Melita Djurić

Dealing with Situations of Positive and Negative Washback

Abstract

The article deals with the complexity of a washback phenomenon in language testing. It focuses on its positive effects within an institution as well as on the situations of negative washback. Washback is presented as a stimulus for a change and as a bridge for efficient communication between teachers and testers. Certain changes as a result of positive washback point at the opportunities which a testing institution has when it organizes, designs and administers criterion-referenced tests.

The complexity of washback is confirmed when the teachers' perspective is discussed. Teacher-tester relationship and the lack of teacher insight into testing may contribute to negative washback. Within this frame, concrete situations are described and lessons learned are summarized. In the conclusion, the teachers and testers are reminded of professional and ethical standards and of their responsibility towards their clients, students and test takers.

Keywords: washback, STANAG tests, criterion-referenced tests, test validity.

1. Introduction

The concept of washback covers both teaching and testing situations. Washback is generally known as the effect of testing on teaching. Alderson and Wall (1993, as cited in Fulcher and Davidson, 2007: 224) described it as "a complex phenomenon" and in their washback hypotheses they assumed that teachers and learners "do things they would not necessarily otherwise do because of the test". Alderson and

Wall presented numerous elements which create positive or negative washback and emphasized the need to further investigate the nature of washback. Hughes (2003) in his second edition of the book writes[1] about "greater interest in backwash than was previously the case and admits its importance in language testing (2003: 53).

Zavašnik and Pižorn (2006) examined the situation in Slovenia and found that there were no empirical studies of washback although external examinations and proficiency testing in a foreign language were introduced at the national level by the National Examination Centre in 1996. Another unfortunate fact is that testing as a field of applied linguistics is not included in the undergraduate or postgraduate studies of Slovenian language teachers. Consequently, this has resulted in an absence of empirical studies and research into the current language testing situation in Slovenia.

It is not only the National Examination Centre which organizes language testing in Slovenia. The Ministry of Defence also needs to organize proficiency testing that follows the NATO Standardisation Agreement (STANAG), which under the number 6001 describes language levels set for the international military community. The Slovenian Ministry of Defence established the School of Foreign Languages (SFL) in 1999 and STANAG testing for English started the same year.

This paper deals with washback at the institutional level as a result of ten years of testing experience at the SFL. The text begins with the theoretical background of three issues which are closely connected, criterion-referenced testing, test validity and washback. The presentation of English STANAG testing follows as an illustration of a specific testing situation. Practical experiences are described together with some problems and solutions which both show the complexity and importance of washback. In the end, managers of language institutions, language teachers and language testers are reminded of their responsibilities towards their clients, students and test takers.

2. Criterion-referenced tests

Brown and Hudson (2002: xiv) claim that criterion-referenced testing is most useful to classroom teachers and curriculum developers because "criterion-referenced tests are specifically designed to assess how much of the content in a course or program is being learned by the students". This is why criterion-referenced tests are good measures of students' strengths and weaknesses considering the goals and objectives of a particular course or programme.

Hughes (2003: 21) explains the purpose of criterion-referenced tests as "to classify people according to whether or not they are able to perform some task or set of

[1] Hughes uses the expression *backwash*.

tasks satisfactorily". He emphasizes two positive sides of criterion-referenced tests: set standards in terms of what people can do and motivation of students to attain these standards. Hughes pays attention to test specifications which must make clear what candidates have to be able to do and with what degree of success. Only then can students have a clear picture of what they have to achieve. "What is more, they know that if they do perform the tasks at the criteria level, then they will be successful on the test, regardless of how other students perform" (2003: 55).

Criterion-referenced testing, as well as testing in general, is not widely known among language teachers. The majority of teachers did/do not have many opportunities to learn about various types of language tests during their studies. The exams that teachers come across during their studies are more or less achievement tests and during their teaching practice, they usually design progress or achievement tests based on textbooks. Hughes (2003) emphasizes the importance to base achievement tests on objectives instead of on detailed teaching and textbook content to get a truer picture of what has actually been achieved but teachers need to be taught how to do this. Lack of information and skill puts teachers into a compromised position. The fact that their students who they know very well from courses need to take a criterion-referenced test after the course leads to affectionate feelings towards students and rejecting feelings towards testers. However, student motivation in criterion-referenced tests should bring teachers and testers together rather than apart. The washback perspective may help both groups understand the broader picture of testing and challenge a need for research and consequently a beneficial washback.

2.1. Test validity

This part has no ambition to cover the complete theoretical background of test validity. It deals with some aspects of test validity which support the connection between test validity and washback especially considering the issues developed later in the paper.

Alderson and Wall (1993, as cited in Fulcher and Davidson, 2007: 223) claim that washback cannot be related directly to a test's validity and criticize the statement of some writers that a test's validity should be measured by the degree to which it has had a beneficial influence on teaching. Alderson and Wall reject the concepts "washback validity" because "this form of validity has never been demonstrated, or indeed investigated, nor have proposals been made as to how it could be established empirically rather than asserted" (Ibid.).

Messick (1996) emphasizes two elements of test properties, authenticity and directness, because they are likely to produce washback. He classifies both properties under construct validity. Looking at the broader concept of "validity framework, washback is seen as an instance of the consequential aspect of construct validity" (1996: 242). To encourage positive and reduce negative washback, testers should minimize construct under-representation and construct-irrelevance in the

assessment. According to Messick washback is not simply good or bad teaching or learning practice that might occur with or without the test, but rather good or bad practice that is evidentially linked to the introduction and use of the test.

> "If a test's validity is compromised because of construct under-representation or construct-irrelevant variance, it is likely that any signs of good teaching or learning associated with the use of the test are only circumstantial and more likely due to good educational practices regardless of test use. Similarly, signs of poor teaching or learning associated with the use of a construct-validated test are more likely to reflect poor educational practices regardless of test use. [...] Although there may be exceptions requiring careful scrutiny, negative washback *per se* should be associated with the introduction and use of more valid tests because construct under-representation and construct-irrelevant variance in the test could precipitate bad educational practices while minimizing these threats to validity should facilitate good educational practices."
> (Messick, 1996: 247)

Positive washback is according to Messick linked to authentic and direct assessments and to the need to minimize construct under- representation and construct-irrelevance in the test.

Hughes (2003) agrees that direct testing implies the testing of performance skills with texts and tasks as authentic as possible. "If we test directly the skills that we are interested in fostering, then practice for the test represents practice in those skills" (2003: 54). He is very explicit in promoting direct testing:

> "If we want people to learn to write compositions, we should get them to write compositions in the test. If a course objective is that students should be able to read scientific articles, then we should get them to do that in the test."
> (2003: 54)

For the needs of this paper the concept of washback will be dealt with as consequential validity at the level of institutional organization first and at the level of curriculum later.

2.2. Washback

Alderson and Wall (1993, as cited in Fulcher and Davidson 2007) include different factors in their washback hypotheses. If teachers use tests to get their students to pay more attention to lessons and to prepare more thoroughly, it is positive washback. If teachers fear poor results and the associated guilt which might lead to the desire for their students to achieve high scores in tests, it might be a reason for teaching to the test. Consequentially, teachers narrow the curriculum and produce

negative washback. In their Sri Lankan Impact study Wall and Alderson (1993) came to an important conclusion that "tests have impact on *what* teachers teach but not on *how* they teach" (1993: 68).

Bachman and Palmer (1996) place washback within the scope of impact. They understand learning and teaching as two processes which "take place in and are implemented by individuals, as well as educational and societal systems, and society at large" (1996: 30). Understanding washback as an intended outcome of the test, Bachman and Palmer expect "the specific components (for instance, teaching method, curriculum, materials) to be affected and the degrees to which they are affected" (1996: 137).

Shohamy, Donitsa-Schmidt and Ferman (1996) agree that policy-makers are aware of the power of tests and may use them "to manipulate the educational system and to control curricula and new teaching methods" (1996: 316).

Bailey (1996) refers to *washback to the learners* and *washback to the programme*. The latter includes judging students' language in relation to the expectations of the curriculum, to determine whether the school as a whole performs well or whether teaching methods and textbooks are effective tools for achieving the curricula goals.

Being aware of the complexity of washback, Hamp-Lyons (1997) is concerned about its ethical side and points out the responsibility of language testers to accept responsibility for all those testing consequences they are aware of. She emphasizes a need to develop "a conscious agenda to push outward the boundaries of our knowledge of the consequences of language tests and their practices" (1997: 302). Wall and Alderson (1993) also pointed at that aspect when they warned "to guard against oversimplified beliefs that good tests will automatically have good impact. Washback needs to be studied and understood, not asserted" (1993: 68).

In her review of empirical studies of washback, Spratt (2005) identified the areas of teaching and learning which could be affected by washback: curriculum, materials, teaching methods, feelings and attitudes, learning. A teacher is the most important and influential agent in the process of introducing the effects of washback into teaching and learning. Spratt sees teachers facing "a set of pedagogic and ethical decisions about what and how best to teach and facilitate learning if they wish to make the most of teaching towards exams" (2005: 27).

Wall and Horak (2008) focus on the role of communication in creating positive washback. They found that teachers usually do not understand the nature of tests and encourage testers to communicate their intentions so that teachers and learners can prepare for new kinds of assessment. They also call for dissemination of the principles embodied by the tests and the provision of teacher and learner support and conclude "Much advice is available from exam designers and teachers, if only someone could collect it and organize it effectively" (conference presentation, 2008).

Washback (and its communication) shows itself as a gap or a bridge between teachers and testers as well as an indicator for a need for change. If teachers are

not isolated from testing and if they recognize and respect ethical principles in the classroom, their awareness process works towards positive washback and they will promote good practices. The complex nature of washback allows broad expectations in different areas. Consequently, washback can be understood as a powerful tool to introduce changes not only in teaching and testing but also in educational policy if it is supported by evidence and/or research.

Situations of positive and negative washback will be described in STANAG testing situations.

3. English language tests STANAG 6001

STANAG proficiency levels were introduced in 1976 (Edition 1) for English and French languages and updated in 2003 (Edition 2). Three purposes of the document NATO Standardisation Agreement, STANAG 6001 were:

- to meet language requirements for international staff appointments;
- to compare national standards through a standardized table;
- to record and report, in international correspondence, measures of language proficiency (if necessary by conversion from national standards).

Language proficiency levels and standards classify STANAG tests among criterion-referenced tests. Language standards are described for four language skills and as such form a constituent part of test specifications.

The descriptions of five levels[2] give definitions of language proficiency in four language skills: oral proficiency (listening and speaking) and written proficiency (reading and writing). A language proficiency profile (Standard Language Profile, SLP) is recorded by four digits indicating the specific skills in the following order: Listening, Speaking, Reading, and Writing (SLP 3321)[3].

STANAG 6001 tests are language tests for the military but Green and Wall (2005) found in their study that the tests were not necessarily ESP tests in nature. Some country testing teams practice a general English approach while others add more or less a specified number of military texts. Testing teams usually know SLP requirements within NATO but they have little or no information about what candidates should do with the language. Testers and teachers need to know what constitutes adequate language performance. It is an important issue for test constructors, as it has ”implications for how much they can be expected to contribute to the design and running of tests and how much candidates from different backgrounds and levels of the hierarchy can be expected to handle” (2005: 382).

[2] Level 1- Survival, Level 2 - Functional, Level 3 - Professional, Level 4 -Expert, Level 5- Highly-Articulate Native Speaker. These new labels are in process for a STANAG 6001, Edition 3.

[3] SLP 3321 means level 3 in listening, level 3 in speaking, level 2 in reading and level 1 in writing.

The Slovenian Armed Forces included the STANAG language levels among the criteria for working positions from the lowest to the highest military ranks. The same criteria were upheld for the civilian employees of the Ministry of Defence. To meet the needs of STANAG levels for military personnel, the SFL used to organize three testing sessions per year and test 200 – 250 candidates in all four language skills for levels 1-3.

The teaching staff of the SFL was small and nobody was specialized in testing. This situation was typical in other countries as well and not just in Slovenia. In his survey of modern language testing, Bachman (2000) writes that language testing as a subfield within applied linguistics "evolved and expanded in a number of ways in the past 20 years or so" (2000: 3). Foreign contract testers and teachers who worked temporarily in the SFL were mostly American and English. They organized internal workshops to familiarize Slovenian language teachers with basic testing principles. At the same time international networking of military testers started. Testing workshops and seminars were organized by the British Council Peacekeeping English Project and the Bureau of International Language Co-ordination (BILC)[4].

Testing sessions were reduced to two per year in 2003 when a systematic approach to organization and quality of proficiency testing started. As of 2008, the SFL has three full-time testers who are continuously perfecting and upgrading not only their skills but the overall quality of the tests. They are responsible for the organization of testing work, test design and piloting. They compare STANAG standards with the levels of Common European Framework (CEF) since CEF standards are recognized as national language standards in Slovenia. There is close co-operation between testers and a group of teachers who are included in item writing, moderation and test administration processes.

3.1. Experiences of positive washback

STANAG tests are high stakes tests, which significantly affect the lives of those who take them. Within this context, washback represents a constant pressure to respond to test results with appropriate actions. In the SFL, it was not before 2005 that certain changes started being introduced. The situations will be described first and the actions as examples of positive washback will be presented later.

Situation 1
Following the test results from 1999, the SFL testers were faced with significant differences in language proficiency on a scale within the same STANAG level. For example, the STANAG threshold level 2 differed in the quality of language knowledge a lot from level 2 at the upper part of the scale. When the candidates from both extremes on the scale applied for a higher level course (STANAG 3), the teachers complained that some of the students who had reached level 2 did not show

[4] BILC is an advisory language body to NATO.

adequate language knowledge and as a result the students did not form a homogenous learning group. The need arose for levels within a level.

Situation 2
The Slovenian STANAG requirements for writing skills were/are a level lower[5] than for the other three language skills. In 2005 Slovenia became a NATO member and suddenly a demand for and the awareness of the importance of writing skills increased. There was not much that teachers could do because both the teaching and testing staff of the School were aware that weak writing skills in English may be correlated with weak writing skills in the mother tongue. In the past, the experience of learning and teaching reading techniques in English improved the reading skills in the mother tongue. The challenge to introduce the same practice for writing skills was in the air.

Situation 3
Test results showed that younger generations of military personnel (20-30 years of age) were achieving better STANAG results without any participation in language courses than the older candidates (40-50) who were regulars in language courses from STANAG 1 to 3. However, the English communicative abilities of younger candidates were not enough when they needed higher levels, especially levels 2 and 3. To fill the gap between longer courses (beginner, intermediate, upper-intermediate)[6] and to give younger employees an opportunity to improve their language accuracy, a need to introduce new types of courses appeared.

Situation 4
Candidates have to re-take the STANAG exam if they do not reach the required language level. This is a requirement in the contracts for participation in a language course as well as a requirement for promotion and nomination for positions abroad (international military delegations, peacekeeping missions). In many cases re-takers wanted to reach the level they were striving for but they did not ask for feedback or explanation as to why they had previously failed. Their attempts to achieve higher levels were mostly unsuccessful.

The actions and changes introduced referred to the programme and the policy of the SFL:

Descriptive marks within levels
Three descriptive marks within each STANAG level were introduced: *threshold*, *good, excellent.* Two main purposes were to better place students in language courses and to give more specific information about the reached language level to the personnel department and to the candidates themselves. Descriptive marks explain that at the level of *threshold* a candidate's language is too weak for a higher-level course, the *good* language level means that a candidate is ready for a higher-level course and the *excellent* language level allows a candidate to take a test without any course.

[5] 1110, 2221, 3332.
[6] Beginner course – 300 hrs, Intermediate course – 300 hrs, Upper-Intermediate course – 450 hrs.

These descriptors were explained to the personnel department of the Ministry of Defence and have been applied since 2005.

Writing courses
New courses were developed aiming at improving writing skills. During the course design phase the element of mother tongue was also considered and the content was aimed at improving writing skills and organization of writing in both languages. The Writing 2 and Writing 3 curricula included contact time with teachers and self-study time. The SFL has organized two courses, Writing 2 and 3 since 2006 but there has been little interest in the Writing 2 course and no interest in Writing 3. Our assumption is that military employees do not perceive English writing skills as lacking in their professional communication (yet).

Refresher courses
Two new courses were developed, Refresher Basic and Intermediate, aiming at refreshing and improving the existing language levels STANAG 1 and 2. Two additional purposes were to fill the gap between the existing courses and to offer those with *threshold* marks an opportunity to raise their language level up to *good* so that they would be able to apply for higher level courses. The Refresher Basic has proven to be a very attractive course and the SFL has organized 6 from 2006. There have always been more candidates than free slots which is not the case with the Refresher Intermediate. It was organized once and was not full. Our assumption is that the language level STANAG 2 is the realistic language level for functional purposes of military employees.

Exam-related material
A solution as to how to reach the military population with basic and simple information about proficiency testing was to publish information about tests in a military newsletter. After it was published, students made little effort to find out more about the test, although some made inquiries by phone. As a result testers produced two booklets; Frequently Asked Questions and a Self-Study Guide with instructions on how to learn and study for the test. Also included were test-taking strategies, different test methods and the test content. Test samples for all skills were added. The editor of the Ministry of Defence home page agreed that both booklets should be made accessible through the Intranet Information Point where a special slot was created for the SFL. After three years of disseminating testing information more widely, there are still too many candidates who do not know what is available and where. Additional strategies have to be considered and introduced.

3.2. Experiences of negative washback

It has been mentioned above that teachers often bring negative feelings into the issue of testing especially in the cases of high-stakes tests. Hamp-Lyons wrote about "fears of teachers and their intuitive avoidance of formalized testing for their students" (1997: 295) and Bailey (1996) introduced the concepts of washback to the programme and washback to the learner. Spratt's survey (2005) also reported about

teachers' rather negative attitudes and feelings generated by exams. Following Spratt's areas, the SFL course content and the SFL testers' and teachers' feelings/attitudes may have been affected by negative washback. Neither of these issues has been investigated yet.

Course content - Test content

Evidence about the course content was gathered from final course reports, evidence about the test content was taken from feedback forms which test takers and test administrators filled in after STANAG exams.

Test takers were from two different level courses, STANAG 2 and 3. They perceived bi-level listening tests as (completely) different from the course content. The same feedback appeared repeatedly after testing sessions. Test administrators were teachers in both courses and were involved in testing only as test administrators for reading, listening and writing. Test administrators did not notice the difference between the content of the listening test and the course content, however, they commented on the quality of listening test tasks, sound quality and insufficient time for test-takers' answers (test-takers' comment too). The difference in interpretation between test takers and test administrators was noticed, but testers did not pay special attention to it.

Final course reports reflected the opinions of students' about the course content after the course and before the exam. Listening skills came out repeatedly as the skills not taught enough during courses and students' perception was that they did not progress in listening as much as they could. The same information appeared after several courses and the course directors and course teachers were expected to discuss the issue and suggest a change whether in test method or in the organization of the listening module. Unfortunately, little has been changed if at all. This fact can be understood as confirmation of Wall and Alderson's statement (1993: 68) that tests do not have an impact on *how* teachers teach especially because testers did not highlight the difference in interpretation between the course content and the test content.

Feelings/attitudes

Feelings/attitudes were not observed or noted in an organized way. We are going to assume them from test administrators' feedback being aware that there are different factors which might contribute to our assumptions (subjective perceptions, tense atmosphere resulting from test anxiety, etc.).

Test-takers' feedback for reading, writing and speaking tests did not differ significantly from final course reports. Still, the teachers/administrators' tone of their perfectly legitimate comments appeared repeatedly more negative than neutral and referred to the type of writing tasks not being authentic[7], typos in reading tests, more than one correct answer in reading and listening tasks. On the other hand, the testers assumed from speaking tests that teaching to the test was a part of the

[7] Testers' purpose was to elicit a certain language sample from *a memo* and *a letter* as two test tasks. However, a memo does not represent a common practice in Slovenian official correspondence and letters are not perceived as a standard way of communication any more.

course content because test-takers showed different language knowledge in different speaking tasks; some portions of speaking tests did not sound spontaneous and were told as if memorized or did not correspond to testers' sub-questions asked during the test. After exams, washback workshops were organized for course teachers to inform them about test results and the mistakes most frequently made by the students. Testers perceived negative feelings from the teachers which frequently resulted in communication misunderstandings between teachers and testers.

It can be assumed that listening tests and the listening module of the course are both affected by the lack of information about the target language situation. Not being able to specify an authentic listening situation has resulted in inefficient teaching and learning practices. After testers find an accurate as possible description of the target language, it will be realistic to organize listening training sessions and discuss the appropriate teaching methods with teachers. Teachers will then be able to re-organize the listening module and perhaps suggest more communicative and authentic test tasks to be introduced into the listening tests. A common interest of the SFL to improve the course listening module and the listening test may contribute to wider discussion about teaching to the test and its ethical aspects. One would hope that fair discussion minimizes the negative feelings on both sides.

3.3. Lessons from washback experiences

In the SFL, washback is observed during each testing session. So far more action has been taken concerning the educational policy and curriculum organization than teaching methods and specific teaching situations. It seems that changes which can be interpreted as positive washback are more efficiently introduced into the system than into the thinking process of teachers and testers.

The actions of positive washback placed new tasks on testers and teachers:

- Descriptive marks required the adaptation of cut-off scores in listening and reading tests. A rating scale for writing was supplemented by more analytical descriptions within each level.
- Standard placement tests in reading and listening skills were supplemented by a short writing composition. The results affect the decision whether or not to place a candidate into the writing course.
- Writing courses have encouraged teachers to learn and adopt the principles of e-learning. The writing course combines teacher contact time and distance learning time which makes it a blended course.

Considering the difference between course content and test content, Wall and Alderson (1993) reported that "teachers cannot tell by looking at the exam how they should teach" (1993: 66) a certain skill and many teachers are "unable, or feel unable, to implement the recommended methodology" (1993: 67). They found that "an exam on its own cannot reinforce an approach to teaching the educational system has not adequately prepared its teachers for" (Ibid.). Among elements of the educational system Wall and Alderson included insufficient exam-specific teacher

training. Exam-specific teacher training reveals itself as an opportunity to start or/and improve communication between teachers and testers. It also shows itself as a need to deal with negative washback, to discuss its reasons and consequences.

Similarly, testers need course-specific training. Firstly, they need to understand that *a course* includes teachers, students, a curriculum, a syllabus and a textbook. The test purpose and the course objectives cannot be pulling in different directions. Communication between teachers and testers and observation of the classroom situation are to be planned and controlled. Secondly, feedback from teachers – test administrators needs to be taken into consideration very seriously; if it is noted and discussed in details, it may produce constructive advice. Thirdly, test validity may be compromised if there is no control of tests and testers. Teachers should certainly represent a kind of control. They will contribute to positive washback if they discuss with testers how authentic and direct the tests and test tasks are. Constructive discussions should lead to minimizing construct under-representation and construct-irrelevance in the test. Finally, tests are not independent and testing cannot be separated from teaching and vice versa. As such, testers need to communicate with teachers and teachers need information and training from testers.

Negative washback has raised new issues for teacher training: designing courses aiming at specific language levels instead of at test tasks or a test in general; classroom assessment according to the objectives of the course instead of the textbook content; training in methods to teach individual language skills.

The areas remaining open for research are the following:

- How efficient are new supplemented descriptive marks within the STANAG levels? Being applied does not necessarily mean that they are efficient. How do they correlate with students' progress in higher level courses?
- What affects the interest in writing courses? Is a new teaching method sufficient enough to make the course more attractive for military employees? Won't an element of distance learning time contribute to less interest from military employees instead of making them more responsible for their own learning?
- How to disseminate exam-related material to potential candidates? How to make exam feedback interesting so that military employees understand its value before re-taking an exam?
- Will encouraging teachers to learn a new teaching method turn them away from professional development because a new method is time-demanding? Since e-learning requires a very different perspective on teaching and learning it may present a challenge for teachers but it may also be counter-productive if teachers do not see any changes (better testing results, more interest from students).
- Where is the point at which communication between teachers and testers stops being efficient and honest?

A number of questions that are and will be difficult to answer.

4. Conclusions

Washback needs to be planned, observed, studied, and communicated. The process of producing positive washback includes testers and teachers, their training, communication and consistency. The management of a language/testing institution needs to inform teachers and testers how influential their roles are when introducing changes at the institutional, programme or classroom levels. These changes present positive washback when teachers know how to introduce a change and when testers and teachers are aware of their professional responsibility and ethical aspects. The culture of sharing teaching and testing information and further discussion on these professional issues will contribute to the awareness of and a need for professional development and life-long learning.

Negative washback does not necessarily have negative effects. As soon as negative washback is noted it can be addressed. Considering its complex nature it is difficult enough to identify it but responding to it professionally and timely is the responsibility of testers, teachers and institutions towards their clients – students and test takers.

Teachers help testers improve their tests, testers help teachers improve their teaching and both need to accomplish a common mission i.e. help students and test takers reach the course objectives during a course and reach the required language level by valid tests. Changes as results of washback should be introduced to improve teaching and testing processes primarily for the sake of students and test takers.

Acknowledgements

I would like to express thanks to both reviewers for their detailed reading and constructive suggestions.

References

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

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing,* 17 (1), 1-42.

Bailey, K. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing,* 13 (3), 257-278.

Brown, J. D. and Hudson, T. (2002). *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

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

Greene, R. and Wall, D. (2005). Language testing in the military: problems, politics and progress. *Language Testing,* 22 (3), 379–398.

Hamp - Lyons, L. (1997). Washback, impact and validity: ethical concerns. *Language Testing,* 14 (3), 295-303.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

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

NATO Standardization Agreement, STANAG 6001. (2003). Edition 2.

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

Spratt, M. (2005). Washback and the classroom: the implications for teaching and learning of studies of washback from exams. *Language Teaching Research,* 9 (1), 5–29.

Wall, D. and Alderson, J.C. (1993). Examining washback: the Sri Lankan Impact Study. *Language Testing,* 10 (1), 41–69.

Wall, D. and Horak, T. (2008). The Role of Communication in Creating Positive Washback. Presentation at EALTA Conference, Athens.

Zavašnik, M. and Pižorn, K. (2006). Povratni učinek nacionalnih tujejezikovnih preizkusov: opredelitev pojma in posnetek stanja v svetu. *Sodobna pedagogika*, 57 (1), 76-89.