

Review of washback in language testing: How has been done? What more needs doing?

Dina Tsagari

Lancaster University, UK

25 July 2007

Abstract

The review presented and discussed in this paper explores the theoretical underpinnings and research findings of the washback of high-stakes tests in the field of language teaching and testing as well general education and suggests areas and ways of researching the phenomenon in the future.

Definitions and scope

Labels used

In the educational measurement literature, ‘washback’ is variously known as ‘*test impact*’ (Baker, 1991), ‘*consequential validity*’ (Messick, 1989, 1996), ‘*systemic validity*’ (Frederiksen & Collins, 1989), ‘*measurement-driven instruction*’ (Popham, 1987) or ‘*curricular alignment*’ (Madaus, 1988; Smith, 1991a).

Various labels are attached to the concept of exam influence in the field of language testing and teaching, too. Some of the best-known are ‘*backwash*’ (Hughes, 1989), ‘*washback*’ (Alderson & Wall, 1993) and ‘*impact*’ (Wall, 1997).

Definitions of terms in language testing

The terms ‘*washback*’ and ‘*backwash*’ are used interchangeably in the field. ‘... *to clarify the distinction between the terms backwash and washback*’, Alderson says ‘*there is none*’ (2004:xi). Nevertheless, ‘*washback*’ is the preferred term in British applied linguistics (Cheng & Curtis, 2004:5).

In their simple definition, ‘*backwash*’ or ‘*washback*’ refer to the influence of testing on teaching and learning (e.g. Hughes, 1989; Alderson & Wall, 1993; Bailey, 1996; Saville, 2000; Cheng & Curtis, 2004). However, numerous explanations of the term ‘*washback*’ can be found throughout the published research and literature on language testing with various meanings, which reveal differences in scope and intentionality.

These are presented below categorised under common themes:

- ***The nature and extent of washback***
- Washback is seen as a consequence of high-stakes exams (e.g. Alderson & Wall, 1993; Hamp-Lyons, 1997).

- Washback is seen as the link between testing, teaching and learning (e.g. Shohamy et al., 1996; Hamp-Lyons, 1997).
- Washback is seen as a potential instrument for educational reform (e.g. Pearson, 1988; Shohamy, 1992).
- Washback can have an influence on various aspects, e.g. teaching and learning (Buck, 1988; Alderson & Wall, 1993; Prodromou, 1995), teachers and learners (Alderson & Wall, 1993; Bailey, 1999).
- Washback can make teachers and learners do things '*they would not necessarily otherwise do because of the test*' (e.g. Alderson & Wall, 1993; Messick, 1996).
 - ***The direction of washback***
 - Washback is seen as being potentially positive (beneficial), negative (harmful) or neutral (e.g. Buck, 1988; Heaton, 1990; Bachman & Palmer, 1996; Bailey, 1996; Messick, 1996; Shohamy et al., 1996; Davies et al., 1999).
 - Washback can be intended and unintended (e.g. Andrews, 2004; Qi, 2005).
 - There is a direct and linear relationship between the stakes of a test and the strength of washback: the higher the stakes, the stronger the washback (e.g. Alderson & Wall, 1993; Alderson & Hamp-Lyons, 1996; Shohamy et al., 1996).

'Washback' and 'Impact'

Language testers consider '*washback*' as one dimension of '*impact*'. The latter is used to describe effects on the wider educational context. For example, Wall (1997), who discusses in detail the relationship between '*impact*' and '*washback*', suggests that '*washback*' is '*frequently used to refer to the effects of tests on teaching and learning*' whereas '*impact*' refers to '*any of the effects that tests may have on individuals, policies or practices, within the classroom, the school, the educational system, or society as a whole*' (ibid: 291).

A number of authors support Wall's view that 'washback' should be seen as a form of 'impact'. For example, McNamara (1996; 2000) and Shohamy (2001) place 'washback' within the scope of 'impact'. Hamp-Lyons (2000) helpfully suggests that the term washback refers to '*influences on teaching, teachers, and learning (including curriculum and materials)*' whereas the '*wider influences of tests*' are '*codified under the term 'impact'*' (ibid: 586). She also suggests that

We must see washback as one form of impact (as suggested by Wall: 1996), and impact as pervading every aspect of our instruments and scoring procedures (Hamp-Lyons, 1997:299)

Bachman and Palmer (1996) also '*feel that washback can be best considered within the scope of impact*' (ibid: 30). The writers refer to issues of test use and social impact as '*macro*' issues of impact, while washback is seen to take place at the '*micro*' level of participants, mainly learners and teachers (see also Bachman, 1990).

Washback and validity

An important feature of washback that merits specific consideration is its relationship to test validity.

The precise nature of the relationship between washback and validity has been debated. On the one hand there are authors (e.g. Morrow, 1986; Frederiksen & Collins, 1989; Weir, 1990; Shohamy et al., 1996) who support Messick's views (1989; 1996) that the effect of a test on teaching and learning is a major aspect of its validity. More specifically, Messick locates washback within the theoretical notion of consequential validity in which the social consequences of testing are seen as part of a broader, unified concept of test validity:

In the context of unified validity, evidence of washback is an instance of the consequential aspect of construct validity, which is only one of six important aspects or forms of evidence contributing to the validity of language test interpretation and use (1996:254-255)

Messick sees that the consequential aspect of validity:

includes evidence and rationale for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning (1996: 251)

Conversely, proponents of the opposite view argue that '*validity is not a property of the test or assessment as such, but rather the meaning of the test scores*' (Ferman, 2004:245) and, therefore, have difficulty in seeing a direct connection between washback and validity. They argue that how test scores are interpreted seems quite different and actually quite removed, for instance, from how teachers teach or how learners learn before an examination takes place.

Alderson and Wall (1993), in favour of this argument, point out that:

Whereas validity is a property of a test, in relation to its use, we argue that washback, if it exists - which has yet to be established - is likely to be a complex phenomenon which cannot be related directly to a test's validity (ibid: 116)

The authors argue that there might be other factors at work, in addition to the design of a test that could influence the presence or absence of washback. For example, there might be many consequences which can be caused by misuses of a test or by a teacher's linguistic ability, training, motivation or course hours, class size, extra lessons and so on which are beyond the control of the test. Therefore, Alderson and Wall argue, washback should not be considered a standard for judging the validity of a test. Davies (1997:335) shares the same view stressing that '*the apparent open-ended offer of consequential validity goes too far. I maintain that it is not possible for a tester as a member of a profession to take account of all possible social consequences*'.

However, Messick (1996) is not completely opposed to Alderson's and Wall's argument about the need to differentiate between washback effects and other effects.

He recognises that teaching is a complex activity and it is likely to be influenced by various effects arguing that to analyse washback effects is likely to be complicated as it requires the separation of washback effects from other effects operating in the educational environment. Messick also stresses that for washback effects to be seen as part of a test's validity, evidence of washback needs to be linked directly to the test:

washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test and not of other forces operative on the educational scene (1996:242)

In this way, he suggests, only effects that can be clearly and directly related to the test should be regarded as washback. However, he cautions that evidence of washback should not be confused with evidence of poor teaching. He notes the need to separate:

...test-linked positive 'washback' from good teaching regardless of the quality of the test and negative 'washback' from poor teaching (Messick, 1996:243)

Nevertheless, providing evidential grounds for claiming that washback is present in a given context is a challenge for any study of washback, particularly when this is mediated by other factors such as exam-oriented materials or the teacher. Evidence from washback studies shows that it is difficult to separate other influences on teaching and not to mistake them for washback (see Gosa, 2004; Wall & Horak, 2006).

Finally, Messick's advice to language testers to accomplish positive washback is '*... rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback*' (Messick, 1996: 252) by enhancing test tasks and content so that they adequately represent test constructs. However, it could be counter-argued that a test might not be valid and might still cause effects that are related to it. Under the circumstances, it seems to be problematic to treat washback as a form of validity.

In conclusion, despite the conflicting views between the two schools of thought, there appears to be very little disagreement over the necessity of investigating effects and consequences of tests (Alderson & Wall, 1993; Alderson, 1995; Messick, 1996).

Theoretical models of washback

It is probably true to say that the start of what is now termed ‘washback research’ can be traced to a major landmark in the field at the beginning in the early 1990s: the publication of the seminal paper by Alderson and Wall (1993) with the title ‘*Does Washback Exist?*’. The article, considered a classic, has greatly influenced all major recent research reports and literature reviews in the area of washback in language testing. The reason for this is that the writers were the first to look critically at the notion of test ‘*washback*’, and the way in which washback had been treated as a ‘*unique hypothesis*’, e.g. a ‘poor’ test would have negative washback while a ‘good’ test would have positive influence (Morrow, 1986; Hughes, 1988; Pearson, 1988; Khaniya, 1990).

The writers posited fifteen possible washback hypotheses relating to various behaviours, attitudes, test consequences, and the different effects on different persons:

- 1) A test will influence teaching
- 2) A test will influence learning
- 3) A test will influence what teachers teach
- 4) A test will influence how teachers teach
- 5) A test will influence what learners learn
- 6) A test will influence how learners learn
- 7) A test will influence the rate and sequence of teaching
- 8) A test will influence the rate and sequence of learning
- 9) A test will influence the degree and depth of teaching
- 10) A test will influence the degree and depth of learning
- 11) A test will influence attitudes to content, method, etc. of teaching/learning
- 12) Tests that have important consequences will have washback
- 13) Tests that do not have important consequences will have no washback
- 14) Tests will have washback on all learners and teachers
- 15) Tests will have washback effects for some teachers and some learners, but not for others. (1993:120-121)

However, Alderson and Wall's intention was not to argue for or against any one of these hypotheses, but to '*lay out the territory*' so that they could inform the questions that studies of washback might address in the future.

Other than working towards a re-conceptualisation of washback, the article is important in other ways too. First of all, it discussed the possibility that there might be factors other than test design which needed to be taken into consideration in attempts to investigate how washback operates, e.g. teacher lack of understanding of materials, inadequate training opportunities, school management problems, difficulties in resourcing, etc. It also considered the methodology that should be used when investigating washback. The writers advocated '*a more ethnographic approach to the topic than has been common to date*' (ibid: 127) for obtaining accounts of participants' perspective of the phenomenon e.g. including direct classroom observation and other methods of data collection for triangulation reasons. In addition, it discussed the importance of accounting for what occurs in the classroom, rather than just describing it. It also argued that researchers should specify the kinds of features that they wish to look for when deciding whether the desired washback has occurred. Finally, it concluded by stressing that researchers who wish to research washback need to take account of the educational context and the nature of the test as well as research literature in at least two related areas: motivation and performance and innovation and change in educational settings.

A few years later, Alderson and Hamp-Lyons (1996) revisited and refined the Washback Hypotheses in Alderson and Wall (1993), as follows:

<p>Tests will have different amounts and types of washback on some teachers and learners than on other teachers and learners. The amount and type of washback will vary according to</p> <ol style="list-style-type: none">(1) the status of the test (the level of the stakes);(2) the extent to which the test is counter to current practice;

(3) the extent to which teachers and textbook writers think about appropriate methods for test preparation ... ; and
(4) the extent to which teachers and textbook writers are willing and able to innovate ... (Alderson & Hamp-Lyons, 1996: 296)

Hughes (1994), on the other hand, questioned Alderson and Wall's views and pointed out that it was necessary to be more precise about what constituted washback. He suggested that a general definition – '*the test's effect on any aspect of teaching and learning*' (ibid: 1, underlined in the original) – was adequate for most purposes and introduced his own way of categorising the types of effects that might occur. More specifically, he made a distinction between washback on three constituents: the '*participants*', the '*processes*' and the '*products*' of an educational system.

According to Hughes, '*participants*' are classroom teachers and students, educational administrators, textbook developers and publishers '*all of whose perceptions and attitudes towards their work may be affected by a test*' (ibid: 2). '*Process*' refers to '*any actions taken by the participants which may contribute to the process of learning*' (ibid), such as materials development, syllabus design, changes in teaching methodology, the use of test-taking strategies, etc. Finally, '*product*' refers to '*what is learned and the quality of the learning*' (ibid).

However, Hughes advised that at least five conditions have to be met before all of the possible washback effects can occur:

- Success on the test must be important to the learners,
- Teachers must want their learners to succeed,
- Participants must be familiar with the test '*and understand the implications of its nature and content*',
- Participants must have the expertise which is demanded by the test (including teaching methods, syllabus design and materials writing expertise), and
- The necessary resources for successful test preparation must be available (ibid: 2-3)

Combining reviews of the literature as well as Alderson and Wall's (1993) Washback Hypotheses and Hughes' (1994) distinction between participants, process and products, Bailey (1996:264) put forth her own '*basic model of washback*' (see Figure 1 below)

Bailey specified a number of different participants, including researchers, and the types of products that might be affected by an examination. She also illustrated how these products might affect other products as well, e.g. research results can feed into materials, curriculum design and teaching. She then suggested a distinction between '*washback to the learners*', which is the result of supplying '*test-derived information*' to the test-takers, and '*washback to the programme*', which is the result of supplying information to all of the other participants in the education system.

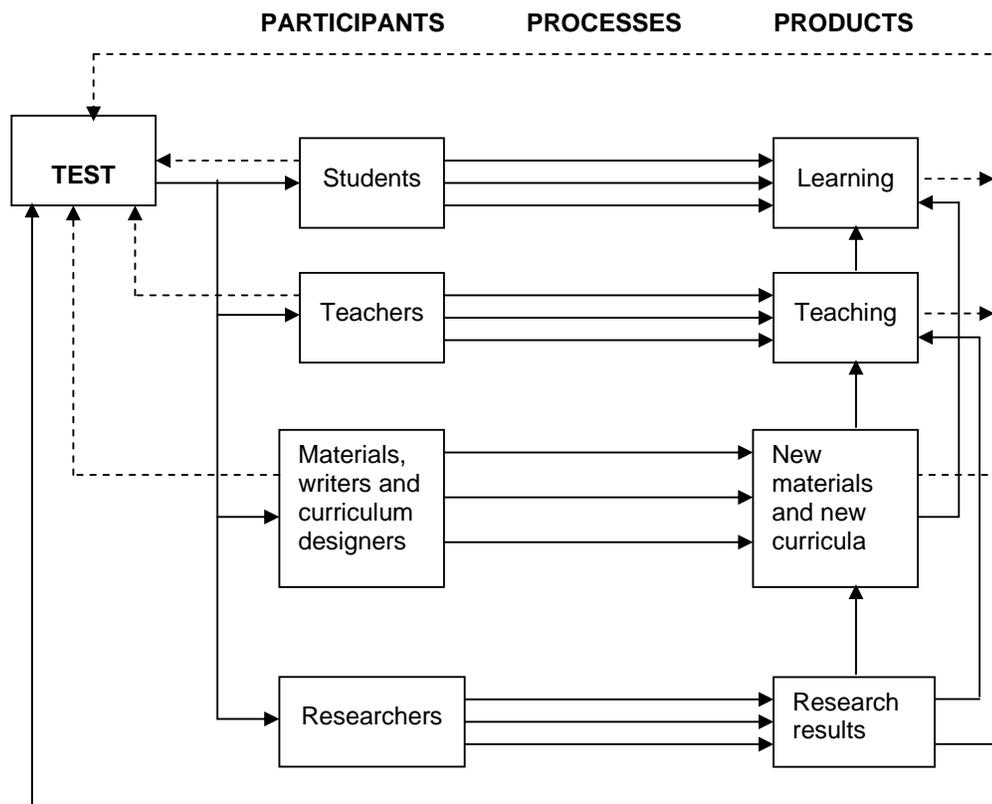


Figure 1 A basic model of washback

She further suggested that five of the Alderson and Wall hypotheses (2, 5, 6, 8 and 10) fit under the '*washback to the learners*' heading and provided ten examples of the processes that learners might engage in when preparing for important tests. These range from practising items similar in format to those in the test, to practising test-taking strategies, to enrolling in test-preparation courses and to skipping language classes to study for the test (ibid: 264-265). She also stated that six of the hypotheses (1, 3, 4, 7, 9 and 11) fit under the '*washback to the programme*' heading; however, she did not specify what kinds of processes the participants (e.g. the teachers) might participate in. She only stated that there is room here for future research.

Bailey also discussed the difficulties of investigating washback, which include working in '*naturally occurring settings*', using a '*non-random sample of subjects*', employing classroom observation and triangulation and collecting baseline data before the introduction of new tests. Like Alderson and Wall (1993) and Messick (1996), she also drew attention to the fact that attempting to sort out how much of what happens in classrooms needs to be evidentially linked to the introduction and use of the test if this is to count as washback.

Research studies

The following sections look at findings from empirical research studies and summarise empirical research into washback of both language and general education.

The findings are organised under Hughes' (1994) headings of '*process*', '*product*', and '*participants*', with '*process*' further broken down into '*content*' (curriculum and teaching materials), '*methodology*', and '*classroom assessment*'. In addition, these are sub-divided into several themes followed by short summaries which highlight the main findings and point out where further research is needed.

For ease of reference, the following table provides background information for the most frequently quoted research papers in terms of the educational context, exam type and research methods used.

Table 1 Overview of the research literature

Authors	Context	Exam	Methods
Alderson and Hamp-Lyons (1996)	USA	TOEFL exam	<ul style="list-style-type: none"> • Individual and group teacher and student interviews • Observations • Fieldnotes
Andrews (2002)	Hong Kong	Oral component of the Revised Use of English (RUE)	<ul style="list-style-type: none"> • A neutral testing instrument • Videotapes of simulated oral tests
Andrews (1994a,b)	Hong Kong	Oral Component of the Revised Use of English (RUE)	<ul style="list-style-type: none"> • Two parallel questionnaires to the working party members and teachers
Cheng (1997a,b)	Hong Kong	Revised Hong Kong Certificate of Education Examination (HKCEE)	<ul style="list-style-type: none"> • Questionnaires to teachers and students • Observations • Interviews
Cheng (1998)	Hong Kong	Revised Hong Kong Certificate of Education Examination (HKCEE)	<ul style="list-style-type: none"> • Student questionnaires
Cheng (1999)	Hong Kong	Revised Hong Kong Certificate of Education Examination (HKCEE)	<ul style="list-style-type: none"> • Observations
Ferman (2004)	Israel	National EFL oral matriculation test	<ul style="list-style-type: none"> • Students' structured questionnaire • Document analyses
Gosa (2004)	Rumania	English component of the Romanian school-leaving exam (<i>Bac</i>)	<ul style="list-style-type: none"> • Student diaries (10 students, retrospective use)
Glover (2006)	Hungary	English language section of the Hungarian school-leaving examination (<i>érettség</i>)	<ul style="list-style-type: none"> • Observations (2 teachers, 12 lessons) • Field notes, teacher interviews and background information about the context (supportive data)
Hawkey (2004)	UK	CPE (Cambridge ESOL)	<ul style="list-style-type: none"> • Textbook Analysis (revised IATM)
Hughes (1988)	Turkey	University entrance test	<ul style="list-style-type: none"> • Test scores • Questionnaire to lecturers
Kiss-Gulyas (1999)	Hungary	New School-leaving exam	<ul style="list-style-type: none"> • Teacher interviews
Lam (1993,1994)	Hong Kong	New Use of English (NUE) (end of secondary school)	<ul style="list-style-type: none"> • Questionnaire to teachers • Textbook analysis • Analysis of test scripts and scores
Li (1990)	China	Matriculation English Test (MET)	<ul style="list-style-type: none"> • Questionnaires to teachers and local officers (and student 'discussions')
Lumley and Stoneman (2002)	Hong Kong	Graduate Student Language Proficiency Assessment (GSLPA) – tertiary level	<ul style="list-style-type: none"> • Questionnaire • Interviews
Nikolov (1999)	Hungary	Secondary School leaving exam	<ul style="list-style-type: none"> • Structured interviews with teachers • Observations
Qi (2004,2005)	China	National Matriculation English Test (NMET)	<ul style="list-style-type: none"> • Interviews and questionnaires with NMET constructors, inspectors, teachers and students • Observations
Read and Hayes (2003)	New Zealand	IELTS	<ul style="list-style-type: none"> • Interviews • Questionnaires • Observations • Pre- and post-English tests
Saif (2006)	Canada	<ul style="list-style-type: none"> • Institutional needs-based spoken language test designed 	<ul style="list-style-type: none"> • Teachers' and students' interviews • Observations and follow-up

		for International Teaching Assistants (ITA test)	interviews • Analysis of test scores
Shohamy (1993)	Israel	<ul style="list-style-type: none"> • Arabic as a Foreign language test (ASL), • English Foreign Language Oral test (EFL), • L1 Reading test 	<ul style="list-style-type: none"> • Student questionnaire • Classroom observations • Interviews • Analysis of documents
Shohamy et al. (1996)	Israel	<ul style="list-style-type: none"> • Arabic as a Foreign language test (ASL), • English Foreign Language Oral test (EFL) 	<ul style="list-style-type: none"> • Questionnaires to students • Individual interviews with teachers and inspectors • Materials and document analysis
Stecher et al. (2004)	USA	Washington Assessment of Student Learning (WASL tests)	<ul style="list-style-type: none"> • Principal and teacher questionnaires
Wall (1999, 2005)	Sri Lanka	O-Level Examination in English	<ul style="list-style-type: none"> • Individual and group interviews with teachers • Questionnaires to teachers and teacher advisers • Materials and test analysis • Observations
Wall and Alderson (1993)	Sri Lanka	O-level, English as an International Language (at the end of the 11th year of education)	<ul style="list-style-type: none"> • Questionnaires to teachers and teacher advisers • Observations • Follow-up interviews with teachers • Materials and test analysis
Wall and Horak (2006)	Central and Eastern Europe	TOEFL	<ul style="list-style-type: none"> • Teacher interviews • Classroom observations
Watanabe (1992)	Japan	University entrance exams	<ul style="list-style-type: none"> • The Strategy Inventory of Language Learning (SILL)
Watanabe (1996,1997)	Japan	University entrance exams	<ul style="list-style-type: none"> • Questionnaires • Interviews with students and teachers • Observations
Watanabe (2001)	Japan	University entrance exams	<ul style="list-style-type: none"> • Student interviews
Wesdorp (1982,1983)	The Netherlands	Multiple-choice language achievement and final exams in Dutch secondary schools	<ul style="list-style-type: none"> • Scores on essay tests • Analysis of tests • Teacher and students' questionnaires

Process – washback on content

Curriculum

Reports of the washback effects on teaching and learning curricula were contradictory indicating that washback operates in different ways in different situations, and that in some situations it may not operate at all.

Studies that found washback on the curriculum

The following studies found washback on curricula from new and revised exams. For example, in Li's study (1990), teachers and officers reporting on changes brought about after the introduction of the MET test, indicated that, along with

traditional language skills (e.g. phonetics, grammar and vocabulary), increased attention was paid to *'the practice skills: reading, listening, writing and speaking'* which were tested in the exam (ibid: 400). Reading received the most attention because it had the heaviest weighting on the MET. Li saw this as a sign of positive washback: *'underneath the changes recorded by the survey a tendency can already be seen - a shift from formal linguistic knowledge to practice and use of the language'* (ibid: 402).

Alderson and Wall (1993) also concluded that the new O-Level exam introduced in Sri Lanka *'has had a demonstrable effect on the content of language lessons'* (ibid: 126-127). Through classroom observations, the researchers found that teachers spent more time on writing and reading skills, which were the skills tested in the exam. However, the researchers saw this as negative washback because teachers, after the introduction of the exam, did not pay equal attention to listening and speaking (Wall & Alderson, 1993: 66-67) (for similar findings see also Wall, 1999, 2005).

Similarly, Lam (1993; 1994) reported an emphasis in teaching those parts or subsections of the exam carrying the most marks. He also found evidence of washback on the number of periods devoted to exam preparation within the school timetable. The researcher concluded that such effects can be both positive (e.g. when teachers use authentic materials) and negative (e.g. when they use class time to teach the objective parts of the test). Cheng (1997b; 2005) also noted that the content of teaching had changed after the introduction of the revised HKCEE exam in public schools in Hong Kong. Classroom observations and teacher questionnaires indicated that reading aloud, for instance, was replaced by role-play and group discussion activities which reflected the new exam content.

Concern about curriculum-narrowing is also mentioned in a study of ‘*assessment-driven reform*’ conducted by Stecher et al. (2004). Teachers reported changes in their allocation of time and emphasis placed on writing leading to positive effects, e.g. the replacement of ‘*multiple-choice tests with more performance-based assessment has led to a dramatic increase in the amount of writing students do in school*’ (ibid: 68-69). However, there were concerns about curriculum narrowing. The principal and teacher surveys indicated narrowing of the writing curriculum by focusing only on the writing genres tested in the WASL tests.

Examination effects on lesson content were also reported for exams that have been in operation for a longer period of time. Nikolov (1999) observed 118 lessons in secondary schools in Hungary and found that the most frequent task types in the lessons were ‘*question–answer, translation, reading aloud and grammar exercises in the form of substitution drills*’ (ibid: 243). She concluded that these practices were typical language examination techniques of the school-leaving exam in operation at the time and ‘*therefore, they indicate a washback effect*’. (ibid: 233)

Shohamy et al. (1996) presented a slightly different picture. Based on questionnaires and interviews, the researchers reported that the low-stakes, Arabic exam had little effect on the content of teaching whereas the high-stakes EFL exam had greater effect on the curriculum. With regard to the latter, teachers claimed that they focused their teaching exclusively on the oral skills and activities to be tested in the exam (e.g. interviewing, asking questions, engaging in debates, speeches, etc) and found that more curriculum time was given to exam preparation.

However, variation was evidenced in the next two studies. For example, Alderson and Hamp-Lyons (1996), via classroom observations, looked at the time spent on different activities and found that TOEFL did exert an influence on the

content of the lessons but this varied between the two teachers observed. However, variation was also noted in terms of time devoted to exam classes, e.g. some institutions offered extra time to TOEFL classes while others did not. The researchers also discussed class size, pointing out that there were many more students in TOEFL classes than in 'regular' classes.

Variation was also found by Read and Hayes (2003) whose results indicated that washback on the curriculum depended on two courses observed: Course A, a short intensive IELTS preparation course and Course B, an extensive one, focusing on general and academic English skills as well as familiarization with IELTS. For instance, on Course A, more time was spent on procedural matters, on aspects of language compared and on giving students tasks under test conditions. On Course B, the different language skills were addressed in a more balanced way and greater use was made of integrated skills work. Read and Hayes' study also noted that time allocated to exam preparation may be greater or lesser depending on the school.

In the general education field, research, based on student surveys, reported by Paris et al. (1991), revealed that high-stakes standardised achievement tests in the US encouraged teachers and schools to spend valuable class time on what was measured at the expense of other school outcomes that were not tested.

Based on interviews with head teachers, staff, classroom observations and, occasionally, informal talks with pupils in an African country, Dockrell (1991) found that the effect of the high-stakes primary school-leaving examination under investigation was '*by common consent disastrous*' (ibid: 45). Primary heads stated that for the last two years of primary schooling and, in some cases even longer, the schools concentrated exclusively on preparation for the examination narrowing the curriculum

to the two subjects tested, e.g. Arithmetic and English. As for the content of their lessons, this was reduced to drilling items from previous test papers.

Hargreaves (1997), based on classroom observations and interviews with teachers, students and other stakeholders, generated substantial evidence of the domination of the secondary leaving examination in Egypt (the thanaawiya aama) on curriculum and pedagogy which produced a focus on examination subjects, pressure at home and extensive use of private tuition. The researcher stressed that the curriculum, the textbook and the examination syllabus became one. In effect tertiary level learning and teaching were negatively affected, too. Hargreaves commented that the examination system was so deeply entrenched, that the Ministry of Education had difficulty in reforming it to promote the goals of learning despite coordinated attempts. Hargreaves proposed that a much more fundamental reform of society is necessary than mere reform of the assessment system if the quality of education is to be improved (for similar results see also Morrison & Tang, 2002).

Studies that found no washback on the curriculum

Wesdorp (1982) investigated whether the introduction of new multiple-choice language tests would lead to an impoverishment of the curriculum, resulting, in particular, in neglect of the writing skill and a decline in reading comprehension. The questionnaire data gathered showed no evidence of curriculum narrowing through the use of multiple-choice testing.

Watanabe's (1997; 2000) findings are somewhat different. He speaks of teachers not necessarily teaching listening or writing even though the target university exam contained these skills. The researcher attributed this lack of positive and negative washback to lack of test specifications, teachers' guilt feelings about test coaching,

teachers' beliefs about the best teaching method, school atmosphere and cultural tradition, e.g. students as passive listeners in the exam classes (2000:45).

Summary

The review of the literature of washback on the curriculum revealed that

- Research reports have come up with conflicting results. There are studies that have found washback on the curriculum and studies that have not or found that washback on the curriculum varied.
- Research findings are also inconclusive as to whether high-stakes exams create narrowing of the curriculum. In some studies, teaching exam content presented a shift from teaching formal aspects of language to using performance-based activities while in others this led to a narrowing of the range of language skills taught to only those that were tested. In other cases the effect was not so straightforward as it led to both positive and negative effects (e.g. Lam, 1993, 1994; Stecher et al., 2004).
- The studies examined a variety of aspects: exam skills taught, activities, classroom events such as turn-taking, amount of metalanguage, laughter, etc, as well as time devoted to test preparation and class size.
- A range of research methods was also used: e.g. classroom observation, teachers' and other stakeholders' questionnaires and interviews. However, no major differences were observed in the results across methods or types of exams (old and new).
- In conclusion, it can be said that further research is needed in the area to provide a clearer picture of the effects of high-stakes exams on curricula and resolve the differences observed.

Teaching materials

In the available literature, the impact of high-stakes tests on teaching materials, known as '*textbook washback*' (Lam, 1993), is seen to invariably lead to the publication of exam-oriented materials designed explicitly to cater for the needs of students (and their teachers) preparing for such tests (e.g. Pierce, 1992; Alderson & Hamp-Lyons, 1996; Read & Hayes, 2003; Wall & Horak, 2006). The effect is noticeable in the context of the implementation of a new examination or a change in an existing exam (Li, 1990; Raimes, 1990; Fullilove, 1992; Lam, 1993; Shohamy, 1993; Andrews, 1994a, 1994b; Shohamy et al., 1996; Cheng, 1997b; Qi, 2004).

The issue of textbook washback has been addressed through a small number of empirical research studies. The research studies available discuss washback on materials in terms of their content, classroom use, and their effect on test performance.

Research on the content of exam-preparation materials

The general assumption among writers in the field had been that high-stakes tests have a direct impact on the content of teaching materials which they saw as evidence of washback (e.g. Johnson & Wong, 1981; Fullilove, 1992; Pierce, 1992; Shohamy, 1993; Shohamy et al., 1996).

Initial attempts to analyse the content of exam-oriented materials have supported textbook washback on teaching materials. For example, in the Hong Kong context, Lam (1993) examined the content of two sets of textbooks (18 texts in total) designed for a public proficiency exam (NUE) using a specially-designed grid. Samples of textbooks were analysed for their range of skills, authenticity of materials and task types contained. Lam concluded that most of the textbooks analysed '*are just exam crammers with lots of exercises following the exam format published by the HKEA*' (ibid: 86). Watanabe (1996) in Japan, also analysed teaching materials used to prepare

students for university entrance examinations. The materials analysed '*consisted of past exam papers and materials which were constructed by the instructors ... on the model of past exam papers*' (ibid: 325). Analysis of the content of these materials was based on calculation of the percentage of task types reflected in the exam. The results showed that '*washback did exist on materials*' (ibid: 326).

However, the following recent research studies suggest that exam-preparation materials can have an uneven relationship to the exam they are preparing for. For example, Hilke and Wadden (1997) aimed at finding out how accurately 10 TOEFL test-preparation textbooks widely used in Japan reflected the exam. The researchers looked at the ratio of question types, topic areas and their placement within the exam textbooks based on taxonomies derived from recent versions of the test. The results of their analysis indicated that the textbooks were indeed influenced by the exam but they varied '*...considerably in the accuracy with which they represent the exam*' (ibid: 28) with some textbooks reflecting the exam requirements more faithfully than others.

Hamp-Lyons (1996; 1998) also carried out a small-scale study by looking at the content of five TOEFL test-preparation textbooks. To analyse the materials, the researcher designed a framework of about 19 criteria (based on Mehrens & Kaminski, 1989; Popham, 1991; Alderson & Hamp-Lyons, 1996) used impressionistically. The results of the study revealed that '*the skills promoted by the textbooks generally consist of (a) test-taking strategies and (b) mastery of language structures, lexis and discourse semantics that have been observed on previous TOEFLs*' (Hamp-Lyons, 1996:6). The researcher observed that the materials would have negative washback on teaching and learning as teachers and learners would find themselves teaching and learning discrete chunks of language rules and vocabulary items without context or

even much co-text and might lead to ‘*curricular alignment*’ (Madaus, 1988; Cooley, 1991). However, the researcher stressed that there were several omissions in the textbooks, e.g. lack of guidance for the teachers, preceding material to teach the point tested, help to students and teachers after taking a practice test, etc.

Other than TOEFL, teaching materials that were used to prepare for the IELTS exam were also analysed with equally interesting results. For example, Wang (1997) used a specially-designed instrument, the Instrument of Analysis of Textbook Materials (IATM), designed by Bonkowski (1996), to analyse sample units of two IELTS exam-preparation textbooks in detail. The results indicated that the IELTS test did have an influence on the content and format of the preparation textbooks in terms of content, e.g. skills, item types and item content but had little impact on textbook methodology. Wang also noted that the textbooks demonstrated certain omissions, too, e.g. the diagnostic aspect of the test (scoring profiles) was not well-reflected in the textbooks. Wang saw that such an omission was a sign of negative washback because students would find it impossible to self-score and get an IELTS equivalent band score on the subtests or on the whole test. Consequently, students would not be in a position ‘*to monitor their own progress and where to put more effort when using these textbooks*’ (ibid: 44-45).

The analysis also revealed that the extent to which the IELTS test influenced the content of preparation textbooks varied from book to book and from aspect to aspect.

Wang explained:

... differences between textbooks play an important role in determining the extent and quality of that influence. Textbooks vary in different aspects in their reflection of the test and are not a 100% accurate reflection of the specifications (ibid: 50)

(see also Smith, 2004; Hawkey, 2006 for analysing IELTS textbooks).

Hawkey (2004a; 2004b) and Hawkey and Turner (forthcoming) used a revised version of IATM (based on an impressionistic approach to materials evaluation than detail analysis, see Saville & Hawkey, 2004) to examine 10 CPE (Cambridge Proficiency in English - Cambridge ESOL) textbooks. The results, based on the independent ratings of two evaluators (Cambridge ESOL staff specialists), showed that the exam exerted strong washback on the evaluated textbooks in their treatment of language skills, micro-skills, task types, language elements and topics. However, the researchers stressed that certain aspects of the exam were omitted or received insufficient coverage in the books, e.g. lack of pronunciation practice, limited work on the marking criteria and on how to approach the speaking test.

Research on classroom use of exam-preparation materials

Lam (1994), based on teachers' questionnaire, speaks of Hong Kong RUE teachers as '*textbook slaves*' and '*exam slaves*' with large numbers of the former relying heavily on the exam textbooks and of the latter relying even more heavily on past papers in exam classes rather than using materials that aim '*at maximising students' language learning*' (ibid: 99). Lam saw teachers' adherence to exam-preparation materials as '*a sign of negative washback because instead of introducing more authentic materials, they prefer to use commercial textbooks*' (ibid: 90). He also reported that teachers do this as '*they believe the best way to prepare students for exams is by doing past papers*' (ibid: 91).

In the same context, Andrews' study (1994b), based on questionnaires to members of the exam working party and teachers, showed that teachers relied on exam textbooks for the content of their teaching for an estimated two thirds of class-time. The most popular materials were highly exam-specific, focusing in detail on the format of the Oral exam, and on the precise skills required. The researcher concluded

that this tendency is likely to limit the focus of teachers and learners rather than broaden their horizons (ibid: 80).

In Sri Lanka, Wall and Alderson (1993) also found, via classroom observations, that teachers, for a large part of their teaching, relied on textbooks for their content of teaching but not for their methods. Follow-up interviews with teachers revealed that this was mainly due to teachers' lack of understanding of the approach and philosophy embodied in the materials on which the new examination was based as well as their lack of awareness of the new exam.

Shohamy (1993) also found that in the three language tests she examined:

... testing materials and methods became an integral part of 'normal' teaching as many teaching activities became testlike, mostly as a result of the new textbooks, which were strongly influenced by the test (ibid: 15).

Much like Wall and Alderson, the researcher attributed this state of affairs to insufficient teacher training and inadequate understanding of the exam.

In their study of TOEFL test preparation classes, Alderson and Hamp-Lyons (1996) describe teachers' reliance on exam materials in preparing students for the exam. Based on observations of classes and interviews with both teachers and students, the researchers found that '*most teachers just seemed to do what the book says and what they claim the students want*' (ibid: 286). The researchers attributed teachers' reliance on exam materials to their negative attitude towards the exam which discouraged them from creating their own materials. They also stressed that familiarity with the exam or teacher experience was not a variable as many of the teachers, independently of their amount of experience of teaching towards the exam, made heavy use of exam materials.

Cheng (1997b), via teacher questionnaires and classroom observations, found that her teachers in Hong Kong relied on the exam textbooks, too. The teachers

followed the syllabus of the exam by adherence to the textbooks. The researcher concluded that this was an indication of obvious washback on the content of teaching. However, she suggested that the changes made were changes of *'form'* rather than of *'substance'*, and that teachers were more influenced by writers and publishers' understanding of the new exam as reflected in the textbooks than by their own.

Teachers' reliance on exam-preparation textbooks was also reported in the work of Read and Hayes (2003), who, based on teachers' questionnaires, also talked about the reliance of teachers on exam-preparation materials and stressed that in 90% of cases in their New Zealand IELTS study, exam preparation books were employed.

The above findings are similar to those of Wall and Horak (2006) who, through extensive interviews with teachers in Central and Eastern Europe, found that the TOEFL preparation textbooks were at the heart of the majority of the courses investigated. The textbooks were used as a source of information about the exam while TOEFL courses were *'shaped by the format and content of the coursebooks used'* (ibid: 74). The researchers also stressed that in many cases, the exam books acted as the syllabus for teaching, influencing the content and sequence of teaching. The researchers felt that this was a sign of negative washback:

The coursebooks that made up the syllabus of the courses seemed to be based more on the notion of an accumulation of language and practice rather than a progression from less to more difficult (ibid: 112)

Wall and Horak commented that the reasons why teachers adhered to their exam textbooks was because of student expectations, textbooks' adequate coverage of exam preparation, teacher training and lack of other resources.

In the literature reviewed, teachers seem to favour traditional methods of teaching despite the quality of the textbooks. Nikolov (1999) showed, through teachers' observations and reports of effects of an existing school-leaving examination

in Hungary, that although the majority of teachers ‘...used *British communicative coursebooks as core syllabuses, these materials were exploited traditionally and eclectically*’ (ibid: 243), using, that is ‘... *techniques of the grammar-translation and audio-lingual method*’ (ibid: 238)

Finally, unlike the above studies where teachers are portrayed as textbook-bound and traditionally-oriented in terms of methods, Saif (2006) reports that the teacher she observed modified and adapted the test materials in favour of the students’ needs and the test objectives. As Saif explains, the teacher

... did not go through the prescribed textbook chapter by chapter and paid less or no attention to the sections (like those discussing cultural topics) that did not practice the oral skills evaluated by the test. On the other hand, she routinely covered and expanded on selected exercises practicing common pronunciation problems, complex structures, organizational methods, and communication strategies (ibid: 28)

Use of other exam-oriented materials in class

Other than exam textbooks, supplementary materials, mostly geared towards exam requirements were often used by teachers in examination preparation to meet their needs as well as their students’. For example, Nikolov (1999) found that the supplementary materials used for exam preparation were mostly local publications ‘*focusing on grammar and exam preparation, representing an exam washback effect*’ (ibid: 243). Wall (1999) came up with similar findings. Her teachers were using supplementary material books to compensate for lack of grammar in the coursebook.

Watanabe (2000) also reported that his teachers used ‘*a variety of organisation patterns and self-made materials (e.g. revised listening tasks taken from past exam papers, hand-outs explaining cultural background to English words)*’ (ibid: 44).

Use of non-exam oriented materials in class

Lam (1993; 1994) reported some innovative use of materials generated by the introduction of the RUE exam, e.g. the use of teacher-produced and authentic

materials based on mass media by a small number of teachers. Lam considered this practice a sign of positive washback. Andrews (1994b) also found that a small percentage of the teachers (e.g. 16.7%) preparing for the oral component of the RUE made '*extensive use of their own materials*' (ibid: 78).

A teacher questionnaire item asking what material, other than a main textbook, teachers used for the preparation of IELTS in Hawkey's impact study (2006) revealed that teachers, other than test-related materials, also used additional materials targeting specific language skills and components from a variety of sources outside the exam, e.g. '*from the press, TV and radio, video and audio and the Internet, in-house or teachers' own materials*' (ibid: 109-110). This was also confirmed by classroom observations (ibid: 112). Hawkey reported the same teacher practices in the PL2000 impact study where teachers tended to use additional materials from a variety of sources beyond the coursebook, e.g. '*cut-out photographs, self-designed spider-grams, information-gap hand-outs, audio-cassettes, wall charts*' (ibid: 143).

Effect of exam textbooks on test performance

The only study that investigated the effectiveness of exam-preparation materials on learning was conducted by Andrews et al. (2002). The researchers examined the presence of functions and forms taught in the exam preparation textbooks in students' speech. Based on the frequencies of these language features and the contexts in which they appeared in the oral performance of two cohorts of students over two administration periods, Andrews et al. came to the conclusion that there was evidence of exam washback on student performance and that this appeared to be linked to the mediation of exam materials. However, the sort of washback observed was negative. As Andrews et al. explained, this was represented at

a very superficial level of learning outcome: familiarisation with the exam format, and the rote-learning of exam-specific strategies and

formulaic phrases ... the inappropriate use of such phrases by a number of the students ... seems indicative of memorisation rather than meaningful internalisation ... the students appear to have learnt which language features to use, but not when and how to use them appropriately (ibid: 220-221)

Summary

- Textbook washback has only recently been recognised as an important area of test washback with major effects on teaching and learning.
- However, research in the area has come up with varying results. There are studies that have found a direct match between the teaching materials and the exams they are preparing and studies where exam materials do not necessarily faithfully reflect the exam requirements, e.g. exams may have little impact on textbook methodology, exam influence varies from book to book while the materials fail to represent important exam features.
- To analyse exam materials, researchers have employed different methods ranging from taxonomies derived from exam papers to specially-designed checklists and grids. However, the criteria of analysis are either limited to a few features of the exam, are impressionistic or used on a small sample of the materials.
- Clearly further research of teaching material is needed that relates directly to exam specifications, examines teaching materials in-depth and looks at full sets of exam materials in order to clarify the nature of textbook washback. To gain a more detailed perspective of textbook washback, further research also needs to look into the features of general language materials and see in what ways they differ from exam materials.
- The study of how teachers use exam-preparation materials is a relatively unexplored area. Whatever studies have been conducted so far, show that when working towards exams, teachers use exam materials to different degrees. In some studies teachers tend to teach extensively with the exam textbook and accept it as the major and usually

only source of content and activities but not methods. Studies also show that teachers prefer traditional teaching methods irrespective of the quality of their materials while some other studies show that teachers do actually adapt and modify the content of their exam materials to the needs of their students and the exam. Finally, there are studies where teachers make use of non-exam materials, too, e.g. authentic materials taken from a variety of sources.

- Teachers' reliance on the exam materials is largely considered negative and believed to narrow the focus of teaching and learning.
- Studies on how teachers use exam materials are mainly based on indirect research methods, e.g. teacher questionnaires and interviews while, occasionally, researchers use classroom observations.
- Evidence of the effect of exam preparation materials on test performance is hard to come by. Whatever research there is, has pointed to negative influence (e.g. students used the language taught in the exam textbook inappropriately).
- Further research is needed to find out the exact ways with which teachers use exam materials and whether these can be directly related to exam requirements using refined methods of data collection over long periods of time.

Process – washback on teaching methods

In the following sub-sections, I review studies of washback on teachers' methodology. By 'methodology' I refer to 'how' teachers teach (e.g. use of techniques such as explaining, talking time, interaction, metalanguage, feedback, etc) rather than 'what' they teach or their attitudes about teaching and examinations (discussed in later sections).

Studies that found washback on how teachers teach

a. Using indirect research methods

Via teachers' questionnaires, Lam (1993) found that teachers used the new approach to teach for the NUE. He also observed that the teachers who had worked under the old and new systems were '*much more examination-oriented than their younger counterparts*' (ibid: 91). The latter were more likely to work with authentic materials and use activities which required student participation or an integration of the skills rather than isolated skills work. The researcher concluded that it is not sufficient to change exams to bring about the desired results:

The challenge is to change the teaching culture, to open teachers' eyes to the possibilities of exploiting the exam to achieve positive and worthwhile educational goals (ibid: 96)

Similarly, Shohamy (1993), summarizing the findings from all three language test studies she conducted, emphasized that teaching methods became 'test-like'. On revisiting the impact of the same tests a few years later, Shohamy et al. (1996) found that the low-stakes Arabic exam involved '*virtually no change from normal teaching*' (1996: 304), whereas teaching towards the high-stakes EFL exam continued to lead teachers to teach through simulating the exam tasks or through carrying out other activities that directly aimed at developing the exam skills and strategies. The researchers also noted that these activities became more prevalent as the exam dates got closer. The researchers also noted that there were significant differences between the experienced and novice teachers. The former '*turned to the test as their main source of guidance for teaching oral language and used only material to be included in the test*' (ibid: 301), while the latter used '*a variety of additional activities in the teaching of oral language*' (ibid: 301).

Finally, Stecher et al. (2004), using teachers' questionnaires, investigated the effect of the Washington Assessment of Student Learning tests (WASL) on the

methods teachers used to teach writing. The results showed that teachers changed their methods as a result of the examination incorporating the processes and strategies of the tests into their existing teaching practice.

b. Using direct research methods

The following research studies based mainly on classroom observations of teachers have found that tests affect different teachers in different ways.

Alderson and Hamp-Lyons (1996), using empirical data (classroom observations) from TOEFL and non-TOEFL classes, looked at a range of lesson features (e.g. teacher and student talking, number of turns, pair work, etc). The researchers found influence on how teachers teach, but note that *'the effect is not the same in degree or kind from teacher to teacher'* (ibid: 295).

In a completely different context (e.g. Japan) and using a similar research design, Watanabe (1996; 1997) obtained results similar to those found by Alderson and Hamp-Lyons. The researcher looked in detail at empirical classroom data to see how translation and grammatical explanation were used in two types of university exam preparation lessons. Even though he found evidence of washback on teaching, he also found differences between how teachers taught and used translation. Watanabe concluded that it is too simple to expect that an examination will affect all teachers in the same way. He considered that the personal characteristics of the teachers, e.g. educational background, beliefs about teaching and attitudes towards the exam, and, possibly, the proximity of the exam in terms of time have an important role to play in how teachers conduct their lessons.

Read and Hayes (2003) also found evidence of IELTS washback on how teachers teach (through observations) manifested in heavy use of practice tasks, homework, the nature of the feedback given, presence of laughter, explanation of test-

taking strategies and students' consideration of their own strategies. However, the researchers stress that these effects vary between teachers and schools and pointed that the nature of the course is a strong variable (see also Hayes & Read, 2004).

Burrows (2004), through classroom observations, looked at the application of a new classroom assessment system in Australia, and its relationship with methods, teacher discourse, explanations, instructions and interaction. As in the previous studies, her research revealed washback on some of the teachers but not on others.

The results obtained in Saif's study (2006) are somewhat different from those of previous studies in that the researcher found clearer connection between the test investigated and how her teacher taught. The writer concluded that '*... the teacher's methodology and the choice of class activities were, to a large part, adapted to the contents and goals of the test*' (ibid: 28). Saif attributes teacher's behaviour in class to the fact that

... in this particular context, the teacher's enhanced awareness of the test caused by her involvement in the test administration process, interaction with other raters, understanding of the rating process, and the ability components of the rating instrument were partially responsible for the changes she made to her teaching later during the program (ibid: 29)

Further exemplification of the range of ways in which teachers choose to teach towards an exam comes from the field of general education. Smith (1991a), based on interviews and classroom observations, reported the role of external testing in elementary schools in the USA. The researcher identified and categorized eight types of exam preparation practices operating in classes ranging from no special preparation to cheating. Although she watched subjects other than English language being taught, the categories she proposes may prove helpful in facilitating our understanding and awareness of the range of activities used to teach towards exams in language exam classrooms (ibid: 526-537).

Finally, Sturman (2003), looking at the effect of national tests in three core subjects (e.g. English, mathematics and science) in England, with data collected via teachers' questionnaire, found that for most teachers, preparation replaced other activities, e.g. test-taking skills were widely taught, and specific item types were commonly rehearsed. However, he found variation in the time spent on preparation, approaches to revision and resources used. Some outcomes were related to background variables of attainment, size of school, composition of class and respondents' professional responsibilities.

Studies that found no washback on how teachers teach

Several other studies concluded that there was no evidence of washback on how teachers teach.

a. Using indirect research methods

The study in the Netherlands investigated claims that multiple choice items limited the '*teaching practices*' (Wesdorp, 1982:47). Using questionnaires teachers were asked to assess the time spent each week on various types of teaching, activities, and language curriculum components. The investigation into methodology in schools with and without multiple-choice final tests failed to reveal any clear difference in the way teachers taught. The researcher concluded that no washback was present in aspects of how teachers teach, and concluded that

The so-called backwash effects are a myth. If they do exist, they must be so weak or small that our research methods cannot detect them (ibid: 130)

b. Using direct research methods

Wall and Alderson (1993), using classroom observations, found no evidence for any change in teachers' methodology before and after the introduction of the new English school-leaving examination in Sri Lanka despite teacher reports who believed that the examination influenced their methods (see also Wall, 1999, 2005). The

researchers concluded that there are other factors that prevent the implementation of the washback of an exam, e.g. resources, management practices, lack of communication between test designers and users and teachers' beliefs, training, commitment and other obligations.

Using similar methods, Cheng (1997a; 1999; 2005) reported that teachers only adapt their methodology slowly, reluctantly and with difficulty. Cheng compared classroom behaviour before and after the introduction of a revised examination and noted changes in teaching content as a result of the exam but found no change in teaching methods. Teachers made greater use of discussions and role-plays rather than reading aloud after the introduction of the revised exam but there was no significant change in the amount of teacher talk. The interaction patterns had not changed much either, and the lessons were overall conducted similarly before and after the introduction of the new examination syllabus.

Qi (2004; 2005) examined the reasons why the NMET test failed to bring about the intended washback in ELT in China. She found that the NMET had considerable impact on materials and learning activities but not on teaching methods as envisaged by its designers. She found that the reason for this was that the two functions of the test (e.g. selecting candidates for tertiary education and making changes in ELT) were in conflict with each other making the test ineffective for changing teaching and learning. Inadequate communication between test makers and test users was among the factors that hindered the operation of intended washback.

Finally, Glover (2006) sought to identify washback on how teachers teach by analysing teacher talk from both examination and non-examination classes in Hungary. He found only scant empirical evidence of washback on how teachers

teach. The researcher concluded that there were many differences between teachers and stressed that

how teachers teach in examination lessons may be different because the teachers have a different pedagogic purpose: teaching the examination, not the language. (ibid: 324).

From general education, of interest is the point raised by Hargreaves (1997) regarding the relationship between examinations and teaching practices. Based on classroom observations conducted in Egyptian secondary schools, she suggested that while the secondary leaving examination dominated teaching methods it did not dictate them. Reform on one was not dependent on the other. Other factors such as poor quality teacher education and a general lack of resources contributed to low quality pedagogy and assessment.

Similar conclusions were proposed by Chapman and Snyder (2000). Using classroom observation and interviews, they reported that teachers in Uganda could not adjust to the changes in the new national examination and did not change their instructional practices. The researchers put the blame on a lack of understanding of *'the intermediate conditions that had to be met for changes in test content, format, or use to have the desired impact on teachers' classroom practice'* (p. 457). In their model of testing and classroom practice they specified that the most difficult condition to meet was getting teachers to understand what changes were needed to raise student performance. The researchers stressed that success depended on

the government's political will in the face of potentially stiff opposition and the strategies used to help teachers make the transition to meet the new demands (ibid: 462)

Summary

- About two-thirds of the washback studies reviewed deal with the methodology that teachers use in the classroom. However, empirical evidence suggests that

washback on how teachers teach is unclear and complex. The studies follow a cline from indicating heavy washback to no washback.

- It is also interesting to note that the studies that found evidence of washback on teaching also found large differences in the way teachers teach towards the same exam, with some adopting much more overt ‘teaching to the test’, while others follow more creative and independent approaches.
- Researchers have looked at a variety of different aspects when investigating teachers’ ways of teaching. This could perhaps be one of the reasons that makes this aspect of washback so complex.
- In terms of research methodology, there are studies which used classroom observation, and studies which relied on interview reports or questionnaires. Comparing the two types of studies, there seems to be a tendency for claims about evidence of washback to be found in studies that used reports while those that used classroom observation highlight the variation of ways with which teachers teach.
- Overall, washback on teaching methods is not an inevitable or universal phenomenon. The studies reviewed demonstrate the need to be clear about which features of classroom behaviour to study that relate to how teachers teach.
- Research needs to employ additional ways of looking at the influence of tests on teachers’ methods that could also help explain why teacher differences occur.

Process – washback on classroom assessment

Wall and Alderson (1993) have in many respects set the standards for work in the area of impact on classroom assessment. Whilst various rounds of classroom observation produced the core data for their findings, they stressed the importance of complementary data from other sources, e.g. classroom tests prepared by teachers for

their classes. Without these, Alderson and Wall state, they would have missed out on a whole series of insights not obtainable via observation alone.

The need for examination of classroom assessment practices in studies of washback has also been stressed by Watanabe, e.g.

To observe the nature of washback from high-stake entrance exams, it would be valuable to compare this washback with the washback from smaller-scale tests (e.g. in-class tests, practice tests, placement tests) employed by the same population (2000:46)

Wall and Horak (2006) in their TOEFL Impact Study note:

Essential to any study of classroom practices are questions concerning assessment: what are the functions of assessment in a particular setting, what is assessed, and how does the assessment take place? (ibid: 72)

Nevertheless, despite the importance of studying the impact of high-stakes tests on classroom assessment, only a very small number of research papers have looked into it. However, even the ones that did so have come up with mixed findings.

Studies that found no washback on classroom assessment

Wesdorp (1982), using teachers' questionnaires, looked at the frequency of use of multiple-choice questions (which were the main task type of the tests introduced) in classroom tests across all languages taught in Dutch schools. The researcher found limited evidence of the effect and concluded that '*multiple-choice tests only have limited influence on classroom test format*' (ibid: 51).

Studies that found washback on classroom assessment

In their early reports on the Sri Lankan impact study, Alderson and Wall (1992) and Wall (1994) found that the new examination was having a strong impact on classroom test design, mainly in the emphasis given to exam skills (e.g. reading and writing), and in the abundant use of certain testing techniques associated with the new examination (e.g. short-answer questions, true/false, matching, etc). Wall and

Alderson found this practice positive but negative when certain types were over-used, and when passages and questions were copied straight from past papers.

In later studies, the researchers (Wall & Alderson, 1993; Wall, 1999; 2005) found that the exam had little effect on how teachers marked classroom tests because they were not familiar with the marking criteria used on the exam and had not received official exam-support materials which explained the marking system. The researchers also noted that the Teacher's Book was not helpful in this direction as it lacked the necessary advice.

Wall and Horak (2006) found that TOEFL preparation lead to the existence and administration of several types of classroom tests in the schools they visited, e.g. screening, diagnostic, and practice tests. The tests were either in the original or adapted format of commercial TOEFL tests, based on in-house TOEFL tests or taken from the TOEFL preparation coursebooks being used at the institutions observed.

Wall and Horak noted that the general assumption among teachers and students was that classroom tests were parallel in form to the real TOEFL test. However, the researchers questioned the validity of these saying that:

They may have served a useful function in terms of raising consciousness (or confidence building), but there is no evidence (other than student report) that the tests were truly parallel to the real TOEFL (ibid: 77)

Given that the tests were administered under a range of conditions unlike the TOEFL test, the researchers also expressed concern about their reliability:

The tests seemed to play a useful role as a means of test familiarization, especially if they were done on computers. We stress, though, that they can only give the students some indication of their level of preparedness for the TOEFL test rather than a true picture of their language proficiency (ibid: 112).

The researchers also looked at the use of the scales for marking the Writing section of the TOEFL test and reported that there was evidence that these were used

by some teachers, but in a variety of ways. Wall and Horak concluded that the teachers needed more guidance before they could competently and confidently use the scales with their students.

Summary

The review of the literature of washback on classroom assessment showed that

- Very little research has been conducted so far in the way high-stakes exams influence classroom assessment. Whatever research has been done has yielded contradictory results: some studies found washback on classroom assessment while others did not.
- Researchers stress the need for detailed analysis of the classroom tests and research into their reliability and validity.
- Further research is also needed that will clarify the ways in which it mediates the washback effect of high-stakes exams in the classroom.
- Finally, research studies need to take into account students' attitudes and feelings towards classroom assessment in high-stakes contexts as they are considered to be major participants in the washback process (e.g. Bailey, 1999).

4 Product – washback on student learning

We come now to another key question about washback: '*Does washback from exams have an effect on learning, and, if so, how?*'. Yet again, the review of the literature has come up with mixed results.

Studies that found washback on learning

Hughes (1988), following the introduction of a new university exam in a Turkish university, argues that students' performance increased. Evidence for this included data on students' performance on the Michigan Test and teacher's perceptions of the

gains of the first cohort of students to pass the new test (collected through a survey). Hughes attributed the test effect to the fact that the test was criterion-referenced and that it was based on the needs of the undergraduate students. However, Hughes' approach is not convincing. There is no discussion as to why the introduction of the new proficiency test is compared to students' results on a test with which it had no resemblance, neither is there enough information about the way teachers taught when preparing students for the new test.

Lam (1993) also argued that there is evidence that the NUE has brought about improvements in the actual language of the students. Through analysis of exam papers and exam scripts the researcher argued that the new exam was testing a wider range of skills as these were demonstrated by the new examination candidates, particularly in their performance on the Practical Skills for Work & Study subtest. Lam concluded that such improvements in student learning can be interpreted as positive washback.

Finally, the findings of Saif's study (2006), based on analysis of test scores collected at different intervals, e.g. before, during and after a language training program, suggested a positive relationship between preparation for the (ITA) test and learning outcomes.

However, as the researcher stressed, the results cannot be generalised to other settings beyond the context of her study simply because they were gathered at an institutional level.

Studies that found limited or no washback on learning

Wesdorp (1982) demonstrated, through global and analytic judgments of students' essays by 15 judges, that using the multiple-choice technique in the achievement tests administered in Dutch schools, did not lead to a decline in writing

abilities. However, the researcher also admitted that no improvement was observed either.

In Shohamy et al. (1996) teachers reported that the low stakes Arabic exam may have promoted learning at lower levels but not at upper levels as the students were committed to learning the subject anyway by that stage. In relation to the oral EFL exam, they believed that it had undoubtedly brought a focus on oral proficiency but the Reading component had not affected reading in class, as this part of the exam was considered to be poorly designed.

Cheng's Hong Kong study (1998), based on student questionnaires, came up with negative conclusions:

The washback effect of this exam seems to be limited in the sense that it does not appear to have a fundamental impact on students' learning. For example, students' perceptions of their motivation to learn English and their learning strategies remain largely unchanged (ibid: 297).

In order to measure students' oral performance on the RUE, Andrews et al. (2002) conducted simulated oral tests with three groups of candidates, matched for their ability over a three year period. The results showed that the nature of the washback varied across the student groups: only a small improvement in performance between the first and the third group was indicated, leading researchers to conclude that the washback effect of the test was delayed, e.g. it was '*more noticeable in the second year of the test than the first*' (ibid: 220) (the issue of time before washback takes effect is also discussed in Li, 1990; Cheng, 1997b, 2005).

Another study that attempted to measure learning outcomes is that of Read and Hayes (2003). The researchers had two small groups of students (total 17) take retired versions of the IELTS exam as pre- and post-tests to two IELTS courses (intensive and general). The results did not show any significant improvement overall (with the exception of the listening tests at one of the two schools investigated), nor between the

groups of students. As in previous studies, the researchers concluded that time is needed for washback to occur:

It is generally recognized that students need an intensive and usually extended period of study to achieve any substantial increase in their score on a proficiency test like IELTS (ibid: 110)

Summary

- The findings on washback studies focusing on student learning were disparate and too mixed to provide a definite conclusion.
- In fact, of the washback studies reviewed only one study has documented any demonstrable gains in student learning that can be tied to the use of a test (Saif, 2006). The remaining studies have either used dubious approaches, did not find any considerable gains or found negative results.
- It is not always clear what was meant by '*student learning*'. Some researchers looked at actual student performances, others at students' perceptions (of their motivation or learning strategies), and others at the comprehensiveness of skills tested in the exam.
- The majority of the research was based on small numbers of participants making it difficult for researchers to generalise their findings.
- Researchers have used a variety of methods to investigate the influence of exams on students learning, e.g. teacher and student questionnaires, examination of test papers, performance on external or retired versions of the target test or performance on the target test at different intervals of the exam preparation period.
- Other than the exam, the stakes of the language tested as well as the design of the test seem to be important variables influencing student learning.

- An important condition to achieve any substantial influence on students' learning is to allow for an extended period of time between the introduction of the test and the collection of data.
- To conclude, investigating washback on 'the product of learning' is still an area in need of research (stressed also in Wall, 2000; Alderson & Banerjee, 2001).
- For instance, further research is needed to clarify what exactly is meant by 'student learning'.
- Future research needs to collect empirical data, e.g. test results that can clearly show whether students have learned better due to their preparation for a particular test, rather than student or other stakeholder perceptions or simply an examination of exam papers.
- Finally, research needs to be conducted not only in contexts where a new exam has been introduced but where exams have been operating for an extensive period of time too.

Participants – washback on feelings and attitudes

Teachers

Many of the studies reviewed indicate that examinations have effects on teachers' attitudes and feelings ranging from positive to negative.

Feelings and attitudes towards newly-introduced or revised exams

Li (1990) found that the introduction of the MET made teachers feel uncomfortable when it was first introduced but a few years later, her survey, revealed that *'the overwhelming majority of the teachers had accepted these subtests along with the whole MET, admitting that the subtests were an effective measure of the*

candidates' ability to use English' (ibid: 402) (see also Lam, 1993 for similar reactions from teachers).

On the contrary, Shohamy (1993), summarizing the findings from all three language test studies she conducted in Israel, emphasized that because tests did not provide information that was useful for future teaching, teachers felt degraded as a result of having tests imposed upon them and having to accede to their demands. More specifically, the Arabic test created tension among teachers while for the L1 reading comprehension test, teachers felt stressed, angered and humiliated because they had not been consulted about the test in advance.

Both negative and positive attitudes were shown about the Israeli tests over time. In a later study, Shohamy et al. (1996) found that teachers still had negative feelings towards the Arabic exam and expressed complaints that the test was of no importance. On the other hand, teachers approved of the EFL exam in as much as they saw it as having brought about an acknowledgement of the importance of communicative oral skills that, they believed, would stand their students in good stead in the future. However, some teachers claimed that the exam forced them to teach in ways '*dictated by the examination*', and felt they could not teach creatively at times because they were '*pressured by the exam*' (ibid: 308-309). In addition, the exam was reported to generate anxiety among teachers because they felt '*pressure to cover the materials for the exam*' and because '*the success or failure of their students reflects on*' (1996:309-310). Others were more positive, appreciating the EFL oral test, as without it '*there would be no motivation to teach oral proficiency*' (ibid: 308-309) (see also Ferman, 2004:204).

In the Sri Lanka study, Wall and Alderson (1993) reported that the introduction of the new exam made teachers '*anxious to cover those parts of the textbook they feel*

are most likely to be tested' (ibid: 67) while Cheng (1998) speaks of the pressure and motivation felt by teachers of the revised HKCEE. In her study, teachers appeared worried about how the shy or less outspoken students would fare in the new exam. One teacher admitted she would feel guilty if she did not familiarize her students with the test formats.

Finally, Kiss-Gulyas (2001) reported that teachers expressed fears in the face of the new school-leaving examination in Hungary related mainly to students' performance and teaching:

they were afraid that more students would fail than at present; they were worried about the achievement of lower ability students ... their hopes were related to the prestige and acceptability, or currency of the new examination, and its positive effect on classroom practices (ibid: 45)

Feelings and attitudes towards exams operating for longer periods of time

Negative feelings were reported by Alderson and Hamp-Lyons (1996) who found that the majority of the teachers had a negative attitude towards the exam and teaching for TOEFL. The researchers also mentioned teachers' feelings of guilt and frustration at *'being unable to make the content interesting or to ensure improved scores for their students'* (ibid: 292). Two teachers, however, were more positive. They *'enjoyed the teaching and felt they could help students cope with something important'* (ibid: 285).

Kiss-Gulyas (2001) noted an influence on teachers' attitudes of the existing English language school-leaving examination in Hungary. The washback of the exam was thought by teachers to be widespread and negative.

However, Watanabe (2000) reported that the atmosphere in the university exam-preparation classes he observed *'was not necessarily tense. It seemed to depend on the teacher's attitude towards exam coaching'* (ibid: 44) while Read and Hayes (2003)

found generally positive feelings about IELTS amongst their teachers. Wall (2005) also presented a picture of mixed but, overall, positive reactions from her teachers:

The teachers were generally positive about this examination, although there were some differences of opinion regarding its difficulty level and some feeling that students should be given choices within the reading and writing components to play to their individual strengths (ibid: 241)

In a similar vein, Wall and Horak (2006) found that their TOEFL teachers were, on the whole, not as negative about the work they were doing in the TOEFL classroom as the teachers in the Alderson and Hamp-Lyon's study (1996) and they took their job very seriously.

In the general education field, Smith (1991b) conducting a qualitative investigation, reported negative findings on the way external testing influences teachers. The results, based on interviews and classroom observations, revealed that teachers experienced feelings of '*shame, embarrassment, guilt and anger*' as a result of the publication of test scores and were determined to do what was necessary to raise low scores in order to avoid such feelings in the future. Teachers also believed that the scores were used against them which created feelings of '*dissonance and alienation*' despite the perceived invalidity of the tests themselves. In addition, teachers felt that during the testing sessions young children experienced negative emotions which generated feelings of anxiety and guilt among the majority of teachers.

Herman and Golan (1993) also reported a study undertaken in nine different American states that revealed serious effects of standardised tests on teachers. Teachers believed that testing created a great deal of pressure on them to improve test scores and placed equal pressure on their students as well. As a result, teachers adjusted the sequence of their curriculum based on what was included on the tests.

Finally, Johnstone et al. (1995) undertook a study that used interviews in order to investigate the perceptions of US teachers towards external testing. Their study revealed that teachers felt pressure and '*powerful feelings of being overwhelmed and of insecurity, guilt, frustration, and anger*' (ibid: 359).

Summary

- Evidence of washback on teachers' attitudes was evident in the studies reviewed but these actually followed a cline. On the one hand there are studies that noted a range of rather negative attitudes and feelings generated by exams, e.g. stress, anger, frustration, humiliation, pressure, worry, fear, and guilt. These negative feelings often produced a conflict between how teachers feel they would like to teach and how they feel they are forced to teach for examinations.
- Exam influence occasionally led to positive attitudes. Teachers were motivated to teach towards exam skills and enjoyed teaching and exam preparation. Some other studies also indicated that exams generated mixed feelings in teachers.
- In some studies, feelings of worry and fear were mainly related to the performance of students.
- However, what the studies do not explore so far is whether teachers' positive or negative attitudes and feelings generate more or less effective teaching or learning, and, if so, how.

Students

The purpose of this section is to see whether and how the learners' perspective is described in research studies in terms of attitudes, beliefs, learning strategies, self-

esteem, anxiety and motivation towards high-stakes exams, commonly known in the literature as '*washback to the learner*' (Bailey, 1996).

Positive attitudes towards exams

Li (1990) reported on students' (and teachers') positive attitudes towards the exam and motivation to study. The researcher explained that there seemed to be a new enthusiasm for learning English outside the classroom on the part of students, which included more after-class learning and led to high sales of simplified English readers. However, it is not clear whether what Li reported is based on the teachers' opinions about their students' attitudes only or whether she got direct access to the students' views.

Read and Hayes (2003) also reported, via students' questionnaire, positive feelings about the IELTS exam and motivation among learners (and teachers).

Negative attitudes towards exams

Shohamy et al. (1996) investigated students' perspectives using student questionnaires. The results showed that students (and teachers) expressed negative feelings towards the low-stakes ASL test and complained that the test was of no importance (ibid: 306). As for the high-stakes EFL oral exam, the results showed that 82% of the students regarded the exam as very important while 84% of the students stated that it was '*of considerable importance to them to succeed in the oral exam*' (ibid: 311). Also, like their teachers, 96% of the students reported being '*quite anxious about the test*' (ibid: 310).

The studies so far have shown that students' attitudes towards the exam are not homogenous as they can be both positive and negative and were mostly in agreement with those of their teachers'. However, the following studies show that students' and teachers' practices and views can differ.

Discrepancies between students' and teachers' views and attitudes

Wesdorp's study (1982) indicated that students' habits and views did indeed differ from what their teachers believed them to be. Student questionnaires showed that students' study habits did not change much after the introduction of multiple-choice questions despite teachers' beliefs.

Perrin (2000) also found that students showed a preference for multiple-choice questions contrary to teachers' beliefs.

Alderson and Hamp-Lyons (1996) interviewed students at three different institutions in TOEFL preparation courses in the United States and found that there were discrepancies between the students' views and their teachers' regarding methods and materials in the exam preparation classes:

... most teachers claimed that it was students who drove the methodology, who insisted on practice tests and on work on TOEFL-like items. However, ... in our discussions with students we did not find these claims borne out (ibid: 286).

Wall (1999) described student attitudes towards a newly-introduced exam as revealed in teacher interviews. The teachers believed that it was their students who wanted them to focus on what was on the exam and aimed at good exam results to get ahead in life:

... teachers talked about the students' desire for predictability: they wanted to know what type of questions they would get..... and they were upset if their teachers were not able to prepare them for what was coming. ... (ibid: 253)

Wall and Horak (2006) also reported that one of the reasons teachers adhered to the exam coursebooks was because this was what they claimed their students expected. As one of the teachers interviewed said:

I know they just want to do exercises for the TOEFL. They just want to go through the book and get as much practice with what's going to be on the test as possible (ibid: 82)

Finally, Lumley and Stoneman (2000) found a mismatch between the attitudes of teachers and students. Via students' and teachers' questionnaires, the researchers explored teacher and student attitudes towards the new teaching materials designed to prepare for the GSLPA exam and found that students are much more exam-oriented than their teachers.

Students show mixed feelings towards exams

Cheng (1998), using student questionnaires, found that students had mixed feelings towards the HKCEE exam, recognizing on the one hand that the exam made them work hard to achieve good scores but at the same time they considered that exams were not an accurate reflection of all aspects of their study.

Students demonstrate individual differences towards exams

Students' differences were found in the work of Hahn et al. (1989). The researchers conducted a small-scale study of the effects of grading on oral performance in the first six months of instruction of beginning students of German. Although no effects on developing oral proficiency were found, student attitudes were different: those who had been graded considered the experience stressful and unproductive, whereas the group that had not been graded wished it had been graded.

The idea that there might be individual differences among students in the way they perceive and react to exams was shown in the work of Shohamy (1993). Via student questionnaires, she established that 62% of the students claimed that the ASL test affected them positively, e.g. experienced an increase in motivation, while 38% reported that they were affected negatively, e.g. experienced fear, pressure, and anxiety, felt that the test did not reflect real learning etc. (ibid: 9). By contrast, 96% of the students reported being quite anxious about the EFL oral test and believed that the test results can affect their overall matriculation score to a large extent. 70% of all

students believed that the results can affect their success in future studies. In later discussions, Shohamy et al. (1996:314-315) explained that two of the reasons for the differential washback effect reported towards the two tests was that the ASL was a low-stakes test while the EFL test was high-stakes and that the differences may also reflect the language status of each (low vs. high).

Ferman (2004) also examined the washback effects of the EFL oral test. The researcher concluded that the test resulted in differential washback among learners. In particular, Ferman found that average ability level students were significantly different from other students: their anxiety level was the highest and they were most adversely affected by potential failure in the test. Therefore, the researcher suggested that to ensure the desired washback, individual differences among students need to be taken into account.

Gosa (2004) sought to identify possible washback effects that took place inside and outside classrooms as experienced by her Romanian students using student diaries. The analysis revealed that the personal environment of the students was affected by test washback *'to a greater extent than their classroom one, not only regarding the tasks practised but in almost every respect'* (ibid: 226). She concluded that students' expectations, feelings, attitudes, perceptions, beliefs, learning styles, and anxiety should be taken into account when trying to promote positive washback as they are likely to interact with the test and, therefore, intervene in the washback process.

Students' motivation towards exams

Watanabe (2001) attempted to cast some light on the relationship between motivation and test impact as part of a larger project (see Watanabe, 1997, 2000). Through interviews with university students about their test preparation practices,

Watanabe found that the relationship between students' test preparation and their motivation was complex. More specifically, he found that students' attitudes to test preparation varied and that the impact of the university exams was far from uniform: it depended on the importance and the difficulty of the exam.

The researcher concluded that a test can be motivating and have a positive effect on students' test preparation if it is of the appropriate difficulty to the learner (see also Moeller & Reschke, 1993). In discussing the complexity of washback, Watanabe stressed that it is not the test alone that causes washback but the learners' perception of the difficulty of the test, e.g. its 'face validity', which the researcher considers to be a rather neglected area of research.

The following two research studies report on the use of learning strategies and the extent to which motivation to learn the language can be sustained after the exam. Watanabe (1992) investigated possible differences in learning strategy use of first and second year students who entered college through entrance examination and those who entered college through recommendation. His results showed that the exam students used more learning strategies than the recommended groups and that their strategy use did not change over the period of the two years studied. Watanabe attributed students' strategy use not only to the influence of the exam, but also to language proficiency, the effect of supplementary preparatory classes and students' motivation. The researcher concluded that the washback effects of the examination drove students to learn the language only in order to pass the examination.

Finally, Berwick and Ross (1989) assessed attitudes and motives for learning English at the beginning and end of the freshman year at a public university in Japan (90 students) using pre- and post-tests of English proficiency and a motivational index was drawn after 150 hours of instruction. Their findings concurred with Watanabe's in

that '*the intensity of motivation to learn English hits a peak in the last year of high school*' (ibid: 206). The researchers also found that the students' overall intensity of motivation of the college students was low:

Once the university examinations are over, there is very little to sustain this kind of motivation, so the student appears in freshmen classrooms as a kind of timid, exam-worn survivor with no apparent academic purpose at university (ibid: 206).

Other affective factors

Research in the general educational field showed that students of various age groups are equally affected by evaluation processes. Smith (1991a), reporting on research that she and other colleagues conducted (Haladyna et al., 1991), found that young children suffered in the pressurized environment of high-stakes testing. Teachers reported that children experienced frustration, physical symptoms, loss of self-esteem, saw themselves as unable and gave up rather than try to meet the challenge. In consequence, teachers worked on test preparation activities to inoculate these pupils against emotional paralysis in the face of the tests and against feelings of stupidity that the tests seemed to engender.

Paris et al. (1991), based on student surveys, concluded that adolescent students, as opposed to younger students, are less likely to do their best on standardized achievement tests especially when they believe that such tests are unrelated to what they are learning or are used to serve political purposes. Adolescent students were also more likely to use inappropriate test taking strategies, e.g. to cheat, to become nervous, to have difficulty concentrating, to guess and to look for answers that matched the questions without reading the passage. All of these strategies, the researchers commented, are designed to avoid personal effort and responsibility, and thus are detrimental to higher order thinking and intrinsic motivation.

Paris et al. (1991) also stressed that low achievers, in their efforts to decrease personal anxiety and increase the protection of their own self-esteem in the face of an important exam, abandoned effort and appropriate strategies. However, the researchers found that it is not just low-achieving students who experience test anxiety; students of all achievement levels suffer from worry and preoccupation about not being able to do well on tests.

Hargreaves (1997), using classroom observations and interviews with students, found that students tended to rely on extrinsic prompts rather than motivating themselves to learn and that teachers sometimes used the pending examination as extrinsic motivation, especially when faced with discipline problems. Overall, the enormous importance of future good grades on the examination closely associated with employment became the driving force behind schools (see Harlen & Crick, 2003 for a review of relevant studies).

Summary

- Even though there are a number of studies acknowledging the importance of the learners' views in the study of washback, their number is still fairly limited while findings are contradictory, indicating a complex relationship between exams and students' perspectives.
- Some studies have shown that students' and teachers' attitudes towards the exam were in agreement, be they positive or negative while in others students' practices and views differed from what their teachers believed them to be.
- There are studies which have shown that the students can differ in the way they experience exam influence.
- Students were some times presented as being more exam-oriented than their teachers.

- There is also evidence that students play a major role in the presence or absence of washback.
- The effects of factors such as motivation, language proficiency, previous education, current context of education, and time are also worth taking into consideration when investigating washback to the learners. The stakes of the test, the status of the language taught, the difficulty and importance of the test as perceived by the students are important variables, too.
- Furthermore, studies have considered various affective and cognitive factors such as students' attitudes, study habits, views, anxiety, personal environment, perceptions, beliefs, learning styles, motivation, etc. with conflicting results. This could be one of the reasons that investigating washback on learners is so complex.
- From a methodological point of view, questionnaires and, occasionally, interviews were the most frequent methods that have been used so far in investigations of 'learner washback'. In some of the research studies that have looked at the students' perspective, researchers have consulted teachers about their students rather than students directly. It would be problematic to assume that what the teachers believed about their students corresponded entirely to the students' real views and attitudes.
- More evidence-based support for claims of positive and negative washback on students is needed that would help resolve conflicting results.
- For instance, studies of test anxiety and its facilitating and debilitating effects during exam preparation would merit further research.
- That exams impact on feelings and attitudes seems clear but how these in turn impact on teaching and learning is much less clear.
- It is surprising to note that since Alderson and Wall's seminal paper (1993) only a very small number of studies has dealt with the issue of learners' motivation and its

relation to exams, even though a number of claims assert that exams motivate learners (e.g. Wall, 1999; Read & Hayes, 2003). The relationship between motivation and second language learning is complex (e.g. Gardner, 1985; Dornyei, 2001; Csizer & Dornyei, 2005; Dornyei, 2005), while the relationship between these and high-stakes exams is likely to be even more complex. Much more research needs to be done on how high-stakes exams motivate students to learn and whether they can help sustain students' motivation for learning after the exam. This would be an equally useful addition to the field.

- Students' perspectives concerning the relationship between teaching, learning and high-stakes exams needs to be studied longitudinally and directly using research methods that will capture the aspects under investigation more clearly.

Concluding remarks

From the above literature review several findings have emerged with regard to washback.

First of all, we can now see more clearly from the studies reviewed that rather than there being a direct, and automatic effect, washback is actually complex and elusive; while studies showed that there can be washback from exams onto a variety of teaching and learning areas, the same studies indicated that this washback was not always present and, very often, it varied in '*form*' and '*intensity*' (Cheng, 1997b, 2005).

The literature review also showed that washback is broad and multi-faceted and can be brought about through the agency of a great many independent and intervening variables beside the exam itself. Some of the factors which seem to have affected the form that washback can take include teacher and student factors (e.g. beliefs, attitudes,

experience, education, training, personality, teaching and learning style, etc), textbook writers and publishers (e.g. their interpretation of exam requirements), the status of the subject being tested, resources and classroom conditions, management of practices in the schools, communication between test providers and test users, the socio-political context in which the test is put to use, etc. (see also Spratt, 2005).

In addition, various methods have been employed in the washback studies reviewed. Actually, the post-1993 era brought with it a diversification of methods including methods that had not been widely used before like the analysis of teaching materials and documents, classroom observations, individual and group interviews and analysis of teacher talk (see also Table 1).

An important tendency in the more recent work is that researchers have not limited themselves to describing washback only but have also attempted to provide the reasons why washback appeared or not and why it had taken on the form it finally did. Some of the washback studies used a single method (e.g. Andrews, 2004) while others adopted more than one (e.g. Shohamy et al., 1996; Cheng, 2005). Language testers have also developed various instruments for measuring washback and impact, and tried to evaluate the degree to which these may be considered positive or negative (e.g. Saville & Hawkey, 2004).

Such diversification has undoubtedly made possible the discovery of new and important issues in the study of washback. However, it could also be argued that even more methods could be employed to help researchers probe deeper into the less observable factors related to the individuals involved.

On the basis of the literature considered so far, I would like to make some additional points which entail further research on the influence of high-stakes tests.

As was seen in this review, it is still the case that more research is needed in the area of test washback, if only to confirm how generalizable the results of these studies are to other populations and situations, and to follow up on issues they raise. Spratt stresses the same point:

There is a need for more studies to be carried out in different learning contexts. Use of parallel methodologies for studies in different contexts might also allow researchers to investigate some of the apparent contradictions in the findings to date (2005:27)

In addition, although external factors such as teacher or student factors have been identified, insufficient research has been done to reveal how they function and interact with a test to bring about the results observed. An important step forward for the field of language testing would be to construct a model of washback which would not only take account of these factors but also be able to demonstrate the interaction between them. Alderson and Banerjee (2001:215) argue that this would be a valuable addition in the field:

More sophisticated conceptual frameworks, which are slowly developing in the light of research findings and related studies into innovation, motivation theory and teacher thinking, are likely to provide better understanding of the reasons for washback and an explanation of how tests might be developed to contribute to the engineering of desirable change.

One feature that the majority of studies focusing on the washback of a revised exam have in common is that they investigated washback practices shortly after the introduction of the revised exam and, therefore, failed to capture the intended effects clearly. It is necessary to study the intended washback effects of a test that has been in existence for quite a long time so as to confirm that the absence or presence of such effects in the classroom are not caused by the time factor. Spratt agrees by saying:

It would be interesting to see if similar findings emerged from a study conducted once the exam's contents and standards had become

familiar to teachers; that is, how much were these results a fruit of uncertainty about the exam on the teachers' part? (2005:11)

With regard to methodology, it is preferable if more than one method be used to increase the validity of the research. To get a more comprehensive picture of test washback, it is desirable to conduct studies which look at washback of a specific test from different perspectives (including at least the two central participants involved: teachers and students) in order to investigate the influence it exerts on classroom teaching and learning in depth.

References

- Alderson, J. C. (1995). *Ideas for Research into Impact, Washback and IELTS*. Unpublished manuscript, Ingleton.
- Alderson, J. C. (2004). Foreword. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. ix-xii). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Alderson, J. C., & Banerjee, J. (2001). Language Testing and Assessment (Part 1). *Language Teaching*, 34(4), 213-236.
- Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL Preparation Courses: A Study of Washback. *Language Testing*, 13(3), 280-297.
- Alderson, J. C., & Wall, D. (1992). *The Sri Lankan O-Level English Language Evaluation Project Fourth and Final Report*. Lancaster: Department of Linguistics and Modern English Language / Institute for English Language Education.
- Alderson, J. C., & Wall, D. (1993). Does Washback Exist? *Applied Linguistics*, 14(2), 115-129.
- Andrews, S. (1994a). The Washback Effect of Examinations: Its Impact Upon Curriculum Innovation in English Language Teaching. *Curriculum Forum*, 4(1), 44-58.
- Andrews, S. (1994b). Washback or Washout? The Relationship between Examination Reform and Curriculum Innovation. In D. Nunan & R. Berry (Eds.), *Bringing About Change in Language Education* (pp. 67-81). Hong Kong: Department of Curriculum Studies. University of Hong Kong.
- Andrews, S. (2004). Washback and Curriculum Innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 37-52). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting Washback—a Case-Study. *System*, 30(2), 207–223.
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: OUP.
- Bachman, L. F., & Palmer, A. (1996). *Language Testing in Practice : Designing and Developing Useful Language Tests*. Oxford: OUP.

- Bailey, K. M. (1996). Working for Washback: A Review of the Washback Concept in Language Testing. *Language Testing*, 13(3), 257-279.
- Bailey, K. M. (1999). *Washback in Language Testing*. TOEFL Monograph Series. Report Number: RM-99-04, TOEFL-MS-15. Princeton, NJ: Educational Testing Service. Retrieved 11/3/2001, from <http://www.ets.org/>.
- Baker, E. (1991). *Alternative Assessment and National Policy*. Paper presented at the National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement. Washington, DC.
- Berwick, R., & Ross, S. (1989). Motivation after Matriculation: Are Japanese Learners of English Still Alive after Exam Hell? *Japan Association for Language Teaching Journal*, 11, 193-210.
- Bonkowski, J. F. (1996). *Instrument for the Assessment of Teaching Materials*. Unpublished manuscript, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, England.
- Buck, G. (1988). Testing Listening Comprehension in Japanese University Entrance Exams. *JALT Journal*, 10, 15-42.
- Burrows, C. (2004). Washback in Classroom-Based Assessment: A Study of the Washback Effect in the Australian Adult Migrant English Program. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 113-128). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Chapman, D. W., & Snyder, C. W. (2000). Can High Stakes National Testing Improve Instruction: Re-Examining Conventional Wisdom. *International Journal of Educational Development*, 20(6), 457-474.
- Cheng, L. (1997a). How Does Washback Influence Teaching? Implications for Hong Kong. *Language and Education*, 11(1), 38-54.
- Cheng, L. (1997b). *The Washback Effect of Public Examination Change on Classroom Teaching*. Unpublished PhD thesis, The University of Hong Kong, Hong Kong.
- Cheng, L. (1998). Impact of a Public English Examination Change on Students' Perceptions and Attitudes toward Their English Learning. *Studies in Educational Evaluation*, 24(3), 279-301.
- Cheng, L. (1999). Changing Assessment: Washback on Teacher Perspectives and Actions. *Teaching and Teacher Education*, 15(3), 253-271.

- Cheng, L. (2005). *Changing Language Teaching through Language Testing: A Washback Study*. Cambridge: CUP.
- Cheng, L., & Curtis, A. (2004). Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 3-17). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Cooley, W. W. (1991). State-Wide Student Assessment. *Educational Measurement: Issues and Practice*, 10, 3-6.
- Csizer, K., & Dorneyi, Z. (2005). The Internal Structure of Language Learning Motivation: Results of Structural Equation Modelling. *Modern Language Journal*, 89(1), 19-36.
- Davies, A. (1997). Demands of Being Professional in Language Testing. *Language Testing*, 14(3), 328-339.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: CUP/UCLES.
- Dockrell, B. (1991). The Effect of System Wide Testing: Issues Raised by a Case Study in a Developing Country. *Studies in Educational Evaluation*, 17, 41-49.
- Dorneyi, Z. (2001). *Teaching and Researching Motivation*. London: Longman.
- Dorneyi, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Ferman, I. (2004). The Washback of an EFL National Oral Matriculation Test to Teaching and Learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 191-210). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Frederiksen, J. R., & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher*, 18(9), 27-32.
- Fullilove, J. (1992). The Tail That Wags. *Institute of Language in Education Journal*, 9, 131-147.
- Gardner, R. C. (1985). *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*. London: Edward Arnold.

- Glover, P. (2006). *Examination Influence on How Teachers Teach: A Study of Teacher Talk*. Unpublished PhD thesis, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, England.
- Gosa, C. M. C. (2004). *Investigating Washback: A Case Study Using Student Diaries*. Unpublished PhD thesis, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, England.
- Hahn, S., Stassen, T., & Deschke, C. (1989). Grading Classroom Oral Activities: Effects on Motivation and Proficiency. *Foreign Language Annals*, 22(3), 241-252.
- Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising Standardized Achievement Tests Scores and the Origins of Test Score Pollution. *Educational Researcher*, 20(5), 2-7.
- Hamp-Lyons, L. (1996). *Ethical Test Preparation Practice: The Case of the TOEFL*. Unpublished manuscript.
- Hamp-Lyons, L. (1997). Washback, Impact and Validity: Ethical Concerns. *Language Testing*, 14(3), 295-303.
- Hamp-Lyons, L. (1998). Ethical Test Preparation Practice: The Case of the TOEFL. *TESOL QUARTERLY (Forum Commentary)*, 32(2), 329-337.
- Hamp-Lyons, L. (2000). Social, Professional and Individual Responsibility in Language Testing. *System*, 28, 579-591.
- Hargreaves, E. (1997). The Diploma Disease in Egypt: Learning, Teaching and the Monster of the Secondary Leaving Certificate. *Assessment in Education: principles, policy & practice*, 4(1), 161-176.
- Harlen, W., & Crick, R. D. (2003). Testing and Motivation for Learning. *Assessment in Education: Principles, Policy & Practice*, 10(2), 169-207.
- Hawkey, R. (2004a). *Cambridge ESOL CPE Textbook Washback Study: Full Report*. Cambridge: University of Cambridge ESOL Examinations.
- Hawkey, R. (2004b). The CPE Textbook Washback Study. *Research Notes*, 20, 19-20.
- Hawkey, R. (2006). *Impact Theory and Practice: Studies of the IELTS Test and Progetto Lingue 2000*. Cambridge: CUP.
- Hawkey, R., & Turner, R. (forthcoming). Washback of the Cambridge ESOL Proficiency in English Exam on Preparation Course Books: An Empirical Study. *Language Assessment Quarterly*.

- Hayes, B., & Read, J. (2004). IELTS Test Preparation in New Zealand: Preparing Students for the IELTS Academic Module. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 97-112). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Heaton, J. B. (1990). *Classroom Testing*. Harlow: Longman.
- Herman, J. L., & Golan, S. (1993). The Effects of Standardized Testing on Teaching and Schools. *Educational Measurement: Issues and Practices*, 12(4), 20-25.
- Hilke, R., & Wadden, P. (1997). The TOEFL and Its Imitators: Analyzing the TOEFL and Evaluating TOEFL-Prep Texts. *RELC Journal*, 28(1), 28-53.
- Hughes, A. (1988). Introducing a Needs-Based Test of English Language Proficiency into an English Medium University in Turkey. In A. Hughes (Ed.), *Testing English for University Study (ELT Documents #127)* (pp. 134-146). London: Modern English Publications in association with the British Council.
- Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: CUP.
- Hughes, A. (1994). *Backwash and TOEFL 2000*. Unpublished manuscript, commissioned by Educational Testing Service (ETS). University of Reading.
- Johnson, F., & Wong, C. L. K. L. (1981). The Interdependence of Teaching, Testing and Instructional Materials. In J. A. S. Read (Ed.), *Directions in Language Testing* (pp. 277-302). Singapore: Regional Language Centre.
- Johnstone, P., Guice, S., Baker, K., Malone, J., & Michelson, N. (1995). Assessment of Teaching and Learning in Literature-Based Classrooms. *Teaching and Teacher Education*, 11, 359-371.
- Khaniya, T. R. (1990). The Washback Effect of a Textbook-Based Test. *Edinburgh Working Papers in Applied Linguistics*, 1, 48-58.
- Kiss-Gulyas, J. (2001). Experiencing the Examination Design, Content, Materials and Procedures. In J. G. Egyud, I. A. Gal & P. Glover (Eds.), *English Language Education in Hungary, Part Ili: Training Teachers for New Examinations* (pp. 40-58). Budapest: The British Council.
- Lam, H. P. (1993). *Washback-Can It Be Quantified? A Study on the Impact of English Examinations in Hong Kong.*, Unpublished MA dissertation. University of Leeds, UK.
- Lam, H. P. (1994). Methodology Washback - an Insider's View. In D. Nunan, R. Berry & V. Berry (Eds.), *Bringing About Change in Language Education: Proceedings of the International Language in Education Conference 1994* (pp. 83-102). Hong Kong: University of Hong Kong.

- Li, X. (1990). How Powerful Can a Language Test Be? The MET in China. *Journal of Multilingual and Multicultural Development*, 11(5), 393-404.
- Lumley, T., & Stoneman, B. (2000). Conflicting Perspectives on the Role of Test Preparation in Relation to Learning. *Hong Kong Journal of Applied Linguistics*, 5(1), 50-80.
- Madaus, G. F. (1988). The Influence of Testing on the Curriculum. In L. N. Tanner (Ed.), *Critical Issues in Curriculum: 87th Yearbook for the National Society for the Study of Education* (pp. 83-121). Chicago: University of Chicago Press.
- McNamara, T. (1996). *Measuring Second Language Performance*. Harlow: Longman.
- McNamara, T. (2000). *Language Testing*. Oxford: OUP.
- Mehrens, W. A., & Kaminski, J. (1989). Methods for Improving Standardized Test Scores: Fruitful, Fruitless or Fraudulent? *Educational Measurement: Issues and Practice*, 8(1), 14-22.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: ACE and Macmillan.
- Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing*, 13, 241-256.
- Moeller, A. J., & Reschke, C. (1993). A Second Look at Grading and Classroom Performance: Report of a Research Study. *Modern Language Journal*, 77(2), 163-169.
- Morrison, K., & Tang, F. H. J. (2002). Testing to Destruction: A Problem in a Small State. *Assessment in Education: principles, policy & practice*, 9(3), 289-317.
- Morrow, K. (1986). The Evaluation of Tests of Communicative Performance. In M. Portal (Ed.), *Innovations in Language Testing: Proceedings of the IUS/NFER Conference* (pp. 1-13). London: NFER/Nelson.
- Nikolov, M. (1999). Classroom Observation Project. In H. Fekete, E. Major & M. Nikolov (Eds.), *In English Language Education in Hungary: A Baseline Study* (pp. 221-246). Budapest: The British Council Hungary.
- Paris, S. G., Lawton, T. A., Turner, J. C., & Roth., J. L. (1991). A Developmental Perspective on Standardized Achievement Testing. *Educational Researcher*, 20(5), 12-20.

- Pearson, I. (1988). Tests as Levers for Change. In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the Classroom: Practice and Evaluation. ELT Documents Volume 128* (pp. 98-107). London: Modern English Publications.
- Perrin, G. C. (2000). *The Effect of Multiple Choice Foreign Language Tests of Listening and Reading on Teacher Behaviour and Student Attitudes*. Unpublished PhD thesis, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, England.
- Pierce, B. N. (1992). Demystifying the TOEFL Reading Test. *TESOL QUARTERLY*, 26(4), 665-689.
- Popham, W. J. (1987). The Merits of Measurement-Driven Instruction. *Phi Delta Kappan*, 68, 679-682.
- Popham, W. J. (1991). Appropriateness of Teachers' Test-Preparation Practices. *Educational Measurement: Issues and Practice*, 10(4), 12-15.
- Prodromou, L. (1995). The Backwash Effect: From Testing to Teaching. *ELT Journal*, 49(1), 13-25.
- Qi, L. (2004). *The Intended Washback Effect of the National Matriculation English Test in China : Intentions and Reality*. Beijing: Foreign Language Teaching and Research Press.
- Qi, L. (2005). Stakeholders' Conflicting Aims Undermine the Washback Function of a High-Stakes Test. *Language Testing*, 22(2), 142-173.
- Raimes, A. (1990). The TOEFL Test of Written English: Causes for Concern. *TESOL QUARTERLY*, 24(3), 427-442.
- Read, J., & Hayes, B. (2003). The Impact of IELTS on Preparation for Academic Study in New Zealand. *IELTS International English Language Testing System Research Reports*, 4, 153-206.
- Saif, S. (2006). Aiming for Positive Washback: A Case Study of International Teaching Assistants. *Language Testing*, 23(1), 1-34.
- Saville, N. (2000). Investigating the Impact of International Language Examinations. *Research Notes*, 2, 2-7.
- Saville, N., & Hawkey, R. (2004). The IELTS Impact Study: Investigating Washback on Teaching Materials. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 73-96). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

- Shohamy, E. (1992). Beyond Proficiency Testing: A Diagnostic Feedback Testing Model for Assessing Foreign Language Learning. *Modern Language Journal*, 76(4), 513-521.
- Shohamy, E. (1993). *The Power of Test: The Impact of Language Testing on Teaching and Learning*. National Foreign Language Center Occasional Papers. Washington, DC: National Foreign Language Center.
- Shohamy, E. (2001). *The Power of Tests : A Critical Perspective on the Uses of Language Tests*. London: Longman.
- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test Impact Revisited: Washback Effect over Time. *Language Testing*, 13(3), 298-317.
- Smith, I. (2004). IELTS Impact: A Study on the Accessibility of IELTS GT Modules to 16-17 Year Old Candidates. *Research Notes*, 18, 6-8.
- Smith, M. L. (1991a). Meanings of Test Preparation. *American Educational Research Journal*, 28, 521-542.
- Smith, M. L. (1991b). Put to the Test: The Effects of External Testing on Teachers. *Educational Researcher*, 20(5), 8-11.
- 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.
- Stecher, B., Chun, T., & Barron, S. (2004). The Effects of Assessment-Driven Reform on the Teaching of Writing in Washington State. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 53-72). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Sturman, L. (2003). Teaching to the Test: Science or Intuition? *Educational Researcher*, 45(3), 261-273.
- Wall, D. (1994). Measuring Examination 'Washback': The Sri Lankan Evaluation Project. In R. H. Flavell (Ed.), *ELT Policy and Its Impact: A Case Study. Review of English Language Teaching 1* (pp. 103-122). London: Modern English Publications in association with The British Council.
- Wall, D. (1997). Impact and Washback in Language Testing. In C. Clapham & D. Corson (Eds.), *Encyclopaedia of Language and Education. Vol. 7. Language Testing and Assessment* (pp. 291-302). Dordrecht: Kluwer Academic.
- Wall, D. (1999). *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*. Unpublished

PhD thesis, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, England.

Wall, D. (2000). The Impact of High-Stakes Testing on Teaching and Learning: Can This Be Predicted or Controlled? *System*, 28, 499-509.

Wall, D. (2005). *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*. Cambridge: CUP.

Wall, D., & Alderson, J. C. (1993). Examining Washback: The Sri Lankan Impact Study. *Language Testing*, 10(1), 41-69.

Wall, D., & Horak, T. (2006). *The Impact of Changes in the TOEFL® Examination on Teaching and Learning in Central and Eastern Europe: Phase 1, the Baseline Study*. TOEFL Monograph Series. Report Number: RR-06-18, TOEFL-MS-34. Princeton, NJ: Educational Testing Service. Retrieved 29/6/2006, from <http://www.ets.org/>.

Wang, Y. W. (1997). *An Investigation of Textbook Materials Designed to Prepare Students for the IELTS Test: A Study of Washback*. Unpublished M.A. dissertation, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, UK.

Watanabe, Y. (1992). Washback Effects of College Entrance Examination on Language Learning Strategies. *JACET*, 175-194.

Watanabe, Y. (1996). Does Grammar-Translation Come from the Entrance Examination? Preliminary Findings from Classroom-Based Research. *Language Testing*, 13(3), 318-333.

Watanabe, Y. (1997). *The Washback Effects of the Japanese University Entrance Examinations of English-Classroom-Based Research*. Unpublished PhD thesis, Department of Linguistics and Modern English Language, Lancaster University, Lancaster, England.

Watanabe, Y. (2000). Washback Effects of the English Section of Japanese University Entrance Examinations on Instruction in Pre-College Level EFL. *Language Testing Update*, 27, 42-47.

Watanabe, Y. (2001). Does the University Entrance Examination Motivate Learners? A Case Study of Learner Interviews. In Akita Association of English Studies (Ed.), *Trans-Equator Exchanges: A Collection of Academic Papers in Honour of Professor David Ingram* (pp. 100-110).

Weir, C. J. (1990). *Communicative Language Testing*. New York: Prentice Hall.

Wesdorp, H. (1982). Backwash Effects of Language-Testing in Primary and Secondary Education. *Journal of Applied Language Study*, 1(1), 40-55.