CSCanada
Studies in Literature and Language
Vol. 13, No. 4, 2016, pp. 6-14
DOI:10.3968/8877

ISSN 1923-1555[Print]
ISSN 1923-1563[Online]
www.cscanada.net
www.cscanada.org

An Overview of Studies Conducted on Washback, Impact and Validity

Forough Rahimi[a]; Mohammad Reza Esfandiari[b],*; Mansour Amini[c]

[a]English Language Department, School of Allied Medical Sciences, Shahid Behehsti University of Medical Sciences, Tehran, Iran.
[b]Department of Foreign Languages, College of Humanities, Shiraz Branch, Islamic Azad University, Shiraz, Iran.
[c]Department of English Language and Literature, Azarbaijan Shahid Madani University, Iran.
*Corresponding author.

Received 15 July 2016; accepted 22 September 2016
Published online 26 October 2016

Abstract

This article aimed at presenting a comprehensive overview of three interrelated concepts of washback, impact and validity in language testing and a myriad of studies conducted at different places to investigate the influence of testing on teachers and teaching, textbooks, learners and learning, attitudes toward testing, test preparation behaviors, etc.. Some of these studies present the results of various investigations on the influence of a national English examination on the local English language teaching and learning due to its high-stakes nature in particular countries such as Brazil, China, Hong Kong, Iran, Israel, Japan, Romania, Sri Lanka, and Taiwan. Some others cover a wide range of worldwide investigation on English testing such as the IELTS, TOEFL, and MECC. Moreover, there is a complete report of several important projects appointed by major testing agencies such as Cambridge ESOL and Educational Testing Services (ETS) on washback and impact studies. The article proceeds by reviewing the relevant literature on test validation which is a key concept in language testing domain since it is concerned with test interpretation and use. This domain is characterized and enriched by studies of washback and impact.

Key words: Washback; Impact; Validity

Rahimi, F., Esfandiari, M. R., & Amini, M. (2016). An Overview of Studies Conducted on Washback, Impact and Validity. *Studies in Literature and Language*, *13*(4), 6-14. Available from: http://www.cscanada.net/index.php/sll/article/view/8877
DOI: http://dx.doi.org/10.3968/8877

INTRODUCTION

Tests are increasingly used throughout educational systems of most countries as a basis to make important score-based decisions about test takers. Testing tends to induce consequences for its participants because it is a way of differentiating among the individuals. Stobart (2003) believes that testing is not a neutral process and always has consequences for test takers.

Researchers are principally concerned that tests and their results may be used improperly to make interpretations and decisions, which may lead to unfair consequences to different groups of test takers. These concerns have increased social and educational demands, warranting the need to carry out rigorous research. Conducting research in this domain needs a deep understanding of the correlation that exists among the concepts of washback effect, test impact and test validity.

As a result, this review was targeted to accumulate the most prevalent and significant studies conducted on these three issues in order to highlight their interrelationship.

1. STUDIES ON TEST WASHBACK AND IMPACT

The concept of washback rooted in the notion that tests or examinations can and should drive teaching and hence learning (Pophem, 1987). The idea that testing influences teaching is familiar in the educational and applied linguistics literature. Many researchers have worked on the influence of examinations over the classroom practices. Cheng (2008) maintains that there is a set of intended and unintended, positive and negative relationships between testing, teaching and learning. Pearson (1988) believed that "public examinations influence the attitudes, behavior, and motivation of teachers, learners, and parents" (p.98). This influence is often seen as negative. Swain (1985) recommended that "test developers bias for test and work for washback" (p.42), while Alderson (1986) argued for "innovations in the language curriculum through innovations in language testing" (p.104).

Washback and impact of language testing is, however, a relatively new concept. The concept of measurement-driven instruction requires that testing should drive instruction. It focuses on the relationship between the content of tests and courses, which may lead to narrowing down the course instruction by teaching to the test. Tests may introduce intended or unintended and positive or negative aspects of instruction, students, teachers, and the school.

A great body of research has been conducted in language testing since the late 1980s (Alderson & Wall, 1993; Bailey, 1996; Wall, 1997). Wall (1997) describes 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 maintains that "washback is sometimes used as a synonym of impact, but it is often used to refer to the effects of tests on teaching and learning" (p.291). Some scholars suggest that washback is one dimension of impact (Bachman & Palmer, 1996; Hamp-Lyons, 1997). Hamp-Lyons (1997) believed that test influence would fall between the narrow one of washback and the all-encompassing one of impact.

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 believes that there are two threats to test validity, construct under-representation and construct-irrelevant variance, which affect the consequences that a test can have on teaching and learning.

Bachman (2005) suggests a framework consisting of some principles and procedures that connects test scores and score-based inferences to test use and its consequences. Moreover, the effects of testing on teaching and learning are studied from the standpoint of critical language testing (Shohamy, 2001) and ethics and fairness in language testing (Elder, 1997; Hamp-Lyons, 1997; Kunnan, 2000). Shohamy (2001) highlighted the political uses and abuses of language tests and claimed that there is a need to evaluate the hidden agendas of the testing. Kunnan (2000) discussed the role of tests as instruments of social policy and control. He also pointed to research in ethics which links validity and consequences and created a test fairness framework (Kunnan, 2004). Hamp-Lyons (1997) argued for a comprehensive 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 these created a Code of Ethics for the International Language Testing Association (Davies, 2003). The work of Alderson and Wall (1993) and Wall and Alderson (1993) promoted the constructs of washback studies for the field of language testing. Alderson and Wall (1993) 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 regarding the potential influence of language testing on various aspects of language teaching and learning. Wall and Alderson (1993) conducted the first empirical research on the nature of washback of a new national English examination in Sri Lanka by observing what was happening inside the classroom.

A review of the literature shows there are two major types of washback studies: those relating to traditional, multiple-choice, large-scale standardized tests, which have negative influences on the quality of teaching and learning (Shepard, 1990) and those studies where a specific test or examination has been modified and improved upon (Wall & Alderson, 1993) in order to employ a positive influence on teaching and learning (see also Cheng, 2005).

Bailey (1996, p.268) contended that any testcan have either negative or positive washback "to the extent that it promotes or impedes the accomplishment of educational goals held by learners or personnel." Her argument indicated that washback effects (positive or negative) might differ for different groups of stakeholders. Wall (1997) emphasized that it is difficult to show how tests influence teaching.

In their study of washback on Test of English as a Foreign Language (TOEFL) preparation courses, Alderson and Hamp-Lyons (1996) found that the TOEFL test affects both what and how teachers teach, but the effect is not the same in degree or kind from teacher to teacher, and the simple difference of TOEFL versus non-TOEFL teaching did not explain why the teacher taught the way they did. Watanabe (1996) investigated the effect of the university entrance examination on the prevalent use of the grammar-translation method in Japan. His analyses of the past English examinations, classroom observations and interviews with teachers showed very little relationship between the test content and the use of this particular teaching methodology. Rather, teacher factors, including personal beliefs, past education, and academic background, seemed to be more important in determining the teaching methodology a teacher employs. Shohamy, Donitsa-Schmidt, and Ferman (1996) contended that test impact may be due to several other factors such as the status of the subject matter tested, the nature of the test (low or high stakes), and the uses to which the test scores are put. Additionally, the washback effect may change over time and may not last in the system. In summary, testing may be only one of those factors that "affect how innovations succeed or fail and that influence teacher behaviors" (Wall & Alderson, 1993, p.68).

It is clear in the literature that the main contribution of research in this area is the 10 year span since the 1996, when the special issue in Language Testing reported empirical studies which investigated different tests in different teaching and learning contexts.

These studies have investigated the influence of testing on teachers (including teaching assistants) and teaching (Borrows, 2004; Cheng, 2005; Ferman, 2004; Hayes & Read, 2004; Nazari, 2005; Scaramucci, 2002; Saif, 2006; Wall, 2005), textbooks (Read & Hayes, 2003; Saville & Hawkey, 2004; Yu & Tung, 2005), learners and learning (Andrews, Fullilove, & Wong, 2002; Chen & He, 2003; Robb & Ercanbrack, 1999; Watanabe, 2001), attitudes toward testing (Cheng, 2005; Jin, 2000; Read & Hayes, 2003), and test preparation behaviors (Stoneman, 2005). Some of these studies investigated the influence of a national English examination on the local English language teaching and learning due to its high-stakes nature in a particular country such as Brazil (Scaramucci, 2002), China (Qi, 2004, 2005; Zhao, 2003), Hong Kong (Andrews, 1995; Andrews, Fullilove, & Wong, 2002; Cheng, 2005), Iran (Nazari, 2005; Nemati, 2003), Israel (Ferman, 2004; Shohamy, Donitsa-Schmidt, & Ferman, 1996), Japan (Watanabe, 1996), Romania (Gosa, 2004), Sri Lanka (Wall, 2005), and Taiwan (Chen, 2002; Shih, 2006). Some of these studies investigated worldwide English testing such as the International English Language Testing System (IELTS) (Green, 2003; Hayes & Read, 2004; Nguyen, 1997), TOEFL (Alderson & Hamp-Lyons, 1996; Robb & Ercanbrack, 1999), and the Michigan Examination for Certificate of Competency (Irvine-Niakaris, 1997).

Stecher, Chun, and Barron investigated the influence of tests on school practices as a result of the introduction of test-based reform efforts at the state level in the USA, Saville and Hawkey looked at the impact of IELTS on the content and nature of IELTS-related teaching materials in the context of the UK and Hayes and Read reported their study of the impact of IELTS on the way international students prepare for academic study in New Zealand.

Cheng investigated the impact of the Hong Kong Certificate of Education Examination in English (HKCEE), a high-stakes public examination, on the classroom teaching and learning of English in Hong Kong secondary schools. The findings indicate that although the new examination was specifically designed to bring about positive washback effects on teaching and learning in schools, the washback effect of the new examination on classroom teaching is confined.

Cheng, Watanabe, with Curtis's Washback in Language Testing: Research Context and Methods (2004) is a cornerstone collection of washback studies—an area of research, which attracted the initial attention of the field of language testing about 20 years ago.

This book has tried to focus on the nature of washback by collecting washback studies from around the world. The first section of this book highlights the concept and nature of washback by providing a historical review of the phenomenon by Cheng and Curtis, the methodology to guide washback studies. The second section showcases a range of studies conducted in the USA, the UK, New Zealand, Australia, Japan, Hong Kong, China, and Israel. This book brings together washback studies on various aspects of teaching and learning conducted in many parts of the world and constitutes an extensive body of research that has contributed to our understanding of test washback and impact.

The study on wachback and impact is in progress. Recently, several important projects appointed by major testing agencies such as Cambridge ESOL and Educational Testing Services (ETS) have promoted washback and impact studies. These studies are conducted in many countries around the world on the same test, for example, TOEFL or IELTS. These studies tend to be large-scale, multi phased, and multifaceted, and offer important directions for future research. Impact (including washback) is a key focus of the Cambridge ESOL research and validation program, which is designed to ensure that all ESOL assessment products meet acceptable standards in relation to the four essential test qualities of validity, reliability, impact and practicality.

In addition to such fairly large-scale impact studies, 65 projects under the joint IDP Education Australia/ British Council IELTS funded research program which is managed jointly with Cambridge ESOL, have included, since 2002, around 20 studies directly investigating test impact and washback (see further details at www. Cambridge ESOL.org/rs_notes). These studies have been conducted in different parts of world with test-takers taking IELTS, and have investigated aspects such as candidate identity, their learning and performance, ethnographic study of classroom instruction, the impact of IELTS on receiving institutions, perceptions of the IELTS skills modules, and the impact of computer versus pen-and-paper versions.

These studies investigated the consequences of IELTS on a wide-ranging factors as well as the effects of IELTS on classroom teaching and learning (washback). It is obvious that international high-stakes language tests such as IELTS powerfully influence large numbers of language learners and teachers. Similar to IELTS is the TOEFL test. With the introduction of the Next Generation TOEFL (TOEFL iBT) in 2005, ETS has funded a series of studies, two of which aim at examining the impact of the TOEFL test (see Hamp-Lyons & Brown, 2007; Wall & Horak, 2006).

Hamp-Lyons and Brown (2007) conducted a study with three phases: the first phase established and validated instruments for the impact study; the second phase collected and analyzed data on TOEFL preparation in the USA, China and Egypt before the introduction of the TOEFL iBT in 2005. In the third phase, data was collected and analyzed in the same countries and institutions in order to identify change and constancy in beliefs, attitudes, methods, and the content of instruction under the influence of the significant changes to the existing TOEFL. Within these three countries, university-based and commercial institutions have been studied, and subjects from both TOEFL-taking and non-TOEFL-taking contexts have been included.

The results revealed alterations in perceptions and attitudes between TOEFL teachers and their students. However, there were few differences between the views of students who are preparing for TOEFL and those who are not. Some differences emerged across the three countries.

The TOEFL Impact Study in Central and Eastern Europe (Wall & Horak, 2006) investigated whether the new version of TOEFL iBT, which contributed to changes in teaching and learning after its introduction.

Shih (2006) investigated stakeholders' perceptions (including department heads, teachers, students, and their partner/ spouse) of the Taiwan General English Proficiency Test and its washback on school policies and teaching and learning. Stoneman (2005) examined and compared the nature and extent of test-preparation of university students.

2. STUDIES ON TEST VALIDATION

Test validation methods are at the heart of language testing research. Validity is a theoretical notion that defines the scope and the nature of validation work, whereas validation is the process of developing and evaluating evidence for a proposed score interpretation and use. The way validity is conceptualized determines the scope and the nature of validity investigations and hence the methods to gather evidence. Validation frameworks specify the process used to prioritize, integrate, and evaluate evidence collected using various methods. In general, developments of validity theories and validation frameworks in language testing have paralleled advances in educational measurement (Cronbach & Meehl, 1955; Cureton, 1951; Kane, 1992; Messick, 1989). Validation methods have been influenced by three areas in particular. Developments in psychometric and statistical methods in education have featured prominently in language testing research (Bachman, 2004; Bachman & Eignor, 1997). Qualitative methods in language testing (Banerjee & Luoma, 1997) have been well informed by second language acquisition (Bachman & Cohen, 1998), conversation analysis, and discourse analysis (Lazaraton, 2002). Research in cognitive psychology has also found its way into core language testing research, especially that regarding introspective methodologies (Green, 1997) and the influence of cognitive demands of tasks on task complexity and difficulty (Iwashita, McNamara, & Elder, 2001).

The validation of the discrete-point language tests popular in the 1950s and 1960s, including language aptitude tests, was mostly couched in the validity conceptualization by Lado (1961). The 1970s witnessed a trend toward more direct and communicative language tests, yet the focus still centered solely on face or content validity and predictive or concurrent validity (Clark, 1975, 1978).Clark proposed that direct and indirect language proficiency tests begged for different validation techniques because of their different characteristics (Clark, 1975).To summarize, earlier conceptualizations of validity, represented by Lado and Clark, focused on a few limited types of validity that support primarily score-based predictions, rather than theoretically and empirically grounded explanations of scores that provide the basis for predictions. Treating validity as different types invited researchers to select only one type as sufficient to support a particular test use. Further, test-taking processes and strategies, and test consequences were not examined. In keeping with how validity was conceptualized from the 1950s through late 1970s, the validation methods were limited to correlational analyses and content analyses of test items.

Another fairly common line of validation research in the 1960s and1970s employed factor analytic techniques to test two competing hypotheses about language proficiency, that is, whether language proficiency is a unitary trait or made up of several divisible competences (Oller, 1983). During the 1980s, there was a shift of focus from predictive or concurrent validity studies to explorations of test-taking processes and factors affecting test performance. These studies attested to the growing attention to score interpretation based on empirically grounded explanations of scores.

As validity theories in educational measurement advanced in the1980s and culminated in Messick's explication of validity (1989), different types of validity became pieces of evidence that supported a unitary concept of construct validity, highlighting the importance of combining different types of evidence to support a particular test use. Messick also formally expanded validity to incorporate social values and consequences, arguing that evaluation of social consequences of test use as well as the value implications of test interpretation both "presume" and "contribute to" the construct validity of score meaning (p.21). Messick's unitary validity model quickly became influential in language testing through Bachman's work (1990) (Cumming & Berwick, 1996; Kunnan, 1998a). However, although theoretically elegant, Messick's model is highly abstract and provides practitioners limited guidance on the process of validation, that is, how to prioritize validation research and gauge progress. To make Messick's work more accessible to language testers, Bachman and Palmer (1996) proposed the notion of test usefulness. They discussed six qualities: validity, reliability, authenticity, interactiveness, and impact, as well as practicality, which functions to prioritize the investigations of the six qualities. Due to its value in guiding practical work, this framework quickly came to dominate empirical validation research and became the cornerstone for language test development and evaluation (Shohamy, 2001). Nevertheless, this formulation of test usefulness does not provide a logical mechanism to prioritize the six qualities and to evaluate overall test usefulness. Since the trade-off of the qualities is dependent on assessment contexts and purposes, evaluations of overall test usefulness are conveniently at the discretion of test developers and validation researchers.

Following the shift in focus of validity investigations to score interpretation for a particular test use (rather than the test itself), theories of validity, impact, ethics, principles of critical language testing (Shohamy, 2001), policy and social considerations (McNamara, 2006), and fairness (Kunnan, 2004) have been formulated to expand the scope of language test quality investigations (Bachman, 2005). Although some aspects of their work contribute to the validity of test score interpretations or uses, others address broader policy and social issues of testing, which may not be considered as qualities of particular tests (Bachman, 2005).

During this period, empirical validation research flourished to address more aspects of validity including factors (test, test-taker, and processes and strategies) affecting test performance, generalizability of scores on performance assessments, and ethical issues and consequences of test use (Bachman, 2000; Cumming & Berwick, 1996; Kunnan, 1998a).

Furthermore, the maturity of sophisticated methodologies, both quantitative (Kunnan, 1998b, 1999) and qualitative (Banerjee & Luoma, 1997), and triangulation of different methodologies (Xi, 2005b) took place.

The search for a validation framework that is theoretically sound but more accessible to practitioners continues. The major development of an argument-based approach to test validation in educational measurement (Kane, 1992; Kane, Crooks, & Cohen, 1999) has recently inspired parallel advancements in validation frameworks in language testing, represented by Bachman (2005). The notion of a validity argument is nothing new to the field of educational measurement. Nearly two decades ago, Cronbach (1988) started to think of validation as supporting a validity argument through a coherent analysis of all the evidence for and against a proposed score interpretation. Kane and his associates have taken up on this and formalized the development and evaluation of the validity argument by using practical argumentation theories (Toulmin, 2003). They see validation as a two-stage process: constructing an interpretive argument, and developing and evaluating a validity argument. They propose that for each intended use of a test, an interpretive argument is articulated through a logical analysis of the chain of inferences linking test performance to a decision, and the assumptions on which they rest. The assumptions, if proven true, lend support for the pertinent inference. The network of inferences, if supported, attaches more and more meaning to a sample of test performance and the corresponding score, so that a score-based decision is justified. The plausibility of the interpretive argument is evaluated within a validity argument using theoretical and empirical evidence. Their approach also allows for a systematic way to consider potential threats to the assumptions and the inferences and to allocate resources to collect evidence to discount or reduce them. This argument-based approach to test validation has motivated the development of a validity argument for the new Test of English as a Foreign Language (TOEFL) to organize and evaluate a whole program of validation research (Chapelle, Enright, & Jamieson, 2008).

The literature highlights a high demand which has motivated and increased a need for conducting research on validity evidence, because stakeholders were concerned about the purpose, quality, and quantity of testing in schooling (Lederman & Burnstein, 2006; Supon, 2008). These stakeholders (teachers, parents, students, and businesses) were concerned to ensure that the purpose of the testing was explained and defined clearly. Based on this clear statement of the objective of an assessment, the validity of the scores was evaluated. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the 'adequacy' and 'appropriateness' of 'inferences' and 'actions' based on test scores or other modes of assessment" (p.13). From this definition of validity, the purpose of the assessment was known prior to the evaluation of the consequent inferences and actions. This need for a clearly defined purpose of the assessment was also conveyed in the 1999 Standards for Educational and Psychological Testing. It suggested the use of validation to develop scientifically sound evidence to support the proposed interpretation of test scores and their intended use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Goodwin & Leech, 2003). Based on the above-mentioned standpoint of the purpose for test validity, research direction shifted toward the process by which validity evidence may be established. Kane (2006) described validation as the process of evaluating the credibility of interpretations and uses. This understanding of validity led to several faces of validity that allowed the researchers to establish evidence of the appropriateness of inferences and actions of a test. These types of validity evidence were categorized by several authors as; criterion validity, content validity, and construct validity (Angoff, 1988; Cureton, 1951; Kane, 2006; Messick, 1989; Pellegrino, 1988). During the beginning of the twentieth century, criterion validity was defined by Cureton (1951) as the correlation between the actual test scores and the 'true' criterion score and was considered the gold standard. The concept of criterion validity was recently divided into two schools of thought; concurrent validity and predictive validity. Two tests given at the same time with a high correlation between their scores can be thought of as having concurrent validity; while predictive validity involved the ability of the test scores to predict future performance (Kane, 2006). The interpretation of validity as defined by Cureton (1951) was later extended to include content validity; which was used to validate academic measures. The idea behind content validity was to provide evidence that the content of the measure was relevant and appropriate for the inferences from and uses of the test score (Messick, 1989). The final extension of the concept of validity, specifically construct validity, was used to validate measures of a psychological nature or theoretical attributes. Cronbach and Meehl (1955) described construct validation as the process to follow when criterion and content measures are unavailable. As time progressed in the field of validity studies, the construct validity approach was widely accepted as a general model for validation of a measure (Anastasi, 1986; Embretson, 1983; Guion, 1977; Messick, 1980, 1988, 1989). Applications of the construct validity model required researchers to clearly define the interpretation and use of the test scores. In the case of NCLB, the federal government required the student scores in grades 3-8 and high school to be aggregated to the subgroup level on the statewide assessments (No Child Left Behind Act, 2002). Knowing this level of aggregation was used to make inferences about the type of education students are receiving and whether or not all students received tutoring services, it was imperative to validate the inferences using subgroup level data. In particular, it was important to provide evidence of subgroups being measured on the same latent trait. Identification of the same latent trait for all subgroups, such as gender and ethnicity, was identified by researchers as validity evidence in the form of measurement invariance (Mellenbergh, 1989).

REFERENCES

Alderson, J. C. (1986) Innovations in language testing. In M. Portal (Eds.), *Innovations in language testing* (pp.93-105). NFER Nelson.

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.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Andrews, S. (1995). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, V. Berry, & R. Berry (Eds.), *Bringing about change in language education* (pp.67-81). University of Hong Kong, Hong Kong.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback: A case study. *System, 30,* 207-223.

Angoff, W. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.19-32). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

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

Bachman, L. F., & Cohen, A. D. (Eds.). (1998). *Interfaces between second language acquisition and language testing research*. New York, NY: Cambridge University Press.

Bachman, L. F., & Eignor, D. R. (1997). Recent advances in quantitative test analysis. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education*, *Volume 7: Language testing and assessment* (pp.227-242). Kluwer Academic, Dordrecht, The Netherlands.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

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

Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education*, *volume 7: Language testing and assessment* (pp.275-287). Kluwer Academic, Dordrecht, The Netherlands.

Borrows, 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.113-118). Mahwah: Lawrence Erlbaum Associates.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language™*. Mahwah, NJ: Lawrence Erlbaum.

Chen, L. (2002). *Washback of a public exam on English teaching* (Unpublished PhD dissertation). The Ohio State University.

Chen, Z., & He, Y. (2003). Influence of CET-4 on college students and some suggestions. *Journal of Technology College Education, 22,* 40-41.

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

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In *washback in language testing: Research contexts and methods* (pp.3-18). Mahwah, NJ: Lawrence Erlbaum Associates.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Volume 7: Language testing and assessment* (2nd ed., pp.349-364). New York: Springer Science and Business Media LLC.

Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In S. Jones & B. Spolsky (Eds.), *Language testing proficiency, center for applied linguistics* (pp.10-24). Arlington, VA.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cumming, A., & Berwick, R. (Eds.). (1996). *Validation in language testing*. Multilingual Matters, Clevedon, Avon.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp.621-694). Washington, DC: American Council on Education.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing, 14,* 261-277.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

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 contexts and methods* (pp.191-210). Mahwah, NJ: Lawrence Erlbaum Associates.

Gosa, G. (2004). *Investigating washback: A case study using student diaries* (Unpublished PhD dissertation). Lancaster University, UK.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. (Assessment in Action). *Measurement and Evaluation in Counseling and Development, 36*(3), 181- 191.

Green, A. (1997). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Green, A. (2003). *Test impact and English for academic purposes: A comparative study in backwash between IELTS preparation and university pre-sessional courses* (Unpublished PhD thesis). Centre for Research in Testing, Evaluation and Curriculum in ELT, University of Surrey, Roehampton.

Guion, R. M. (1977). Content validity—The source of my discontent. *Applied Psychological Measurement, 1*(1), 1-10.

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 (2003-5): Entering innovation.* Submitted to the TOEFL Research Committee, Educational Testing Service.

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 contexts and methods* (pp.97-112). Mahwah, NJ: Lawrence Erlbaum Associates.

Irvine-Niakaris, C. (1997). Current proficiency testing: A reflection of teaching. *Forum, 35,* 16-21.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 51*(3), 401-436.

Jin, Y. (2000). Washback of college English test-spoken English test on teaching. *Foreign Language World, 80,* 56-61.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement* (4th ed., pp.18-64). Washington, DC: American Council on Education/ Praeger.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.

Kunnan, A. J. (1998a). *Validation in language assessment*. Selected Papers from the 17thLanguage Testing Research Colloquium. Mahwah, NJ: Long Beach, Lawrence Erlbaum.

Kunnan, A. J. (Ed.). (1998b). Special issue: Structural equation modeling. *Language Testing, 15*(3).

Kunnan, A. J. (1999). Recent developments in language testing. *Annual Review of Applied Linguistics, 19,* 235-253.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp.1-14). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testingin a global context: Proceedings of the ALTE Barcelona conference* (pp.27-48). Cambridge, UK: Cambridge University Press.

Lado, R. (1961). *Language testing*. New York, NY: McGraw-Hill.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral tests*. Cambridge: Cambridge University Press.

Lederman, L. M., & Burnstein, R. A. (2006). Alternative approaches to high-stakes testing: Mr. Lederman and Mr. Burnstein propose a novel way to increase student engagement and counter the pressures of high-stakes testing. *Phi Delta Kappan, 87*(6), 429.

McNamara, T. F. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31-51.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127-143.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.33-45). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13-103). New York: American Council on Education and Macmillan.

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

Nazari, A. (2005). Washback effects on TEFL: A case study from Iran. *IATEFL Voices, 185,* 9-10.

Nemati, M. (2003). The positive washback effect of introducing essay writing tests in EFL environment. *Indian Journal of Applied Linguistics, 29,* 49-62.

Nguyen, P. (1997). *Washback effects of international English language testing system at the Vietnam national university* (Unpublished PhD thesis). University of Melbourne.

No Child Left Behind Act (Reauthorization of Elementary and Secondary Education Act). (2002). Public Law 107-110 § Section 1202(c)(7)(A)(IV)(2).

Oller, J. W. (1983). *Issues in language testing research.* Newbury House, Rowley, Mass.

Pearson, I. (1988) Tests as levers for change. In D. Chamberlain & R. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (pp.98-107). ELT Document 128. London*:* Modern English.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.49-60). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Qi, L. (2004). Has a high-stakes test produced the intended changes? In A. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*, (pp.171-190). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Read, J., & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. In R.Tulloh (Ed.), *International English language testing system research reports 2003, volume 4.* IELTS Australia, Canberra.

Robb, T. N., & Ercanbrack, J. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *TESL-EJ, 3,* A2. Retrieved from http://tesl-ej.org/ej12/toc.html

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing, 23,* 1-34.

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 contexts and methods* (p.73, 96). Mahwah, NJ: Lawrence Erlbaum Associates.

Scaramucci, M. V. R. (2002). Entrance examinations and TEFL in Brazil: A casestudy. *Revista Brasileira de Lingüística Aplicada, 2,* 61-81.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement Issues & Practice, 9*(3), 15-22.

Shih, C. M. (2006). *Perceptions of the general English proficiency test and its washback: A case study at the two Taiwan technological institutes* (Unpublished PhD dissertation). University of Toronto.

Shohamy, E. (2001). *The power of tests: A critical perspective of 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,* 298-317.

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

Stoneman, B. (2005). *An impact study of an exit English test for university graduates in Hong Kong: Investigating whether the status of a test affects students' test preparation activities* (Unpublished PhD thesis). Hong Kong Polytechnic University.

Supon, V. (2008). High-stakes testing: Strategies by teachers and principals for student success. *Journal of Instructional Psychology, 35*(3), 306-308.

Swain, M. (1985). Communicative competence: Some roles of comprehensive input and comprehensible output in its development. In *Input in second language acquisition.* Cambridge, MA: Newbury House Publishers.

Toulmin, S. E. (2003). *The use of argument* (Updated edition). Cambridge: Cambridge University Press.

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.

Wall, D. (2005). The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory. *Studies in Language Testing* (Volume 22). Cambridge: Cambridge University Press.

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

Wall, D., & Horak, T. (2006). The TOEFL impact study: Phase 1. *TOEFL Monograph* (p.34). Educational Testing Service.

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

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). Author, Adita, Japan.

Xi, X. (2005b). Do visual chunks and planning impact performance on the graph description task in the SPEAK Exam? *Language Testing, 2*2(4), 463-508.

Yu, G. K. H., & Tung, R. H. C. (2005). The washback effects of JCEEEs in the past fifty years, *proceedings of 22nd conference on English teaching and learning* (pp.379-403). Normal University, Taipei, Taiwan.

Zhao, L. (2003). College English teaching evaluation system in China: Major problems and corresponding counter measures. *Indian Journal of Applied Linguistics, 29,* 85-98.