A re-evaluation of washback for learning and testing language in aeronautical communication

Neil Bullock, ICAEA Research Group, August 2017

Available to download from
http://commons.erau.edu/icaea-workshop/2017/monday/19

Abstract

Washback has been a concept of learning and assessment for almost 30 years. The notion dictates that a test will have an effect on the learning process linked to it. This effect can be both positive and negative depending on the affecting factors. What is less clear is on what are the principles that underlie this concept based.

In addition to this, it seems somewhat unusual in an environment where learning would be the principle activity, that a test should dictate what is learned. Theory would seem to indicate that assessment and testing are systematic ways of determining the extent to which a learner has learned a given subject.

Using the domain of learning and testing English for aeronautical communication, this paper will show that, if learning and assessment are aligned correctly as part of an ongoing learning process, washback is simply an integral part of this process and not a mechanism working in isolation. It will also demonstrate that an integrated process of learning and testing with a common core objective can go a long way towards reducing the challenges of assessing language proficiency in this specific purposes domain.

1. Background

It is widely agreed that washback as a concept requires a test system to have an influence on the learning that precedes the test (Shohamy et al, 1996; Messick, 1996; Alderson & Hamp Lyons, 1996; Fulcher, 2010; Green, 2014). What is not so clear is how washback can be defined in terms of a systematic procedure and, indeed, what empirical evidence there is to support such an idea (Shohamy et al, 1996; Alderson & Hamp Lyons, 1996).

One reason for this 'influence' theory was the potentially negative effect of the large-scale general language testing on any learning process that preceded it (Buck, 1988; Green, 2014). However, while clearly understandable, this pre-supposes that it is testing and not learning which is the driving force. This is highlighted in specific purpose language testing in aeronautical communication, where a system of testing plain language was introduced in 2004 with an associated 149-page manual, of which only four and a half pages were dedicated to language training in this domain. It took a further 5 years before a separate document, the 80-page *Guidelines for Aviation English Training Programmes* (ICAO, 2009), was published. Whilst laudable in its safety-led objectives, the domination of this testing system, almost in isolation, has created many tensions and challenges for all stakeholders: Poor test quality, ignorance over testing practice, a system that appears to pre-suppose L1 speakers are the most proficient communicators and increased stress amongst test takers over potential job-loss are just some of the many challenges (Bullock and Westbrook, 2017). It is not therefore fanciful to suggest that testing as the dominating factor may well be the root cause of challenges such as these. Furthermore, a great deal of the research related to language in aeronautical communication to date has focused on either learning or

testing as individual and separate entities (Douglas, 2004; Kim, 2009; Alderson, 2009, 2010; Sarmento, 2011: Paramasivam, 2013; Kukovec, 2008 and Yan, 2009) or from a purely Applied Linguistics angle (Breul, 2013; Ragan, 1997). Few, like Uplinger (1997) and Farris et al (2008) have tackled how both learning and assessment can work together. Would not a congruence between learning and testing, if correctly calibrated around a clear learning objective and associated test construct, with no one element the dominating influence, better define and drive the learning and testing process and at the same time reduce the effect of the challenges mentioned above?

Specific purpose domains should, because of their often regulated lexis, syntax and referential meanings, make the defining of target language use (TLU) easier, and, along with this, learning and testing objectives. The case study presented here will use the domain of language used in aeronautical communication between pilots and air traffic controllers – often erroneously referred to as *aviation* English (Bullock, 2015) – to question some of the ideas behind washback and argue that if such a congruence between learning and testing is appropriately designed then the concept of washback becomes a *de facto* element of the whole learning and testing process, and no longer remains a stand-alone item requiring a defining theory to support its existence. Furthermore, the research will suggest how such a change would help reduce tensions among stakeholders and increase the validity of the testing system as a whole.

2. Washback, the origins

Searching for a clear rationale for washback is made difficult by the fact that, while many authors agree on the idea of washback as a guiding principle (Shohamy et

al, 1996; Messick, 1996; Alderson & Hamp Lyons, 1996; Fulcher, 2010; Green, 2014), empirical evidence is not so easy to source, with some even calling for just such empirical evidence to strengthen the case (Alderson & Hamp Lyons, 1996).

An overriding common ground states that washback, as a concept, requires a test system to have an influence on the learning that precedes the test. Hughes (1993:2) states that items found in a test 'will affect learning outcomes', Douglas (2000) believes that tests should mirror materials and methodology used in learning and reduce where possible any disparity between the assessment process and what is taught, while Shohamy et al (1993:298) simply refer to the 'impact tests have on teaching and learning'. Thus, the assumption is that if such effects are good then we can refer to positive washback and conversely, if bad, such washback would be negative. The issue here, however, is that the test is the driving force, with learning, where acknowleged, reduced to secondary interest.

Other authors, however, are not as clear as to what washback constitutes. Messick (1993:241) cryptically talks about how a test 'influences teachers and learners to do things they would not otherwise do', while at the same time suggesting that washback is only 'linked to the introduction and use of the test'. Green (2014:86) similarly narrows this down to what 'teachers and learners do in the classroom when preparing for (a specific) assessment'. Thus, it is not so difficult to see that when some even doubt the existence of any real empirical evidence, the reality of washback as something tangible becomes increasingly unclear. Alderson and Hamp Lyons (1996:281) claim that not only is such empirical evidence not available, but that

hypotheses about washback are too ‘naive’ to be of great use and that the effect of testing on learning is much more complex than examined beliefs allow’.

What is notable is that the introduction and domination of large-scale testing methods may have led to fears about what has become referred to as ‘teaching to the test’. Alderson and Lyons (1996:280) suggest, that exams such as TOEFL exert ‘an undesirable influence on language teaching’ through ‘inappropriate learning strategies’ and ‘unnatural teaching’, where clearly the aim of ‘passing’ the test prevails over any tangible learning outcomes. Buck (1998) re-enforces this belief by suggesting that testing affects and drives the learning of foreign language while, Shohamy (1993) believes that the need for washback comes from the authoritarian effect of external testing and how it impacts on the lives of those taking the test. Thus we may suppose that the idea of washback is a reactive concept, rather than a tangible theoretical framework.

It is not difficult to understand how such fears arise about the domination of large scale testing over tangible learning. Nevertheless, if the result of any such domination of large scale external testing over learning is simply to teach to the test (Hughes 1993), or seen as almost a dishonest activity by Hamp-Lyons (1998), then to pre-suppose that simple adherence to an idea of positive washback as a get-out principle would seem strange, even naïve. Furthermore, when Messick (1993:241) posits that a test could induce ‘curricular and instructional changes that foster development of cognitive skills that the test is designed to measure’, it simply reinforces the principle that testing is the driver of learning. Using washback as a gatekeeper to maintain learning objectives and practices does nothing to prevent

testing and assessment dominating and means that teaching to the test is seen as being the only realistic, albeit regrettable, outcome.

3. Perception based theory

The potential lack of any solid theoretical basis for washback may be influenced by how the effects of learning and testing are determined from outside a strict pedagogical framework. We may see implicit factors such as teachers preparing more comprehensive lessons or getting learners to engage in the homework process, which may produce an effect albeit indirectly linked to any testing. When Alderson & Wall (1993:117) suggest further that bad tests may increase work and that this work would be 'better than nothing at all' whilst good tests could have a negative effect increasing learners' anxiety, it is not difficult to see the concept of washback as anything other than a very nebulous perception.

Stakeholders' perceptions on testing may also create an additional influence that is less about empirical evidence and more about individual experience. Shohamy (1992:514) theorises that testing is often seen as an 'authoritative tool, dictated from above' and if teachers are not asked to be and do not choose to be involved in the testing process, he concludes that it is difficult to believe that anything positive can occur. Of course authorities and legislators would argue that they have an obligation to ensure the test is taken, which is certainly the case in many professional domains. However, it is not hard to imagine that the perception of the test taker may not be quite so accommodating when the test is seen as the difference between keeping a job or not. Even a valid and justified test, with tasks set to represent required skills, could

meet with resistance if test takers are not aware why it is being introduced or if teachers resist a new approach in favour of old ideas (Green, 2014).

There are strong arguments for teachers and other educators being involved in the assessment development process, not least to maintain a valid and appropriate link between what is taught and what is tested. Hughes (2003:2) calls this 'proper relationship' one of a 'partnership'. So, a valid test may suffer from learners not receiving the instruction they need in order to understand the tasks and the TLU involved in the testing process, if the teachers are not experienced in either understanding the test construct or a specific purpose technical domain. Hughes' (2003:2) assertion that good testing may well have a 'corrective influence' on bad teaching thus seems rather difficult to comprehend and a more thorough assessment of the teaching/testing process may be needed before making such assertions. Any attempt to harmonise learning to a testing system with even the most tenuous case for its validity and reliability, will hold little credence for the teacher who simply wants to get learners through the exam, or for the learner who simply wants to keep a job.

Finally, Messick (1993) points out that evidence about effects on learning can only be *washback* once the test has been introduced and in use, a sort of *a posteriori* evidence. Thus, such a reliance on the test only seems to confirm that any relationship between learning and testing is driven by the system of testing and not learning. Surely a valid and reliable test is born out of an appropriate assessment process of what was learned with similar objectives in both domains. The two need to be developed in congruence and not in isolation or through dominance.

The lack of any real co-ordinated approach towards a common learning and testing objective can only make it harder to find a coherent and justifiable theory regarding washback. As Shohamy et al (1996) recognise, any impact of testing on learning is therefore certain to be complex and the value of any impact will only be a subjective perception, varying from case to case, and from stakeholder to stakeholder. Little wonder therefore that few have managed to fully explain a coherent and tangible theory for washback.

4. Towards a re-alignment of learning as the driver

If testing initiates a learning process, then there is every possibility that learning will be devalued to that which is needed to simply pass a test. There are few tangible benefits that the student will learn in connection with real world authentic communication.

With possible specific exceptions such as diagnostic evaluations and pre-screening, it is hard to disagree with Hughes (1993) when he describes teaching, and, by default, learning, as the 'primary activity'. Of course, anything, including testing, that has a positive influence on learning should be encouraged and indeed Green, (2014) suggests that even if a system is driven by testing then as long as the test tasks reflect authentic language, there will be some positive outcome, if learning about the same tasks follows. This, however, accepts that testing is still driving learning, and thus any positive outcome would be fortuitous, rather than explicitly designed for the learning of authentic communicative language skills. Indeed, as Bailey (1993) posits in his survey of Japanese students, non-authentic tasks lead to cramming where students master only the tasks to pass a test and few tangible skills are learned.

On the other hand if we know that such influences from testing are not dominant, but simply the result of a well-calibrated learning continuum, similar to Morrow's feedback loop (1991:112), where there are diverse set of variables interacting in a learning activity, with clear learning objectives as the core driver, we can start to see that whole process as one of congruence between the content of learning and the performance the learner gives in the final test (Fulcher 2010). If such a congruence is indeed well calibrated towards learning, then fears that domination of testing would be unfounded. The basis of the whole process

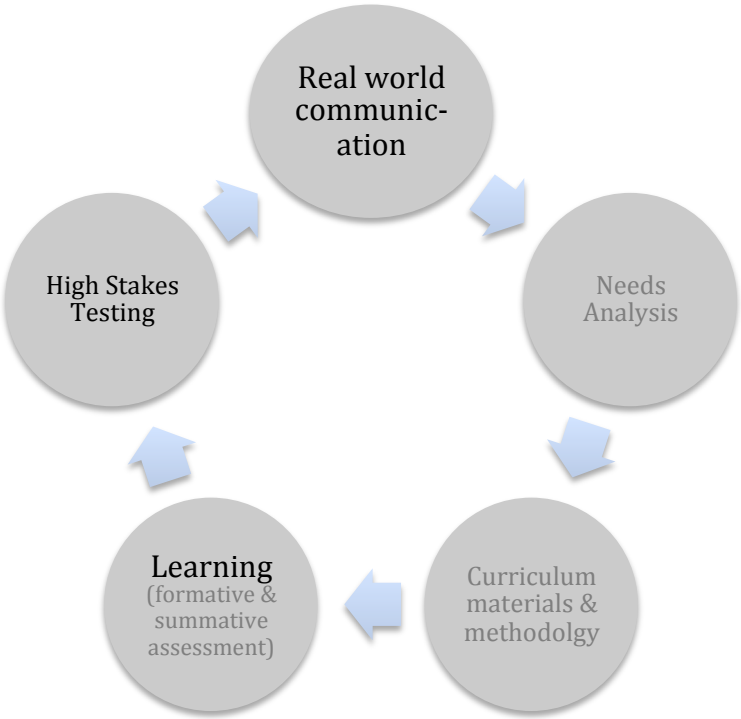

Fig.1 The Learning and Assessment continuum

would come from a clear understanding of the target language use, diagnostic evaluations of learners' abilities, focused and meaningful curriculum development and adaptive teaching methodology (see Fig. 1).

5. Defining learning and assessment objectives from the TLU

A clear process driven by learning of the TLU is primordial when narrowing the field of learning to that of specific purpose language (LSP), such as required in aeronautical communication between a pilot and an Air Traffic Controller (Farris et al, 2008; 2012; Bullock, 2015). Regular evaluation of the achievement of learning goals would thus form the testing element of this process, but would remain driven by learning. As LSP learning should have a clearly defined TLU, a congruence between learning objectives and testing outcomes should be easier to achieve. It theoretically alleviates the need for any form of washback, as washback thus becomes a *de facto* element in the process.

It may, of course, seem easy to suggest that the TLU should form the foundation of the task-based learning objectives and, ultimately, the test construct, nonetheless, such language and use in a multi-faceted communicative process must be correctly identified and specified for the continuum to achieve *a priori* construct and face validity. Learning and testing should reproduce real-life situations in order to 'examine the student's ability to cope with it' (Doye, 1991). As a way of reaching such goals, Messick (1993) and Green, (2014) both suggest test developers should strive to minimise two key elements: *construct under representation* (elements missing an identified construct) and *construct irrelevant variance* (elements included but not required in the construct). As noted by Moder & Halleck (2009), Alderson (2009) and Read & Knoch (2009), there are sadly tests of English proficiency in an aviation context where such elements are all too often evident. These include no real-world communicative tasks and general purpose Oral Proficiency Interviews (OPIs) unrelated to any aspect of the TLU.

In aviation some of the major features of discourse include a highly restricted lexis (standard phraseology), specialized referential plain language lexis, restricted syntax and specific interactional characteristics (Breul, 2013; Rubenbauer, 2009 and Yan, 2009). Studies of transcripts of unexpected situations indicate that pilots are more likely to use plain language to supplement phraseology in problematic or emergency situations (Linde 1988; 4-Bühlmann, 2005). One can also see commonly occurring plain language functions, including 'greetings, sign-offs, politeness markers, and questions' (Moder, 2013). Furthermore, as Bullock (2015) suggests, 'The operational specificities of pilot/ATCO communication mean that it is not sufficient either to be simply offered lists of aviation specific vocabulary' by a teacher, but to be given 'the ability to produce, receive and process language in a "highly technical and safety specific context"'.

We can thus pre-suppose that if test and curriculum developers already exist in their respective technical field, then learning and testing development should be a relatively simple task. However, the testing of language in aeronautical communications has not been without its challenges. In April 2017 a pre-conference survey conducted by the International Civil Aviation English Association (ICAEA) among delegates showed a clear disparity between the perceptions of different groups of people. Delegates represented a cross-section of the industry including pilots, ATCOs, language trainers, test developers, legislators and Air Navigation Service Providers (ANSPs).

The questionnaire was organised to source opinions of delegates on the theme which was to look at the 10 years since the testing system was set up by ICAO.

Questions were based on recurrent themes from research articles, earlier conferences, ICAEA's own Linked-in forum, and problematic parts of the system that are widely acknowledged by stakeholders. Responses were taken from a 5-part Likert scale, which ranged from 'completely agree' with the statement to 'completely disagree'. It included 22 questions divided into four key themes. The survey was completed by 81 out of 116 registered delegates (n=81, 0.70 participation). Such high participation from delegates was seen to be very encouraging.

Not all of the questions in the survey are directly related to the theme of this paper so the data here concerns only those areas concerning the impact of the Language Proficiency Requirements on both language testing and training and those where inherent differences of opinions between participants were relevant to the subject of this paper.

If we first look at the statement: "ICAO LPR language tests in your region adequately assess the communication needs of pilots and controllers in air-ground communication contexts", overall 42% of respondents gave a positive response (fully agree & agree) and only 24% a negative. But when we break participants down into groups, responses were somewhat different. Here we saw that amongst trainers and test developers the response remained in line with the group as a whole (43% v 26%) however the responses of non-L1 English speaking pilots and ATCOs, i.e. those likely to be affected the most by the LPRs, only 25% agreed with the statement whereas 50% disagreed. This indicates that test developers and trainers largely believe they are doing a good job, whereas those actually being tested do not. This may well have

quite serious implications for face and consequential validity of any tests and certainly supports earlier criticisms of the system (Alderson, 2009; Kim & Elder, 2009).

If we look at the responses relevant to teaching/training and the statement: “The introduction of the ICAO LPRs has led to a meaningful increase in the amount of language training”, the group response was clearly positive with 55% who (strongly) agreed v 19% who disagreed (strongly). However, again if we look at the breakdown between trainers & curriculum developers and Non-L1 speaking pilots and ATCOs, the response again is somewhat disjointed. 56% of the first group agreed or strongly agreed with the statement, and only 17% disagreed, whereas of the second group 50% disagreed or disagreed strongly. This indicates a difference of opinion between those responsible for the training and teaching and those who are or who should be receiving the training.

Looking further at an additional statement: “Attention is primarily given to test preparation, focusing on practising possible responses, rather than meaningful language training that promotes learning and maintains & improves proficiency and communication skills”, the group of trainers and curriculum developers agreed with 50%, however 75% of the Non-L1 speaking pilots and ATCOs, agreed with this statement. This not only shows yet another disparity between learning and testing service providers and the test taker / learner population, but supports Alderson’s (2009) fears, that testing may often not meet international standards for high-stakes language testing.

6. Towards a congruence between learning and testing

The effects of such doubt and scepticism amongst test takers as to the validity of testing and learning systems can be put in the context of a test which was developed in Switzerland to test pilots' language proficiency in English. A team of experienced English Language Experts (ELEs) and Subject Matter Experts (SMEs) worked together to ensure that contextually valid tasks from the real-life TLU were included which clearly identified the construct in the elements being tested. Subsequent post-testing feedback from Test Takers (TTs) (n=233 ie: 56% of 557 TTs replied) gave a positivity response co-efficient of 0.85. This showed that TT responses to the test developers' statements about the tests' various elements of validity were 85% in agreement or fully in agreement, thus largely demonstrating face and consequential validity and going a long way to supporting construct, context and content validity as well.

Such data certainly suggests that the inclusion of real-life authentic TLU in context leads to greater test validity, and vice versa when in inverse proportions. This matches observations that there must be a congruency between test tasks and real life (Doye, 1991), while Bailey (1993), suggests that a congruency is necessary between authentic language situations and test tasks. Realistic settings and close simulations, parallel to the real world, enable the learner and the test taker to 'perform the task as freely as he would do in real life' (Messick 1993: 243). So, in a system designed to improve safety in aeronautical communications, it is remarkable, 8 years after Anderson first voiced his concerns, to still see evidence that shows that the skills required for the safe communication between pilots/ATCOs are not being appropriately tested.

If we go even further and look at how the functioning of a systemic process can be ensured, then learning tasks must also include this congruence with real life authentic tasks in order to achieve a clear match between the construct to be learned and that which is being tested. Furthermore, as language does not exist in isolation, any learning or testing process must include the communicative concepts, contexts and processes of the human interactions for which it is to be ultimately used for it to be considered valid.

As we saw in Fig. 1, any system aimed at achieving and maintaining language proficiency must start from an analysis of real-world communication. The inclusion of linguists (ELEs) and technical specialists (SME) in the learning and assessment process is primordial (ICAO, 2010 & Knoch, 2009). Both parties working together help in conceptualising the TLU and it can even be advocated that both SMEs and ELEs can learn much more about the construct and language used by working together (Bullock, 2015). Such work should include focused discourse analysis that allows both groups of experts to identify and fully understand the contextual use of language in the specific purpose domain.

Another part of calibrating the learning process with formal testing can be, as Goh (2013) suggests, listening to and carrying out discourse analysis on authentic texts. This can enhance the contextual learning of real-life language in the LSP communication and highlights the importance of not simply focussing on lexical and grammatical forms, but on the language and its use as a communicative tool. A focus on communication through language used as a lingua franca in cross-cultural communication, as well as in more micro-, socio- and inter-cultural settings, can also

be suggested for discourse analysis. Communication in such contexts should also include those expected irregularities in authentic contexts such as interruptions, technical deficiencies, and background noises, so that learning (and assessment) targets all the communication processes for their functional importance. The inclusion of non-linguistic features in assessment also helps mirror the entire communicative process and increases the need for learners and test takers to replicate the cognitive processes involved. This, as Weir (2005) and Field (2013) suggest, helps to correlate cognitive activity from the real world with learning and test tasks, thus increasing the cognitive validity in assessment.

One final point worth noting in ESP discourse is cited by Farris et al. (2008) who describe, in an aeronautical context, 'how Controllers and pilots work under various workload conditions and may be required to perform several tasks concurrently' which require memory and processing demands in terms of cognitive workload. It can be attested that such complex cognitive workload must also be attributed to Air Traffic Controllers. Such cognitive communication load becomes higher for all those involved in the communicative process in critical stages of flight operations, because of the need to coordinate procedures and information quickly and accurately. This is even more intense and complex in unexpected and non-routine situations and of course is resultant on many human-factor based events.

7. From theory into practice

In order to demonstrate the large divergence that exists between teaching and testing of language used in aeronautical communications, the author conducted a workshop experiment at the above mentioned ICAEA conference in April 2017.

Participants were offered the chance to look at six test tasks chosen from tests of English language proficiency for aviation and were asked to suggest why they were good or bad tasks. They were then asked to identify what effect they could have on learning and the use of target language of potential learners and test takers. Tasks were taken from a wide variety of publicly available tests around the world. Participants were a mixture of teachers, testers, administrators, pilots and ATCOs. The majority of each group used English as a lingua franca, with only a minority having English as a first language. The results are shown in Tables 1 & 2. The tasks were deliberately chosen for two reasons:

i) to elicit why certain available tests were of poor quality and failed to offer a valid testing platform for the intended TT population.

ii) to demonstrate what good and valid tests should be including and how test tasks can be constructed to foster learning and assessment of the TLU.

The rationale for these two reasons was to elicit important issues associated with testing and learning in this domain:

- How learning objectives can be associated with test tasks.
- How key elements constituting test validity can be identified, such as construct, content and context validity as well as cognitive, face and consequential validity.
- How test and curriculum developers can focus on real-world tasks.

From the responses of the participants, the following elements in terms of positive and negative washback were suggested as to why the tasks were not valid and what effect they would have on learning and subsequent contextual use of the target language:

Speaking Task	Positive attributes	Negative attributes
1) a face-to-face oral proficiency interview (OPI) where the pilot Ttst taker (TT) talks about his life uninterrupted for 10 minutes. No visual prompts.	• allows a variety of grammatical structures to be demonstrated.	• no interaction • restricted and inappropriate range of language (content) • allows personality rather than communicative skills to dominate. • able to be rehearsed easily • output language learned at very early stage. • non contextual
2) a voice-only classroom interaction between a pilot test taker and an interlocutor role-playing an ATCO. TT has a list of tasks he must complete in the air and on the ground including non-routine and routine situations using standard phraseology plain language as appropriate. The ATCO has a script but may deviate where necessary.	• contextually valid • content appropriate to TLU • allows most elements of construct to be demonstrated • task based items from real-world events. • allows SME and ELE input	• classroom-based so could lack context if raters are untrained or inexperienced.
3) OPI where TT has to recount the events of a video showing a news report of an aircraft accident. The video is publicly available on YouTube	• content language from real-world tasks.	• material readily available so not a true reflection of skills. • limited range of vocabulary to that one situation. • limited interaction

and was well documented in the media.		• does not test the construct. • does not replicate cognitive processes of construct.

Listening Task	**Positive attributes**	**Negative attributes**
1) TT (ATCO) listens to one short recording of a simulated Pilot / ATCO exchange taken from a commercially available aviation English book and answers questions on it.	• contextually valid	• limited context • limited content • known content • does not test construct • no interaction with TT • promotes teaching to the test.
2) TTs (ATCOs) listen to a series of pilot / ATC exchanges in non-routine situations through headphones and must answer questions on what they heard. Answers are a mixture of multiple-choice and free response. TT can choose whether to enter answers in a computer or write on paper.	• contextually valid • cognitively valid • content valid • construct valid • real-world relevant • promotes real-world language learning • practical • promotes face and consequential validity • promotes and allows learning in situational awareness	• no interaction
3) TT (pilot) listens to pre-recorded prompts on a PC and responds accordingly. The responses are recorded and sent to another assessor for later assessment.	• partially contextually valid • independent rating	• no real life interaction • no read back hear back • limited cognitive validity • reduced face validity • prompts unnatural language. • unable to ask back or clarify.

Feedback from the participants shows that there is awareness of what constitutes construct in aeronautical communication and of what constitutes relevant and authentic learning objectives and test items. Given the challenges seen in the survey mentioned earlier, however, it would seem that there is still a gap between ensuring this knowledge is transferred to both learning methodology and test tasks.

I would argue that in such a safety related domain, we must redress the balance of this to ensure a learning process takes place which teaches those skills we want to use in real-life communication and as such those we want to test. Even if we do pre-suppose that the test will dominate, it is not a fanciful wish to ensure that our tests are, as a minimum, valid and reliable and reflect authentic real-life communication. As Green (2014:87) states, learning is for life 'beyond the test' while Bailey (1996) suggests that the test itself should come from the classroom. The move from learning exercises to testing tasks should be seamless (Messick, 1996) and there should be little difference between learning and being tested. Knowing what an authentic task is ensures we target the right skills. Taking that into the classroom means that we have authentic task and skills learning for use in real life, the test being simply a measure to calibrate the process of skills acquisition. Test tasks should encompass authenticity, practicality, interactivity from the content and the context in which the language necessary for the communicative process will be used.

Learners and test takers alike will surely be more intrinsically motivated and confident if they see a tangible link between the testing apparatus and their operational duties. Additionally, by addressing the needs of all stakeholders, literacy

will not just be about assessment but about underpinning a serious attempt to provide a valid and reliable system of maintaining and improving language proficiency. Moreover, good language test tasks can be shown to share characteristics with genuine language use in situations outside the test, engaging both test takers' language knowledge and their knowledge of relevant content and procedures (Douglas, 2004).

Finally, in remembering that the language is only a specific part of the communication, learning and testing tasks should also take into account the extraneous features of the communication as far as is practicable. Only when we identify and focus on such elements will we have achieved the congruence that is so important to both learning and testing.

8. A contextual congruence between language and communication

As was shown in the workshop results above, the bringing together of learning and testing into one entity should not be seen a forced collusion of bi-polar elements, but be a calibrated re-alignment of skills-based learning and assessment, to demonstrate that such skills have been learned. Achieving this goal should encompass a blended approach from SMEs and ELEs alike and not be seen as the domain of one or the other. The closer the fusion of skills and competencies the more accurate the construct taught in the classroom will be and the better will be the chances that learners will acquire the skills they need for real-life communication.

To underpin this theory, I will offer five examples here from my own professional environment where an integrated approach helps to ensure authentic learning.

1. The understanding of communicative language learning supports documentation and training material for raters. All the raters I train have access to an 'Assessor Handbook' which includes additional advice and support information on the language areas to be tested. It also includes information on how to give relevant and appropriate feedback and learning advice for test takers.
2. Basic discourse analysis with learners, item writers and raters, as well as with curriculum designers and material writers enables all stakeholders to see exactly what functions the language is forming in the process of communication. Furthermore seeing the juxtaposition of plain language and standard phraseology, both SMEs and ELEs can help each other develop an awareness of exactly where the language fits into the operational communicative process.
3. Helping SMEs value linguistic input and ELEs know exactly where and why it is needed at certain times in an operational context aids in matching language levels to those to be tested. Such skills awareness helps item writing for test tasks and focussed learning on the skills required in real-life.
4. Training also goes into the administration of the testing system and the training of raters in helping and assisting test takers with questions they may have about learning and being tested.
5. Having the chance to explain to learners where and why the language is used in an operational context adds face validity to the language teacher. It aids in

justifying his / her teaching methodology. Such face validity, as mentioned above, creates motivation and brings about confidence that what is being learned is valid and relevant. The test is then certainly less daunting than it might otherwise be, where there is a low level of acceptance from the test taker population.

ELEs can enhance their own operational awareness by making visits to operational locations, by using their students as technical matter experts and by seeing the benefit of language, not purely on its linguistic merits, but in real-life task based learning. SMEs likewise can take on board the linguistic knowledge that underpins language proficiency in a context that is familiar to them. It becomes less about an isolated view of language and more about successful communications. As noted by Douglas (2004) test development, and we could also arguably add materials and curriculum development in LSP learning, involve a wide spectrum of stakeholders in the test design process.

9. Conclusion

This paper set out to look at washback as a nebulous concept. It may not be ill-conceived, but could well be seen as a reactive ideology to something which threatened the evolution of communicative language teaching. Seeking not washback by design, but clearer valid objectives that align learning and testing as a likely basis for a harmonised learning process will mean thus washback becomes a *de facto* part of the process; a integrated theory.

By simply re-aligning the concept into a more logical and learner-focussed concept we are even able to suggest that the idea of washback need not exist at all. A well calibrated congruency of learning and testing based on real-life language and tasks can be foreseen as a systematic process. This process would form part of a continuum offering the ability to continue learning long after the test has been passed. It is also one way of ensuring that washback remains a concept acting as a safety harness integral to a systematic process. Fostering skills in an aligned process that targets real-life skills in authentic communication must surely be a better way of looking at learning and assessment.

REFERENCES:

Alderson, C., & Hamp-Lyons, L. (1996). TOEFL Preparation Courses: A Study of Washback. *Language Testing 13*(3), 280-297. Breul, K. 2013. 'Language in aviation: The relevance of linguistics and relevance theory' *LSP Journal*, Vol.4, No.1 71-86.

Alderson, C.J. (2009). 'Air safety, language assessment policy, and policy implementation: the case of aviation English' *Annual Review of Applied Linguistics*, 29, 168–187.

Alderson, C.J. (2010). 'A survey of aviation English tests' *Language Testing,* 27(1) 51-72.

Alderson, J. Charles. (1993). Does washback exist?, *Applied Linguistics*, 14, p.115

Bailey, K (1996). Working for washback: a review of the washback concept in language testing. Language Testing, 13/3, 257-279.

Breul, K. (2013), 'Language in Aviation: The Relevance of Linguistics and Relevance Theory' *LSP Journal*, Vol.4, No.1 71-86Ragan, 1997

Buck, G. (1988): Testing listening comprehension in Japanese university entrance examinations. *JALT Journal 10*, 15-42.

Bullock, N & Westbrook, C. (2017). *Testing in ESP: Approaches and challenges in Aviation and Maritime English*, Powerpoint slides from IATEFL TEASIG PCE, Glasgow, UK, April 3 2017.

Bullock, N. (2015). Wider considerations in teaching speaking of English in the context of aeronautical communications, *IATEFL ESPSIG Journal*, 45, 4-11.

Douglas, D. (2000). *Assessing language for specific purposes*, Cambridge: CUP.

Douglas, D. (2004). Assessing the language of international civil aviation: Issues of

validity and impact. Proceedings from the *International Professional Communication Conference*, IEEE Professional Communication Society (pp. 248-252). Minneapolis: IEEE.

Doye, P. (1991): Authenticity in foreign language testing. In Anivan, S., editor, Current developments in language testing. Anthology Series 25, Singapore: Regional Language Centre, 103-10.

Elder, C., Iwashita, N. and McNamara, T. (2002). 'Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?' *Language Testing* 19 (4) 347–368.

Farris, C., Trofimovich, P., Segalowitz, N., & Gatbonton, E. (2008). Air traffic communication in a second language: Implications of cognitive actors for training and assessment. *TESOL Quarterly*, 42 (3).

Field, J. (2013). *Cognitive validity* in Geranpayeh, A. and Taylor, L. (ed.) *Examining listening – Studies in language testing.* Cambridge: Cambridge University Press, 77-151.

Fulcher, G. (2010) *Practical Language Testing.* London: Hodder Education.

Green, A. (2014) *Exploring Language Testing and Assessment*, London: Routledge.

Hughes, A. (1993): *Backwash and TOEFL 2000*. Unpublished manuscript, University of Reading.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

International Civil Aviation Organisation. (2007). *Doc 9432*, *Manual of radiotelephony* 4th Edition. Montreal: ICAO.

International Civil Aviation Organisation. (2009). *Cir 323*, *Guidelines for Aviation English Training Programmes*, ICAO.

International Civil Aviation Organisation. (2010). *Doc 9835*, *Manual on the implementation of ICAO language proficiency requirements* 2nd Edition. Montreal: ICAO.

Kim, H. (2013). Exploring the construct of radiotelephony communication: A critique of the ICAO English testing policy from the perspective of Korean aviation experts. *Papers in Language Testing and Assessment,* 2 (2), 103-110.

Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca - perceptions Of Korean aviation personnel. *Australian Review Of Applied Linguistics*, 32 (3) 23.1-23.17.

Kukovec, A. (2008). '*Teaching aviation English and radiotelephony communication in line with the newly established International Civil Aviation Organization language proficiency requirements for pilots'*. Inter Alia, 1, 127-137. http://www.sdutsj.edus.si/InterAlia/2008/Kukovec.pdf. (Accessed 22nd June 2014).

Linde, C. (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. *Language in Society*, *17* (3), 375-399. In Moder 2013

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

Messick, S. (1996). Validity and washback in language testing. Language Testing 13(3),241- 256.

Moder, C. (2012). Aviation English. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for Specific Purposes* (pp-). Chichester, UK: John Wiley & Sons.

Moder, C. (2013). Aviation English. In B. Paltridge (Ed.), The Handbook of English for Specific Purposes. West Sussex: Wiley Blackwell.

Moder, C., & Halleck, G. (2009). Planes, politics and oral proficiency: Testing international air traffic controllers. *Australian Review of Applied Linguistics, 32* (3), 25.1-25.16. DOI: 10.2104/aral0925

Morrow, K., (1991): Evaluating communicative tests. In Anivan, S., editor, *Current developments in language testing. Anthology Series 25*, Singapore: Regional Language Centre, 111-18. 1991.

Paltrdige, B. & Starfield, S. (2013). *The Handbook of English for Specific Purposes*, Wiley-Blackewell.

Paramasivam, S. (2013). 'Materials development for speaking skills in Aviation English for Malaysian air traffic controllers: Theory and practice'. *Journal of Teaching English for Specific and Academic Purposes,* 1 (2), 97-122.

Read, J., & Knoch, U. (2009) 'Clearing the air: Applied Linguistic perspectives on aviation communication'. *Australian Review of Applied Linguistics,* 32 (3), 21.1-21.11.

Rubenbauer, F. (2009) *Aspects of Oral Communication in Aviation*. Aachen: Shaker Verlag.

Sarmento, S. (2011). What makes a good aviation English teacher? *Aviation in Focus*, (2).

Shohamy, E. (1992): Beyond proficiency testing: a diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal 76*, 513-21.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13* (3), 298-317.

Uplinger, S. (1997). 'English–language training for air traffic controllers must go beyond basic ATC vocabulary'. *Flight Safety Foundation* Airport Operations, *23*(5), 1-5.

Weir, C.J. (2005) *Language Testing and Validation – An Evidence Based Approach.* London: Palgrave Macmillan.

Wyss-Buhlmann, E. (2004). *Variation and co-operative communication strategies in air traffic control english*. Bern: Peter Lang.Goh (2013)

Yan, R. (2009). *Assessing English language proficiency in international aviation.* Saarbrücken: VDM.