Apr 24th, 2:30 PM - 2:45 PM

How does test design influence training? Washback effects of LPR tests

Michael Kay
President, ICAEA

How does ICAO LPR test design influence training?

The Washback Effect of ICAO LPR Aviation English Tests

Michael Kay

June 2017

Introduction

In the field of language testing, especially high-stakes language testing, where test results can have serious consequences for test-takers – for example, the results being used to make decisions on career pathways, entry into universities, permission to immigrate, or in the case of aviation with the ICAO LPRs – approval to operate as ATCOs or pilots, the tests have a lot of power and influence on individuals, organisations, and in some cases, society. Unfortunately, however, often proficiency tests are designed and developed without much consideration of the ripple effects they have and test developers can be unaware of the significant effects their tests have on lives of the people directly and indirectly affected, including their perceptions and attitudes towards language proficiency – and language training.

There are a number of criteria which affect the quality of a language proficiency test instrument, including all the various facets of validity (construct, content, face and consequential validity), reliability, practicality and the test's impact. In the established mainstream language testing industry, much attention and effort is given to ensuring the test is an effective evaluation tool. Rigorous piloting and reviews are conducted to check the test format and design is effective. This is followed by test validation trials, reviews and adjustments to ensure the test content, tasks/items and levels of difficulty can be categorised or aligned with 'grades' so that the results the test generates are meaningfully able to differentiate between the established target proficiency levels for each version of the test. While these processes are essential best practice processes, the fundamental starting point, irrespective of these criteria, is to ensure at the outset, that a language proficiency test is built on a solid foundation – that is the test tasks, content and format (delivery) are appropriate. This design consideration is fundamental and the foundation for any quality test. Only once this has been established and that the test construct and how the test works as a measure of language proficiency, can other aspects of test quality be considered in a worthwhile way.

In aviation English ICAO LPR testing, many tests of varying standards have been developed for use in different countries. While many ICAO LPR proficiency tests have been scrutinised in terms of their quality, there has been less scope, within a relatively short period since the introduction of the LPRs, to consider the effect these high-stakes tests have on downstream language training.

Test developers with expertise in language testing are more likely to be aware of the effects their tests can have on language training or attitudes to testing, training or even the aims of the ICAO LPRs. More importantly, test developers who produce poor quality tests (most likely as a result a lack of expertise and an awareness that language testing is a highly technical and specific field requiring dedicated expertise) are much less aware of the negative effects their tests can have on the pilots and controllers who sit these tests. These test developers are less aware of the effects poorly designed or developed tests have on the perception and attitude towards English testing, training and its role in communications and safety among airlines, air traffic control providers and licensing authorities (aviation regulators).

It is for these reasons that it is the responsibility of test developers to develop and implement good quality tests to minimise the negative effects their tests can have on the people and organisations who are affected.

What do we mean by the washback effect?

It is often not recognised or understood that the structure, content, skills focus, methodology, test task types and delivery style of a test can strongly influence training programmes which are developed in response to a proficiency test.

Therefore, while a test may not be directly related to a training programme, it can significantly affect the way it is taught and students' attitudes towards learning. As a result, these features of the test do have a strong influence on the outcome and success of the language training programmes in terms of how well they develop proficiency and equip students with the required language skills.

Test Washback

Washback is typically referred to the extent to which the test influences language teachers and learners to do things they would not otherwise necessarily do (Alderson and Wall, 1993).

The concept of washback can be extended to not just refer to the effect language tests have on language training and learning (the curriculum, teaching and learning styles), but also the wider consequences, within Messick's framework of consequential validity (Messick, 1989). This includes the effects and influences tests can have on attitudes and values towards policies, training and even the use of the language among students, teachers and organisations where the results of these tests serve a purpose and have consequences.

The test washback effect is the influence that a test has on the way students learn and how they are taught. Positive washback occurs when the design, content and implementation of a language test leads to meaningful and useful language development. Positive test washback means the test has positive effects on the curriculum, teaching, language development and learners' values and attitudes towards learning.

Negative washback occurs when minimal meaningful or useful language development occurs because teaching and learning focuses exclusively on preparing for a poorly designed and developed test at the expense of developing required language proficiency and skills needed in real-world situations. Clearly, the direction of the washback – positive or negative – is determined by the quality of the language test. Negative washback arises when the test construct fails to align with the target language usage situations. As a result, test results are in fact not a valid means of evaluating language proficiency in the context of the original purpose for the implementation of the assessment. Typically this occurs when the test construct is based on a narrow concept of language ability or suffers from construct under-representation by not adequately covering a sufficiently wide array of language usage situations or contexts, therefore constrains the teaching/learning (Green, 2014).

Similarly, language tests which are flawed as a result of insignificant attention given test construct and alignment with real-world language needs are likely to have negative washback effects. These flaws can manifest themselves in the test instrument either through poor test task design and/or test content (forms of language and the contexts in which language is used) that do not reflect the required target-language use situations. In other words, tests which lack authenticity and/or are too narrow in how they define the language proficiency they aim to assess – in the types of language and formats in which the language is used as representations for the assessment (the test content and test task types) are likely to have negative washback effects.

Shohamy (1993) notes that external proficiency tests play a powerful role in modifying the behaviour of those affected by the results of these test – beyond just the teaching and learning context, to also include administrators and agencies. In such cases the authority the test imposes can influence curricula, teaching methods and attitudes to learning.

Organisations and individuals who are responsible for supporting test-takers who need to sit high-stake tests feel a pressure to enable their students to succeed and pass the test. It is only natural that teachers and organisations who are affected by the test want their students to succeed.

Does this influence the test has on training always undermine the effectiveness of the training and throw into question the validity of the test? It depends. Language tests which are designed to reflect the real-world communicative language needs of test users and assess language use in contexts indicative of real-world target language use situations are more likely to lead to positive washback. The results of these language tests are more meaningful to all stakeholders because there is confidence that the test results are more likely to be aligned to real-world communication needs, therefore reflect how well the test takers are able to perform and use the language in real-life. On the other hand, poorly designed tests which contain content or task types which bear little resemblance to the real target language use situations in which the test takers need to be assessed are more likely to result in negative washback.

Below are two completely fictitious examples of language tests demonstrating how washback can occur, highlighting the relationship between real-world target language use needs, test design and language training programmes. The first example relates to a positive washback scenario and the second provides an example of a negative washback effect.

Hypothetical case study 1: A test with positive washback effects

Imagine a situation where the Canadian medical industry employs doctors from all over the world. Many of these doctors may have English as their second or third language and have done their medical training overseas in non-English speaking countries. Imagine hospitals need to ensure the doctors are competent in English so that these doctors can:

1. Listen to talks about medical procedures;
2. Read and understand drug and medical related texts accurately;
3. Communicate with patients successfully (diagnose and explain treatments);
4. Communicate with nurses and other staff successfully in hospital environments;
5. Write patient reports in English.

To ensure doctors are proficient in English and can work successfully in Canada, suppose there is a requirement that overseas qualified doctors must pass an English test – The Medical English Proficiency Test (MEPT).

Because the doctors need to demonstrate proficiency in specific skills, using specific language in medical and hospital situations, this special test has been developed. The test is developed by a university at their educational measurement school. A team of language test experts developed the test and continue to monitor and upgrade the test. There are now 20 versions of MEPT. Versions that are more than three years old are retired and new versions are added. The test is now delivered at five specialised test centres across the country.

Table 1: *MEPT Test overview*

Test part and duration	Skills assessed	Content	Delivery and task-types
1 (10 minutes)	Listening comprehension to: Understand the main and specific ideas in a detailed medical lecture	3 short medical talks on new developments in medicine	Paper-based: Three 4-5 minute recordings with 10 short answer questions (paper-based)
2 (15 minutes)	Listening comprehension to: Understand patient needs and symptoms Understand doctor-patient relationship Infer patient feelings and moods Understand treatment methods proposed	5 short patient and doctor conversations	Paper-based: Take short notes to complete information in tables
3 (15 minutes)	Reading to comprehend: Key information in a range of technical medical texts	4 short medical extract texts related to drug delivery or treatment plans	Paper-based: 5 short answer questions for each text
4 (15 minutes)	Report writing	2 short reports: Candidate listens to a 2 minute recording comparing two medical treatments (case studies) and is then given a set of 3 questions to respond to	Paper based: Candidates write responses to 3 topical questions and give opinions, compare treatments or describe the advantages/disadvantages of each
5 (15 minutes)	Speaking and comprehension. (communicative ability)	2 short roleplays with patients	Roleplays with interlocutor pretending to be a patient (4-5 minutes each) Performance is evaluated live by interlocutor and recorded for rating by two raters later.
6 (5 minutes)	Speaking (communicative ability)	Interview roleplay: Candidate listens to a short explanation of a typical problem described at a hospital	Interview: 8 questions are asked requiring a summary of the situation then asking for opinions and problem solving ideas

Washback effects of the MEPT Test

Imagine, in this scenario, if doctors do not achieve BAND A on the test (based on a 4-band rating scale, with BAND A the top level), they are not permitted to receive a medical licence in Canada. This is because the government considers it a risk to safety to not have highly proficient English speaking doctors employed in hospitals. This means the careers of doctors is at stake. If they achieve BAND A, they can apply to be a doctor in Canada. If not, they are barred from practicing medicine in Canada. A lot is at stake for these test-takers so this can be referred to as a high-stakes test.

As such, some universities and language schools provide specialised language training programmes to help foreign doctors achieve a good standard of English before they take the MEPT.

The training programmes vary but they focus on developing reading, writing, speaking and listening skills related in a range of medical situations. Vocabulary, grammar, reading, listening comprehension and writing skills as well as pronunciation and conversation skills are central to the curricula.

One popular programme is a 12 week course (4 hours per day in a class situation with an instructor and 1 hour of online self-study). The training programme focuses on mirroring the kind of content and tasks in the MEPT.

Student motivation is high and most students who take the course are positive about their learning. Teachers must have some knowledge in medicine and the school selects highly qualified instructors who are able to use communicative teaching methodology and develop curriculum materials that meet the students' needs.

Overall, we can say the washback is positive because of the effect the MEPT has on the training programmes: the MEPT test design, delivery and content reflects the real-world language usage situations in which doctors use English for their work. And, because of this, the training programme also develops the skills required to develop English – both to achieve BAND A on the test but also to successfully use English in their jobs in hospitals. The programme contains a lot of variety, many activity types, and focuses on developing all the skills that allow students to participate and complete tasks on the MEPT successfully. The test and programme are well respected and popular among migrant doctors. It is considered rigorous and tough but fair. The Canadian government and hospitals trust the results of the test as doctors who achieve BAND A go on to integrate into the Canadian medical industry confidently and successfully.

This would be a clear case of positive washback where the test reflects real job needs and so the training also reflects these needs when preparing students to sit the MEPT test. In effect, there is no difference in teaching to prepare students for the test and teaching students to become proficient users of English for their future jobs in Canada.

Hypothetical case study 2: A test with negative washback effects

In this example, imagine Korea has changed its banking laws. As a result foreign banks are now allowed to operate in Korea. International banks start making applications to the government to open offices and branches in Korea. However, the Korean Ministry of Finance makes a requirement that the banks must employ Korean citizens. The banking industry agrees but states that all staff working in the foreign banks opening in Korea need to have English speaking staff. As a result, the banks agree that a new English test is needed to select new staff to be employed in international banks. The Korean Government Ministry of Finance approves this idea but takes it further and makes a new requirement: all bank staff must now be able to speak English and therefore must pass a proficiency test.

The bank industry management teams are not exactly sure what the purpose of using English in the job might be, but decide the test they use must assess English for listening, speaking and reading related to finance.

The bank industry management team and the Korean Ministry of Finance then decide they need to commission a new English test to assess all existing and new bank staff.

A local college in Seoul wins the contract. The college asks some of their teachers to develop a language test. They develop four versions of the test. The college launches the test and calls it the iBanK English Test. The test looks good and is high-tech (using online delivery and scoring processes).

Table 2: *iBanK English Test overview*

Test part and duration	Skills assessed	Content	Delivery and task-types
1 (6 minutes)	Vocabulary	20 multiple choice questions related to finance and banking from newspapers and text books	Computer delivered: 20 sentences with multiple choice options (x4). Candidates select the best word to complete each sentence.
2 (5 minutes)	Grammar knowledge	20 sentences pairs (each pair contains one sentence with a grammatical error). Sentences are based on general topics.	Computer delivered: Candidates select which sentence is correct.
3 (10 minutes)	Listening comprehension to understand radio broadcasts and news reports	5 short extracts from radio reports (e.g. BBC)	Computer based: Summary which appear on the screen after the listening and require candidates to type in missing words to complete the summaries.
4 (12 minutes)	Reading to comprehend: Key information in a range of technical medical texts	6 articles from popular magazines e.g. The Economist, The Financial Times etc about world economics	Computer based: 12 multiple choice questions. 2 for each text.
5 (15 minutes)	Speaking	Interview	Interviewer asks a range of questions about hobbies, background, experience in banking and problems experienced.

Washback effects of the iBanK English Test

New and existing Korean bank staff are required to achieve 70% on the computer based component of the test and be rated as Level 3 or above on a 5-band rating scale. Existing staff who do not meet these requirements are told they are allowed to retry the test three more times and if they do not achieve 70% and Level 3 on the speaking test will be moved to another position with a lower salary in the banks. Because of these consequences, the iBanK English Test is a high-stakes test.

After the test launches a lot of Koreans attempt the iBanK English Test when, but soon fail the test. As a result a few language schools advertise language training programmes to help

Koreans improve their English and pass the iBanK English Test. Many Koreans enrol in the programmes.

These training programmes vary at different schools but they are mostly only short 1 or 2 week courses. The curricula mostly reflect the test content. As a result, teachers spend a lot of time developing practice multiple choice vocabulary activities (as in Part 1 of the iBanK English Test). Also, in the classes, students are required to write sentences so the teachers can identify their errors. Listening activities are mostly based on BBC news reports where students listen and then read a short summary of the recording, adding missing words (similar to Part 3 of the iBanK English Test).

Teachers also provide one-to-one classes where they ask students exam-style questions about their experiences, hobbies and what they would do if problems occurred at a bank.

After the iBanK English Test has been available for six months, a few existing and new bank staff have passed the test after two or three attempts. However, many other staff have not been able to pass the test.

Enrolments at the schools which advertise “English for Banking classes” or “iBanK English Test preparation courses” increase.

Student and teacher motivation is low. Classes are considered dull and mostly repetitive. There are few speaking activities and teachers mostly just lecture or provide answers to practice questions. Students feel their English has not really improved after attending these courses but they do not realise the significance of this. Their goal is to just pass the test. In fact, they do not pay any attention to their language progress. They fully believe that success on the iBank English Test is what matters, and that if they can pass the test it must demonstrate their worthiness for working in the Korean banking industry. They continue to believe that if they attend the training programmes it will help them pass the test.

After one year the four test versions of the iBanK English Test become well known and the training programmes start to incorporate the exact same BBC news reports, newspaper articles and vocabulary lists into their curricula. Classes become only 2 hours in the evenings. Eventually most existing Korean staff pass the iBanK English Test. The test becomes established as a key prerequisite to employment in Korean banks.

After the test has been used for a few years it becomes well known, however, when the Korean bank staff are required to participate in meetings with overseas staff, communicate with foreign bank customers or write short reports in English they cannot use or understand English effectively. Some of the senior management in the banking industry start to believe the English training programmes the bank staff attend are not useful and are a waste of time and money. They blame the training programmes the staff have attended for not preparing their staff to use English in their jobs effectively. The banking industry and Ministry of Finance continue to believe the iBanK test serves a useful purpose. The iBanK test does not come under scrutiny, as it is perceived has having authority and power. On the contrary, it is the training programmes which are criticised as failing to deliver on equipping students to ‘pass the test’.

This is a clear case of negative washback. The iBanK English test does not reflect real job needs and as a result, neither does the training. It is the training which is ultimately blamed for the deficiencies of a flawed test. Further, student motivation in English is low and their overall improvement in English proficiency after attending the training programmes is minimal, yet they can still pass the iBanK English Test. The teachers also have little creativity and feel the training programmes are dull. There is little incentive to develop the curriculum as the iBangk English Test is the basis for why these programmes were established. The banking industry in Korea develops an increasingly negative attitude towards the training programmes and eventually the iBanK English Test just becomes more of a bureaucratic requirement which all staff are required to achieve, and do. Still, nobody questions or challenges the quality of the iBanK English test despite the fact it is poorly designed and maintained. Nobody recognises the negative washback this test has and so the test remains in use. What is the cause of this mess? It all began with the implementation of an ill-conceived test low in authenticity and with a poorly considered test construct which fails to reflect the real-world target language needs of the banking industry. This occurred because the test was designed by teachers without insufficient expertise in language test design or awareness of the power or influence their high-stakes test would have.

Washback - the responsibility of test providers

While both of these examples are completely fictitious, situations like these really do exist. In Case study 2, the problem started because of a lack of understanding about language testing and what needs to be assessed by the test. Insufficient thought went into what needed to be tested and how. The test did not adequately reflect real-world target language use situations in banking. The test lacked authenticity. Unfortunately, many individuals and organisations in charge of commissioning or developing and delivering high-stakes tests may not realise specific language testing expertise is required. They are often unaware of how much of an effect a poorly designed test can have and the negative consequences it can produce.

As you can see in these examples, it is the test developers who are ultimately responsible for the washback effects their tests create, even if they are not aware of this. Poor quality tests which are not well designed and do not reflect real-world language use situations but are high-stakes in nature lead to negative washback. Similarly, tests implemented with an insufficient number of versions can promote negative washback – impact negatively on the consequential validity of the test (McNamara, 2000). The limited content can become well known and predictable, encouraging teaching and learning to focus on preparing responses to the known content, at the expense of developing broader language competence. Good quality tests which reflect the real-world language needs of test-takers and are properly developed and maintained so there are many versions. As a result, they are more likely to lead to positive washback.

Negative washback has serious consequences. Training may be ineffective and not result in any real or meaningful language development and can also cause negative perceptions towards the training and proficiency levels and even the language itself. The effects can be long lasting and damaging to students, teachers, and work cultures and, in some cases society – especially when language proficiency is needed to help or protect society.

The ICAO LPRs and testing and washback effects

The ICAO LPRs have resulted in the development and implementation of language tests which are high-stakes in nature. Pilots and controllers who do not achieve ICAO Level 4 may lose their ability to work in international operations. The outcome of ICAO LPR tests is not just high-stakes for these pilots and controllers but also for their organisations, especially in cases where staff shortages occur or in situations where replacing staff is complex, difficult and takes time. Most organisations cannot afford to lose staff if they do not achieve ICAO Level 4.

In the aviation field ICAO LPR tests need to reflect the real-world language usage situations of pilots and controllers. This is firstly because the aim of the ICAO LPRs is for tests and the results they generate to allow valid inferences to be made about test-takers' communicative abilities in air-ground communication contexts when phraseology alone is not sufficient to convey meaning (in non-routine situations) (ICAO, 2004). Secondly, ICAO LPR tests need to have high authenticity and reflect operational conditions in their content and task types so that the test-takers value and respect the tests. Finally, this is also important because of the washback effects. If ICAO LPR tests reflect real-world communication needs associated with air-ground communications this is likely to positively influence training curricula and teaching practices – directly developing the very language knowledge and skills pilots and controllers actually need in order to be effective and safe in their jobs.

The field of aviation is unique because unlike the two case study examples above, the ICAO LPRs require pilots and controllers to demonstrate and be assessed in their ability to communicate in unusual situations – situations they may never have experienced. This is because non-routine situations in aviation are rare, but if they occur, English is essential to allow effective communication to occur to manage the situations for safe outcomes. In other words ICAO LPR tests need to assess language in communication contexts which test-takers have little or no experience in, but which are essential in the event they have to deal with a non-routine or emergency situation. Language training is also, therefore contingency based in that it aims to prepare controllers and pilots to be able to communicate in situations they are unlikely to ever face. And, as they do not get exposure to this language in their jobs, because they are so rare, they can only develop proficiency for communication in these situations by attending language training programmes – specific to aviation communications where the curriculum includes content and task types which reflect real-world non-routine operations.

The relationship between aviation job-needs and test construct, design, delivery and content

In the two case study examples above we saw the case where positive washback occurred in Case study 1 because the English language needs of the foreign doctors in real-world situations was reflected in the test. This, along with other factors related to test quality, resulted in positive washback. In Case study 2 we saw that the test developed to assess English proficiency of Korean banking staff had little or no connection to the way they would use English in their jobs. This is a major contributor to negative washback in poorly designed tests.

The same principles apply to the aviation field. Pilot and controller communication needs need to be reflected in the ICAO LPR proficiency tests they take in order for the tests to have construct validity. Table 3 shows some of the ways pilots and controllers might use English in their jobs.

Table 3: *English job needs for pilots and air traffic controllers*

Skills	Pilots	Air Traffic Controllers
Listening	Understand ATC instructions in routine situations	Understand pilot requests and reports in routine situations
		Recognise accuracy in pilot readbacks
	Understand ATC and ground unit (emergency services) questions and instructions in non-routine situations	Understand pilot requests, information report and intentions in non-routine situations
	Understand information relayed by other pilots	Understand information provided by other units or in neighbouring FIRs over the telephone
Speaking	Ask questions and provide information to ATC in routine situations	Ask pilots questions and provide instructions to pilots in routine situations
	Ask questions and provide information to ATC in non-routine situations	Ask pilots questions and provide instructions in non-routine situations
	Provide information to other traffic or crew in non-routine situations	Ask questions and provide information to other units or neighbouring FIRs over the telephone
	Participate in professional development training courses where English is used	Participate in professional development training courses where English is used
	Make passenger announcements	
	Participate in ground briefing sessions (e.g incident debriefs)	
	Communicate with other on-board crew during non-routine situations	
Reading & Writing	Read aircraft manufacturer manuals and checklists	Read and understand ATC documents and manuals
	Type and understand data-link messages, including free text	Type and understand data-link messages, including free text
	Write incident reports	

The shaded areas in the table relate to the English requirements for communication over the radio. It is these listening and speaking requirements which ICAO identified as the skills that need to be assessed by ICAO LPR tests because these relate to communications in safety-related situations. ICAO requires all pilots and controllers to demonstrate competence in radio communication, in both routine and non-routine situations in order to be considered as able to manage flights in non-routine and emergency situations.

Table 4 shows what tests should reflect according to ICAO (ICAO, 2004).

Table 4: *The ICAO Language Proficiency Requirements*

In addition to the ICAO Rating scale (six levels with six criteria), as outlined in the ICAO holistic descriptors, ICAO also requires that Aviation English tests used for licencing purposes need to be able to assess how well pilots or controllers can:

a) Comprehend information and communicate effectively in voice-only telephone and (radiotelephone) and face-to-face situations;

b) Communicate on common, concrete and work-related topics with accuracy and clarity;

c) Use appropriate communication strategies to exchange messages and to recognise and resolve misunderstanding (e.g. to check, confirm or clarify information) in work-related contexts;

d) Handle successfully and with relative ease the linguistic challenges presented by an unexpected turn of events that occurs in the context of flight/work operations in air-ground situations or communicative task with which they are otherwise familiar; and

e) Use and comprehend accents which are intelligible to the wider aeronautical community.

Proficiency tests used for licencing purposes for the ICAO LPRs should:

- Focus only on speaking and listening skills;
- Assess the ability of the six skills included in the rating scale in a range of aviation-related communication contexts: pronunciation, vocabulary, structure, pronunciation, fluency, interactions and comprehension;
- Assess the ability to resolve and repair communication breakdowns;
- Include task types that require use of voice-only communications;
- Assess communicative ability in unusual/unpredictable operational contexts;
- Be able to assess listening comprehension;
- Contain content that reflects on the type of communication in air-ground communications.
- Assess how well test-takers can communicate in real-work situations using English, including the in radiotelephony based communication contexts;
- Expose test-takers to non-routine situations which require them to use language to communicate in non-routine situations (i.e. plain English in radiotelephony communication contexts).

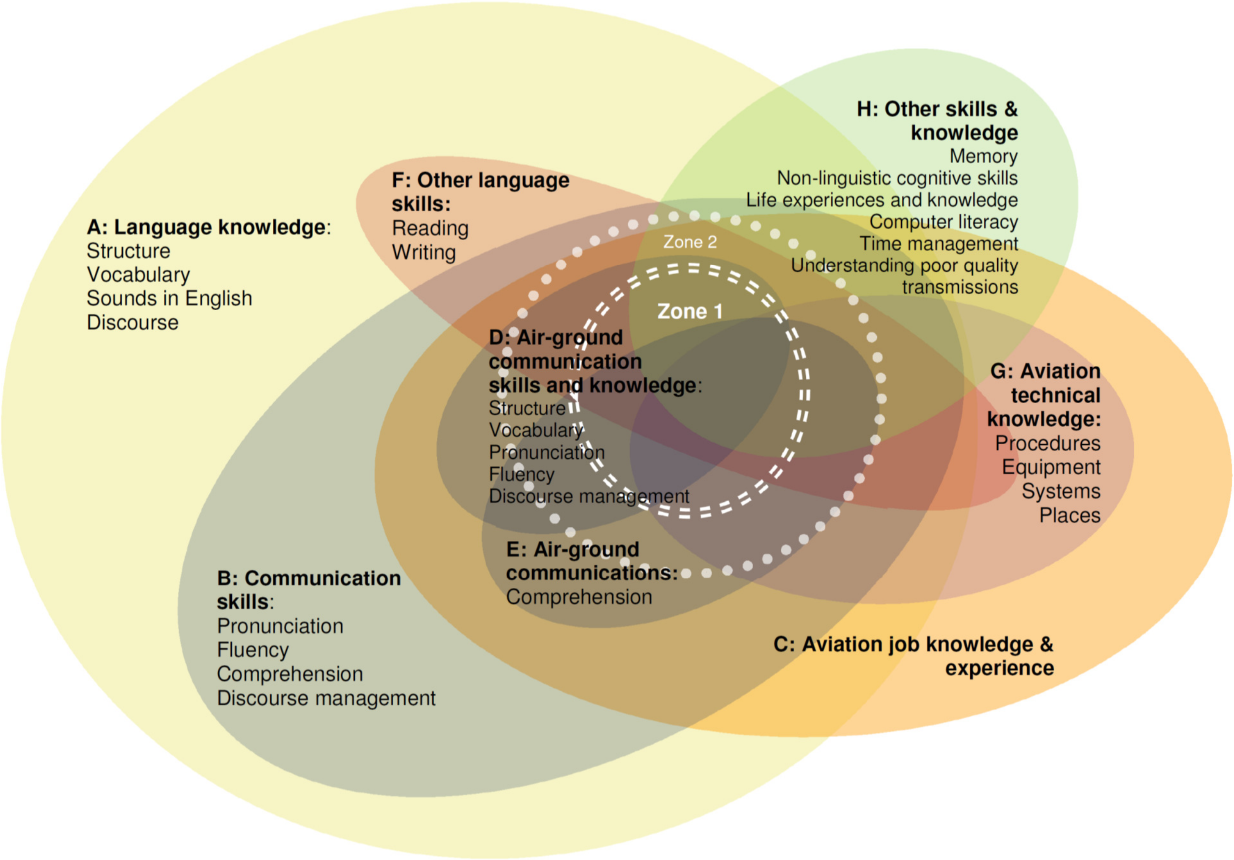
If we take all these requirements and consider how an ideal ICAO LPR test might be designed, we could expect that it should:

- Assess speaking and listening (ideally separately);
- Be based on communications in air-ground contexts where the test-takers communicate over the radio in radiotelephony communication contexts;
- Contain content which requires test-takers to produce a range of vocabulary, structures to effectively communicate in a range of routine and non-routine flight situations;
- Evaluates test-takers ability to communicate using pronunciation which is understandable and at a rate which is acceptable (not too slow or too fast);
- Interact with pilots/air traffic controllers in air-ground communications to collect and provide information in order to resolve situations
- Recognise and overcome communication breakdowns in air-ground communication contexts;
- Place test-takers in non-routine work-related situations to simulate the ways in which English would be used to communicate in air-ground situations.

In addition to these criteria, ICAO also requires test-takers to demonstrate English proficiency in face-to-face situations. This is most likely because if tests only contain voice-only communication test tasks there is a risk that the tests may not provide sufficient opportunities to have test-takers to demonstrate a sufficient range of complex language. It is much easier to design a test that makes test-takers use complex language in face-to-face assessment situations than in voice-only situations alone.

Basically, ICAO LPR tests should contain both speaking and listening components and tasks that mirror real air-ground communication situations and include unusual situations which require test-takers to understand and use complex language. In other words, the test construct needs to include an ability to evaluate proficiency in non-routine air-ground communication contexts, as in Zone 1 of Figure 1.

Figure 1: A schematic way of depicting the language construct associated with the language knowledge and skills required for communication by operational pilots and controllers. Zone 1 represents the ideal target language use elements that should inform the basis of a valid LPR test construct. Zone 2 represents secondary elements that may also form part of the LPR test construct, when accompanied by Zone 1 elements.



In order for ICAO LPR tests to effectively assess the language knowledge and skills associated with operational air-ground communication contexts – Zones 1 and 2 in Figure 1, they need to be designed so that the language content, and contexts in which this language is used, is authentic. Similarly, as Bachman (1990), states, “the closer the correspondence between the characteristics of the test method and the essential features of language use contexts, the more ‘authentic’ the test task will be for the test takers” (that is, the test methods and language use contexts fall within Zone 1 and 2 in Figure 1). Further, the closer the test reflects the specific language use tasks which test-takers are likely to encounter in real-world situations, the more valid the generalisations we make about test-taker performance (Bachman and Palmer, 1996).

We might therefore expect ICAO LPR tests to include tasks such as roleplays, and where the language content requires test-takers to recognise and use the type and form of language used in operational settings (pilots communicating with controllers over the radio), and listening tasks to assess comprehension of complex language in non-routine air-ground and other work-based communication situations. There is also a need to include test tasks which involve face-to-face communication. Again, this could be in the form of a roleplay or participating in a discussion with an interviewer in contexts which reflect how communication occurs in real pilot or controller operational situations (for example, pilots participating in a debriefing with a chief pilot, or controllers explaining an incident to a supervisor).

As Messick (1996) states, positive washback is related to the use of authentic and direct assessments and more basically, to the need to minimize construct under-representation and construct-irrelevant difficulty in the test. Authenticity is important in good quality language tests. A test which is authentic reflects the kinds of situations where test-takers use the language in real life. The content and task types simulate real-work communication situations. Tests that have high authenticity are stronger for two reasons: the results on these tests reflect more accurately how test-takers can communicate in real-life situations and the test-takers feel the test is measuring language they really need or use (that is they respect the test). LPR tests designed to sufficiently reflect real-world pilot/ATC communication needs and have high authenticity are more likely to have positive washback effects. Conversely, LPR tests poorly designed and so do not sufficiently reflect real-world pilot/ATC communication needs and have low authenticity, leading to negative washback effects.

Washback from ICAO LPR tests

The content and design of ICAO LPR tests directly affect the way training programmes are designed and delivered. Higher authenticity promotes positive washback. The more disconnected the test content, task-tasks and delivery are from the real-world language use situations, the more negative the washback is likely to be.

Messick (1996) makes the point that tests which promote positive washback are likely to include tasks which are criterion samples - that is, are based on "authentic and direct samples of the communicative behaviours of listening, speaking, reading and writing of the language being learnt", and he adds that the transition from learning exercises to test exercises "should be seamless". Clearly in the case of ICAO LPR testing it is of upmost importance to have high authenticity since the contexts in which language use is narrow (over the radio), and that if tests do not attempt to assess this communication format and content directly, will lack authenticity and be highly susceptible to producing negative washback effects.

ICAO LPR test washback not only has strong influences on training programmes, teachers' and students' attitudes but also influences airlines, ANSPs and regulators in their perception of what language proficiency is, how it should be assessed and indeed it should be taught/learnt. Washback is ongoing; there is a continuous relationship between LPR tests and training programmes. If a test changes, its influence on the programme and attitudes change.

ICAO LPR tests which follow the ICAO LPR guidelines, reflect the contexts in which pilot or controllers use English in their jobs (as shown in Table 3) and which provide many versions (so that test-takers develop language skills rather than just prepare answers to known test content) are more likely to have positive washback effects, as shown in Figure 2. Conversely, LPR tests which are not well designed and do not reflect the ICAO LPRS or real-world language needs in their content, test-task and delivery have negative washback effects, as shown in Figure 3.

Test content used in LPR tests and the washback effects

Test content (the extent to which the topics and situations in which the language is presented and used are reflective of real-world communication contexts) in aviation English tests should be aviation-related. The closer it resembles the content pilots or controllers use in their jobs, the more likely it is to be perceived as relevant and lead to positive washback.

ICAO LPR test content should include (reflecting the content associated with Zone 1 in Figure 1):

- Communication contexts containing or requiring the use of plain language alongside phraseology when flight situations shift from routine to non-routine;
- Other communication contexts between pilots and controllers and also between controllers (e.g. for coordination between ATC units) or pilots (e.g. for communication on the flight-deck) containing or requiring the use of plain English during unusual situations.
- Other work-related communication contexts associated with operational or other work-related situations – containing and requiring test-takers to understand and/or use language in authentic contexts related to their jobs (e.g incident debriefings and training scenarios).

Test content may also include (reflecting the content associated with Zone 2 in Figure 1):

- News reports about aviation concepts (e.g. accidents, incidents or investigations);
- Reports and conversations on general interest-related aviation topics; and
- Routine and non-routine incident pictures and recordings (real or simulated).

Test content has a significant effect on the training programmes. Tests which contain a too-narrow-a-focus on one type of content cause training programmes to focus on this type of content at the expense of other content types. Similarly, test tasks that are insufficiently work-related or authentic can result in training programmes focusing on skills which controllers or pilots do not need for communication in real-life situations.

The content of a test influences the kind of language – vocabulary, structure, functions (e.g. making requests, giving instructions, complaining, showing appreciation, etc) and communication contexts in the test. If the language or communication contexts do not reflect the ICAO LPRs (that is, do not contain language related to commination between pilots and controllers to allow effective communication in non-routine situations), this creates negative washback. And, as a result, the training is also less likely to reflect these requirements.

Tests that have a narrow or undefined scope of content (e.g. focus heavily on tasks to assess grammatical or vocabulary knowledge and accuracy in discrete paper-based items) tend to have negative washback effects on training because those tests do not encourage training programmes to develop use of grammar or vocabulary in authentic types of aviation communication (instead only focus on memory of language knowledge). As a result, a pilot or controller might achieve ICAO Level 4 one of these tests and have a good knowledge of grammar but lack the fluency skills to effectively communicate over the radio in non-routine or less predictable situations. Similarly, tests which have a limited scope of content will have negative washback effects on training programmes by limiting the relevance or range of content in the programme. Indeed these types of tests can promote rote learning or memorisation training techniques among students and teachers because the limited content can be memorised at the expense of developing proficiency for communication in a wider range of content and communication contexts.

Figure 2: *A schematic flowchart showing a positive washback effect of an LPR test and the ongoing continuous relationship between pilot/ATC real-world English language use, the ICAO LPRs, ICAO LPR tests, aviation training programmes, test taker attitudes and organisational perceptions of language proficiency. The positive washback effect occurs when well designed ICAO LPR tests are implemented.*

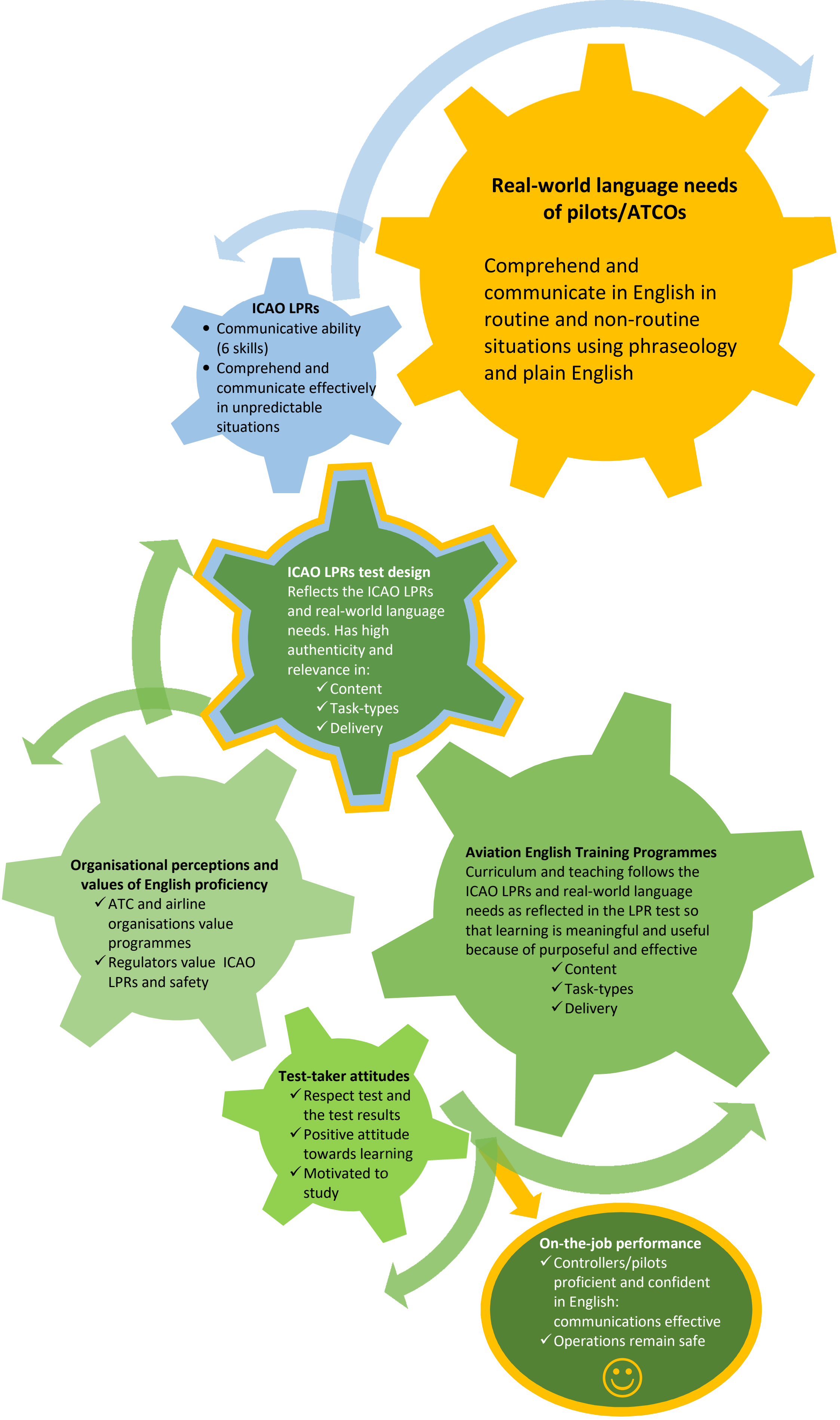
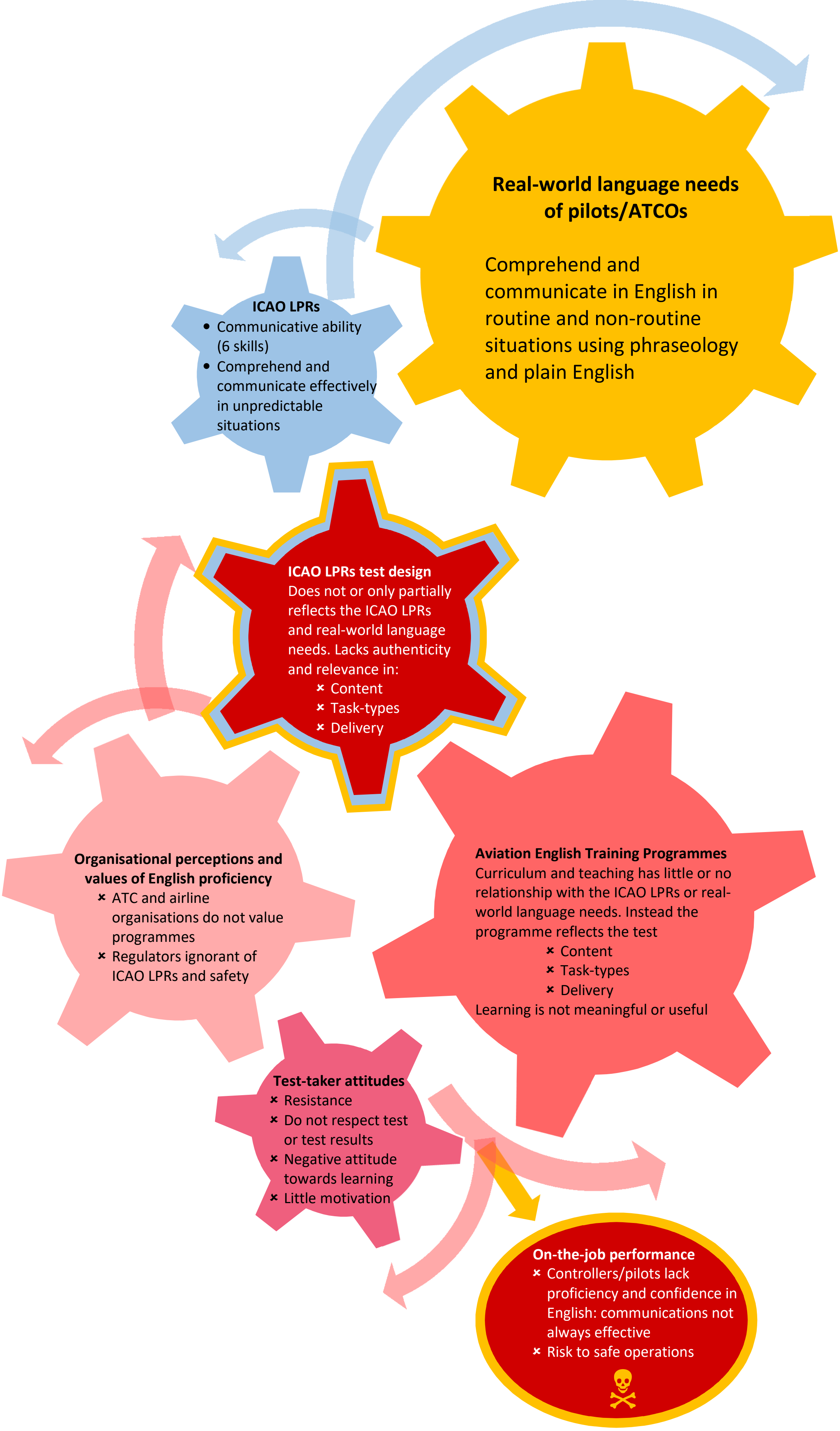

Real-world language needs of pilots/ATCOs
Comprehend and communicate in English in routine and non-routine situations using phraseology and plain English
ICAO LPRs
• Communicative ability (6 skills)
• Comprehend and communicate effectively in unpredictable situations
ICAO LPRs test design
Reflects the ICAO LPRs and real-world language needs. Has high authenticity and relevance in:
✓Content
✓Task-types
✓Delivery
Organisational perceptions and values of English proficiency
✓ATC and airline organisations value programmes
✓Regulators value ICAO LPRs and safety
Aviation English Training Programmes
Curriculum and teaching follows the ICAO LPRs and real-world language needs as reflected in the LPR test so that learning is meaningful and useful because of purposeful and effective
✓Content
✓Task-types
✓Delivery
Test-taker attitudes
✓Respect test and the test results
✓Positive attitude towards learning
✓Motivated to study
On-the-job performance
✓Controllers/pilots proficient and confident in English: communications effective
✓Operations remain safe
☺

Figure 3: *A schematic flowchart showing a negative washback effect of an LPR test and the ongoing continuous relationship between pilot/ATC real-world English language use, the ICAO LPRs, ICAO LPR tests, aviation training programmes, test taker attitudes and organisational perceptions of language proficiency. The negative washback effect occurs when poorly designed ICAO LPR tests are implemented.*


Real-world language needs of pilots/ATCOs
Comprehend and communicate in English in routine and non-routine situations using phraseology and plain English
ICAO LPRs
• Communicative ability (6 skills)
• Comprehend and communicate effectively in unpredictable situations
ICAO LPRs test design
Does not or only partially reflects the ICAO LPRs and real-world language needs. Lacks authenticity and relevance in:
× Content
× Task-types
× Delivery
Aviation English Training Programmes
Curriculum and teaching has little or no relationship with the ICAO LPRs or real-world language needs. Instead the programme reflects the test
× Content
× Task-types
× Delivery
Learning is not meaningful or useful
Organisational perceptions and values of English proficiency
× ATC and airline organisations do not value programmes
× Regulators ignorant of ICAO LPRs and safety
Test-taker attitudes
× Resistance
× Do not respect test or test results
× Negative attitude towards learning
× Little motivation
On-the-job performance
× Controllers/pilots lack proficiency and confidence in English: communications not always effective
× Risk to safe operations
☠

Test task types used in LPR tests and the washback effects

There are various types and tasks used in language testing to assess speaking:

- Indirect computer-assisted testing, where both/either the delivery and/or rating are provided by a computer using conventional and voice recognition technology. While possibly appropriate for benchmark testing, this type of testing system does not offer the required interaction for proficiency in aviation ICAO LPR testing.
- Semi-direct computer-assisted testing involves using pre-recorded prompts to elicit responses from test-takers which are rated later by human assessors. The downside of this format is that the tasks are not interactive and therefore do not involve real communication.
- Interviews which involve conventional question and answer interview techniques, ideally building on topics which allow the test-takers to feel they are participating in a conversation with an interlocutor (interviewer).
- Stimulus and response tasks where test takers are asked to listen to a short recording then summarise it later, describe pictures or create a story based on a series of prompts (e.g. words or pictures).
- Roleplays where the test-taker takes on the role of someone facing a situation in a simulated context and participates in an exchange with an interlocutor.

There are risks that some types of test tasks are not suitable for evaluation of interactions and natural fluency. For example, if a task requires test-takers to produce lengthy speech samples in response to computer-generated or isolated interviewer questions or prompts, it is unlikely these can reflect natural interactions in authentic communications.

Such test types can promote a training with a narrow curriculum which focus more on exam preparation than on natural and authentic speech production for real life communication. This is an indication that this kind of test task can lead to negative washback.

The role of test delivery in test instrument design

There are different variables of test delivery that include:

- Computerised delivery;
- Telephone/video conference-based or face to face;
- Textual, pictorial or oral delivery of test content and items;
- Interlocutors assuming the role in simulated communications (e.g. in roleplays) requiring the test-taker to participate and interact in the communication; and
- Test interviewers asking prescribed questions or delivering test rubric which requires test takers to perform or use specific language/complete test tasks.

All of these may have either positive or negative washback effects depending on how they are managed and presented and how these modes of delivery interact with the test tasks and test content. While these are all valid types of test delivery in themselves or in combination, they all have the potential to affect both the reliability of the outcome and the perception of language proficiency among students, teachers and aviation organisations. For example, if a test relies on a lot of computer-based delivery to prompt test-takers to speak, it is possible this will influence training programmes by removing the focus on real communication, instead focusing on just accuracy of grammar, vocabulary and pronunciation.

Conclusions

The design and delivery of high-stakes tests has a strong but often unrecognised effect on the training programmes test-takers attend before taking these tests. This effect influences not just the programme but also attitudes to teaching and learning as well as perceptions of what language proficiency is among organisations, including licensing authorities. ICAO LPR tests are high-stakes tests. Developers of ICAO LPR tests have a responsibility to ensure the tests they design and implement have positive washback effects. If training is developed in response to robust quality language tests which reflect real-world target language situations then, as Weigle and Jensen, 1997 put it, 'there is no difference between teaching the curriculum and teaching to the test' – providing a clear indication that a positive washback effect is in play. Positive washback is more likely to occur when the test construct of an ICAO LPR test captures the real-world communicative target language use situations (namely air-ground communications in non-routine situations – the objective the ICAO LPRs (ICAO, 2004) and when task types, content and delivery is high in authenticity.

Positive washback means the test results in language tests lead to effective language training programmes which in turn leads to real and meaningful language improvement among pilots and controllers so that they are effectively able to communicate in real-world situations.

ICAO did not develop the ICAO LPRs to simply require pilots and controllers around the world to be tested every three or six years. They developed the LPRs to encourage positive washback so that individuals (pilots, controllers and teachers) and organisations– licencing authorities (regulators), training schools, airlines and ANSPs would implement training programmes which develop and maintain language proficiency. ICAO's aim with the LPRs is to have tests encourage pilots and controllers to reach and maintain a standard of English so they can function in their jobs using English effectively - keeping aviation safe. This was ICAO's intention: implement quality testing systems to promote good training practices. Well-designed ICAO LPR tests which reflect job needs and have high authenticity lead to positive washback.

References

Alderson, J. C., and Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14, 115-129.

Bachman, L. F. (1990). *Fundamental considerations in language testing*, Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Green, A. (2014) *Exploring Language Assessment and Testing: Language in Action*. New York, NY: Routledge.

International Civil Aviation Association (2004). Manual on the Implementation of ICAO Language Proficiency Requirements, Montreal, Canada: Author. Doc. 9835

McNamara, T. (2000). Language Testing. Oxford: Oxford University Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13 (3), 241–56.

Shohamy, E. (1993). A collaborative/diagnostic feedback model for testing foreign languages. In D. Douglas and C. Chapelle (Eds.), *A new decade of language testing research* (pp. 185-202). Alexandria, VA: TESOL Publications.

Weigle, S. C., and Jensen, L. (1997). Issues in assessment for Content-Based Instruction. In M. A. Snow and D. M. Brinton (Eds) *The content-based classroom: Perspectives on integrating language and content* (pp. 201-212). White Plains, NY: Longman.