Applied Linguistics Papers 27/1: 2023, 40–52

How Can Listening Contribute to Aviation Safety? EPLIS Paper 1 under the Spotlight

Paula Ribeiro e SOUZA
Airspace Control Institute, Brazil
E-mail: paulaprs1@fab.mil.br,

Beatriz Faria ARAGÃO
Airspace Control Institute, Brazil
E-mail: beatrizbfa@fab.mil.br,

Abstract: Assessing listening for specific purposes involves understanding the nature of listening and making careful theoretical and practical decisions on the authenticity of the input, the type of tasks, mode of delivery, as well as the characteristics of the target test population. In a context of aviation language testing for licensing purposes, in which listening comprehension is central to aeronautical communications, test developers must guarantee that tests are valid, effective and reliable by complying with best practices. In this paper, the listening test of the Aeronautical English Proficiency Exam (EPLIS) for the Brazilian Airspace System will be appraised. EPLIS item development process will be examined in order to see how it contributes to the test construct validity, in terms of offering empirical evidence and theoretical rationales to ensure that the interpretations of test scores are meaningful for the uses and impact desired by its test developers.

Keywords: language assessment for professional purposes, listening comprehension, aeronautical English, EPLIS paper 1, test construct validity

Introduction

This paper addresses the Language Proficiency Requirements (LPR) established by the International Civil Aviation Organization (ICAO) in 2003 as a licensing prerequisite for pilots, air traffic controllers and aeronautical station operators involved in international flight operations. These professionals have been required to demonstrate their ability to speak and understand the language used in radiotelephony communications through formal evaluation (ICAO 2004, 2010).

ICAO provisions related to language proficiency are published in Annex 1 together with a proficiency six level scale, from Pre-elementary to Expert, to be used in the assessment of these professionals. Operational Level 4 in the ICAO Rating Scale has been considered the minimum level acceptable to ensure safe operations, and professionals at ICAO level 4 and 5 should have their proficiency reassessed in periods no longer than 3 and 6 years, respectively.

In compliance with ICAO LPR, the Airspace Control Department (DECEA), through the Airspace Control Institute (ICEA), has designed the Aeronautical English Proficiency Exam (EPLIS) to assess speaking and listening skills for Brazilian air traffic controllers

and aeronautical station operators (P. Tosqui-Lucks et al. 2016). EPLIS comprises two papers. Paper 1 is a computer mediated listening test, with 30 multiple choice items. Test takers must score 21 to be eligible to sit Paper 2.

Paper 2 assesses integrated listening and speaking skills in an interview format. It is conducted with one test taker at a time and lasts about 15 to 30 minutes. The test taker performance is assessed by two raters: the interlocutor who gives a holistic score and a second rater who employs the ICAO Rating Scale, giving a score to each of the six areas of the scale (Pronunciation, Structure, Vocabulary, Fluency, Comprehension and Interactions). There are concurrent versions of Paper 2 according to test takers' professional profile: a radio/tower, an approach control, an area control centre or an ab-initio version for pre-service professionals.

Approximately 2500 professionals take EPLIS annually. Around 97% of them are air traffic controllers and only 3% are aeronautical station operators. In relation to air traffic controllers, 32% work at control towers, 32% at approach control, 19% at area control centres and 17% in other areas.

Aviation language tests like EPLIS are considered very high-stakes as their results inform important decisions that bring about serious consequences. As P. Souza (2020) points out, professionals may be denied a licence to operate internationally if they do not comply with ICAO LPR. On the other hand, States cannot afford to lose competent professionals for whom significant amounts of money have already been invested in language training and testing. B. Aragão (2018) adds that language proficiency has become one of the items evaluated in the ICAO Universal Safety Audit Program, in which non conformity may result in political and financial losses to a country.

Given the high-stakes nature of EPLIS, test developers have the responsibility to clearly demonstrate that the test is accurate, reliable and fair to test takers. In this paper, EPLIS Paper 1 construction process will be examined in order to show how quality is ensured. Firstly, the concept of construct validity is briefly explained. Secondly, EPLIS construct is described, as well as Paper 1's main characteristics and the rationales behind them. Then, the procedures followed in the development of test items will be explained. To conclude, it will be discussed how EPLIS paper 1 contributes to ensure that the interpretations of its test scores are meaningful for the intended uses and impact proposed by its test developers.

1. Construct validity

S. Messick (1989: 13) proposes a unified view of test validity defined as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on the test scores". In a process of validation, S. Messick (op. cit) argues that evidence from different sources must be collected in order to guarantee that all score-based inferences and uses are meaningful, useful and appropriate. In this unitary concept of validity, reliability, once understood as one of main qualities of a test alongside validity, becomes one type of evidence to be collected, which refers to the degree of test scores consistency. Likewise, decisions regarding practicality are now seen as another type of evidence as well as the social consequences of tests.

S. Messick (op. cit) also warns test developers about the two major threats to the validity of a test: "construct underrepresentation" and "construct irrelevant variance". In construct underrepresentation, the test does not include important aspects of the construct, defined as the ability the test developers intend to measure in a test. Construct irrelevant variance, on the other hand, refers to variance in test scores related to factors other than the construct in question (A. Davies et. al. 1999: 32–33).

From this perspective, it is important to point out that validity is not an all or none concept, but a matter of degree. Its inferential and ongoing nature makes that "the existing validity evidence becomes enhanced (or contravened) by new findings, and projections of potential social consequences of testing become transformed by evidence of actual consequences and by changing social conditions. Inevitably, then, validity is an evolving property and validation is a continuing process. Since evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean" (S. Messick 1989: 1).

Validity must be followed since the very beginning of the test development process. This means that a test must represent the construct to be measured starting from the observed communicative behaviour, the tasks, items, scoring criteria, grades, to decisions and extrapolations of test results, so that it may be considered useful for its purpose.

2. Test specifications

Test specifications is an explanatory document for the construction of a test. G. Fulcher/ F. Davidson (2007) argue that they play an important role not only in the process of creating tasks and tests, but also in explaining the rationales behind the choices made. ALTE (2011) states that the test specifications should include what is tested (the test construct), how it is tested, and details regarding assessment criteria, test format, number and length of papers, item types used, etc.

Aeronautical communications are defined by ICAO (ICAO 2010) as the language used in the exchanges between pilots and air traffic controllers or aeronautical station operators. It includes the use of standardised phraseology and plain language. In international civil aviation, the language used is English. Standardised phraseology is a specialised code for communication within routine situations, characterised by "a reduced vocabulary [...], deletion of function words, auxiliary or link verbs, subject pronouns and many prepositions" (ICAO 2010: 3–4). The plain language, otherwise, is "the spontaneous, creative, and non-coded use of a given natural language" (ICAO 2010: 3–5), mainly required in urgent or emergency situations but also used to share information in everyday situations for which phraseology does not exist.

G. Buck (2001: 51) states that "listening comprehension is a complex, multidimensional process, and a number of theorists have attempted to describe it in terms of taxonomies of sub-skills that underlie the process". G. Buck (2001:54) cites the communicative listening sub-skills proposed by C. Weir: listening for gist; listening for specifics and important details; listening for main idea and supporting details; and listening to determine a speaker's attitude or intention.

In aeronautical radiotelephony communications, listening also involves different purposes, for example: a pilot *requests deviation* (gist); a pilot requests to descend to *flight*

level 180 (specific information); a pilot informs that the *right engine* is shutdown (important detail); a pilot informs that a passenger seems to have a *heart attack* because he is complaining of *chest pains* and *trouble breathing* (main idea and supporting details); the request sounds urgent because of the *pilot's tone of voice* (speaker's attitude).

Therefore, when making decisions on the test development, test developers must be aware of the difficulties imposed by the nature of listening itself and the complexity involved in listening processing: the types of knowledge involved whether linguistic (phonology, lexis, syntax, semantics and discourse structure, length, speed, accent, intonation) or non-linguistic (knowledge of the context and specific facts about the way things happen in a certain context), as well as the restrictions imposed by it.

Language processing itself must also be considered within the context it occurs. Given that aeronautical communications take place mainly via radiotelephony, they are characterised as: requiring listening and speaking skills; being highly dependent of technical knowledge; having an absence of visual and kinetic cues; having interlocutors separated in space and transmitting each one a message at a time; and having poorer acoustic conditions than face to face interactions (ICAO 2010: 3–2).

Moreover, comprehension under these conditions can become particularly difficult and challenging especially when complications and an unexpected turn of events will demand the use of plain English to be dealt with. In addition, interactions in aeronautical communications take place within an international community of users with different levels of proficiency in English and with their own accents and delivery styles.

Those considerations and characteristics are especially important because test developers must have clarity of the ability they want to measure in their tests. The ability we want to measure and its underlying sub-skills are what we can call the construct of the test, and it must be represented in every stage of the test development cycle. By ensuring a test measures the construct and nothing else is added or subtracted, S. Messick (1996) explains that test developers are able to maximize the test positive impacts although this is not, by any means, a guarantee since there are other factors that mediate the washback process, as demonstrated by P. Souza (2020), A. Green (2007) and L. Cheng (2005).

3. EPLIS Paper 1

The ICAO Language proficiency Requirements (ICAO 2010) specify that speaking and listening skills should be assessed in the context of aeronautical communications. Thus, listening comprehension needs to be considered integrated to speaking in a communicative approach to language testing.

However, considering the major role of listening in aeronautical radiotelephony communications, which comprises at least half of the workload of pilots and air traffic controllers, comprehension is assessed in EPLIS in two distinct moments: as an isolated skill in Paper 1 and as integrated to speaking in paper 2. This decision has contributed not only to EPLIS practicality but also to its validity. So, through paper 1, it is possible to assess test takers' ability to understand a wide range of accents, regional varieties and delivery styles, in different topics and domains of aviation covered in Appendix B of Doc 9835, and under unfavourable acoustic conditions.

EPLIS paper 1 is a computer-mediated listening test, with 30 multiple choice items and for each one there is a pilot-controller transmission. Test takers must score 21 out of 30 items to be eligible to sit Paper 2, with 1 mark per question.

The audio files in paper 1 are authentic aeronautical radiotelephony communications, in which the crew or the air traffic controllers give or request clearances, permission/approval, information, reasons, instructions; check, confirm or clarify messages; deny or refuse clearances, permission/approval, etc. (ICAO 2010).

The use of actual spoken texts aims to reflect real life situations and their own characteristics in terms of accent, speed of delivery, distinct levels of proficiency, and acoustic features of radio frequencies. It is also an attempt to foster positive washback in training programs in which scripted and manipulated audio material has been broadly used.

Multiple choice questions (MCQ) have been proved useful in dealing with different types of listening behaviours. Besides, given that MCQ is the most popular and widely used test method in Brazil's Education, the chance a test taker gets a lower score for not being familiar with the test format is minimised. Considering that approximately 2500 test takers sit EPLIS paper 1, MCQ also contributes to EPLIS practicality as it facilitates marking and provides more reliable scores.

On the other hand, MCQ involves some amount of reading. As T. Haladyna et al. (2002) point out, test takers' performance can be affected by the reading demand of the item. In order to minimize the influence of reading ability, a source of construct-irrelevant variance in listening tests, EPLIS paper 1 items are written in simple language and in Portuguese. In addition, test takers are allowed to listen to the sound files twice and the items contain three options only (A, B, C)[1].

4. EPLIS Paper 1 development process

Following R. Green's task development cycle (2017), once the test specifications have been developed, the next stages consist of selecting of appropriate audios, extracting of the information on which the item will be based, and the construction of the item itself. The item is then subjected to a rigorous process of review, editing and pretesting. Following that, the results of the pretesting are analysed statistically to check if the items comply with the requirements. Items that showed good statistics are banked and can be used in a real test. Items with poor statistical features are qualitatively analysed by a test developer.

4.1. Selecting and textmapping sound files

EPLIS test developer is in charge of identifying sound files in accordance with the test specifications: authentic aeronautical radiotelephony communications in which interlocutors had to rely on the use of plain English in order to deal with unexpected, unusual or even emergency situations in different air traffic service facilities.

The next stage is to form a group of three or more EPLIS item writers to ponder on the adequacy of the sound files. EPLIS item writers are EPLIS examiners, language

[1] Recent studies have also argued that three-option items discriminate as well as four-option items, besides taking less time to be constructed (H. Lee, P. Winke 2013).

experts or subject matter specialists, with a vast experience and familiarity with aeronautical communications, and who received training in item writing provided by an EPLIS test developer. These professionals are qualified for judging the audios in terms of their intelligibility, length, speed of delivery, background noises, domain, ATS facility they refer to, phraseology and plain English use, as well as if they are suitable for the listening behaviours they are targeted.

R. Green (2017) proposes a procedure called 'textmapping' to assist test developers in evaluating the appropriateness of the sound files and in exploiting their content. The textmapping is described by R. Green (2017: 57) as "a systematic procedure which involves the co-construction of the meaning of a sound file (or text)". Participants are required to focus solely on the sound file instead of the transcript of the audio. By experiencing the audio as listeners, participants will be able to make a more robust judgement about its difficulty.

In order to conduct a textmapping activity, the steps to follow will depend on the type of listening behaviour targeted by the test developer who first identified the sound file. So, when selecting the audios, the test developer must specify the type of listening that is being targeted.

In case of textmapping for gist, for example, the test developer must first be sure the item writers have a clear understanding of what gist means. Before playing the sound file, a context must be provided. Participants are instructed to listen to the sound file only once and take no notes while listening. Then, they have to synthesise the main idea of the recording in a sentence. R. Green (2017: 61) recommends writing down between 14 to 20 words. Silence is required during all the procedures. At the end, the test developer who is conducting the activity collects the sentences and compares them in order to check if there is a consensus on what they extracted from the audio. In a group of four item writers, three item writers must have had the same understanding in order to have a consensus of opinion. A high consensus in textmapping is defined as n - 1 (n = number of participants) (R. Green 2017: 61).

The rules are not very different if the test developer intends to textmap a sound file for specific information and important details[2]. Instead of writing down a sentence, participants are instructed to make a list containing types of specific information, such as ‘runway in use’, ‘wind speed’, ‘approach procedures’, ‘taxi instructions’, or important details such as ‘flameout engine’. Fig. 1 shows the results of a textmapping procedure conducted to identify specific information and important details.

[2] For more details about how to conduct a textmapping procedure for different types of listening behaviours, check R. Green (2017: 55–84).

Audio's file name: J19A39_02_SIID						
Point mapped	CB	GC	SM	SF	Target	Author
16 NM	x	x	x	x	Q1	SM_NG
Not enough		x	x	x	Q2	SM_NG
Lose altitude		x	x			
3000 feet	x	x	x	x	Q3	SM_NG
4 NM	x					
localizer	x	x				

Figure 1. Textmapping results for audio J19A39_02.

The above table includes the information that was taken away by the participants while listening to the recording. It is noted that the distance '16 miles' as well as the altitude '4000' reached consensus, 4 participants out of 4 understood the same information, which indicates that a test item might focus on these specific pieces of information. The fact that the distance was 'not enough' might also be targeted as it was understood by 3 out of 4 participants. The other textmapped points did not reach consensus and cannot constitute the focus of a test item.

In case a task contains two or more items based on the same recording, it is important to mark the time a particular piece of information occurs in the textmap table to see if there will be enough time to answer both items. If two pieces of mapped information are too close to each other, only one of the items can be used, as R. Green recalls (2017: 72).

4.2. Item writing, peer review and revision

EPLIS item writers must strictly follow a set of multiple choice item construction guidelines in order to standardise the test item writing process. They are closely supervised by EPLIS test developers to ensure the items produced comply with the test specifications.

Firstly, the EPLIS test developer pairs up the textmapping participants, preferably a language expert with a subject matter specialist, and distributes the sound files together with their respective textmap tables. Each pair has to write the item focusing on the information which there was consensus on and following the item writing guidelines to create the stem and the options/alternatives. Although, Paper 1 items have 3 options, at this stage item writers develop 4 options. The option that does not work properly during the pretesting will be eliminated. Item writers also have to estimate the item level of the difficulty taking into consideration the saliency of the information in the audio, the speed delivery, the accent, the level of background noise, the use of phraseology or plain English, and the quantity of information being processed to get to the right answer.

Once the items are ready, an EPLIS test developer sets up a group of reviewers to give feedback on the quality of the items produced. The reviewers are EPLIS item writers who did not take part in the textmapping of the audios the items refer to. It means that reviewers

are familiar with the test specifications as well as with the item writing guidelines, ensuring that the feedback be constructive and useful.

Following ALTE's recommendations (2011: 28–29), reviewers first try to answer the item without reference to the text (written or oral) as this helps to identify whether common sense or background knowledge or even any hint left by the item writer can lead test takers to choose the correct response. After that, they answer the item as if taking the test. They take notes of their answers and later compare them with the key. This will help reviewers identify possible flaws in the items, such as wrong key, more than one possible correct answer, unclear or badly phrased options, ridiculously implausible distractors, or items which are very difficult and probably tap into a different type of knowledge.

Reviewers must check the textmap table to see whether the information targeted on the item refers to the point on which there was consensus. The wording of stems and options must also be revised if they do not conform to the item writing guidelines. Any problem raised must be discussed in detail by the reviewers and a record of all the changes must be clearly kept. Lastly, an EPLIS test developer receives the edited material and is responsible for making final decisions about the acceptance of the items.

Figure 2 shows the item produced and revised by EPLIS item writers based on the textmapping results for audio J19A39_02 (Figure 1.1). The item was numbered as Q791. It is a multiple choice question with 4 options. The key is D. The audio it refers to is named J19A39_02 and was recorded in an approach control facility. Its content is related to routine situations and the items' level of difficulty is estimated as high (a difficult question).

QUESTION 791	KEY	AUDIO	ATS	IF EST	DOMAIN
Para a perda de altitude, o piloto necessita de: a) menos de 6 milhas b) 6 milhas c) 16 milhas d) mais de 16 milhas	D	J19A39_02	APP	D	Routine
English version: **For altitude loss, the pilot needs:** a) less than 6 miles b) 6 miles c) 16 miles d) more than 16 miles					

Figure 2. Item for audio J19A39_02.

4.3. Pretesting new items

Items need to be pretested to see how well they are working. EPLIS Paper 1 pretesting programme is like a live test, with a representative sample of test takers in sufficient numbers to enable various possibilities for statistical analysis.

EPLIS pretesting program has the following characteristics: pretest test takers are

professionals who are preparing to take the live test; live test conditions are observed in terms of alloted time to do the test, the software used to deliver the test and the test method; and test venues and staff are prepared to securely deliver the pretest paper as the items can go into the actual live test.

Despite the complex logistics of the pretesting phase, the information this process can provide is particularly important and valuable in order to judge the item effectiveness and level of difficulty so that informed decisions can be made on the acceptance of the items to the live test.

4.4. Item Statistical analyses

Statistical analysis of pretest scores provides EPLIS test developers very useful information about the quality of the items and is one way of preventing poor items from going to live tests.

Data gathered at the pretesting is analysed using classical statistics through SPSS. The analyses conducted provide information on item facility, item discrimination, distractor performance and test reliability.

Facility refers to the proportion of correct answers to an item. It can be reported on a scale of 0 to 1 or as a percentage. It provides information on how easy the item was for the group of pretest takers.

In a proficiency test, the appropriate level of difficulty is at the mid-point (L. Bachman 2004), but facility values between 30 to 70 percent can also provide good information on test takers and/or on the items. The item shown above had a facility value of 0.278. It means that 27.8% of pretest takers answered the item correctly, which is considered low. Considering that the facility value of this item is very close to the limit range (30%), it is not rejected at once. It goes back to the revision stage and after adjustments is ready to be pretested again.

Distractor analysis shows the proportion of test takers choosing each distractor. Fig. 3 shows the percentage of test takers choosing each option. Distractor C attracts more responses (47.2%) showing it is a good distractor, whereas distractor B attracts very few test takers (5.6%). Because EPLIS live test items contain 3 options, the distractor that is not working very well is probably the one that will be rejected.

Q791

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	5	13.9	14.7	14.7
	B	2	5.6	5.9	20.6
	C	17	47.2	50	70.6
	D = KEY	10	27.8	29.4	100
	Total	34	94.4	100	
Missing	No answer	2	5.6		
Total		36	100.0		

Figure 3. Frequencies for item 791.

Discrimination refers to the extent the item discriminates between weaker and stronger test takers. It is measured on a scale of -1 to +1. An item with a high discrimination index, close to +1, shows that strong test takers are answering the item correctly whereas weak ones are answering it incorrectly. If the index is negative, it means the strong test takers are getting the item wrong. This indicates that there might be a problem with the key. Items with a value of 0.30 are considered suitable.

The item in the example provided had a discrimination index of 0.247. Because this item was considered difficult for this group of pretest takers, its discrimination might have been underestimated as strong and weak groups scored badly. Anyway, the item needs to be revised.

Internal reliability (consistency) of scores is calculated in EPLIS paper 1 using Cronbach's Alpha and refers to the degree the items are measuring the same underlying construct. It is reported on a scale of +1 to -1. The higher the Cronbach's Alpha value is, the higher the internal consistency among the items. Internal consistency values above 0.7 are considered acceptable.

Figure 1.4 shows the Cronbach's Alpha values for the test in which the item in the example was part of. The Cronbach's Alpha for the test as a whole was calculated in .807, which shows a high level of reliability. The table also contains the values for Cronbach's Alpha if Item Deleted (column 5). So, if the item in the example is deleted from the test, the consistency of the test as a whole drops to .805. It means that the item contributes to the test reliability, but does not add much to it as its power of discrimination is weak.

Cronbach's Alpha	N of Items
.807	30

Item- Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Item 767	22.3889	15.844	.542	.794
Item 777	22.3889	16.759	.129	.808
Item 787	22.4722	15.685	.436	.796
Item 790	22.7222	15.806	.271	.805
Item 791	23.0278	16.028	.247	.805

Figure 4. Cronbach's Alpha values.

4.5. Item live administration

The items that did not show good statistics go through a qualitative analysis carried out by the EPLIS test developer who decides whether the item needs to be revised or cancelled. In the case of item 791, an EPLIS test developer opted to mark the item for revision as the stem showed ambiguity. Items that are revised need to be pretested again. On the other hand, the items that showed strong statistics are banked.

Once banked, the item is ready to be used in the next live test with the facility value it obtained during the pretesting. After its first administration, an analysis of its behaviour will be conducted again and its facility value will be recalculated for the next administrations.

Conclusion

Listening is considered a receptive skill as far as it relies on the understanding of an utterance and not on its actual production. But is there a passive way of comprehending and processing information without getting actively involved in it? Specially because there are "different degrees of listening comprehension" which may vary depending on the context, background knowledge, familiarity with the topic and language proficiency, test developers must bear in mind that even in a very specific context like aviation and, more specifically, in aeronautical radiotelephony communication (in which there is a script to rely on and operational procedures to be followed for every situation), listening is highly dependent on the context and specific knowledge. Considering the context in which it occurs, it is important to point out that, in radiotelephony, listening plays a major role and accounts for at least 50 percent of the workload of air traffic controllers. If a controller cannot make sense of an utterance nor is able to negotiate its meaning, safety may be hindered. In this regard, test developers must give the due amount of representativeness and importance while making decisions about the items, tasks and scoring so that they may represent as close as possible the real-life situation.

During a validation process, for every claim or decision that is made based on the test scores, there must be explicit statements and justifications to support them. Likewise, statements that challenge or reject those qualities must be addressed. Especially on high stakes exams, test developers must assume responsibility to constantly address and review the interpretations and uses made on their test scores on a continuing basis, by articulating evidence that support and/or weaken them.

In the case of EPLIS paper 1, the obtained scoring and the interpretation made on it accounts for the claim. If a candidate scores 21 points, the claim made is that he has the ability to understand a variety of accents, different speed and rhythm in a range of topics covered in Appendix B of Doc 9835, and the decision is that he is eligible to continue the assessment process during paper 2. In turn, the support to this claim is that EPLIS paper 1, from its content to format and results, preserves and represents the language used in aeronautical radiotelephony communications in the best way possible. Characteristics of the setting, timing, rubrics, scoring, as well as the characteristics of the input, language used, topical knowledge, expected response and the relationship between input and response (L. Bachman/ A. Palmer 1996) are carefully thought of to refer to the real-life situation.

In addition, statistical analyses conducted during the pretesting and live administration of the test account for EPLIS reliability. The facility value of each item is first estimated by test developers based on the test specifications. After analysing the item behaviour in the pretest, the value is either confirmed or reset. It is only after the real-life test administration that the final level of facility is set for every item in the item bank. After each test administration, test developers analyse and contrast candidates' performance in Paper 1 and 2 regarding candidates' level of listening ability. Along with some other empirical and statistical analyses, test developers can provide quality evidence that justify the use of Paper 1 both in terms of its suitability to the test purpose (test specifications) and to the target population (real test takers, in this case, air traffic controllers and aeronautical station operators).

As test takers participate in the pretesting stage, they are not only contributing to the test development and maintenance process, but they may also benefit by finding similar conditions as in the real test administration in terms of content, length and format, which, in turn, reduces error measurement caused by not being familiar with the test method.

As for practicality, the MCQ format allows a large number of air traffic controllers to take the test as it poses no restraints both in terms of financial and human resources available as well as test correction. The benefits of using this format and method were discussed in the paper. As noted by L. Bachman/ A. Palmer (1996) the practicality of a test development process relies primarily on the relationship between the required and available resources.

By doing this cyclical process, though challenging, the test is constantly being revised and evaluated, so that best practices are pursued. As an ultimate aim, one can claim that a candidate who meets the setting parameters of EPLIS will mostly be able to understand a variety of accents, linguistic or situational complications that he may face at work. Since EPLIS is a high stake exam, candidates that succeed in Paper 1 tend to feel more confident to handle the complications they may face at work and feel more encouraged to proceed to paper 2.

To conclude, EPLIS paper 1 has been constantly evolving and aiming at best practices in test design and construction. Also, it has produced positive consequences in training programs as listening practice sessions have reflected the situations encountered in real life more closely and authentically. Ultimately, by improving test takers' proficiency in understanding aeronautical radiotelephony communications, EPLIS paper 1 contributes to aviation safety.

References

Association of Language Testers in Europe (2011), *Manual for Language Test Development and Examining.* Strasbourg: Council of Europe.

Aragão, B.F. (2018), *O uso de critérios autóctones no contexto aeronáutico: Contribuições para uma nova escala de proficiência para controladores de tráfego aéreo*, (in:) M.V.R Scaramucci/ P. Tosqui-Lucks/ S.M. Damião (eds), "Pesquisas sobre inglês aeronáutico no Brasil". Campinas: Pontes, 243–269.

Bachman, L. (2004), *Statistical Analyses for Language Assessment.* Cambridge University Press: Cambridge.

Bachman, L./ A. Palmer (1996), *Language Testing in Practice: Designing and Developing Useful Language Tests*. Abingdon, Oxford: Oxford University Press.

Cheng, L. (2005), *Changing language teaching through language testing: a washback study*. Cambridge.

Buck, G. (2001), *Assessing listening.* CUP: Cambridge.

Davies, A. et al. (1999), *Dictionary of language testing*, (in:) "Studies in Language Testing 25". Cambridge University Press and Cambridge ESOL.

Fulcher, G./ F. Davidson. (2007), *Language Testing and Assessment.* Routledge: London & New York.

Green, A. (2007), *IELTS washback in context: Preparation for academic writing in higher education*, (in:) "Studies in Language Testing 25". Cambridge University Press and Cambridge ESOL: Cambridge, UK.

Green, R. (2017), *Designing Listening Tests: A Practical Approach.* Palgrave Macmillan Limited.

Haladyna, T. et al. (2002), *A review of Multiple-Choice Item-writing Guidelines for Classroom Assessment*, (in:) "Applied Measurement in Education" 15.3, 309–334.

International Civil Aviation Organization (2010), *Manual on the Implementation of ICAO Language Proficiency Requirements:* Doc 9835 AN/453. 2nd ed. ICAO: Montreal.

International Civil Aviation Organization. (2004), *Manual on the Implementation of ICAO Language Proficiency Requirements:* Doc 9835 AN/453. 1st ed. ICAO: Montreal.

Lee, H./ P. Winke. (2013), *The differences among three-, four-, and five option-item formats in the context of a high-stakes English-language listening test*, (in:) "Language Testing" 30, 99–123

Messick, S. (1989), *Validity*, (in:) R. Linn (ed,), "Educational measurement". 3 ed. New York: Macmillan.

Messick, S. (1996). *Validity and washback in language testing*, (in:) "Language Testing" 13.3, 241–255.

Souza, P.R. (2020), *The washback of EPLIS on teachers perceptions and actions: Implications for reviewing ICAO language policy in Brazil*, (in:) "The Especialist" 41.3, 1–18. (URL https://revistas.pucsp.br/index.php/esp/article/view/47394/33385). [Accessed on 3.30.2023].

Tosqui-Lucks, P. et al. (2016), *Ensino e avaliação de Língua Inglesa para controladores de tráfego aéreo como requisito de segurança em voo*, (in:) "Revista Conexão SIPAER" 7.1, 44–54. (URL http://conexaosipaer.cenipa.gov.br/index.php/sipaer/issue/view/19). [Accessed on 4.25.22].