UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

http://revistas.um.es/ijes

Washback in language assessment

ANTHONY GREEN*
University of Bedfordshire

Received: 30 July 2013 / Accepted: 10 September 2013

ABSTRACT
This paper reviews the progress made in washback studies over the quarter century since Hughes' (1989) placed it at the centre of his textbook *Testing for Language Teacher*s. Research into washback and the development of models of washback are described and an agenda is suggested for test developers wishing to build washback into their programmes. It is recommended that future projects should pay greater attention to test design features and to the outcomes of learning as well as continuing to explore learner motivation and cultural factors that might encourage participants to react to tests in certain ways, but not in others. Washback research itself is seen to be a potentially valuable tool in persuading participants to adopt new practices.

KEYWORDS: language testing, test impact, washback, consequences

RESUMEN
Este artículo revisa el progreso realizado en los estudios relativos al impacto de la evaluación sobre los procesos de aprendizaje y enseñanza de lenguas (*washback*) durante el último cuarto de siglo desde que Hughes colocó este tema como capítulo central de su libro *Testing for Language Teachers* (1989). Concretamente, se describen tanto los estudios sobre *washback* como sus modelos y se sugiere, para los evaluadores que así lo deseen, una agenda para la introducción de *washback* en sus programas. Se recomienda que en el futuro se preste más atención al diseño de las pruebas de evaluación de lenguas así como a los resultados de aprendizaje y que se siga explorando la motivación del alumno y factores de índole cultural que pueden dar lugar a que los participantes reaccionen a los exámenes de determinadas maneras, y no de otras. La investigación sobre *washback* en sí misma puede considerarse como una herramienta potencialmente valiosa para convencer a los participantes de adoptar nuevas prácticas en este sentido.

PALABRAS CLAVE: evaluación de lenguas, impacto de la evaluación, consecuencias

Address for correspondence: Anthony Green, Centre for Research in English Language Learning and Assessment, Room 125, University of Bedfordshire, Putteridge Bury, Hitchin Rd, Luton, Bedfordshire LU2 8LE, United Kingdom. Email: Tony.Green@beds.ac.uk

Print ISSN: 1578-7044; Online ISSN: 1989-6131

1. INTRODUCTION

Washback refers to the impact that a test has on the teaching and learning done in preparation for it. This paper reviews research conducted into washback over the quarter century since the publication of Hughes' standard text *Testing for Language Teacher*s (1989). Hughes presented washback (or *backwash* as he called it) as a key concern for teachers. This prompted researchers to begin to investigate whether and how washback came about in different contexts. This paper first offers an extended definition, and then outlines the research that has been carried out into washback. Consideration is given to how findings have informed the development of theoretical frameworks explaining how washback occurs and features that may influence its course. Finally, these frameworks are used to outline an agenda for language test developers who wish to apply the lessons from washback research to their own practices.

Two related trends in language assessment over recent decades have encouraged growth in interest in washback. The first, reflected in Hughes' (1989), has been a movement in test design towards performance testing involving attempts to create assessment tasks that more closely resemble real-world applications of language related knowledge, skills and abilities. The other has been a shift in views of test validity to embrace the use of tests as instruments of social policy.

2. ASPECTS OF WASHBACK

A distinction has often been made between the *extent* (Bachman & Palmer, 1996) or *intensity* (Cheng, 2005) of washback and its *direction* (beneficial or damaging) (Alderson & Wall, 1993; Hughes, 1989). The importance afforded to a test has traditionally been regarded as the motivating force that drives washback, leading to more or less intense effects. The design of the test and the tasks it includes are seen as a rudder that can guide washback in a beneficial or damaging direction (Bailey, 1996, Hughes, 1989).

Washback intensity (Cheng, 2005) refers to the degree to which participants will adjust their behaviour to meet the demands of a test. Hughes (1993) suggested that washback should only be anticipated where participants are i) motivated to succeed on the test, ii) believe they know how to be successful and iii) believe they have sufficient resources to succeed. If the test is not seen to matter, there is little incentive to prepare for it.

Washback is usually evaluated as taking a beneficial or damaging direction to the extent that it encourages or discourages forms of teaching or learning intended by the test developers or considered to be appropriate on other grounds. Of course, what is considered to be appropriate will depend on the position adopted by those making the judgment and the educational goals he or she espouses (Hamp-Lyons, 1987; Mehrens, 1998). Arguments about

Print ISSN: 1578-7044; Online ISSN: 1989-6131

the direction of washback are an expression of debates between competing theories of learning.

Language testing as a field has traditionally been concerned with issues of test design and has given much less attention to the consequences of the use of tests within educational systems. This is perhaps because matters of design lie much more clearly within the control of developers. There is an argument that washback effects can be adequately addressed within the established approach. Messick (1996) has frequently been quoted as recommending to test developers, "rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback" (p. 252). A well-designed test should encourage good teaching; a poorly designed test will tempt teachers and learners into practices that have limited value in relation to long-term learning goals.

Unfortunately, as Messick (1996) observes, no test can hope to eliminate the twin threats to validity of construct irrelevance and construct under-representation. In spite of the best efforts of test developers, the skills needed to succeed on a test can never fully equate to the skills required for success in a target language use domain (Bachman & Palmer, 2010). The limitations on the time and space available for testing mean that developers have to be selective in what they test and be pragmatic in how they carry out the test. There are restrictions on the types of task they can employ which mean that test tasks can never fully reproduce a 'real life' experience.

Just as it qualifies the inferences that users are able to make on the basis of test scores, construct under-representation associated with limitations on test content and format has implications for teaching and learning. Because of their selectivity, it has been suggested for well over a hundred years that the use of tests tends to 'narrow the curriculum' and that 'what is tested is what gets taught' (see for example Herbert, 1889). Teachers may decide to focus only on the skills and knowledge required for the test, giving practice in test-like activities to the exclusion of anything that does not appear on the test. Construct-irrelevance can encourage training in test taking skills that may have little value for any other purpose. The greater the differences between test taking processes and real-world language use, the greater the risk of damaging washback.

Naturally, when a test is used as part of an educational system, many factors other than the design of the test contribute to the nature of learning outcomes. As Messick (1996) pointed out, "a poor test may be associated with positive effects and a good test with negative effects because of other things that are done or not done in the educational system" (p. 242). Effects brought about by, for example, poor teacher training or ingrained approaches to learning have no implications for the validity of the test. On the other hand, where tests are intended to encourage improvements to education, issues of this nature must be confronted. For some, this implies expanding our definition of validity to encompass the integration of tests with other aspects of the educational system: systemic validity (Frederickson & Collins, 1989). For others these are "sources of adverse consequences that are beyond invalidity"

Print ISSN: 1578-7044; Online ISSN: 1989-6131

(Bachman, 2005: 16), but that should nonetheless be considered in an assessment *use* (rather than just *validity*) argument.

While some commentators have focused on test design, others have contested that negative consequences of tests result from their use in determining test takers' life chances. The imperative to succeed on a test encourages teachers and learners to adopt short-term strategies, prioritising memorisation of large amounts of content over building a deeper understanding of underlying principles. The most deleterious effects come from high stakes tests that control access to opportunities and so are seen as very important to test takers' life chances (Crooks, 1998). The choice of test format and content may have a relatively trivial impact on this behaviour.

3. RESEARCH INTO WASHBACK

Studies of washback effects in language testing contexts began to appear in the early 1990s. These have generally either investigated the ongoing effects of established testing programmes or looked into how changes in systems of assessment affect educational practice. Alderson and Wall (1993) is often cited as a foundational text in washback studies as it set out an agenda for washback research. The authors questioned the assumptions being made about the effects of innovative forms of testing and argued that systematic study was needed in order to confirm the presence and nature of washback in any given context.

To systematise the investigation of washback, Alderson and Wall (1993) suggested a set of hypotheses that involved distinctions between effects on attitudes and effects on the content of teaching and learning and between impacts on methods and impacts on processes. They criticized earlier research into washback from language tests (Hughes, 1988; Khaniya, 1990) for supposedly deterministic assumptions and for a lack of empirical data on actual classroom practices (the early studies had relied on insights from interested participants gathered through questionnaires and interviews).

A further impetus to the emergence of washback studies was provided by a special issue of *Language Testing* in 1996 edited by Alderson and Wall with theoretically oriented contributions from Messick and Bailey and research reports by Alderson and Hamp-Lyons; Shohamy, Donitsa-Schmidt and Ferman; Wall (1996) and Watanabe (1996).

Broadly following the approach laid down by Alderson and Wall (1993), many of the studies that followed combined quantitative data from questionnaires with more qualitative descriptions of educational practices based on interviews and direct classroom observation (Burrows, 1998; Cheng, 2005, Watanabe, 1996). The focus has most often been on teachers and classroom practices, although studies of learners have also begun to appear (Gosa, 2005; Green, 2007; Tsagari, 2010; Xia & Andrews, 2013). Findings tend to underline the variety in whether, how and why participants incorporate test preparation into their practices. The

Print ISSN: 1578-7044; Online ISSN: 1989-6131

majority of the published studies have involved large-scale national and international tests used for university entrance. However, a wide range of contexts have been explored including low-stakes classroom assessment systems.

The available evidence suggests that teachers do tend both to limit the content of instruction to material covered in the test and use tasks in the classroom that reflect test tasks, but that methods of teaching are less obviously affected (Alderson & Hamp-Lyons, 1996; Qi, 2005; Wall, 2005; Watanabe, 1996). However, it also appears from studies involving surveys and interviews with participants, analyses of textbook materials (Saville & Hawkey, 2004, Tsagari, 2010) and classroom observation (Alderson & Hamp-Lyons, 1996; Green, 2007; Watanabe, 2004) that materials writers and teachers are selective in focusing more on certain aspects of a test (or test prep textbook) than on others. Learners have not been as extensively studied as teachers, but the evidence suggests that they also determine for themselves how best to prepare and that washback to the learner does not flow in a straightforward manner either directly from the test or from washback to the teacher (Gosa, 2004; Green, 2007; Mickan & Motteram, 2010; Xie & Andrews, 2013).

A shortcoming of much recent research has been the lack of attention to learning outcomes. Studies that fail to investigate whether test preparation strategies result in improved scores on the test in question must struggle to show that preparation is truly relevant to the test. Messick (1996) insisted that products must be of central importance in washback research. In the context of the TOEFL 2000 initiative, he argued that 'programme practices and individual learner strategies' should be related to 'TOEFL proficiency outcomes' (Messick quoted in Bailey, 1996: 274). Similarly Hughes (1993) recommended that research should start from the identification of the skills intended to be developed, with washback being evaluated in light of the degree to which these skills improve or decline when a test is introduced.

Alderson and Hamp-Lyons (1996) made the point that although teachers may choose to follow the format of a test in their test preparation classes, they may have no solid evidence that this will help their students to improve their scores. It is just that this seems an obvious way to approach the short-term goal of passing the test. Hamp-Lyons (1998) suggested that poor teaching practices associated with TOEFL classes might result from a pervasive culture within the English language teaching profession rather than the format and content of the test itself. Green's (2007) investigation of IELTS test preparation practices suggested that, contrary to teachers' beliefs, there was no substantial benefit in focusing on the test in preference to studying broader English for academic purposes program. In cases of this kind, teacher education might lead to greater benefits than can be achieved through test reform. Factors affecting test preparation include deficits in resources; lack of knowledge of the test, lack of training among participants and conservatism on the part of those who consider the innovations as a threat to their current status. These are all issues that can be anticipated and addressed without adjustments to test design.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

4. DEVELOPMENT OF THEORETICAL MODELS OF WASHBACK

Although simple conceptual models of washback have been developed, washback studies have revealed considerable variability in whether and how teachers, learners and others change their behaviour to address test demands, suggesting that the forces shaping washback interact in more complex ways than is yet well understood. Washback effects would appear to be highly variable and intimately dependent on context. It has proved particularly challenging to separate out and quantify the contributions made by test design, test use and other variables implicated in test preparation. Researchers have struggled to find a suitable theoretical framework to account for the variation they have observed and to uncover the inner workings of the 'black box' of washback.

Hughes (1993) proposed a basic, but influential process model of washback. This made a tripartite distinction between effects on participants (the people affected by the test, e.g. teachers, learners and materials writers), processes (participant actions, e.g. teaching and learning activities) and products (the outcomes of these processes: scores on tests, courses, teaching materials etc.). Bailey (1996) further developed this model, representing the relationships between the elements in the form of a diagram.

Green (2007), building on the work of Hughes (1993) and Bailey (1996), expanded on this model to outline the relationships between i) test design considerations as a key determinant of washback direction mediated by both ii) participant values, motivations and resources as the major determinants of washback variability and iii) the perceived importance and difficulty of the test as key determinants of washback intensity. As ii) and iii) are governed by social and individual differences, participants in the same general context may be affected by a test in different ways.

Criticising a lack of attention to test design in some washback studies and a lack of explicit statements of intended washback on the part of test providers, Green (2007) underlined the importance of detailed analysis of the test instrument and an evaluation of its congruence or (adopting a term used by Resnick & Resnick, 1992) *overlap* with the planned curriculum.

While language testing as a field has always been concerned with matters of test design, researchers have had to look elsewhere for insights into the attitudes and motivations of teachers and learners in order to uncover how individual differences affect washback. A good deal of the washback research into participants has been descriptive and exploratory. Some researchers have proposed models (e.g. Burrows, 2005, and Shih, 2007) that map out the flow of influences that bring washback effects about. Others have looked to established theories of learning and motivation to predict and account for washback phenomena.

Wall (2005) pioneered the use of innovation theory (Henrichsen, 1989, Fullan, 2001) to account for the ways in which teachers respond to new tests, integrating washback with evaluative research into curricular change. Wall's findings pointed to the need for testing

Print ISSN: 1578-7044; Online ISSN: 1989-6131

innovations to take account of local conditions, to allow time for adaptation, and to recognise that new ideas would be assimilated and interpreted in many different ways by participants. Investigating the effects on learners and learning, Green (2007) located washback within a model of second language learning based on Skehan (1989), incorporating phenomenographic theories of deep and surface approaches to learning (Entwistle, 1988) and measures of test anxiety (Horwitz, Horwitz & Cope, 1986). His findings suggested a relatively limited role for test preparation in determining score outcomes. Xie and Andrews (2013) employed expectancy-value motivation theory (Jacobs & Eccles, 2000) to develop a statistical model tracing how test takers' perceptions of the design and use of a test impacted on their preparation practices. They found that learners perceptions of both test design and test use influenced (self-reported) test preparation practices, but were unable to identify which played the greater role.

The insight that washback is shaped by participants has brought increasing integration of washback into more general theories of teaching and learning. This is an important step towards both a fuller understanding of how and why participants engage in test preparation and a greater appreciation for the role of assessment in language learning.

5. AN AGENDA FOR LANGUAGE TEST DEVELOPERS

Where test developers build tests for use in educational settings, they must take account of the likely and actual consequences of their use: including their washback. This section suggests ways in which developers may take account of washback when designing and validating a test.

5.1. Test design

Washback research often begins from the intentions of the test developers or policy makers. Research questions focus on whether the effects associated with use of the test are in line with what was intended. Qi (2005), for example, interviewed policy makers to learn what kinds of teaching and learning they hoped would characterise test preparation classes. This can help to establish the direction of washback (at least from the policy maker's perspective): positive washback is found where teachers and learners behave in ways that are considered desirable. However, there is no theoretical basis for assuming that the intended effects will be the most likely effects of a test.

An appropriate starting point for test developers would therefore seem to be an explicit statement of the relationship between what is tested and what, considering the purpose of the test, it is intended should be taught: overlap. What forms of test preparation would be appropriate and how can they be encouraged? What inappropriate forms of preparation might be expected and how might these be discouraged. Methods for investigating overlap would

Print ISSN: 1578-7044; Online ISSN: 1989-6131

embrace most traditional forms of validity enquiry including content analysis and quantitative and qualitative post hoc approaches. These will need to be supplemented by investigation of what has been called 'face validity' or the ways in which key participants view the test and interpret test demands. This is because it is the participants' perspective rather than the test developer's that is more likely to determine washback. Discrepancies between perceptions of overlap on the part of different stakeholders may help to explain washback findings and inform the development of procedures for improving the effects of a test on related educational systems.

From the test provider's standpoint, sound test design is the starting point for encouraging behaviours that are compatible with the aims of the test. Wide and unpredictable sampling of target skills in the test may, as Hughes (1989) suggested, encourage teachers and learners to cover a wide range of skills in their classes. This will be more likely to come about if the test provider informs participants about test content, publicises the theoretical basis for the test and trains teachers in effective forms of preparation. Provision of feedback on test performance with suggestions on ways of developing targeted skills can also help teachers to focus on developing these abilities in their students. There would seem to be particular benefits in involving teachers in test design and development and in communicating the aims of the test to their colleagues.

5.2. Washback variability and intensity

As apparent preconditions for washback, participant knowledge of test demands, beliefs regarding the value of success and assessments of the level of challenge posed must be investigated, whether as a preliminary phase of validation research or as necessary background to a testing innovation. What is needed is a rich understanding of the role of the test in its social context or contexts, especially where a test is to be used in different settings. Issues of particular relevance may include

- Setting: Who are the key participants in the context where the test will be used? What investment do they have in the decisions associated with the test? What roles do tests perform within the local culture?
- Test use: Is the test equally valued by participants? What stakes are associated with test success? How difficult is the test perceived to be? Are alternatives available to test takers?
- Beliefs about teaching and learning: What do teachers and learners believe to be effective strategies for learning a language? Do they see these beliefs as compatible with the demands of the test? What pressures exist to encourage test preparation practices? What local precedents exist for approaches to test preparation?
- Knowledge of the test: How much do the participants actually know about the test? What misconceptions do they have?

Print ISSN: 1578-7044; Online ISSN: 1989-6131

- Resources: What resources do participants have to prepare for the test? What resources are they prepared to commit to bring about success? Do teachers have enough training in the requisite language skills and teaching methodologies? What materials exist to support test preparation?
- Beliefs about testing: What other tests and assessments are participants familiar with? How do they respond to the use of tests? What part have tests played in their lives?
- Interactions between participants: How do participants learn about the test? What information do they pass on to other participants? How do other participants encourage them to prepare for the test?

Evidence relating to issues of this kind accessed from documentary sources, surveys and interviews should help test developers to form a picture of how washback is likely to manifest itself in a local setting, probable sources of variability and differences between settings. There are many possible theories that may provide useful frameworks for this aspect of washback research, although these will inevitably employ theoretical tools beyond those traditionally used in language testing research.

5.3. Accessing participant attitudes and processes

Although surveys and interview methods can certainly provide insights into how participants *believe* they have been affected by a test, Wall and Alderson (1993: 65) argued that direct observation of behaviour in the classroom is also needed to inform interview and questionnaire design and contextualise otherwise incomprehensible responses. Analysis of documents such as textbooks, teacher devised materials, assessment records and student portfolios can provide further evidence of teaching and learning practices.

The timing of any investigation is another important consideration. Washback effects are likely to be greatest when the test date is approaching. Effects may also develop over time as a test becomes embedded in n educational system. Longitudinal research is challenging, but a long term view is likely to be particularly useful in improving our understanding of washback and is crucial where innovation and educational change are involved.

Hughes' (1993) distinction between participants, processes and products offers a basis for deciding which aspects of washback should be the main focus for inquiry. Wall and Alderson (1993) argued the need for clarity in defining dependent variables in washback research and their fifteen washback hypotheses suggest predictions regarding content ('what'), methods ('how'), rate, sequence, degree and depth of teaching and learning as potential process variables for investigation. The distinction between *what* and *how* teachers teach and learners learn has been particularly influential, perhaps because it is easier to observe than rate, sequence, degree or depth. When faced with a test, do teachers teach the same content in different ways or different content using the same methods as in their non-test

Print ISSN: 1578-7044; Online ISSN: 1989-6131

classes? However, in future research a clearer basis needs to be found for deciding which aspects of participant behaviour should be categorised under each heading.

A promising line of inquiry that should shed light on the role of cultural difference might be to conduct washback studies across settings.

5.4. Accessing outcomes

Hughes (1993) argued that the 'ultimate washback objective' of an English language test will be 'the English skills that candidates develop' (p. 5). The measure of washback of greatest interest will be the extent to which criterion abilities improve as a result of test preparation. Wall (2000) acknowledges that "what is missing [in washback research] are analyses of test results which indicate whether learners have learned more or learned better because they have studied for a particular test" (p. 502). The reasons for the lack of consideration given to test results include the problems of comparing non-equivalent, often temporally distant groups and the selection of alternative outcome measures. Improvements in scores may imply no more than test wiseness. Robust designs will therefore include the use of at least one alternative measure of the skills under investigation. However, new tests are inevitably accompanied by other changes, making it difficult to establish the contribution made by the test. It can also be very challenging to find alternative measures of skills that do not suffer from the same limitations as the test under study.

When test data is combined with descriptions of test preparation practices, comparisons can more readily be drawn between those practices which result in increased test scores and those which do not. Where test scores improve in line with criterion abilities, judged by other measures, positive washback is implied. Where test scores improve, but criterion abilities do not, the washback is likely negative. Where preparation practices fail to boost either test scores or criterion abilities, we might look to other variables such as participant beliefs or availability of resources to explain the outcomes.

6. CONCLUSIONS: IMPLICATIONS OF WASHBACK RESEARCH

Washback research has suggested that the issue facing educators is not so much the influence of tests *on* teaching and learning, as an interaction *between* tests, teaching and learning. This interaction has as many implications for educational administration, text book development, teacher training and resourcing as for test development and revision.

The identification of needs in relation to communication between test providers and other stakeholders is one likely outcome of researching washback. As research reveals the ways in which participants understand a test and their beliefs about what is required to perform well, test developers and other stakeholders can work to address the issues that emerge. Greater involvement of administrators, textbook writers, teachers, and even learners

Print ISSN: 1578-7044; Online ISSN: 1989-6131

in test development processes may help to improve the coherence and integration of testing and teaching.

Better understanding of how washback occurs in teaching and learning processes can help to inform targeted intervention. If, for example, teachers are failing to integrate speaking activities into their classes in response to the introduction of speaking tests, causes can be sought in test design (is too little weight given to speaking skills? is the speaking section too easy?) or in pedagogic systems (do teachers lack training in teaching speaking?). If causes are correctly identified, suitable changes can be introduced (test revision, teacher training).

Research evidence can be a powerful tool for encouraging participants to reconsider their current practices. If it can be demonstrated that a fixation with test formats not only leads to tedious and repetitive classroom activities, but also that these are less effective than more interesting and engaging activities at improving test scores, teachers and learners may be less resistant to adopting new approaches. If, on the other hand, teaching to the test is educationally problematic, but successful at improving scores, this would call the test's validity into question and would suggest to the developers and users that the test may need reform.

Washback research has given us some new insights into how tests are used and how they are accommodated in a wide range of educational settings. Major projects instigated by large testing organisations such as IELTS and TOEFL have established washback research as an important element in building arguments to support assessment use. As part of a growing concern for the consequences of testing and social impacts, the investigation of washback is now well embedded among routine validation activities.

It is very clear that washback, like other forms of evidence in our field, has to be considered in relation to specific contexts of test use. This is because local factors can interact with tests to bring about very different effects. We have learned a good deal about teacher perceptions and practices and are beginning to understand some of the reasons for differences between individual teachers in the kinds of effects they experience. On the other hand, we still understand rather less about the roles of other participants such as course leaders, policy makers, textbook writers and even learners (perhaps the most important participants of all). Clearer lines of evidence are now needed linking particular practices to test characteristics and linking what is done in class to test scores and other learning outcomes.

REFERENCES

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. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

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

Burrows, C. (1998). Searching for Washback: An investigation of the impact on teachers of the implementation into the Adult Migrant English Program of the assessment of the Certificates in Spoken and Written English. Unpublished PhD thesis, Macquarie University.

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 Contexts and Methods* (pp. 113-128). Mahwah, NJ: Lawrence Erlbaum Associates.

Chapman, D. W., & Synder, C. W. (2000). Can high-stakes national testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development, 20*, 457-474.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study. Studies in language testing, 21*. Cambridge: Cambridge University Press.

Cheng, L., Watanabe, Y., with Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L.J. (1963). Course improvements through evaluation. *Teachers College Record, 64*, 672-683.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 43-481.

Entwistle, N. (1988). *Styles of Learning and Teaching*. London, David Fulton

Fullan, M. (2001). *The New Meaning of Educational Change* (3rd ed.). London: Cassell.

Green, A. (2007) *IELTS Washback in Context: Preparation for academic writing in higher education. Studies in Language Testing 25.* Cambridge: Cambridge University Press.

Henrichsen, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956–1968*. New York: Greenwood Press.

Herbert, A. (Ed.). (1889). *The Sacrifice of Education to Examination: Letters from All Sorts and Conditions of Men*. London, Williams & Norgate.

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). Modern English Publications in association with the British Council.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

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

Khaniya, T. R. (1990). The washback effect of a textbook-based test. *Edinburgh Working Papers in Applied Linguistics, 1*, 48-58.

Madaus, G.F. (1988). The influence of testing on the curriculum. In L. Travers (Ed.), *Critical Issues in Curriculum* (pp. 83-121). Chicago: Chicago University Press.

Mehrens, W.A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13), 1-30.

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

Mickan, P. & Motteram, J. (2008) An ethnographic study of classroom instruction in an IELTS preparation program. In IELTS Research Reports, Volume 8. Canberra: IELTS Australia.

Mickan, P. & Motteram, J. (2010) The preparation practices of IELTS candidates: Case studies In IELTS Research Reports, Volume 10. Canberra: IELTS Australia

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in Language Testing* (pp. 1-13). London: NFER/Nelson.

Pearson, I. (1988). Tests as levers of change (or 'putting first things first'). In D. Chamberlain & R. Baumgartner (Eds.), *ESP in the classroom: Practice and Evaluation* (pp. 98-107). London: Modem English Publications in association with the British Council.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stake test. *Language Testing, 22*(2), 142-173.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In Gifford, B.G. and O'Conner, M.C. (Eds.), *Changing Assessments: Alternative views of Aptitude, Achievement and Instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

Shih, C.-M. (2007). A new washback model of students' learning. *Canadian Modern Language Review, 64*(1), 135-162.

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

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*(3), 334-354.

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*. *Studies in Language Testing 22*. Cambridge, UK: Cambridge University Press.

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. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 129-146). Mahwah, NJ: Lawrence Erlbaum Associates.

Print ISSN: 1578-7044; Online ISSN: 1989-6131