AN ANALYSIS OF THE QUALITY OF ENGLISH TESTING FOR AVIATION PURPOSES IN FINLAND

***Ari Huhta**, University of Jyväskylä*

Ari Huhta works as a researcher in the Centre for Applied Language Studies at the University of Jyväskylä, Finland. He specialises in foreign and second language assessment, and has participated in several national and international research and development projects in the field, notably in DIALANG. He has published in international journals such as Language Testing.

Correspondence to Ari Huhta: ari.huhta@jyu.fi

This article describes and analyses the development of a new test of aviation English by the Finnish Civil Aviation Authority (FCAA), as well as the overall situation in Finland as regards the testing of aviation English. The article describes the FCAA development project and evaluates the strengths and weaknesses of the new test and the whole testing system, often with reference to the framework of test usefulness proposed by Bachman and Palmer (1996). The quality of the overall system in Finland appears to be quite variable as it is based on the principle of decentralization, in which the FCAA evaluates and approves different tests to be used for certifying the English language skills of aviation personnel. On the other hand, the FCAA commissioned the creation of a test of its own, which appears to have certain strengths, but also has some flaws, such as a lack of systematic double rating of speaking and very little centralized monitoring of overall quality.

INTRODUCTION

This article gives an account of the overall situation in Finland as regards the testing of English for aviation purposes and, in particular, of the development of a new aviation English test by the Finnish Civil Aviation Authority in 2007–08 (henceforth, the FCAA test). The article describes the different phases of the development project and analyses the strengths and weaknesses of the new test and the entire testing system. In the evaluation, reference will often be made to the framework of test usefulness by Bachman and Palmer (1996).

Whenever possible, two aspects are distinguished in the analysis of the usefulness of the FCAA test: *a priori* and *a posteriori* types of validation and evidence (see Weir, 1993, 2005). *A priori* validation includes all non-empirical work to ensure high quality by test designers during the development of the test, before it is piloted or used for real; thus,

it can also be called theoretical validation (see also Messick, 1989). *A posteriori* validation uses empirical data gathered when the test is piloted and used operationally. The data typically consists of test scores and ratings, but also of qualitative evidence such as interview responses.

Bachman and Palmer (1996) tie usefulness to the purpose of the test: a test and the ways it is used may be useful, to some degree, in terms of its intended purpose(s) (Bachman and Palmer 1996, p. 18). The FCAA test has clearly one main purpose: to certify that the test taker has achieved the minimum English language proficiency specified by the International Civil Aviation Organization (ICAO). More specifically, the test aims to find out if the test taker has achieved at least level four on the ICAO scale of language proficiency.

According to Bachman and Palmer (1996, pp. 18–24), several factors contribute to the usefulness of a test for its intended purpose(s): reliability, construct validity, authenticity, interactiveness, impact, and practicality. It is essential to consider the purpose of the assessment to be able to decide if all the aspects of quality are equally important. In fact, the overall usefulness of the assessment is what really counts, and that cannot be evaluated without taking into account the balance of all the factors above that contribute to usefulness (op cit. p. 18).

FCAA AVIATION TESTING PROJECT

The ICAO requirement that air traffic controllers and pilots should have a certificate attesting to their proficiency in the language(s) used for aeronautical communication by 5 March 2008 (ICAO, 2004) prompted the Finnish Civil Aviation Authority into taking action on the matter in 2007. The FCAA decided to develop their own test of aviation English, rather than using any of the available commercial tests, and appointed one of its inspectors to set up and lead a special task force for the purpose in the summer of 2007. Following ICAO recommendations, the task force was to include both aviation and language and testing experts.

The task force responsible for test development consisted of five members. It was chaired by an aviation inspector with a long career in both military and civilian aviation in Finland and abroad. He also represented the middle management level and the licensing section at the FCAA, assisted by a junior member of staff in the same section, who had flown commercial airplanes in the US for several years. The third member was a senior air traffic controller with extensive experience in teaching international aviation terminology. During the project he also became an accredited tester for ELPAC (English Lan-

guage Proficiency for Aeronautical Communication), which is a European-wide language test for air traffic controllers. The other members represented language education: an experienced English language teacher with a background in specific-purpose language teaching, item writing, and administration of a major language examination in the institutions she had formerly worked for. The fifth member was a professional language test developer and researcher, i.e., the present author.

The head of the task force was responsible for recruiting its members. As the FCAA had not previously worked with language testing experts, he contacted language teachers in the Helsinki area where the FCAA is located and soon identified the university and the department of the present author as the likely partner in the project. The fact that the author's department is responsible for the development of a major national language examination in Finland was the most likely reason for this.

The task force started its work by discussing the general quality requirements for language testing, with reference to the guidelines for good practice of organizations such as the International Language Testing Association (www.iltaonline.com) and the European Association for Language Testing and Assessment (www.ealta.eu.org), and by charting a tentative format and content for the test. It turned out that the FCAA had already made decisions about certain aspects of the test and the overall testing system, which will be discussed below in more detail. Probably the key practical issue, which was to have major implications to the other aspects of test usefulness, was the extremely short timescale available to render the test operational. The entire development process was to be completed between August and November 2007, so that the test would be ready for use at the beginning of 2008, before the ICAO deadline. The FCAA had taken quite a long time to put the 2004 recommendation from ICAO into action, and it was not until the spring of 2007 that the decision to set up the task force was made. This, then, left the task force only a few months to accomplish its mission.

Some of the implications of the short timeline were obvious from the beginning. For example, it was not possible to carry out needs analyses: the task force had to rely on the knowledge and experience of its aviation industry members concerning pilots' needs. One of the members had also given English proficiency ratings to aviation personnel under the old language certification system. The drafting of tasks and items had to start immediately after the first version of the test specifications was put together in co-operation with the members of the task force; the present author provided the framework for the specifications based on the recommendations by Alderson et al. (1995) and Douglas (2000).

CONTENT AND FORMAT OF THE FCAA LANGUAGE TEST

The FCAA test comprises three separately scored sub-tests: knowledge of the international aviation phraseology, comprehension, and speaking. The comprehension test is divided into listening and reading but the two are scored together as one test. All three tests must be passed in order to be awarded a certificate.

The sub-test on the knowledge of aviation phraseology differs from the other two tests in that it is a simple pass / fail test. The candidate has to achieve 75 per cent of the maximum score to pass. Test takers who fail this part will automatically fail the whole test. The ICAO proficiency level of the candidates who pass the phraseology test is determined solely on the basis of the oral and comprehension tests, so the phraseology score is not taken into account when the level is determined.

The current phraseology test contains 40 items that each require candidates to translate into English a short sentence presented on paper in Finnish. The items (sentences) are presented in a sequence that follows the phases of a typical flight from the departure to the landing of an aircraft and they include phrases that concern unusual or even emergency situations. Each item is rated on a 0–2 scale for the use of the correct phrase (one point) and for overall comprehensibility (one point). Spelling errors are ignored unless they change the meaning or render the phrase incomprehensible.

The reading test consists of three authentic short texts dealing with aviation, each followed by about ten short-answer or multiple-choice questions. The texts do not require professional-level knowledge in order to be understood. The time allowed is twenty minutes.

The listening test takes about half an hour, including preparation time and pauses for answering questions and has about 30 items. There are two parts in the test, each played twice, first as a whole and then in sections with pauses. As in the reading test, the topics relate to aviation but do not require expert knowledge – although knowledge of aviation obviously helps. The texts are not original recordings but are often based on genuine recordings of, for example, discussions between pilots in the cockpit of an airplane. The texts are recorded with the help of native and non-native English-speaking actors. The input is made as authentic as possible by adding relevant background noise to the recordings, such as the sound of engines. In addition, the actors who speak on the tapes come from different first language backgrounds; typically, at least three non-native accents (sometimes quite strong ones) are included, in addition to one native variety of English.

The speaking test is conducted face-to-face between the tester and one candidate at a time, and it comprises three parts: a discussion / interview, and two picture-based tasks. This sub-test takes about 15 minutes but it can last longer if the tester needs more evidence of the candidate's proficiency. In the first part, the tester asks questions about topics like the test taker's background in aviation, work, hobbies, and future plans. In the picture-based tasks the candidate chooses one of the aviation-related pictures shown by the tester and describes it, after which the two discuss what the picture shows. The picture tasks differ in that one set of pictures depicts typical aviation scenes and situations, such as aircraft landing, taxiing or taking off, whereas the other set presents a range of unusual aviation-related situations, including small mishaps or accidents. Pictures of serious accidents are not, however, included since they might be too distressing. All speaking tests are recorded so that the tester can check his or her rating, as well as for possible double rating by another tester, or for possible complaints and subsequent re-rating by the FCAA. Recordings are to be stored by the tester for six years.

QUALITY CONTROL PROCEDURES

Large-scale piloting of the FCAA test was not possible, given the short timeline for its development. The task force members obviously reviewed each other's test items, and a very small-scale pilot session was organised with five members of the FCAA staff who had an aviation background, which helped in matters such as setting realistic time limits for the tests. Since that was not clearly enough, a decision was made to change the nature of the first three training sessions conducted in November 2007. The sessions were to serve three linked purposes. First, they were to train testers who would later administer the FCAA tests across the country to pilots who needed to renew their language proficiency rating. Second, the 80 trainee testers were to take the brand new FCAA tests themselves to find out who among them met the FCAA language requirement for testers (i.e., ICAO level five). And thirdly, it was decided to make this first large-scale test administration the main piloting of the tests. To achieve this dual aim of piloting the tests and determining tester proficiency, the test results were not issued to the trainees before the completion of item analyses, standard setting, and the cross-checking of the ratings for the speaking test within the task force members. All performances had been double rated and about a third of them were viewed and discussed by all the members.

The training sessions lasted two days. They included lectures on legislation, language testing, and the new FCAA test in particular. After this briefing the trainees took the test themselves. On the second day, the trainees were presented with a training video prepared by the task force on how to conduct the speaking test, because as accredited FCAA

testers they would need not only to rate but also to administer the speaking test. The video first showed an example of an extremely poor interview acted out by two members of the task force – in fact, a caricature of an interview – followed by a re-run of the same interview, but now conducted more or less at it should be. The examples were discussed in depth. Finally, the candidates were given rater training with the help of benchmark recordings distributed by ICAO. The training followed the format of the tried and proven rater training procedures in a national language examination familiar to the present author. It involved listening to about ten benchmarks, rating them individually, collecting the ratings, counting the different ratings, and discussing them.

The approach to standard setting to decide on the cut-scores for different ICAO levels for the comprehension test was based on the examination of the score distributions, rather than on expert judgements of test items or test takers. It could be described as a simple version of one of the empirical-judgemental methods of standard setting described in Kaftandjieva's (2004, p. 15) survey, namely a version of the direct judgment method described in Hambleton et al. (2000, p. 363). In addition, the test takers' self-assessments against the ICAO scale and the speaking proficiency ratings were taken into account in setting standards for the comprehension test, as it could be expected that their level of speaking would be at or below their comprehension level but would probably not exceed it.

The FCAA test became operational in January 2008. Currently, the right to assess the English language skills of aviation personnel by means of the test is restricted, according to whether first time assessment or renewal assessment is involved. The right to administer the FCAA test to those who are getting their first flying license belongs only to the five members of the FCAA task force and to two or three other persons selected from among the trained testers. On the other hand, the right to administer the test to pilots renewing their flying license belongs to a much wider group, i.e., to most of those who were trained in November 2007.

In late 2008, a second version of the FCAA test was created by the same task force that had designed the first one, but only the comprehension and phraseology sub-tests were new. The speaking test remained the same, except for some minor clarifications in the instructions to the testers. It is envisaged that further versions of the test will be needed in the future, and this will obviously be a major challenge to the quality of the system.

To conclude this section on quality control procedures, it should be noted that the FCAA aims to keep the number of false positives (test takers with insufficient proficiency who nevertheless pass the test) at a minimum, a candidate first has to pass the phraseology

test, which has the cut-off at 75 per cent, and then obtain at least level four in both the speaking and comprehension tests. If the results of the latter two tests disagree, the lower one determines the overall grade. Minimizing false positives is also evident in the ICAO guideline that in speaking the lowest of the six analytical ratings (pronunciation, fluency, vocabulary, etc) determines the test taker's overall speaking grade. The need for a specific phraseology test and its relatively high cut-off point are based on the FCAA's view that it is important to make sure the test takers master the international aviation phraseology. In comparison, the cut-off point for level four in the comprehension test is about 50 per cent in the first FCAA test version. This relatively low threshold is due to the difficulty of the first listening test (e.g. sections with fast or unclear speech and a lot of information, and the use of open-ended main questions that minimize the effect of guessing). The second test version turned out somewhat easier and consequently its cut-offs are higher.

THE OVERALL SYSTEM OF AVIATION ENGLISH TESTING IN FINLAND

The development of the FCAA test was just one aspect of the overall picture with regard to certifying proficiency in English for aviation purposes in Finland. The FCAA test is not in fact the only official aviation test in the country. Following a traditional approach to language testing in Finland (e.g., the civil servants' language examination), the overall structure of the system is a mixture of centralization and decentralization. The FCAA has the authority to accredit qualification systems such as language tests, as well as the power to develop its own tests. Therefore, there was not to be only one, nationwide aviation language test: in principle and in practice, anybody can apply for permission to conduct such testing by presenting their tests for inspection to the FCAA.

The FCAA thus potentially will allow numerous organizations or institutions to have their language tests accredited for use in the country. So far, there are only two alternatives to the FCAA language test. One is the international ELPAC test intended for air traffic controllers, which practically all Finnish air traffic controllers take nowadays. The other accredited test was developed by the biggest commercial airline company in the country, Finnair. This is in fact a modified version of the FCAA test, conducted as part of the flight simulator training that the company provides for its pilots. So far, no other organizations have developed alternative language tests for this purpose.

There is, however, another aspect of decentralization that is a graver threat to the reliability of testing. This is the fact that a considerable number of testers are allowed to administer the FCAA test to pilots who already have a license. As mentioned above, the right to test new pilots is restricted to the members of the FCAA task force – apparently

on the recommendation of the FCAA representatives on the task force. However, the FCAA decided that 'renewal' testing could be done by all pilots who were trained in the November 2007 training sessions and then applied to be accredited. The reason is partly practical: the testing of all license holders in the country would have been an overwhelming task for the five members of the task force. Although the FCAA comprehension and phraseology tests are fairly straightforward to mark, the testers also conduct the speaking test, in which reliability is much more of an issue. The real problem with this is the decision by the FCAA not to make double rating of speaking compulsory, on the basis that such procedures were not in place in the other FCAA tests and examinations either.

GENERAL TEST DESIGN PRINCIPLES SET BY THE FCAA

At the time of setting up the task force, the FCAA had already decided on a number of principles that the new FCAA test had to comply with, based on their reading of the ICAO guidelines and on considering the needs of Finnish pilots. These principles were decided at the higher levels of management at the FCAA and introduced into the work of the task force by its FCAA members.

The first principle concerned test content: the focus should be on oral rather than written proficiency. Thus, a test of oral interaction was to be one of the key parts of the whole test. In line with this, listening was to be given more emphasis than reading in the comprehension test. However, reading was not to be totally disregarded because the pilots were known to need to read in English. To avoid the problems with reliability related to testing reading with very few items, it was decided to combine the two comprehension tests and score them as if they were one test. Consequently, the listening test contains about 75 per cent and the reading test about 25 per cent of the comprehension items.

Another general principle imposed by the FCAA was that the test should measure both general and aviation-specific language. Thus, the content of the new comprehension tests, in particular, relates to rather general aviation contexts, and includes, for example, discussions between pilots meeting at the airport or accounts by passengers of some problems when traveling. The content of the oral interaction test is also general in nature, although the topics mostly relate to flying and airplanes. However, aviation specific language is also included in these sub-tests. The speakers on the listening tape sometimes use specific terminology and the picture description tasks almost invariably elicit precise references to, for example, parts of an airplane.

However, the FCAA considered it necessary to have a specific test of international phraseology as one of the three tests to be passed because they wanted to ensure that the

test takers would have an adequate command of those expressions, no matter how good their general English might be. To further stress the importance of a command of the phrases, the cut-off for passing that test was set fairly high, i.e., 75 per cent of the maximum score.

The third general requirement for the new test was that it should measure how well the pilots could cope with unexpected situations and turns in the discourse. There was considerable discussion in the task force about what this meant and how it might be operationalised in the test. An obvious place to operationalise 'an unexpected turn of events' is in the test of oral interaction. Indeed, the test specifications and the instructions to the testers specifically address this requirement. First, the whole structure of the oral test aims to ensure that not everything is predictable from what was discussed previously. The two different picture description tasks automatically introduce changes in the discussion. Additionally, one of the two picture-based tasks includes pictures of unusual or unexpected happenings such as a bird strike or a minor mishap, some of which can be quite amusing. Secondly, the interviewer is instructed to introduce a sudden change of topic in the first part of the test, in which the test taker and tester discuss general matters. Such a change of topic should not be an extreme one but it could be an aviation-related question or topic that is fairly removed from the discussion that preceded it. For example, the tester can ask the candidate's opinion about an aviation matter that has been in the news recently and can thus be assumed to be known by the candidate.

THE STRENGTHS AND WEAKNESSES OF THE FCAA APPROACH

The overall usefulness of a language test is a combination of the different aspects of test usefulness: reliability, construct validity, authenticity, interactiveness, impact, and practicality (Bachman and Palmer, 1996). Weighing different aspects to determine the overall usefulness of a test is not straightforward but it should be judged against the purpose of the test. In the case of the FCAA (and the ELPAC and Finnair) language tests the purpose is to certify the proficiency in English of aviation personnel. This is clearly a very high-stakes decision, which places very high demands on the test in terms of all aspects of quality (or usefulness).

When we consider the available information and evidence of the usefulness of the FCAA testing system, we should distinguish the FCAA test and the overall testing system set up for Finland. It is clear that the overall usefulness of the two is different and we have much more information about the FCAA test than the whole system on which to

base our evaluation (for reviews of aviation tests in general, see Alderson, 2008; Alderson and Horák, 2008).

The usefulness of the overall system for testing aviation English in Finland is not satisfactory because of its decentralized nature and the lack of systematic, ongoing monitoring and planning. It is quite likely that the different tests accredited now (or in the future) are not equal in terms of quality. The ELPAC test has been carefully developed and it is administered under extremely standardized conditions (see Alderson et al., 2007). The new FCAA test has also gone through a serious development process but it has certain problems, discussed below, that reduce its overall usefulness. For lack of information, it is impossible to say much about the Finnair version of the FCAA test. If new tests are proposed for accreditation, the FCAA is in no way obliged to consult language testing experts, but in the light of experience over the past two years, it is likely that their advice will be sought. At the moment, fortunately, it is unlikely that new tests will emerge because the three existing tests seem to be able to meet the demand, and thus this potential threat to the overall usefulness of the system is diminished.

What is the overall quality or usefulness of the new FCAA language test? The reliability, construct validity, authenticity, and interactiveness of the test are very much dependent on the *a priori* work such as the test specifications, item writing, review and piloting, and rater and interviewer training. The main intended impact of the test obviously related to the goal of ensuring that the English proficiency of the Finnish aviation personnel could be certified as fairly and validly as possible. The task force responsible for the development of the test did not consider its practicality to be the most important factor – on the contrary, the present author is convinced that the task force fully realized that high stakes were involved and they made a genuine effort to meet as many quality standards as possible. However, the practical constraints of the project in terms of timetable, funding, and the availability of pilots as test takers hampered development.

On the whole, the test content and format designed at the *a priori* stage of the FCAA testing project seem reasonably useful, given the short time span during which the test was developed. In particular, the listening test appears to have construct validity and authenticity thanks to its range of authentic accents, the use of background noise, and inclusion of various types of discourse, such as dialogues, weather broadcasts, and messages from air traffic controllers. The speaking test provides the test taker with an opportunity to display both general and aviation-specific proficiency; it also tests the candidate's flexibility in moving quickly from one topic to the next. The phraseology test, too, is valid in terms of content coverage, which is in fact rather easy to achieve given the well-specified domain.

In certain respects, the reliability of the FCAA testing is also quite high: the marking of the comprehension and phraseology tests is quite straightforward, with the help of the detailed scoring keys that have been developed. The speaking ratings produced by the task force members can also be expected to be very consistent because the members form a very coherent group that worked together extensively and reviewed dozens of performances to standardize their ratings. However, unless regular re-standardization is provided both for the task force and for the other accredited testers, it is likely that the reliability of the speaking ratings is in jeopardy in the future.

The above, cautiously positive conclusions are supported by the empirical *a posteriori* evidence gathered from the questionnaire survey of the participants in the main training sessions in November 2007 and by interviewing the candidates tested by the present author in 2008–09. Practically all test takers who were surveyed considered that, overall, the test gave them a fair chance to display their command of aviation English. The speaking test was singled out by many as a particularly good part of the test.

Classical item analyses based on the performances of 157 test takers who took the first version of the FCAA test during the 2007 training (about 80 candidates) and during the first months of 2008 indicate that the comprehension test was performing quite consistently. Only two multiple-choice listening items proved to be problematic; they were turned into short-answer items, which seems to have addressed the problem. The reading test turned out to be quite reliable given its short length (predicted alpha for a 40-item version would have been .90).

There are, however, clear problems and potential issues with the FCAA test. The most obvious problem is the lack of systematic double rating in the speaking test. This was strongly recommended by the task force to the FCAA but the organisation declined to include that in the regulations on language proficiency testing on the grounds that double inspection or rating was not conducted in any their other tests and examinations. This was perhaps the prime example of a clash between a very traditional approach to examinations and a more modern testing culture. Interestingly, even the non-language experts in the task force embraced the recommendations of international guidelines on good practice in testing during the project but this change in attitude obviously did not extend to the organization at large. The best that could be done by the task force was to strongly recommend double rating to the accredited testers, especially in borderline cases.

The interviews with test takers revealed an issue with timing in the phraseology test. The time available was considered somewhat too short by some candidates, considering the number of items and the amount of writing required. That can easily be addressed

by adding some more time to the test but the phraseology test also has other issues. The phrases are normally used in oral communication but the FCAA test assesses them through writing, which is not ideal from the point of view of authenticity and the interaction between the task and the test takers' skills. However, oral administration of the test would be rather impractical. It might be possible to design oral tasks that simulate real-life flying and that elicit the use of aviation phrases but such a test would probably not achieve the coverage and representativeness of a long, straightforward paper-and-pencil test. Thus, what would be achieved in increased authenticity and interactiveness of the test would be lost in terms of construct validity, especially coverage of content.

Interestingly, the testing of aviation phraseology in the Finnair version of the FCAA test may in fact be better in terms of authenticity and interactiveness, because it is integrated into the regular training of the pilots in flight simulators, where the language of communication is mostly English.

Some interviewed candidates complained about the background noise and that some speakers on the FCAA listening tape spoke too fast and not always clearly. However, the rate and clarity of speaking, and background noise such as that in a crowded airport or a plane's cockpit, are considered part of the construct that needs to be measured. Another more justified complaint about recording quality came not from the candidates but from the trainee testers. Many of the benchmark recordings provided by ICAO and used in rater training were poor in technical quality. Although the task force had excluded the worst recordings when selecting material for the training sessions, the quality of the recording of even the best examples often left much to be desired.

Finally, it should be mentioned that the key part of the FCAA test can be considered to be the oral sub-test, because the ICAO scale focuses on defining relevant speaking skills and the organization has made available illustrative benchmark performances that help test designers align their ratings of speaking with the levels. Obviously, there is room for improvement in the ICAO scale descriptors (see Knoch's 2009 study for a number of such recommendations) and especially the benchmarks, but they are much more concrete than what ICAO has to offer to those wishing to link their comprehension tests or tests of international phraseology to the scale.

The FCAA project, with limited time and resources, decided to focus on the development of the speaking test and the training of the oral tester-interviewers. The development of the comprehension and phraseology tests was not taken lightly, but it was always clear that the setting of cut-off points for these tests was going to be much less precise than the rating of speaking. The definition of comprehension in the ICAO scale is superficial, and there is even less material available to help place cut-off points on the highly

specific phraseology. This also means that it is much more difficult to ensure comparability between different versions of comprehension tests, in particular, than between speaking tests, as long as the test format is not radically altered.

Given the key role of the speaking test in the whole system, it was obviously extremely frustrating to learn that the FCAA would not agree to a systematic double rating of speaking performances. It is also not known if and how the FCAA is going to monitor the consistency of the testers that it has accredited.

The experience of the task force members has been that when test takers fail they usually do so in the phraseology test. This suggests, but obviously does not prove, that the cut-off for passing the phraseology test is adequately high in order to minimize false positive test results. When a failure occurs in the other two tests, it tends to be in the speaking test rather than in the comprehension test, and in general, when the results of these two tests disagree, the level awarded from the speaking test tends to be lower than the level based on the comprehension test. This is quite a believable trend as it is likely that learners' comprehension is at least at the same level as their speaking skills and may in fact often be more advanced.

To conclude, the FCAA testing project was an educational experience for everybody concerned. The aviation experts and testing / language experts learned a lot from each other, and there was a genuine desire to jointly develop a maximally useful and defensible test. That this aim was not fully achieved was due to factors that the test designers could do very little about, such as the limited time and, to some extent, resources allocated to the development and trialling of the first FCAA tests in 2007 and their new versions in 2008. Underlying traditional assumptions in the contracting organization about what is sufficient for a testing system to work clearly posed challenges to the development work and they are also likely to affect the way the system will be maintained, reviewed and further developed in the future.

REFERENCES

Alderson, J. Charles. (2008). *Final Report on a Survey of Aviation English Tests*. Lancaster: Lancaster University.

Alderson, J. Charles.; Al-Zadjali, Rima.; Banerjee, Jayanti.; Horák, Tania.; Papageorgiou, Spiros.; van Moere, Alistar. (2007). *ELPAC Final Validation Report*. Lancaster: Lancaster University.

Alderson, J. Charles.; Clapham, Caroline.; Wall, Diane. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. Charles.; Horák, Tania. (2008). Report on a Survey of National Civil Aviation Authorities' Plans for Implementation of ICAO Language Proficiency Requirements. Lancaster: Lancaster University.

Bachman, Lyle F.; Palmer, Adrian S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Douglas, Dan (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Hambleton, R; Jaeger, Richard; Plake, Barbara; Mills, Craig(2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24(4), 355–366.

ICAO. (2004). *Manual on the Implementation of the ICAO Language Proficiency Requirements* – Doc 9835-AN/453. Montreal: International Civil Aviation Organization.

Kaftandjieva, Felianka. (2004). *Standard setting*. In Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Section B., pp 1–43). Strasbourg: Council of Europe.

Knoch, Ute(2009). Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale. *Spaan Fellow Working Papers in Second Language Assessment*, 7, 21–46.

Messick, Samuel. (1989). Validity. In Linn, Robert. L. (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Weir, Cyril (1993). Understanding and Developing Language Tests. New York: Prentice Hall.

Weir, Cyril (2005). Language Testing and Validation. An Evidence-Based Approach. Basingstoke: Palgrave Macmillan.

Cite this article as: Huhta, Ari. (2009). 'An analysis of the quality of English testing for aviation purposes in Finland'. Australian Review of Applied Linguistics 32 (3), 26.1–26.14. DOI: 10.2104/aral0926.