote assess, unknown to the candidates. Confidentiality of all system procedures is guaranteed. The protection, security and confidentiality of results and certificates, and data relating to them, is guaranteed In line with current data protection legislation, and candidates are informed of their rights to access this data.
- Test is computer generated with different items. Tests are protected by security codes. Personnel carefully s
- individualized test numbers are assigned to examinees, who are required to present 2 valid photo i.d.'s to enter the testing area. 2) examinees are not allowed to bring anything to the testing area - all the materials (pens, paper, headsets, tape-players and test tapes) will be provided by the company. 3) one test procter is present for every 10 examinees. 4) the test set to be administered will be randomly chosen from 5 different test sets generated on the day of the test. 5) no one is allowed to enter or leave the test area for the duration of the test.
- Verification is done at the beginning of the test. Employers coordinate test times and are responsible to Provide the phone line for their location when they call into the call centre.
- Positive photo ID required and test is supervised by appointed supervisors
- The test taker must prove his / her identity with an ID card or passport etc that has his/her picture. Most test are administered individually or in very small groups, so it is easy for the tester to monitor the test taker. So there is no need for extra persons than the tester. The first / trial test of 80 aviators was an exception in this respect, so please consider the answers to 2, 3 and 4 in this light. The number of authorised testers is only 7 at the moment (till about 2011) so bribing is probably not an issue yet. However, this is an interesting point that we must think about in the future.
- General guidelines for the conduct of tests. Regulations prohibiting cheating.
- We have published Guidelines for administrators and invigilators and test security procedures. Invigilators are trained by National administrators. See point 4 below - Nat Admins write reports on test administration.
- Extreme security measures are in place. Upon cheating candidates are disqualified and blacklisted. Tests are proctored and administered in secure environments by XXX organisation. Identification checks are carried out. Each test is randomly generated on the fly. Items are stored centrally behind a secure firewall. Scoring is performed by YYY organisation's machines and returned directly to XXX organisation.
- The number of applicants is small enough that there is no need to check for cheating even in the listening test, which is usually administered to multiple candidates. The interview test is conducted in a one-to-one situation. The test is administered by the airlines whenever the pilots have time between their flights, and so the numbers taking the test at any one time is very small. The national test is administered six times a year, and average number of applicants is about 20
- The tester is responsible for identifying a person with the photograph in his/her passport. Availability of passports is a requirement to all test-takers. Test database contains, inter alia, information on candidate's full name, date and place of birth, passport number, date of testing, date of rating, test-taker's code, tester's and raters' codes, etc. Information on all test results is sent to the XXX CAA on a monthly basis.
- Before someone is allowed to take a test: - The TC Administrator has to setup an account (name, ID, Date of Birth, ...) - The test taker has to sign a test sign-off sheet - The test taker has to present an acceptable ID (with recent photograph) - The test taker receives his/her test booklet with his/her name printed on the first page together with the access codes During the test: - The test taker has to confirm that he/she is the person whose name appears on the test booklet - The TC Administrator and/or a proctor monitors the test sessions - A webcam is used to film the test taker - The test items are selected at random from the item pool. After the test: - The test booklets are collected by the TC Administrator
- The raters are able to report cheating (e.g. whispering in the microphone, voices in the background ...) The TC Administrator and the System Administrator can look at the webcam images Bribing has no use because the TC Administrator doesn't know which questions will be selected from the item pool.

2. Are test invigilators trained?

Yes 16 No 1 No response 4 Explanation as follows:

- Yes. Test Administrators and raters are trained. For test administrators, there is a one-day course (XXX system, XXX Test, Test Security, Test Administration, Ethics …). For raters, we organise: - A six-day full-time course - Refresher courses - Re-certification courses (every two years) followed by a test. If a raters fails the test, he/she can no longer be a rater. - Self-training (access to the rated speech samples library)

3. *Is test administration monitored?*

Yes 17 No 0 No response 4 Explanations as follows:

- Yes, monitoring of test administration is highly recommended (test administration usually takes place abroad). TC Administrators (Training Center Administrators) have to follow a one-day course about the XXX system, the test administration procedures and test security issues. In our Operations Manual, we clearly indicate the need for proper monitoring. However, test takers don't have to be monitored all the time because test takers are filmed (built-in feature) during the test.

4. *Do invigilators write a report after each administration?*

Yes 10 No 8 No response 3 Explanations of Yes as follows:

- Yes Our XXX Data Management System allows raters to write a report (only for Part II of the test, which is a subjectively marked test). Raters also have the possibility to report a damaged test and/or report suspected cheating. Sometimes, tests are transcribed for further analysis.

5. *What processes are in place for test-takers to make complaints or to seek scrutiny of marks, and/or re-marking of the test?*

No Responses 5

- Dissatisfied candidates may request test be remarked by senior examiners at XXX organisation within 1 month of certification. Candidates return certificate to centre (for secure retention) & pay full test fee & test remarked at XXX organisation. If senior examiner awards higher mark, new certificate despatched to centre immediately. The previous certificate returned to XXX organisation & appeal fee refunded to candidate.
If no change to mark, centre reissues original certificate & no refund given. If senior examiner awards lower mark, original certificate reissued with the original scores & no refund given.
- There is a global re-mark process which enables the test to be re-marked by a second and often third party from any one of 120 countries around the world.
- The test is reassessed by two raters. A transcript of the test responses is provided to highlight reasons for assigning that particular grade
- A complaints form is available. An appeals form is available.
- An appeals procedure
- Extensive survey form given to each test-taker. Exams are electronically recorded.
- Complaints and requests for re-rating or explanation/justification of the results can be directed to (URL). A response will be given within 15 working days, depending on the individual complaint or request.
- Candidate would make a complaint to the employer and the employer would contact us. The file is brought up the/a rater will relisten to the assessment and reconfirm the score.
- Review can be requested for a nominal fee. Feedback can be provided orally at the end of the test and written complaints are also accepted and actione in accordance with standard company policy and procedures
- They can file a request for re-rating at the XXX Civil Aviation Authority, who will then organise for re-rating. The tester is required to tape the speaking test and keep the answer sheets for 6 years.
The test taker can, and they often do, ask about the problems they had in the test, especially in the section concerning English aviation terminology (see elsewhere for an explanation of the structure of the

whole test).

- General guidelines and a internal review and complaint mechanism.
- National regulations. XXX organisation provides guidelines.
- Please refer to URL
- Test takers can direct complaints to XXX organisation, who are able to conduct their own investigations into how the test was administered. If XXX organisation suspect technical problem, they can (and do) escalate the problem to us. We conduct an internal investigation which involves humans listening to the test audio files, verifying the accuracy of scoring, and responding via XXX organisation. In rare circumstances, replacement tests are made available
- No official processes are stipulated
- A test-taker can appeal against the decision, and his/her test recording will be given to a third rater who would not be aware of the test-taker's complaint. The appeal is dismissed if the third rater's overall level is the same. If not, the decision of a senior rater will be final, and the test-taker will be informed accordingly.
- It is up to the Aviation Authorities to implement procedures for related to complaints, re-marking etc. XXX organisation only deals with Airlines and Air Navigation Service Providers, not with the test takers. However, we do make recommendations to our customers. If the Aviation Authorities allow a test taker to seek scrutiny of marks, the TC Administrator has the option to print all the test details (e.g. individual scores for each category, rater's report …) using his/her XXX Data Management System account. The TC Administrators can request a 'double marking' of a test when necessary. Most Aviation Authorities seem to adopt different approaches that we have suggested to our customers.

6. How often can a test-taker re-take the test?

No response 5

- As many times as they like.
- Weekly and no limit.
- It depends on decisions made by XXX airline3. Usually, a candidate has to show they have at least 100 hours of language training before being allowed to resit.
- They must wait 90 days before resitting.
- After a minimum of four weeks
- As often as he/she wishes because each time the exam changes
- Every 60 days. The cumulative scores are not averaged - the latest or most recent rating will serve as the examinees proficiency level.
- One time only, unless requested by employer for a retest.
- XXX test can only be taken once (no repeats) YYY tests no limit on resits
- Number of retakes is not limited.
- Several times but after an approipriate wait period - allowing for additional training time.
- Decision rests with the national licensing authority. XXX organisation recommends two resits, minimum three months apart.
- 3 months
- This depends on the relationship that XXX organisation has with their clients. Resource factors normally prevent test takers from taking the test again and again. Also, a practice test which can be taken at home or the office will tell a test taker whether they are likely to pass or fail if they take the test, and this prevents opportunistic test taking
- As often as they want
- There are no limitations on the time between tests, but test-takers are informed that any measurable improvement requires about 200 hours of studies
- It is up to the Aviation Authorities to decide how often a test-taker can take a test. We do make recommendations however when the Aviation Authorities fail to produce clear regulations (e.g. show proof of training).

SECTION 6. REVIEW

1. *How often are the tests reviewed and changed?*

No response: 5 N/A (!!) 1 Details as follows:

- Test in infancy. Reviewed continuously. Trials ongoing.
- Each of the 48 tests provided in a year is unique and material is destroyed at the end of each test.
- Every 6 months, but this will now be reviewed every 3 months.
- All the time!!! There are 4 full time staff on test writing and this includes continuous monitoring and reviewing (+rewriting of tasks).
- Tests are reviewed quarterly. When tests are changed the format typically remains the same, but new items are added to the items bank after trialling and minor modifications to the interlocutor frame are made in an effort to increase the clarity and conciseness of the oral rubrics.
- Constantly
- Tests are reviewed every quarter and changed when necessary.
- Early days, but questions are randomly generated from a question bank and each question will be replaced after approx 300 exposures
- There is no specific agreement on this in the Authority. The task force plans to review the current test when it starts designing the new version, later this year. (A point of clarification: there are in fact two versions of some test tasks -- comprehension tasks -- that were developed and that allows some limited re-taking of the tests without having to take exactly the same test. For speaking, there are enough parallel tasks to allow repeated retakes, if necessary, without encountering exactly the same task.)
- We plan to review the tests yearly and amend them as needed.
- Test performance is regularly reviewed and versions are withdrawn to prevent over-familiarisation.
- On a regular basis
- The item pool is to be refreshed on an ongoing basis. (A schedule has been fixed for developing new items.)
- Tests are reviewed and necessary changes are decided at least twice a year by the Council for Assessing and Accrediting Airline English Proficiency (Aeronautical Agency in the Ministry of Land and Transport)
- Annually, from 10 to 12 new variants are developed
- Part I (listening section) – There are over 100 items in the pool. Some of these items are anchor items which will allow us to add new items fairly easy. Part II (speaking section) – For each task, the selection algorithm will select one test item for a pool of items. The number of items per task is approximately 30. Therefore, the number of possible combinations is very high. Items are removed when there is evidence of a security breach (all personnel involved in the administration of the test and all the raters have signed an agreement with XXX organisation. It is part of their obligations to report all kinds of issues with test security). How often test items are reviewed and/or changes will depend on how frequently the items have been used.

2. *What procedures are in place to ensure that the test keeps pace with changes in the ICAO requirements?*

No response 5 N/A 1 (!!) Details as follows:

- There have been no changes in the ICAO LPRs, other than by date. If there were changes to the LPRs, XXX organisation would be one of the first to know having been asked to contribute towards revisions of Document 9835, asked to present at international conferences, etc.
- XXX test keeps constant contact and communication with key global governing bodies to work with the likes of medical councils (ie CGFNS), accounting bodies, immigration authorities where XXX test is a high stakes test.
- Any ICAO amendments are forwarded to the person in charge.
- The ICAO LPRs have not changed to date, so there's not much to keep pace with. I am on the ICAEA Board. I have acess to the ICAO reps and good contacts with the international community on the ICAO LPRs and any modifications (of which there have been none since 2003 - apart from the extension for ICAO member States)
- Close contact with the relevant personal at within the Flight Safety Department at ICAO in Montreal. Regular review of the ICAO website and attendance at all related ICAO seminars and workshops. Review

of ICAO publications. Discussion with external stakeholders etc.

- Constant evaluation by the test designer in conjunction with the XXX DGAC and ICAO documents.
- Test and item developers are required to attend one ICAO seminar every year. 2) Test and item developers undergo periodic training and calibration throughout the year.
- Test will be reviewed as and when ICAO requirements are altered.
- I don't think we have specific procedures for this but obviously the aviation members of the task force, especially its convener must know about any changes and will then convene the group, if necessary, to consider the implications of the changes.
- Amendments to the tests as required.
- The XXX test User Group regularly reviews any changes that might affect test performance. Periodical consultations with XXX organisation ATC experts (instructors) to ensure appropriacy of ATC content, special e-mail address for comments on the test (URL supplied), communications with ICAO offices in Montreal and Paris to make sure we are informed about all requirements, regulations, etc..
- Once we receive new requirements from ICAO, we will make neccessary adjustments.
- One of the responsibilities of the Test Development team is to maintain the tests, respond to technical enquiries, and conduct ongoing validation and investigations. Through XXX organisation and our professional contacts we keep open communication with ICAO and XXX organisation. If the requirements change, the test will be modified as necessary
- Information about any changes in the ICAO requirements are provided by the Aeronautical Agency
- Participation in ICAO meetings, seminars and workshops on language proficiency requirements; availability of ICAO documents on language training and testing; correspondence with members of PRICESG
- Collecting feedback from airlines - Collecting feedback from Aviation Authorities - Collecting feedback from Avi English Experts - Literature studies - Attending forums DA37- Attending conferences

3. *Are there any other procedures which you use to ensure the validity and reliability of your tests of Aviation English which we have not mentioned?*

Yes (with no further details) 7 No 7 No Responses 3
Further explanations as follows:

- Yes: It is important for test takers to be familiar with the telephone delivery of the test and also the test tasks before taking the certification test. Not only is a sample test available on XXX website to listen to, but Practice tests can also be purchased. After taking the Practice test, test takers can retrieve their score online within minutes. Practice tests provide an estimate of the test taker’s score on the ICAO Levels, and therefore helps them decide if they are ready to sit the Certification test.
- Yes: As important as the practical development of the test was the basic research that led to its design. This comprised theoretical integration of contributions made by various schools throughout the 20th Century, empirical discourse analysis and ethnographic observation of simulations and authentic communications. (The main problem was that simply asking two raters to provide global assessments by reference to holistic descriptors does not guarantee that they will. More probably, they will focus on some aspects, and most likely, each one on different aspects. In other words, a good understanding of what it is to be measured is required in order to design a good measuring instrument.)
- Yes: Cross-checking of focus group results of XXX test with the same group results with another test-provider (from an English-speaking country).
- Yes. (e.g. several concurrent validation studies)
- Yes, validation study by XXX University with a lot of recommendations which we follow.

Many thanks for completing this questionnaire. Your responses will be very helpful to us in our research.

If you are willing to provide more details of your responses, we would be very happy to receive the details of any URL that provides further information.

Alternatively, if you could send us copies of any reports that provide more detailed evidence of your quality control procedures, we would be extremely grateful.

Specifically, we would welcome any documentation/ reports/ URLs of the detail on the following:

(List of documents follows)

No Responses: 21

Details as follows:

We are not providing any attachments other than the Validation Manual. Most of the information requested is either on the website or in the Manual. PLEASE return the favour – if you think there is something that XXX organisation should be doing but is not, then let us know. Also, if you are writing a report or presentation based on the information we have provided, we would be very grateful if you would share that with us in advance. Thank you.