Washback, Impact, and Consequences Revisited

Dina Tsagari and Liying Cheng

Abstract

Washback, impact, and consequences refer to the educational phenomenon when testing (often large-scale and high-stakes), specifically the uses of test scores and the decisions made based on those scores, influence those stakeholders associated with such testing. Washback, impact, and consequences are used in different fields of research, and these terms encompass different dimensions of the research undertaken. *Washback* is more frequently used to refer to the effects of tests on teaching and learning at the classroom level. *Impact* refers to the effects that a test may have on individuals, policies, or practices, within the classroom, the school, the educational system, or the society as a whole. Many language testers these days consider *washback* as a dimension of *impact*. The effects of testing on teaching and learning have been traditionally associated with test validity (*consequential validity*) where washback is considered as only one form of testing *consequences* that need to be weighted in evaluating validity. This chapter elaborates the origins and dimensions of these terms by presenting the major empirical studies conducted over the past 30 years. Considering the complexity of this educational phenomenon and increasing importance of the testing effects in education and beyond, the authors present the challenges facing such research and point out the directions that future research in this area could embrace.

Keywords

Washback (backwash) • Impact • Consequences • Teaching • Learning • Consequences of tests • Effects of tests - Large scale high-stakes testing • Tests •

D. Tsagari (✉)
Department of English Studies, University of Cyprus, Nicosia, Cyprus
e-mail: dinatsa@ucy.ac.cy

L. Cheng
Faculty of Education, Queen's University, Kingston, ON, Canada
e-mail: liying.cheng@queensu.ca

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language
and Education, DOI 10.1007/978-3-319-02261-1_24

Examinations • Teaching and learning • Validity (consequential validity) • Measurement-driven instruction • Test-curriculum alignment • Ethics • Fairness

Contents

Introduction 360
Early Developments 361
Major Contributions 362
Work in Progress 365
Problems and Difficulties 366
Future Directions 367
Cross-References 369
Related Articles in the Encyclopedia of Language and Education 370
References 370

Introduction

It is accepted nowadays that "testing has become big business" (Spolsky 2008, p. 297) and that it plays a powerful role in education, politics, and society in general (McNamara and Shohamy 2008). High-stakes large-scale testing in particular "is never a neutral process and always has consequences" for its stakeholders (Stobart 2003, p. 140), intended or unintended, and positive or negative.

In the long and substantial amount of research conducted in general education, researchers refer to the phenomenon as *measurement-driven instruction* (Popham 1987), *test-curriculum alignment* (Shepard 1990), and *consequences* (Cizek 2001). By contrast, in language education, test consequences are a relatively new concept since the late 1980s. The two terms commonly used in the field are *impact* and *washback*. Wall (1997) defines *impact* as "any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole". She also points out that "*washback* (also known as *backwash*) is sometimes used as a synonym of impact, but it is more frequently used to refer to the effects of tests on teaching and learning" (p. 291) at the classroom level. Many language testers these days consider *washback* as a dimension of *impact* (e.g., Hamp-Lyons 1997).

Primarily, the effects of testing on teaching and learning have been associated with test validity (*consequential validity*) where Messick refers to washback as "only one form of testing consequences that need to be weighted in evaluating validity" (Messick 1996, p. 243). He stresses the need for the examination of two threats to test validity, *construct under-representation* and *construct-irrelevant variance*, to decide the possible consequences that a test can have on teaching and learning. Bachman (2005) proposes a framework with a set of principles and procedures for linking test scores and score-based inferences to test use and the consequences of test use. Other contemporary validity theories (Chalhoub-Deville 2015; Chapelle et al. 2010; Kane 2013, 2016) adopting the various argument-based models have established grounds for the inclusion of test consequences and uses within validation studies. These theories require the systematic collection of validity evidence at each validation stage and from multiple stakeholder perspectives to better justify the use of test scores in pedagogical and policy practices.

In addition, the effects of testing on teaching and learning are increasingly discussed from the point of view of critical language testing, including ethics and fairness in language testing, all of which are expressions of social concern. For example, Shohamy (2001) points out the political uses and abuses of language tests and called for examining the hidden agendas of the testing industry and of high-stakes tests. Kunnan (2000, 2004) discusses the role of tests as instruments of social policy and control. He also draws on research in ethics to link validity and consequences and created a test fairness framework. Hamp-Lyons (1997) argues for an encompassing ethics framework to examine the consequences of testing on language learning at the classroom, as well as the educational, social, and political levels. All of the above has led to the creation of a Code of Ethics for the International Language Testing Association (see Davies 2008).

Early Developments

The work of Alderson and Wall (1993) marked a significant development in shaping the constructs of washback studies for the field of language testing. The authors explored the potential positive and negative relationship between testing, teaching, and learning, and questioned whether washback could be a property of test validity. They consequently proposed 15 hypotheses (revisited and refined in Alderson and Hamp-Lyons 1996) regarding the potential influence of language testing on various aspects of language teaching and learning, which thus directed washback studies for years to come. The study of Wall and Alderson (1993) was the first empirical research published in the field of language testing. It investigated the nature of washback of a newly introduced national English examination in Sri Lanka by observing what was happening inside the classroom.

A review of the early literature, as pointed out by Cheng (2008), indicates at least two major types of washback studies. First there are those relating to traditional, multiple-choice, large-scale standardized tests; these are perceived to have had mainly negative influences on the quality of teaching and learning. Secondly there are those studies where a specific test or examination has been modified and improved upon (e.g., assessment with more communicative tasks: see Cheng 2005) in order to exert a positive influence on teaching and learning.

In 1996, a special issue in *Language Testing* published a series of articles that further explored the nature of washback and empirically investigated the relationship between testing, teaching, and learning. In this volume, Messick (1996) suggested building on validity considerations through test design in order to promote positive washback and to avoid construct under-representation and construct-irrelevant variance. Although Messick did not specify how researchers could go about studying washback through test design validation, he pointed out that test washback could be associated with test property. He thus offered a coherent argument to investigate the factors in testing that are related to factors in teaching and learning. Bailey (1996, p. 268), however, argued that any test, whether good or bad in terms of validity, can have either negative or positive washback "to the extent that it promotes or impedes the accomplishment of educational goals held by learners and/or program personnel". Her argument indicated that washback effects (positive or negative) might differ for different groups of stakeholders. Finally, Wall (1996) stressed the difficulties in finding explanations of how tests exert influence on teaching, and turned to innovation theory to offer "insights into why attempts to introduce change in the classroom are often not as effective as their designers hoped they would be" (p. 334).

Three empirical research studies are also reported in the same special issue. Alderson and Hamp-Lyons (1996) found that the Test of English as a Foreign Language (TOEFL) affects both what and how teachers teach, but the effect is not the same in degree or kind from teacher to teacher. Watanabe (1996) found that teacher factors, including personal beliefs, past education, and academic background, seemed to be more important in determining the methodology a teacher employs rather than the university entrance examination in Japan. Shohamy et al. (1996) added that the degree of impact of a test is often influenced by several other factors: the status of the subject matter tested, the nature of the test (low or high stakes), the uses to which the test scores are put and that the washback effect may change over time.

In summary, testing may be only one of those factors that "affect how innovations [through testing] succeed or fail and that influence teacher (and pupil) behaviors" (Wall and Alderson 1993, p. 68). The special issue editors of the volume also call for the "need for co-ordinated research into washback and other forms of impact, and for a theory which will guide testers so that they have the best chance of influencing teaching, education and society in a positive way" (Alderson and Wall 1996, p. 240). Indeed the years since the 1996 special issue in *Language Testing* have seen a flurry of publications ranging from collections of empirical studies, doctoral theses, and research projects investigating different tests within different teaching and learning contexts. These will be presented in the following sections.

Major Contributions

Two edited volumes that have become the cornerstone collection of washback studies and initial attempts to capture the essence of washback have been published in the 2000s. The first one was the publication of Cheng and Watanabe, with Curtis's *Washback in Language Testing: Research Context and Methods* (2004). Through its compilation of washback studies, the book responded to the question "what does washback look like?" – a step further from the question "does washback exist?" posed by Alderson and Wall (1993). In its first section, the volume highlights the concept and nature of washback by providing a historical review of the phenomenon (Cheng and Curtis 2004); the second section showcases a range of studies on various aspects of teaching and learning conducted in many parts of the world, e.g., Australia, China, Hong Kong, Israel, Japan, New Zealand, the UK, and the USA. The book has contributed to our understanding of washback and impact of language tests in that we can no longer take for granted that where there is a test, there is a direct effect.

The second publication was the special issue dedicated to investigating washback in language testing and assessment published in *Assessment in Education: Principles, Policy and Practice* in 2007. The editors, Rea-Dickins and Scott (2007), brought together papers from equally varied contexts that looked at washback areas such as the consequences of large-scale school tests on different groups of learners; the effects of a statutory national assessment curriculum on primary school learners; a specific writing task of a high-stakes test on secondary school students; three different program types on the development of students' writing skills; and the impact of the International English Language Testing System (IELTS) preparation classes on improving student scores on the writing sub-test of the test. The papers included also problematize on the selection of appropriate methodologies for researching washback and on how language tests are used as a mechanism in the manipulation of language education policies and policy control. The volume has added to our understanding of washback as being context-specific, unstable, and difficult to predict, and makes a call for greater dialogue between language and education researchers.

Several major doctoral studies have also made a substantial contribution to the understanding of the complexity of washback and offered methodological implications for washback studies over the years. For example, the longitudinal study by Wall (2005) documents research examining one of the widely held beliefs that change can be created in an education system by introducing or by re-designing high-stakes examinations. Wall analyzed the effects of a national examination in English as a Foreign Language in Sri Lanka that was meant to serve as a lever for change. Her study illustrated how the intended outcome was altered by factors in the exam itself, as well as the characteristics of the educational setting, the teachers, and the learners. Her study, located in the interface of examination impact and innovation in education, provided guidelines for the consideration of educators who continue to believe in the potential of examinations to affect curriculum change.

Through a large-scale, three-phase study using multiple methods to explore the multivariate nature of washback, Cheng (2005) investigated the impact of the Hong Kong Certificate of Education in English (HKCEE) on the classroom teaching and learning of English in Hong Kong secondary schools where the examination is used as the change agent. The washback effect of this public examination change was observed initially at the macro level, including various stakeholders within the local educational context, and subsequently at the micro level, including aspects of teachers' and learners' attitudes, teaching contents, and classroom interaction. The findings indicated that the washback effect of the new examination on classroom teaching was limited despite expectations to the contrary. Her study supports the findings of Wall and Alderson (1993), i.e., that the change of the examination can inform what teachers teach, but not how they teach.

Green (2007) used a variety of data collection methods and analytical techniques to explore the complex relationship between teaching and learning processes and their outcomes. Green evaluated the role of *IELTS* in English for Academic Purposes (EAP) particularly in relation to the length of time and amount of language support needed by learners to meet minimally acceptable standards for English-medium tertiary study. This piece of research is of relevance to a range of interested parties concerned with the development of EAP writing skills.

In her study, Tsagari (2009) explored the washback of First Certificate in English (FCE, offered by the formerly known Cambridge ESOL) on the teaching and learning that takes place in intermediate level EFL classes in Greece. The study followed a mixed-method design to data collection and analysis, e.g., interviews, teaching, exam-preparation materials and student diaries. The findings showed that many other factors beyond the test, the teachers or students (e.g., publishers/authors, the school, and the educational context) need to be taken into account when studying the washback effect of a high-stakes exam to explain why washback takes the form it does in a given context. The study led to a comprehensive model of exam washback and suggestions for teachers, teacher trainers, students, material and test developers, as well as future researchers in the area.

Many more doctoral level washback studies have been conducted over the years, which add to our understanding of the complex nature of washback and impact. These studies have been conducted in various contexts investigating the influence of testing on teachers and teaching, textbooks, learners and learning, attitudes toward testing, classroom conditions, recourse provision and management practices within the school, the status of the subject being tested in the curriculum, feedback mechanisms between the testing agency and the school, and the general social and political context. The studies have also focused on the influence of national examinations in countries such as Brazil, Canada, China, Egypt, Hong Kong, Iran, Israel, Japan, Spain, Taiwan, and the UK. Others have also looked at worldwide English testing such as IELTS, TOEFL, TOEIC, Cambridge Young Learners English test series, and the Michigan Examination for Certificate of Competency. For a review of washback and impact doctoral studies, see Cheng and Fox (2013) and Cheng et al. (2015).

The research output of the above studies shows that washback is a highly complex phenomenon due to the fact that it is an interactive multidirectional process involving a constant interplay of varying degrees of complexity among the different washback components and participants. Also the above studies have shown that simply changing the contents or methods of an examination will not necessarily bring about direct and desirable changes in teaching and learning. Rather various factors within educational contexts are involved in engineering desirable washback, e.g., test factors (test methods, test contents, skills tested, purpose(s) of the test), prestige factors (stakes of the test, status of the test), personal factors (teachers' educational backgrounds and their beliefs), micro-context factors (the school/university setting), and macro-context factors (the specific society in which the tests are used) (Cheng and Curtis 2004). However, questions remain about the nature of factors and stakeholders involved, the interaction between them and the conditions under which beneficial washback is most likely to be generated.

Work in Progress

The interest in test washback and impact continues to grow, as evidenced in major conference presentations such as Language Testing Research Colloquium (LTRC) and publications in journals such as *Language Testing* and *Language Assessment Quarterly* and edited volumes or monographs.

Several important projects have been commissioned by major testing agencies and have increasingly played a major role in producing clusters of washback and impact studies. These studies are conducted in many countries around the world on the same test and tend to be large-scale and multi-faceted. They offer important recommendations for the improvement of the tests under study and directions for future research. For instance, findings of funded research studies, such as those conducted by Cambridge English language assessments, report on the impact of examinations at the micro (teaching and learning) and at macro levels (employability, schools, parents, and decision makers) in countries such as Cyprus, Greece, Japan, Romania and Spain. Long-term research on IELTS has been implemented through the Cambridge ESOL Research & Validation Group. Most of these studies are also collaborative in nature, which indicates the importance of working with local experts, and employ mixed-method designs. The results are published online via Research Notes, RN (http://www.cambridgeenglish.org/research-and-validation/published-research/research-notes/) and other publications (Hawkey 2006). This flurry of research has resulted in a richer and more informative picture of washback and impact, and a more in-depth understanding of the current state of English language teaching, learning, and assessment within the particular contexts.

Educational Testing Services (ETS) has also funded a series of studies examining the impact of the TOEFL test. For example, the TOEFL Impact Study in Central and Eastern Europe (Wall and Horák 2006, 2008, 2011) investigated whether the new TOEFL iBT contributed to changes in teaching and learning after its introduction. This study involved three research stages: Phase 1: a “baseline study” described the type of teaching and learning taking place in commercial language teaching operations before details of the test were released about the content and format of the new test; Phase 2: a “transition study” traced the reactions of teachers and teaching institutions to the news that was released about the TOEFL iBT and the arrangements made for new preparation courses; Phase 3 investigated whether textbooks published accurately reflected the new test and what use teachers make of them in the classroom. Data was collected via computer-mediated communication with informants providing responses to the activities in their classrooms and institutions and reactions to tasks, which had been designed to probe their understanding of the new test construct and format (see also Hamp-Lyons and Brown 2007; Tsagari 2012).

Funded by the Social Sciences and Humanities Research Council of Canada (SSHRC), Cheng and colleagues' large collaborative study on *Test Preparation: Does It Enhance Test Performance and English Language Proficienc*y (http://educ.queensu.ca/test-prep) is a multiphase and multiyear investigation into the relationship of test preparation and test performance (Cheng and Doe 2013). This study is conducted in partnership with major test agencies and various stakeholders: test developers, teachers, students, test preparation center administrators and staff, and university admissions officers. The researchers conducted case studies of test preparation courses in Australia, Canada, China and Iran linking students' test preparation practices to their test performance (Ma and Cheng 2016; Saif et al. 2015). The study findings will provide test-designers and test users with empirical evidence regarding the predominant phenomenon of test preparation and the validity of test scores.

Problems and Difficulties

Although there have been increasing numbers of empirical washback and impact studies conducted since the late 1980s, researchers in the field of language education continue to wrestle with the nature of washback, and to research ways to induce positive and reduce the negative washback and impact of language tests. As indicated above, washback is one dimension of the consequences of the testing on classroom teaching and learning, and impact studies include broader effects of testing (as defined in Wall 1997). However, both assume a causal relationship between testing, teaching, and learning which has not been established up to now. Most of the washback and impact empirical studies have only established an exploratory relationship. In many cases, we cannot be confident that certain aspects of teaching and learning perceptions and behaviors are the direct causal effects of testing. They could well be within certain contexts, but this relationship has not yet been fully disentangled.

Furthermore, apart from the studies on IELTS, TOEFL, and large collaborative studies, e.g., Cheng and colleagues' study on test preparation, where a worldwide test influences teachers and learners across countries and educational contexts, the majority of the empirical studies focuses on the effects of one single test, within one educational context using research instruments designed specifically for that particular study. The strength of such studies is that they have investigated factors that affect the *intensity of washback* (Cheng and Curtis 2004). In fact, many of the factors related with the influence of testing on teaching and learning illustrated in Wall (2000) have been empirically studied. However, not only does little overlap exist among the studies regarding what factors affect washback, but little overlap also exists in researcher reports of the negative and positive aspects of washback (Brown 1997). In addition, there does not seem to be an overall agreement on which factors affect the intensity of washback and which factors promote positive or negative washback. This is a challenging feature of washback and impact studies, since researchers set out to investigate a very complex relationship (causal or exploratory) among testing, and teaching and learning.

This complexity causes problems and difficulties in washback and impact research, which in turn challenges any researcher who wishes to conduct, is conducting, or has conducted such studies. In many ways, the nature of such washback and impact study requires subtle, refined, and sophisticated research skills in disentangling this relationship. Researchers need to understand the specificity, intensity, length, intentionality, and value of test washback/impact and how (or where and when) to observe the salient aspects of teaching and learning that are potentially influenced by the test. They also need to identify their own bias, analyze the particular test and its context, and produce the predications of what washback and impact looks like prior to the design and conduct of the study (see also Watanabe 2004). Washback and impact studies are, by definition, studies of program evaluation, which require researchers not only to understand but also to make a value judgment about the local educational context as well as the larger social, political, and economic factors governing teaching and learning in relation to a test/examination or a testing system. Researchers need to acquire both the breadth and depth of necessary research skills to avoid research based on investigating random factors of teaching and learning, which may or may not have a direct relationship with testing.

Future Directions

It is clear that the future direction of washback and impact studies to investigate the consequences of language testing need to be multi-phase, multi-method, and longitudinal in nature. Washback and impact of testing take time to evolve, therefore longitudinal studies are essential with repeated observations (and measures) of the classroom teaching, including teachers and students as well as policy, curriculum, and assessment documents. Also, researchers need to have very good knowledge and understanding of the test they investigate, work collaboratively with the test developers and be well-immersed in the educational system they investigate, interacting with a wide range of stakeholders. In addition, researchers should pay attention to the seasonality of the phenomenon, i.e., the timing of researchers' observations may influence what we discover about washback (Bailey 1996; Cheng 2005) avoiding potential bias. Examples like the IELTS impact studies and the impact studies on TOEFL iBT across different countries and continents over a few years have a great deal to contribute to our understanding of this complex phenomenon. Studies of a single test within an individual context by a single researcher can still offer valuable insights for that particular context; however, it would be best if groups of researchers could work collaboratively and cooperatively to carry out a series of studies around the same test within the same or across educational contexts. The findings of such research could then be cross-referenced to portray a more accurate picture of the effects of the test avoiding the "blind men and elephant" syndrome. Research studies need also to move from the micro-level of the classroom (washback) to the macro-level of society (impact), to analyze the social factors that lead to assessment practices in the first place, and to explain why assessment practices (large-scale testing) are valued more than others. Such studies also need to link the (mis-)uses of test scores with what happens inside and outside the confines of the classrooms.

In addition, the methodology (and the methods) used to conduct washback studies need to be further refined. For example, researchers need to vary their methods including mixed method explanatory, exploratory, and concurrent design. Also, more sophisticated data collection and analysis methods, e.g., those linking directly with test-takers' characteristics, perceptions of assessments, learning processes, and their learning outcomes (test performance) need to be employed beyond classroom observations and survey methods (interviews and questionnaires) (e.g., Xie and Andrews 2013). Building on the increasing numbers of studies carried out on the same test or within the same educational context, future researchers can replicate or refine instruments and analysis procedures, which was not possible in the past. The replication would allow researchers to build on what we have learned theoretically, conceptually, and methodologically over the years and further our understanding of this phenomenon.

While it would be useful to continue to study the effects of tests on broad aspects of teaching, it is essential to turn our attention to investigate the effects on students and their learning as they receive the most direct impact of testing. In other words, what has not been focused on in previous studies is the direct influence of testing on learners (e.g., their perceptions, their strategy use, motivation, anxiety, and affect), on their learning processes (e.g., what and how they learn, or how they perform on a test including test-taking), and learning outcomes (test scores or other outcome measures). Based on these investigations, it is also important to use the results to do in-depth observations of students. For example, it is crucial to study students' understanding of the test constructs especially in the public examination domains where test-related information may not be directly accessible to students. In addition, it is important to study factors that are likely to be shaped by the learning and wider societal context. Other than students' perceptions of a test, it is also important to examine how they obtain such knowledge. This type of research can directly link the consequences of testing with test validity. It would be also worthwhile to look at the test taker population more closely, e.g., the educational characteristics (in terms of learning and testing) of the students. We know by now that high-stakes testing like IELTS or TOEFL influences students. However, is the impact of the test different on students learning English in one country than in another where the educational tradition (beliefs and values) are different? Without a thorough understanding of where these students come from and the characteristics they bring to their learning and testing, it is unlikely that we can fully understand the nature of test washback and impact.

Research also needs to be directed towards the relationship between high stakes large-scale testing and classroom-based teacher-led formative assessment (Tsagari and Banerjee 2014). Research in this area can better inform teachers for their curriculum planning and instruction and can better support student learning, making ongoing teacher involvement a part of test development and validation process

(Froetscher 2016). Another fruitful area of research is the investigation of "language assessment literacy" – LAL (Fulcher 2012, p. 125) of high-stakes test users, e.g., teachers (Vogt and Tsagari 2014), university admissions officers (O' Loughlin 2013), test writers, and professional language testers (Harding and Kremmel 2016). It is important to understand the degrees of LAL needed for different stakeholder groups in high-stakes test contexts as this can induce positive consequences from tests. This is an exciting research venue that will be able to attest to the urgent and largely unrecognized need in many high-stakes educational and policy-making contexts for increasing teacher development opportunities (Taylor 2013).

An additional area that lacks empirical research is washback on stakeholders outside the immediate confines of the classroom, e.g., parents, who tend to be neglected, but take important instructional decisions about learners outside of their school time. In addition, given that assessment is located in the social context, empirical studies of indirect participants – such as public media, language accreditation systems, employers and policy makers – and their perceptions and understanding of high-stakes tests and use of test scores are needed, as these will add greater importance to the washback phenomenon and unveil different degrees of complexity.

Finally it remains controversial in educational assessment research whether and how consequences should be integrated in test validation or even whether they belong to test validation or not (Messick 1989; Moss 1998; Nichols and Williams 2009). A few studies in the field of language assessment have systematically investigated test consequences within a coherent validation framework to examine evidence for the purpose of evaluating the strength of the validity argument (including consequential validity) of a particular test in a given situation (Chapelle et al. 2010; Chalhoub-Deville 2015). In the end, washback and impact researchers need to fully analyze the test under study and understand its test use. Bachman (2005, p. 7) states that "the extensive research on validity and validation has tended to ignore test use, on the one hand, while discussions of test use and consequences have tended to ignore validity, on the other". It is, then, essential to establish the link between test validity and test consequences theoretically and empirically. It is imperative that washback and impact researchers work together with other language testing researchers, as well as educational policy makers and test agencies, to address the issue of validity, in particular, fairness and ethics of language tests.

Cross-References

▶ Critical Language Testing
▶ Ethics, Professionalism, Rights, and Codes
▶ Language Assessment Literacy
▶ Training in Language Assessment

Related Articles in the Encyclopedia of Language and Education

Linda von Hoene: The Professional Development of Foreign Language Instructors in Postsecondary Education. In Volume: Second and Foreign Language Education

Bonny Norton, Ron David: Identity, Language Learning and Critical Pedagogies in Digital Times. In Volume: Language Awareness and Multilingualism

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A case study. *Language Testing, 13*, 280–297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.

Alderson, J. C., & Wall, D. (1996). Editorial. *Language Testing, 13*, 239–240.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

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

Brown, J. D. (1997). Do tests washback on the language classroom? *The TESOLANZ Journal, 5*, 63–80.

Chalhoub-Deville, M. (2015). Validity theory: Reform policies, accountability, testing, and consequences. *Language Testing*. doi:10.1177/0265532215593312.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

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

Cheng, L. (2008). Washback, impact and consequences. In N. H. Hornberger (Series Ed.), *Encyclopedia of language and education* (Language testing and assessment, 2nd ed., Vol. 7, pp. 349–364). New York: Springer.

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. 3–18). Mahwah: Lawrence Erlbaum Associates.

Cheng, L., & Doe, C. (2013). "Test preparation: A double-edged sword", *IATEFL-TEASIG (International Association of Teachers of English as a Foreign Language's Testing, Evaluation and Assessment Special Interest Group). Newsletter, 54*, 19–20.

Cheng, L., & Fox, J. (2013). Review of doctoral research in language assessment in Canada (2006–2011). *Language Teaching, 46*, 518–544.

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

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*, 436–470.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 23*(3), 1–17.

Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy, N. H. Hornberger (Eds.), *Encyclopedia of language and education*. (Language testing and assessment, 3rd ed., Vol. 7, pp. 429–444). New York: Springer.

Froetscher, D. (2016). A new national exam: A case of washback. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 61–81). London: Continuum.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge, UK: Cambridge University Press.

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

Hamp-Lyons, L., & Brown, A. (2007). *The effect of changes in the new TOEFL format on the teaching and learning of EFL/ESL: Stage 2 (2003–2005): Entering innovation.* Report submitted to the TOEFL research committee, Educational Testing Service.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 413–427). Berlin/New York: Muton De Gruyter.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.

Kane, T. M. (2013). Validating the interpretations and uses of test scores. *Educational Testing Service Journal of Educational Measurement, 50*(1), 1–73.

Kane, T. M. (2016). Explicating validity. *Assessment in Education: Principles Policy and Practice, 23*(2), 198–211.

Kunnan, A. J. (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium Orlando Florida*. Cambridge: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic, C. Weir, & S. Bolton (Eds.), *Europe language testing in a global context: Selected papers from the ALTE conference in Barcelona* (pp. 27–48). Cambridge: Cambridge University Press.

Ma, J., & Cheng, L. (2016). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth and significance. *TESL Canada Journal, 33*(1), 58–79 . http://www.teslcanadajournal.ca/index.php/tesl/article/view/1227.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics, 18*(1), 89–95.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 243–256.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28*(1), 3–9.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363–380.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappa, 68*, 679–682.

Rea-Dickens, P. & Scott, C.: (2007). Investigating washback in language testing and assessment' (Special issue). *Assessment in Education: Principles, Policy and Practice*, *14*(1), 1–7.

Saif, S., Cheng, L., & Rahimi, M. (2015). *High-stakes test preparation programs and learning outcomes: A context-specific study of learners' performance on IELTS*. Paper presented at the 37th annual language testing research colloquium. Toronto, 16–20 Mar 2015.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*, 15–22.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex: Longman.

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

Spolsky, B. (2008). Language testing at 25: Maturity and responsibility. *Language Testing, 25*(3), 297–305.

Stobart, G. (2003). The impact of assessment: Intended and unintended consequences. *Assessment in Education: Principles, Policy and Practice, 16*, 139–140.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412.

Tsagari, D. (2009). *The complexity of test washback: An empirical study.* Frankfurt am Main: Peter Lang GmbH.

Tsagari, D. (2012). *The influence of the Examination for the Certificate of Proficiency in English (ECPE) on Test Preparation materials*. Internal report sponsored by the SPAAN fellowship for studies in Second or Foreign Language Assessment, Cambridge Michigan Language Assessments (CaMLA), Ann Arbor.

Tsagari, D., & Banerjee, J. (2014). Language assessment in the educational context. In M. Bigelow & J. Ennser-Kananen (Eds.), *Handbook of educational linguistics* (pp. 339–352). New York: Routledge/Taylor & Francis Group.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.

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

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (pp. 291–302). Dordrecht: Kluwer Academic Publications.

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: Cambridge University Press.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*, 41–69.

Wall, D., & Horák, 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; MS-15*. Report number: RR-06-18, TOEFL-MS-34. Princeton: Educational Testing Service, http://www.ets.org/Media/Research/pdf/RR-2006-2018.pdf

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL® examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change*. TOEFL iBT research report. TOEFLiBT-05*,* https://www.ets.org/Media/Research/pdf/RR-08-37.pdf

Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL® exam on teaching in a sample of countries in Europe: Phase 3, The role of the coursebook phase 4, describing change*. TOEFL iBT® research report TOEFL iBT-17, http://www.ets.org/Media/Research/pdf/RR-2011-2041.pdf

Watanabe, Y. (2004). Methodology in Washback Studies. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 19–36). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*, 318–333.

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation: Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49–70.