What do we mean by the ‘washback effect’ of testing ?

Philip Shawcross
Aviation English Services
pshawcross@aeservices.net

Abstract

What do we mean by the “washback effect” of testing?

The paper introduces the notion of the “washback” effect of testing with well-known examples in the academic field. The form, content, focus and delivery of a test often determine an academic curriculum and the way it is taught. As a result, these features in turn tend to affect the skills set which is the outcome of the training.

In high-stakes aviation licensing testing, there is a greater need for the test results, and the prior or subsequent training, to be more directly applicable to operational conditions than in academia.

With a view to compliance with ICAO language proficiency requirements, the main concern as regards both testing and training is that they effectively address:

- All six skills identified and defined in the Rating Scale; and
- The appropriate and highly specific language content, namely the use of plain language in radiotelephony.

While test design also affects test validity, the focus of the present paper is on the potentially positive and negative effects of test design and content on the form and content of aviation English language training courseware.

The paper concludes with a brief coverage of the way in which all of:

- the type of test;
- the way in which the test is delivered and rated; and, most importantly,
- the language content of the test

can affect the way training is designed, the language skills which are targeted and, ultimately, the quality of training that results.

Summary

The ‘washback effect’ of testing is primarily the influence of testing on training and learning. It is also the potential impact that the *form* and *content* of an aviation English test may have on regulators’ and administrators’ conception of language proficiency and what it entails. The effect of test focus, type, delivery and content on training and training administration is examined.

Introduction

The **‘**Washback’ or ‘backwash’ effect of testing is a well-documented academic phenomenon common to nearly all institutional learning processes. The washback effect has been described succinctly as:

‘the influence of testing on teaching and learning.’ (Gates 1995)

It is a phenomenon known to us all from our school days:

‘Will this be in the exam, sir?’;

and reflected in the way teachers tend to model their curriculum around the focus areas, form and content of an examination or a test. In an academic world increasingly driven by performance, results and league tables, the highlighting of this phenomenon is ever more tangible.

As G. Buck wrote,

‘There is a natural tendency for both teachers and students to tailor their classroom activities to the demands of the test, especially when the test is very important to the future of the students, and pass rates are used as a measure of teacher success. This influence of the test on the classroom (referred to as washback by language testers) is, of course, very important; this washback effect can be either beneficial or harmful.’ (Buck 1988).

Washback however is not restricted to learners and teachers. Indeed, in this paper, which addresses the washback effect of testing within the context of ICAO language proficiency requirements and licensing testing, our concern will be more with the wider social effects of washback.

Bachman and Palmer (1996) consider washback to be a subset of a test’s impact on society, educational systems and individuals. They believe that test impact operates at two levels:

The micro level (i.e. the effect of the test on individual students and teachers); and

The macro level or the impact the test may have on society and the educational system.

Another consideration to be kept in mind as we briefly consider the washback effect of language testing in the aviation industry is that test design and content directly affect the validity of a test, i.e. the degree to which a test is measuring what it claims to measure. A test’s validity determines the reliability of its results.

In a purely academic environment, in which most research has been performed to date, language testing takes place in a largely closed circuit, where success or failure is the key to the next step in the academic chain. Proficiency testing in an aviation context is truly ‘high stakes’ because its outcome directly engages the real world in terms of safety and career.

Therefore test designers, delivers and raters have a particular responsibility in as much as the testing process may have a considerable impact, either positive or negative, not only upon:

The reliability of the test itself; and

The way we train for the *level* and *breadth* of proficiency required to meet the standards defined by ICAO in the Rating Scale; but also

Administrators' and regulators' awareness of the connection between testing and training and the need to get both elements right to ensure an ultimate improvement in language proficiency. This is not a matter to be resolved by the inexpert or inexperienced.

In this short paper, we will explore four test features which can have a washback effect, for good or ill: test focus, test type, test delivery and test content. This exploration is by no means exhaustive; these are only examples of washback dynamics.

Focus

The *focus*, bias or scope of an aviation English test is possibly the most fundamental of all the features we will consider as it, in turn, determines the choices to be made about others aspects of tests.

If we consider a spectrum stretching from the most specific use of language (i.e. standardised phraseology) to the most general (i.e. everyday language), the drawbacks of extremes becomes apparent.

Too narrow a focus - a test largely restricted to standardised phraseology - will fail to address those aspects of language which are the core target of the ICAO requirements, i.e. the use of plain language in a work-related, radiotelephony environment. This is the aviation-specific but still free-form language required to address non-standard and abnormal operational situations, in other words, those circumstances of flight of high inherent risk that require a well-developed proficiency in language to resolve. In these situations, standardised phraseology is not enough. A decision to focus particularly on standardised phraseology would result in the language targeted by the ICAO Standards being by-passed ; it would fail to address that subset of language that ICAO has identified as most critically requiring improvement and would do a double dis-service by fostering a misconception as to the nature of the language proficiency required. This misconception may well induce a false sense of security amongst both operational staff and regulators, as well as blur the distinction between linguistic and operational skills and training.

On the other hand, too wide a focus (e.g. general English with an aviation 'flavour') will fail to assess the use of the language required in those operational situations that are the subject of radiotelephony exchanges between pilots and controllers. These exchanges call for the specific language functions identified in Appendix B (pp. B1-B17) of ICAO Doc. 9835. They include clarification, confirmation, making requests and offering action, expressing intentions, describing ongoing situations, managing dialogues, resolving conflict situations and responding to emergencies. By casting the net too wide, such an approach to testing may result in an excessively long period of training to reach Operational Level 4 while not focusing specifically enough on the language functions which are operationally critical.

Test Type

There are various *types* of language test:

Indirect computer-assisted testing, whereby both delivery and rating are provided by a computer using conventional and voice recognition technology. While possibly appropriate for benchmark testing, this type of testing system without human intervention does not offer the reliability or scope required for proficiency or licensing testing.

Semi-direct computer-assisted testing in which speech samples are elicited by computer-generated prompts and rated later by human assessors.

Oral Proficiency Interviews which use tried and proven conventional interview techniques, and

Hybrid testing using a combination of computer-assisted and live interviews in various formats.

There are concerns about some types of testing being able to properly assess certain of the six skills in the ICAO Rating Scale, notably interactions and natural fluency. For example, it is legitimate to wonder whether lengthy speech samples in response to computer-generated prompts may fail to replicate natural interaction. This is somewhat reminiscent of those students in school examination situations who fill the time with pre-rehearsed speeches to avoid probing questions.

It is to be feared that once such a test type or format is known, it may attract a too narrowly focused kind of training, geared more to exam preparation than to natural speech production.

It may well be that only a skilled interlocutor is able to probe and conduct an interview in such a way as to generate a situation that genuinely provides for reliable assessment of the test-taker’s ability to interact and manage an exchange in an almost limitless range of unpredictable scenarios.

Test Delivery

Equally, there are different variables of test *delivery* that include:

Computerized or person to person;

Telephone / video conference-based or face to face;

Textual or oral delivery of test items;

Native or non-native interlocutors; and

The use of aviation professionals or qualified language specialists.

All of these may have either positive or negative washback effects depending on how they are managed and presented. 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 within the organisation.

Just to take one of the items above as an example, textual or oral delivery: an inappropriate reliance on textual delivery may not only fail to reflect Rating Scale priorities, but also influence learning habits negatively by not encouraging a more communicative approach to language acquisition. ICAO Document 9835 6.8.3 (p. 6-11) addresses this by noting that ‘a language test for the aviation industry should replicate as far as possible the work-related communicative requirements.’

As Pearson remarked, ‘There is an explicit intention to use tests, including public examinations, as levers which will persuade teachers and learners to pay serious attention to communicative skills and to teaching learning activities that are more likely to be helpful in the development of such skills.’ (1988)

Test Content

Finally, test content can have a very direct positive washback effect upon training curricula.

The systematic, accurately targeted coverage in proficiency tests of the relevant lexis, structure and functions will tend to drive training programs to address the required areas of aviation language expertise.

The lexical domains which need to be addressed to ensure the sort of operational proficiency expected at ICAO Level 4 are surprisingly broad, but also intentionally work-related. Consequently, there are many fields that may and should be included in testing and training design. These include operational actions, airfield activities, movement of animals and birds, principles of flight, human behaviour, security issues, cargo, topography, accident causes, environmental conditions, health, communication, emergencies, modality, perception, problems, rules, movement, distance, time, travel and weather, to name but a few. All of these should be included in both training curricula and proficiency test items.

In aviation English, operational functions such as giving orders and making requests tend to define grammatical requirements. Other operational language functions similarly, by their nature, go a long way towards clarifying the form of language testing and training required: these are outlined in the ICAO LPR Scale and include clarifying, paraphrasing, confirming, describing ongoing actions and intentions, relating events in the recent and more distant past and relaying information.

To have sufficient validity to be effective, proficiency tests need to address these areas, either specifically or in an integrated way through oral prompts, but it is vital that this requirement is met in the context of the specific aviation language sub-set being targeted by ICAO. As discussed in Document 9835, ‘Grammatical accuracy might be considered only so far as it impedes communication, for example, but evaluating an individual’s grammatical knowledge (in itself) would not be the test objective.’ (6.6.6)

Conclusions

At the end of this brief study, it is possible to identify two areas in which the dynamics of the washback effect in aviation English testing are particularly significant:

In the effect that test focus has especially on the awareness of the scope of language proficiency by regulatory and managerial staff;

In the way that test focus and test content help define the *level* and *breadth* of language proficiency required to properly meet ICAO Level 4.

Both these factors will tend to drive training policy and training design within the industry. Finally, like aviation English training, aviation English test design needs to be constantly held up to the specific operational requirements of aviation radiotelephony; this is the litmus test which alone can provide validity.

The mechanism of the washback effect of testing is a valuable means that we have of verifying the appropriateness of our overall response to the industry's call for safe and reliable communication.

References:

Babaii, E. (2004). *Washback in language testing.* Lawrence Erlbaum Associates.

Bachman, L. & Palmer, A. (1996). Language testing in practice. Oxford

Brown, J.D. (2002). *Extraneous variables and the washback effect*. JALT vol. 6 No. 2

Buck, G. (1988) Testing listening comprehension in Japanese university entrance examinations. JALT (10).

Cheng, L. & Watanabe, Y (2004). *Washback in language testing: research contexts and methods.* Mahwah, NJ.

Gates, S. (1995). *Exploiting washback from standardized tests*. In J. D. Brown & S. O.

Yamashita (Eds.), *Language testing in Japan* (pp. 101-106). Tokyo: Japanese Association for Language Teaching.

ICAO Document 9835 (2004)

Khoshsima, H. & Roostami, A. (2006). *The washback effect of alternative assessment techniques on students' performance.* PhD abstract.

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

Pearson, I. (1988). *Tests as levers for change*. ELT Documents # 128. London

Spolsky, B. (1994). *The examination-classroom backwash cycle: some historical cases.* In Bringing about change in language education: proceedings of the International Language in Education Conference 1994. University of Hong Kong.

— — — — — — — —