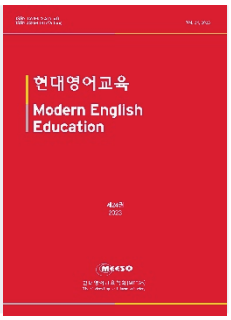
Modern English Education, Vol. 24, (2023) pp. 10-25
doi: https://doi.org/10.18095/meeso.2023.24.1.10



Modern English Education

ISSN: 1598-0782 / e-ISSN: 2586-6141
http://journal.meeso.or.kr

A Washback Effect Model for Investigating the Impact of High-Stakes Language Testing on Teaching and Learning

Han Gil Kim (The Ohio State University)

Received: 11 December 2022
Revised: 27 January 2023
Accepted: 31 January 2023

Kim, Han Gil. (2023). A washback effect model for investigating the impact of high-stakes language testing on teaching and learning. *Modern English Education*, *24*, 10-25.

Keywords
washback, washback model, activity theory
환류효과, 환류효과 모형, 활동이론

Han Gil Kim
PhD Candidate
Department of Teaching and Learning
The Ohio State University
Kim.3208@osu.edu

Abstract

Since the early 1990s, effects of testing on teaching and learning have drawn increased attention. However, most studies in second language assessment have been quantitative. Although a few studies were qualitative, they lacked a conceptual or theoretical framework, thus failing to provide a comprehensive picture of washback. This is also true in the Korean context. An operational model is needed to effectively investigate washback. This paper reviews prior research studies that have introduced seminal washback hypotheses and empirically driven operational models. These studies have identified significant macro- and micro factors such as curriculum, perception, beliefs, students' and teachers' previous experiences, and tests that can impact teaching and learning. Based on a review of existing operational models, this paper proposes Engeström's activity theory as a new and comprehensive model of washback that can be adopted in any context to analyze high-stake language examinations. These findings have theoretical and practical implications for qualitative washback research.

INTRODUCTION

We live in a testing world, inundated with myriad classroom-oriented assessments and high-stakes testing for various social and educational purposes. Educational institutions have come to rely on high-stakes testing in assessing students' value in the eyes of crucial stakeholders (i.e., administrators, parents, the community, and even the greater society). Testing determines students' future as well as their teachers' self-esteem and reputation (Cheng & Curtis, 2004; Qi, 2005; Wall, 2012). This power attributed to testing refers to *washback*, that is, the positive and/or negative consequences of testing in relation to teaching and learning (Cheng, 2008; Green, 2013; Wall, 2012).

Findings from prior empirical research on washback have broadened our understanding of the impact of testing on

teaching and learning. Second language assessment scholarship has consistently revealed that there is little detailed information about washback as a whole (Estaji, 2013; Tsagari & Cheng, 2017; Wall, 2012). A number of macro- and micro-contextual factors such as stakeholders, policy, and ideology have impacted washback. Existing studies have been quantitative, lacking comprehensive analyses of *why* beneficial or harmful impact occurred; thus, they have underestimated crucial stakeholders' motives for their particular attitudes and behaviors (Gosa, 2009; Hughes, 2003).

These results are consistent with the Korean context. In 2018, the Korean Ministry of Education reformed the College Scholastic Aptitude Test (CSAT). The essential change this reform brought about was to shift the CSAT from a norm-referenced to a criterion-referenced test. To date, research methodology used on the impact of the new CSAT on teaching and learning has measured the phenomenon quantitatively; only a few studies have attempted to investigate washback utilizing qualitative methodology. Yet they have explored washback of the old CSAT to teaching and learning or investigated researcher-produced or teacher-produced achievement tests, not the actual CSAT (e.g., D. J. Kim & K.-M. O, 2002; S. Kim, 2014; H. Lee, 2019; K. O, 2001; T. J. Yoon, 2017).

As Gay et al. (2011, p. 400) stated, to answer "descriptive (e.g., what happened?) and explanatory questions (e.g., how or why did something happen?)" about the test's impact on teaching and learning, qualitative studies should be considered along with quantitative research. However to date, qualitative studies on washback have not developed a particular conceptual/theoretical framework or operational model to better capture the phenomenon as a whole.

Thus, the current paper first explores seminal theoretical studies on washback that define its scope, intentionality, and direction. Next, it reviews studies that introduce hypotheses and operational models of washback. The paper poses the following questions: 1) how does washback operate? 2) to what extent can washback be delimited in order to investigate it? and 3) what types of methodological approaches to washback have other researchers used? Finally, the paper proposes a new and comprehensive model of washback based on a review of existing operational models.

THEORETICAL ASPECTS

Since the early 1990s, the conceptual work, "Does Washback Exist?" (Alderson & Wall, 1993), and empirical research, "Examining Washback" (Wall & Alderson, 1993), have prompted an investigation of the potential impact of testing on teaching and learning and vice versa (Cheng, 2008; Green, 2013; Wall, 2012).

As these two articles have established, "testing is never a neutral process and always has consequences" (Stobart, 2003, p. 140). Henceforth, research in language testing (see Alderson & Wall, 1993; Bailey, 1996; Cheng, 2008; Green, 2013; Spratt, 2005; Wall, 2012) has investigated the impact of tests on teaching and learning, widely referred to as *washback* (Alderson & Wall, 1993; Andrew et al., 2002; Cheng & Curtis, 2004), or also known as *backwash*[1] (Hughes, 2003, 2010).

Due to the complexity of the educational system, however, it is generally perceived that the relationship between testing and teaching/learning and vice versa is not straightforward (Cheng & Curtis, 2004; Tsagari & Cheng, 2017; Wall, 2012; Wall & Alderson, 1993). It also involves a number of intervening variables: teachers' and students' attitudes toward the exam, classroom settings, and contexts outside the classroom (Qi, 2005; Saville, 2000; Shih, 2007). In their washback study in Sri Lanka, Wall and Alderson (1993) called for a model of washback and further empirical research to delineate the complex processes of washback. However, only a limited number of qualitative studies on washback have proposed an operational model (e.g., Alderson & Wall, 1993; Alderson & Hamp-Lyons, 1996; Bachman & Palmer, 1996; Bailey, 1996; Burrows, 2004; Green, 2007; Saif, 2006; Shih, 2007; Wall, 2005) for future researchers, and these models were not theoretically grounded (except for Wall, 2005).

Early Discussion of the Impact of Testing

Does a test really impact or even alter teaching, learning, and/or a whole curriculum? Why have researchers conducted studies on washback? The earlier concept of washback is based on Popham's (1987) claim that tests dictate what is taught and learned (i.e., measurement-driven instruction). Since the content of test items can be compelling 'curricular magnets,' tests motivate teachers and learners alike, to meet the test's objectives (Cheng, 2005). As Noble and Smith (1994) suggested, "the most pervasive tool of top-down policy reform is to mandate assessment that can serve as both guideposts and

[1] Hughes (2003, 2010) strongly prefers to use the term *backwash* rather than *washback* because he could find the term *backwash* in dictionaries, but not *washback*. There has been some debate over which term is more appropriate (c.f., Pearson, 1988, p. 98). Alderson explained why they (i.e., Alderson & Wall, 1993) used the term *washback* in their article: "it seems to us that the word washback was commonly used in discussions, in presentations at conferences and in teacher training" (Alderson, 2004, p. xi).

accountability" (p. 1). To achieve these goals, educators have been encouraged to match the content and format of the examinations they give to their students to overlap with or even match the content and format of the curriculum (Cheng & Curtis, 2004); that is, to create a so-called "test-curriculum alignment" (Shepard, 1990).

However, Madaus (1988) noted a problem with this alignment. He observed, "it is testing, not the official stated curriculum, that is increasingly determining what is taught, how it is taught, what is learned, and how it is learned" (p. 83). In other words, "what is assessed becomes what is valued, which becomes what is taught" (McEwen, 1995, p. 42). Pearson (1988) also emphasized the power of the test: "Public examinations influence the attitudes, behaviors and motivation of teachers, learners, and [even] parents" (p. 98). In essence, the test has become a means of control over teaching and learning, and thus plays a significant role in the educational system. In the field of language testing, terms closely related to washback include measurement-driven instruction (Popham, 1987), test-curriculum alignment (Shepard, 1990) or a test's impact (see Baker, 1991), and these terms are synonymously used in the literature (Alderson & Wall, 1993; Bailey, 1996; Cheng & Curtis, 2004; Cheng, 2005; Saville, 2000; Tsagari & Cheng, 2017).

Washback: Scope, Intentionality, and Direction

Scope

In language testing, the term *test's impact* can be interchanged with washback; however, the difference between these two terms delimits the investigation of the boundary of washback related to the scope of washback. For Tsagari and Cheng (2017), the impact is a comprehensive term that 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" (p. 1). Bachman and Palmer (1996) regard washback as part of a test's impact. They consider impact as taking place at the macro level (i.e., more concerned with society and educational system), whereas washback occurs at the micro level (i.e., teachers and learners) (see also Estaji, 2013; Pan, 2008), as seen in Figure 1 below.

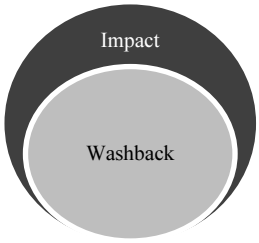


FIGURE 1
Dimensions of Impact and Washback

Other researchers avoided making a distinction between the two terms. Andrews (2004) lumped together both the macro- and micro-level intervening variables in the phenomenon of washback. Tsagari and Cheng (2017) also suggested that washback and impact should not be considered as distinct concepts and that they can also be observed at larger levels, such as within educational institutions, the education system, and beyond. Delineating clear boundaries of washback seems less valuable because in order to capture a complete picture of test impact, washback's all-encompassing impact needs to be considered.

Looking at the scope of washback more specifically, Alderson and Hamp-Lyons (1996) believe that "washback is a consequence of *high-stake*[2] exams" (p. 296, emphasis in original). Others argued that washback can be captured when we view "students and teachers do things they would not necessarily otherwise do because of the test" (Messick, 1996, p. 243). Buck (1988) explained, "washback can have an influence on various aspects such as teaching and learning" He defined *washback* as the indication of direction by stating, "the influence of the test on the classroom … this washback effect can be either beneficial or harmful" (p. 17).

Although researchers have offered different definitions and descriptions, their explanations overlap in terms of the scope

[2] According to Madaus (1988), *high-stakes* exams, such as college entrance exams, are those "whose results are seen – rightly or wrongly – by students, teachers, administrators, parents, or the general public, as being used to make important decisions that immediately and directly affect them" (p. 7).

of washback. However, most researchers define washback broadly and simply as the impact of a test on teaching and learning (Bailey, 1996; Cheng & Curtis, 2004; Green, 2007; Spratt, 2005; Tsagari & Cheng, 2017; Wall, 2012).

Intentionality and Direction

There is some debate on whether washback should be defined as being intentional and/or unintentional (Cheng & Curtis, 2004). Early on, Messick (1989) argued that to measure the consequential validity of tests, it is necessary to evaluate "the intended or unintended social consequences of test interpretation and use. The appropriateness of the intended testing purpose and the possible occurrence of unintended outcomes and side effects are the major issues" (p. 84). Cheng (2005) also recognized not only intended but also "unintended and accidental side effects can occur, because successful curriculum change and development is a highly complex matter" (p. 28), based on her empirical findings. This unintended washback was part of what Alderson and Wall (1993) called "negative work."

Regarding the direction of washback, testing always has an impact, whether it is positive/beneficial or negative/ harmful; it is never a neutral process. Pearson (1988) argued that "good" tests would result in *positive* washback. However, Wall (2000) contested this argument for being too simplistic. According to Cheng (2005), teachers and learners would work willingly to teach and learn the material when they had a positive attitude toward the test (Cheng, 2005).

Other researchers criticize language tests for their negative impact on teaching and learning, known as *negative* washback. Vernon articulated this criticism of testing. He observed that "teachers tend to ignore subjects and activities which are not directly related to passing the exam so that examinations distort the curriculum" (as cited in Cheng, 2005, p. 29). Alderson and Wall (1993) call negative washback the undesirable effect of tests on teaching and learning. In this regard, Messick (1996) added: "A poor test may be associated with positive effects and a good test with negative effects" (p. 242). He believed that such effects are not due to the quality of the test, but to good or bad educational practices. It seems that any test, good or poor, can be judged as having positive or negative washback. Then, it would be useful to evaluate the educational contexts in which a test is carried out (Alderson & Wall, 1993; Cheng, 2005; Green, 2013; Wall & Alderson, 1993).

Alderson and Wall, in this light, stated: "It is surely conceivable that other forces exist within society, education, and schools that might prevent washback from appearing, or that might affect the nature of washback despite the 'communicative' quality of a test" (p. 116). By considering the scope, intentionality, and direction of washback, it is clear that washback may occur not only positively or negatively, but also rapidly or slowly, directly or indirectly, transparently or invisibly to unforeseen stakeholders (Wall, 2012). Therefore, it is important to investigate the impact of language tests on teaching and learning as well as to focus on theoretical and methodological approaches to explore washback.

Washback Hypotheses and Models

Researchers in applied linguistics have emphasized the phenomenon of washback; however, little evidence of washback was presented until the early 1990s (Cheng et al., 2004; Hughes, 2003; Tsagari & Cheng, 2017; Wall, 2012). The advent of the currently termed *washback research* is rooted in a few theoretical models (see Alderson & Wall, 1993; Bailey, 1996; Hughes, 1993). Alderson and Wall (1993, pp. 120-121) have offered 15 possible washback hypotheses based on the literature on language testing and their Sri Lankan study as outlined in Table 1.

The authors have concluded that future researchers must look into motivation and performance, specification of washback hypotheses as well as innovation and change in educational settings, and triangulate their data gathered from direct observation of a classroom using ethnographic lens.

Hughes (2010) suggested three crucial components of backwash: participants, process, and product. The *participants* included students, classroom teachers, administrators, material developers, and publishers whose attitudes and perceptions about teaching and learning could be influenced by a test. He referred to *process* as participants' actions that may influence learning. It includes "material development, syllabus design, changes in teaching methodology, the use of learning and/or test-taking strategies and the like" (Bailey, 1996, p. 262). Finally, Hughes (2010) defined *product* as "what is learned ... and the quality of the learning." He noted that a test could impact the participants' attitudes and perceptions, which would in turn affect the process of teaching and learning, thereby affecting the learning outcomes. Bailey (1996, p. 264) designed his model of washback (see Figure 2) by combining Hughes's 1989 model with Alderson and Wall's (1993) 15 hypotheses.

TABLE 1
Possible Washback Hypotheses

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

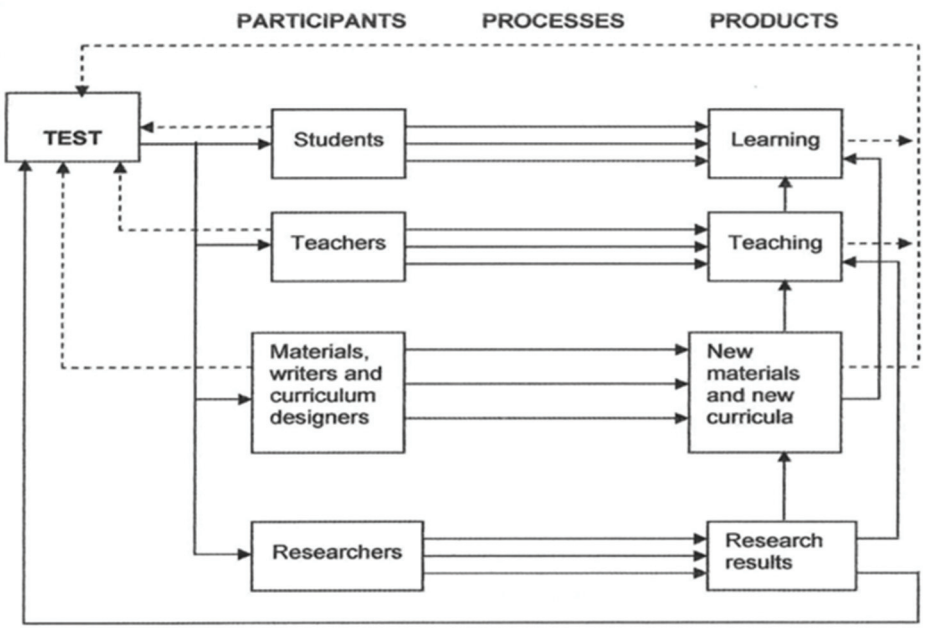


FIGURE 2
Bailey's (1996) Basic Model of Washback

Figure 2 shows Hughes's trichotomy on top of Bailey's model. Solid lines depict a direct connection between a test and the participants: students, teachers, material/curriculum designers, and researchers. The dotted lines illustrate participants' impact on the test. Bailey's model was not based on an empirical study, but on prior literature and models.

Alderson and Wall (1993) suggested a critical examination to find evidence to support the theoretical models of washback. Another publication, Alderson and Hamp-Lyons's (1996) empirical research of impact on the Testing of English as a Foreign Language (TOEFL), also supports this claim. Alderson and Hamp-Lyons (1996) acknowledged that it is first necessary to explore the context and participants' behavior because a "test will have different amounts and types of washback on some teachers and learners than on other teachers and learners" (p. 296). Messick (1996) also warned that not all the features that appeared in the context of washback would really indicate washback. Indeed, theoretical models of washback must be supported by evidence.

EMPIRICAL STUDIES OF WASHBACK

Over the past three decades, researchers have conducted empirical studies on washback across various contexts. Their studies have broadened our understanding of washback in complex educational systems. This section reviews empirical studies on the impact of various language tests on teaching and learning (e.g., TOEFL, IELTS, college entrance exams) and contexts (e.g., Canada, China, Hong Kong, Israel, Japan, Korea, New Zealand, Rumania, Sri Lanka, United Kingdom, United States, and so on). The present paper reviews the central issues addressed in these studies, their methodological approaches, and their own operational models of washback, drawing from both their findings and the primary theoretical hypotheses and models (Alderson & Wall, 1993; Bailey, 1996).

Washback on Teaching

Investigating washback on teaching, three studies (Cheng, 1999; Saif, 2006; Watanabe, 1996) found that washback does exist. Cheng (1999) compared and contrasted three teachers' teaching practices under both the original- and the revised syllabus for the Hong Kong Certificate of Education Examination in English (HKCEE). To be fully aware of the changing behavior of teachers, the researcher used the notion of the baseline study, that is, each participant taught five classes using the original syllabus in 1994 and later in 1995, using the revised syllabus. The baseline notion is that if there is a significant change to a test, then researchers utilize the pilot study. To help those who design and administer the test understand washback regarding teaching practice, researchers tend to compare the contexts before and after a test is released.

Cheng (1999) reported her findings from multiple sources, field notes based on her observations of classes, and transcripts of videos and interviews with teachers. Using findings from both the baseline study (stage one) and the main study (stage two), she triangulated different sources of data and conducted a post hoc analysis of videotaped recordings with field notes to reveal meaningful changes in classroom teaching under two syllabi. Her study illustrated that teacher participants under the new plan successfully "adapt[ed] (not adopt[ed])" (p. 257) their pedagogical knowledge so that students prepared more adequately for their desired outcome: getting a high score on the examination.

Cheng (1999) recognized that these changes in content resulted from the revised HKCEE. She noted that the new exam did not really alter the participants' teaching methods. For example, the content changed from reading aloud to role-playing, while these activities were both taught by drilling. To limit such negative washback, test designers make an effort to overlap a test with the curriculum regarding content and format as much as possible, so that teachers can implement aspects of the new test that they find useful, while also maintaining the aspects of the existing curriculum that they consider to be beneficial (Cheng et al., 2004).

Investigating O-level, which is the new examination in Sri Lanka, Wall and Alderson's (1993) primary finding aligned with that of Cheng (1999). They found that most teachers followed the content of the test in their curriculum and discarded other contents. For instance, teachers highly emphasized writing and reading activities of students' language skills, since the test did not cover listening and speaking. Researchers have termed this phenomenon *negative washback*. Wall and Alderson (1993) concluded teachers were unaware of the test designers' intent and theory behind the test. Therefore, they suggested sufficient teacher training to avoid harmful impacts (see also Alderson & Wall, 1993). Hughes (2003) also pointed out that providing assistance to teachers to adopt the desired objectives of a new test would promote beneficial washback.

Significant evidence of overlap between the curriculum and the test was also evident in Burrows's (2004) findings. Burrows (2004) explored teachers' pedagogical perceptions and practices to examine the washback effects of the new Certificate in Spoken and Written English (CSWE) test in Australia. When CSWE was introduced, dominantly employed pedagogy across teachers was needs-based, that is, determining the needs of each class of students and designing a syllabus to meet those needs. Burrows's (2004) findings showed that three out of the four teachers interviewed reported they had implemented CSWE in their teaching to varying degrees; one did not believe the test was in alignment with the existing curriculum and did not adjust her teaching according to it.

Burrows (2004) also proposed a model to understand teachers' perception of a newly introduced assessment, the curriculum innovation model. This model emphasizes searching for patterns in teachers' responses (i.e., belief systems, assumption, and knowledge) regarding curriculum innovation and impact of a new test. In other words, researchers must not merely focus on individual teachers' responses, but look for common themes for conducting a qualitative analysis (see Appendix A for a detailed diagram).

Using quantitatively oriented mixed methods (Gay et al., 2011; Onwuegbuzie & Leech, 2005), Saif (2006) and Watanabe (1996) explored the washback effect of the exam. Saif (2006) created a new oral proficiency test for international teaching assistants (ITA) based on needs analysis and a baseline study from graduate advisors, administrators, undergraduates, and

ITAs. He wanted to reveal past ITA teaching and testing practices (see Appendix B). After two preliminary studies, Saif (2006) conducted his main study to confirm whether positive washback emerged in the experimental and control groups. Whereas the former group took a semester-based language course, the latter attended the regular ITA program. Saif (2006) used a questionnaire and observation data as the first step to investigate the two ESL teachers and raters and 15 undergraduates.

To explore positive changes in the experimental group, Saif (2006) analyzed data from his classroom observations. For the last step, he searched for evidence of learning outcomes between the two groups. The findings showed the ITA test generated positive washback in the experimental group compared to the control group. The data revealed that owing to the more authentic teaching materials in accordance with the test as well as student-centered activities, ITAs in the experimental group improved their speaking ability. This finding indicated that improving learners' speaking skills requires speaking practice in the classroom. In other words, it is important for test designers to design a direct test that assesses learners' performance with activities and texts as authentically as possible (Bailey, 1996; Hughes, 2003; Wall, 2012).

Shohamy (1993), Shohamy et al. (1996), and Watanabe (1996) presented slightly different perspectives on washback intensity and teachers' perception of the status of the target language. Shohamy (1993) and Shohamy et al. (1996) investigated two tests, Arabic as a second language (ASL) and English as a foreign language (EFL). With ASL being a low-stakes test and EFL a high-stakes test, EFL had more impact on the teaching approaches such as activity, material development, and content and time for exam preparation of the students, which was exclusively devoted to oral skills.

Both studies concluded that the status of a language (high- or low-stakes) likely affects washback more positively in teaching and learning. Shohamy (1993) also highlighted that the findings were influenced by the teachers' negative perceptions of the ASL, that it was of no importance, as well as positive attitudes toward the EFL, and that it was an important communicative tool necessary for students' success. These findings teach us how factors like the perceived status of a certain target language and teachers' beliefs and assumptions about that language can promote either beneficial or harmful washback on teachers and their teaching.

Watanabe's (1996) findings also support this conclusion. He conducted an intervention study, recording two different classrooms and targeting either Grammar Translation (GT)-oriented high-stakes college entrance exams or non-GT-oriented exams. The researcher recorded time spent on the various materials, teachers' classroom practices (e.g., things written on the board, comments & questions), and time spent on learners' focused activities. He wanted to calculate what the teachers chose to highlight in their teaching regarding the GT method since the college entrance exam evaluates grammar proficiency. Watanabe's findings, interestingly, showed that washback existed differently between two Japanese teachers of English: one teacher translated and explained grammar in both classes (GT-oriented), whereas the other teacher alternated between GT and non-GT-oriented teaching in her practice. Watanabe concluded that local factors, such as the teachers' personal educational background, beliefs, and past learning experience, rather than the test, seemed to have influenced their teaching.

In her book-length research, *Washback on High School Classrooms of the English Tests within Korean University Entrance Exams*, O, K.-M. (2001) explored washback of the old CSAT on four Korean teachers of English and 200 12th-grade students. She first looked for a relationship among the CSAT, the national curriculum, and textbooks. She then observed classrooms and conducted interviews with teachers and a few focal students. Her findings revealed that even though the national curriculum and textbooks attempted to embrace all four skills fairly, the test leaned more toward certain skills (e.g., 65% for reading and 15% for speaking and writing). Regarding the intended washback from test designers, O, K.-M. (2001) found that the actual test departed greatly from the theory and design behind it. Surprisingly, the interviews indicated that while the teachers found the CSAT to be "moderately successful," the students were not in favor of it (p. 130).

O, K.-M. (2001) also identified some local factors that hindered positive washback, such as Teachers' workload, Teachers' English proficiency, Teachers' goal, The classroom setting, and Context outside the classroom. Her results support previous findings from empirical studies covered in this paper. Importantly, she revealed that washback differs across different stakeholders and contexts. However, relatively little is known about the new CSAT and its washback effect for different stakeholders.

It is now clear that, in some circumstances, local factors such as individual differences between teachers can have unforeseen effects when they interact with tests. Therefore, washback must be considered within the contexts of examination use (Green, 2013; Spratt, 2005). Regarding local factors and the process of curriculum- and/or examination reform, Wall (1996, 2005) developed an operational model of washback grounded in innovation theory. She combined Fullan's view of the change process to claim reforming a curriculum or an examination as a process rather than an event (as cited in Wall, 2005, pp. 61-63). She also adapted Rogers's notion of the "innovation-decision process" that categorizes innovativeness by adopters' readiness to and speed with which they adopt an innovation: innovators, early adopters, early majority, late majority, and laggards (as cited in Wall, 2005, pp. 63-65). Using this model, Wall (2012) suggested researchers and innovators can understand what portion of the intended reform stakeholders may adopt or reject.

Wall (2005) cautioned that while using the model, innovators and researchers need to consider questions associated with Markee's discussion of the mechanism of washback: "who (participants), adopts (process), what (the innovation), where (the context), when (the time duration), why (the rationale), and how (different approaches to managing change)" (as cited in Cheng, 2005, p. 32). Markee's questions and Wall's (2005) model of washback would help researchers on washback who wish to investigate what to look for and predict what might not be achieved during the process of curriculum- or examination reform.

Washback on Learning

As noted, earlier studies on washback focused on the content and format of teachers' teaching practices, while more recent studies have tended to examine washback on aspects of learners' such as their content and strategy of learning and language skills (e.g., Gosa, 2009; Green, 2007; Qi, 2005; Shih, 2007). These studies also used a variety of methodological approaches to measuring the impact of tests quantitatively and qualitatively. Furthermore, Green (2007) and Shih (2007) proposed the extended model of washback, particularly for, but not limited to, learners and other stakeholders.

Gosa (2009) investigated Romanian pupils' English practice for the English Component of the Baccalaureate (ECB), their graduation examination. Green (2013) and Wall (2012) applauded Gosa's study because it is usually difficult to explore learners' attitudes and behaviors regarding washback; thus, not many studies have investigated them (see also Shih, 2007; Tsagari & Cheng, 2017). Gosa (2009), interestingly, explored learners' learning experiences by examining their journals. In her case study, she found that not only teachers but also students entirely neglected listening and speaking skills, since oral skills were not tested in the ECB. In both contexts, in and out of school, students merely focused on the tasks related to reading and writing, which they regarded as superior skills because they felt the need "to practice extensively the exam tasks" (p. 226).

Qi (2005) conducted a washback study on the National Matriculation English Test (NMET) in China to investigate why the intended positive washback effect failed. She interviewed and a surveyed the NMET designers, raters, teachers, and students. The test builders designed the examination by reducing grammar and vocabulary to encourage teachers and students to perceive English as a communication tool. The school textbooks also introduced many more communicative tasks to reflect the modifications in the test. However, traditional teaching and learning practices (e.g., drilling exercises, isolated linguistic learning of vocabulary, memorization, and practicing with sample test questions) continued to be used. The NMET uses multiple-choice items to measure learners' linguistic ability, and the immediate goal of teachers and students is just to get correct answers and consequently a high score on the test. To counter this tendency, Bailey (1996) argued that avoiding multiple-choice formats in a test would be conducive to good teaching and learning and would promote authentic use of language.

The findings of this paper support those of previous studies on washback that the classroom environment and the goal of the class are significant factors that impact the degree of washback (Gosa, 2009; Qi, 2005). Qi (2005) further argued that matching the curriculum with the test regarding content and format, so that the stakeholders' goal can be maximized. Green (2007) used a questionnaire and interviewed teachers and learners in three different classroom contexts: 1) the composition class for International English Language Testing System (IELTS), 2) regular writing class at an English for Academic Purposes (EAP) course, and a class that combined both syllabi. Not surprisingly, the participants engaged in different tasks, each of which reflected its own goal across the courses. Questionnaire and interview data from students showed that activities in IELTS include writing tasks that closely reflected sample topics from previous IELTS tests. Teachers in the EAP class taught not only writing practices on similar questions, but also different materials beyond the scope of the IELTS test.

Green (2007) provided evidence that some of the washback effect was mediated by the course materials as well as the teachers. This finding highlighted the learners' agency; i.e., the teachers did not determine washback on the learners. For instance, the questionnaire revealed that students' desired outcome or goal, improving their grammar proficiency, did not match that of the teachers, and moved away from the teachers' objectives. Regarding the learners' agency, Gosa (2009) also highlighted that learners' past experiences both in- and out of school may create different washback for learners than that of teachers. Therefore, Gosa suggested investigating washback not only in the classroom, but also outside the classroom because students are "more easily affected by washback than their teachers" (p. 226).

Given the importance of learners' perceptions and characteristics, Green (2007) proposed a model of washback that focuses on learners' values (see Appendix C). Although he emphasized learners' characteristics and values, his model neglected to provide a variety of intervening components that may mediate washback on learners. His model, however, did emphasize the washback intensity on learners; that is, learners' perception of test importance and difficulty. He argued that

participants only adapt a behavior if they believed it would improve their chances of succeeding. On the other hand, if they don't believe they can pass the test, teachers and learners may give up hope and therefore neglect the demands of the test.

Shih (2007) perceived Green's model as a shortcoming. He proposed a washback model for learners that could explain the washback of the General English Proficiency Test (GEPT) in Taiwan (see Appendix D). In addition to students, he interviewed parents, spouses, and friends since he himself had once taken the test and knew that they may have influenced the students in reacting against the test. The interview data corroborated the negative washback he had suspected. He concluded that GEPT had weak washback: the test was not high stakes; the students had sufficient language proficiency to pass it. Shih (2007) also raised a crucial extrinsic factor of negative washback of GEPT; that is, an ethical issue, which he referred to as a "loophole" (p. 149). Both universities offered a makeup examination for the students who failed the GEPT; however, the students could easily obtain the questions from the previous makeup test takers. Spratt (2005) called the issue a "tension between pedagogical and ethical decision[s]" (p. 24). In this regard, Hughes (2003) suggested a solution to the problem would be to "sample widely and unpredictably" (p. 54).

Shih's (2007) model of washback on learning examined both extrinsic and intrinsic factors of learners. He covered the factors that were proved empirically in prior studies. He described these with a solid line, while he used dotted lines to illustrate potential impact, as well as added a time axis (see Appendix D), since washback, including innovation, is not "an event but a process" (Wall, 2005, p. 63) and evolves over time (Shohamy et al., 1996).

Summary of Empirical Studies

The findings of washback studies point to major trends: teaching and learning can be impacted by macro- and micro factors, such as curriculum, perception, beliefs, and previous experiences, as well as a test. Moreover, a large discrepancy exists among stakeholders, which negatively and unintentionally impacts teaching and learning. However, some studies have proved Bailey's (1996, pp. 53-57) ingenious assumption of promoting beneficial washback: 1) Cheng (1999) and Wall and Alderson (1993) argued designers must ensure that the design of the test, not only its content, is clear to students and teachers; 2) Shih (2007) recommended sampling widely and unpredictably, and 3) Saif (2006) and O, K.-M. (2001) suggested using direct testing. Overall, it is unclear how washback operates on teaching and learning (Green, 2013; Tsagari & Cheng, 2017; Wall, 2012).

Researchers have used a wide range of models of washback and methodologies. A majority of earlier studies used quantitative measurements (i.e., questionnaire, test score, and statistical instruments) to investigate washback. Later, due to the complexity of the educational system and the relationship between a test and teaching/learning, as Watanabe (2004) argued, research on washback began to incorporate an ethnographic or qualitative approach to explore the washback effect. Many later researchers (e.g., Burrows, 2004; Green, 2007; Shih, 2007; Wall, 2005) conducted qualitative studies, while others used mixed methods, including a questionnaire, interviews, observations (e.g., Qi, 2005; Saif, 2006). Others creatively used journals to investigate students' learning processes in school and out-of-school (e.g., Gosa, 2009). Empirical studies mostly conducted interviews with significant participants such as teachers, learners, the learners' friends, spouses, parents, test designers, and administrators. Following Alderson and Wall's (1993) suggestion, researchers observed classrooms to find evidence: they examined how participants negotiate the curriculum and the test; they examined participants' perception of the value of the target language and their goal.

Models proposed in washback studies presented a variety of intervening variables in different contexts. However, they are not theoretically grounded (except for Wall, 2005) and they have not extensively reviewed washback models beyond those originally presented by Alderson and Wall (1993), Bailey (1996), and Hughes (1993). In addition, their sample of participants were either only teachers or only students. Therefore, they could not explain conflicting findings, such as individual differences, or offer a comprehensive understanding of washback on both teaching and learning (see Alderson & Wall, 1993, pp. 120-121, Hypotheses 9, 10, & 15).

A POTENTIAL WASHBACK MODEL: ACTIVITY THEORY

This section introduces a potential theoretical framework grounded in activity theory (AT) (Engeström, 1987; Grossman et al., 1999; Lantolf, 2000; Lantolf & Thorne, 2006; Leont'ev, 1978) to capture washback effects comprehensively, especially washback of high-stakes testing in which various macro-and micro-level intervening variables are intertwined.

To understand how a test impacts teachers' and learners' degree and depth of teaching and learning, it is first necessary to understand how teachers and learners behave in different contexts. According to Lantolf and Thorne (2006), AT, built on

Vygotsky's Sociocultural Theory (SCT), is useful for understanding the process of human development. Sociocultural Theory defines human behavior as being socially, culturally, and historically mediated (Lantolf, 2000). *Mediation*, which is central to SCT, refers to the process of culturally created auxiliary means that mediate higher forms of human mental activity (Lantolf & Thorne, 2006). As Thorne (2004) stated, the fundamental goal of AT is "to define and analyze a given activity system, to diagnose possible problems, and to provide a framework for implementing innovations" (p. 18). Indeed, to understand human behavior, it is crucial to explore an activity system, in this case, the context of washback in teaching and learning. Washback research investigates how the test organizes human social and mental activities that are powerful mediating artifacts in an educational system. Figure 3 below presents a diagram of human activity.

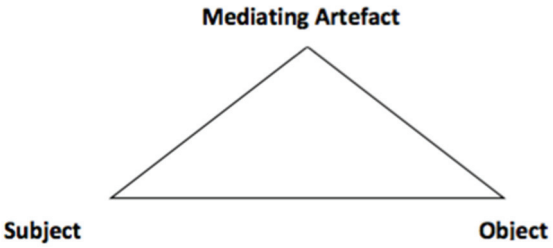


FIGURE 3
Model of Mediated Action (Vygotsky, 1978)

As shown in Figure 3, *mediating artifacts* represent the tools used to achieve the outcome of an activity (Lantolf, 2000; Thorne, 2004). They are categorized into two types: 1) *Conceptual tools* – teachers' philosophy and ideas about teaching and learning (Grossman et al., 1999). E.g., social constructivism; communicative language teaching, students' understanding of a linguistic concept, grammar rule, or rhetorical feature. 2) *Practical tools* – classroom practices, strategies, and resources to implement the concepts. E.g., test materials, textbooks, curricula, language activities, and exercises, etc. *Subject* refers to the unit of analysis (individual or group, here students, teachers, principals, and/or parents) whose agency is selected to be analyzed at the point of data collection. *Object* indicates the orientation of an activity (Leont'ev, 1978) that will be transformed into *outcomes* through mediating artifacts (Johnson, 2009, see Figure 4). Specifically, humans use culturally mediated artifacts (i.e., signs and tools), which serve as a buffer between human beings and their context (Lantolf, 2000; Lantolf & Thorne, 2006; Vygotsky, 1997). Through these artifacts, human beings carry out mediated, so-called object-oriented, actions (Lantolf, 2000).

There are times when the intended positive washback fails in an educational system in general, and a classroom setting in particular, instead indicating negative washback. Thereby, as noted earlier, it is necessary to synchronize conflicting findings, such as explaining individual differences more as pieces of a larger puzzle. In particular, Burrows (2004) suggested harmonizing a single washback response; that is, to search for patterns in teachers' responses to minimize unpredictable variations in their behavior. Although Burrows did not propose her model to explain learners' pattern of behavior, it can be used without risk to investigate learners' aspects. Since AT emphasizes context, it can help to equally consider using teachers and learners as units of analysis within the same study on washback; it can also unify diverse and contradictory research findings. Specifically, AT is concerned with enculturation and the causes and effects that facilitate and impede people's development (Grossman et al., 1999).[3] Activity theory also echoes Alderson and Wall's (1993) suggestion to look into the teachers' and learners' motivation behind their performance in different contexts.

It allows for exploring participants' development of culturally oriented goals (or objects), that is, "culturally defined futures that motivate people's activity and the sort of tools they develop in order to help mediation of one another's progress toward those futures" (Grossman et al., 1999, p. 5). In other words, AT focuses on individuals' motives and connects them with their behaviors.

Engeström's (1999, P. 31) increasingly complex model of an activity system can help to better explain the many contextual local factors that interact to impact an activity, as illustrated in Figure 4 below.

[3] See Lantolf, 2000, p. 8 for Leont'ev's original example of tribal hunting practices.

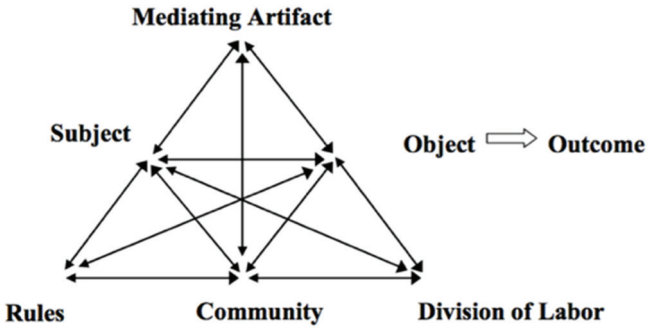


FIGURE 4
An Activity System Model

Engeström (1999) argues that human activity is not only mediated by artifacts, but also by the wider community as illustrated in the triangulation in Figure 4 (i.e., rules, community, division of labor). His model acknowledges the contextualized nature of activity and offers a framework that can put together local human activity with greater social, cultural, and historical structures (Lantolf & Thorne, 2006). This model enabled Engeström (1999) to consider human activities as social and collaborative. *Rules* are both explicit and implicit norms (Johnson, 2009) such as what counts as learning, knowing, assessing, or even failing in an activity system. A *community* consists of groups that share the same or similar objects, such as an English department, a teachers' union, fellow teachers, and so forth. We should note that, depending on the activity system and subject at the point of data collection, a community might shift from fellow teachers in a school to students in a classroom (Johnson, 2009). *Division of labor* includes both (horizontal and vertical) division of status and power (Center for Activity Theory and Development Work Research, 2020). It exists in any system controlled by rules; it determines who does what, how activities get done, and who holds status or power in a community (Johnson, 2009). The concept of division of labor indicates that collaboration is essential in any activity system. To complete an individual's actions, other members must collaborate in the same activity system. For example, language teachers cannot function without their students and vice versa.

Engeström's (1999) concept of rules shows that human activities are bound by explicit and implicit rules (Cole & Engeström, 1995). These rules can dictate one's actions toward achieving a goal. For example, a socially situated rule of a classroom is to actively participate in communicative activities based on the curriculum and the textbook. However, due to the context of high-stakes testing, having to follow rules conflicts with the communicative approach in a curriculum, where teachers and learners are expected to actively interact in the target language. Furthermore, as indicated by the multidirectional arrows in Figure 4, the components of the activity system (subject, rules, community, division of labor, mediating artifacts, object) of high-stakes testing have a direct or indirect impact on one another, which take place among different activity systems.

In the case of the CSAT reform, the Korean government (one activity system) required a new criterion-referenced test that was designed based on the curriculum with the intended positive washback (object) for teachers and learners in a language classroom (another activity system). However, due to their lack of awareness of the exam and of the theory behind the test, as well as the activities given in the textbooks (or for other reasons), tension may arise between the activity systems or between teachers and learners (see $Object_3$ in Figure 5). According to Engeström, $Object_3$, tension, represents the mediational space for potential collaboration between teachers and learners which could result in a change in teaching and learning (as cited in Tasker, 2011).

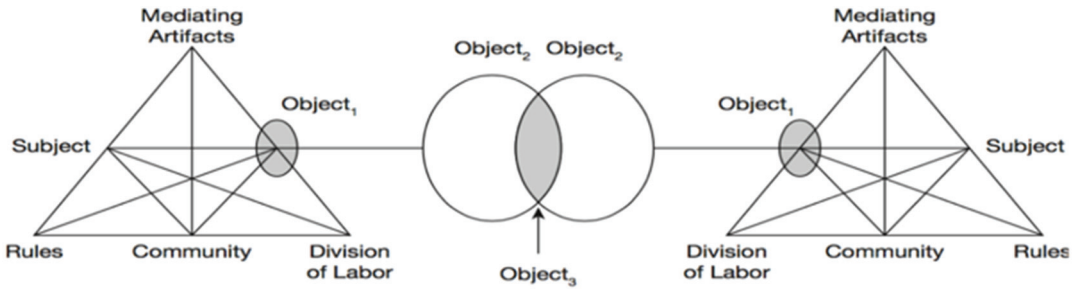


FIGURE 5
Shared Object Space in Two Connected Activity Systems

Seemingly independent human actions are in fact interwoven; they therefore influence and are influenced by each other. In his model, Engeström (1999) suggested that the unit of analysis must be situated in activity systems. A phenomenon (here, *washback*) can be observed in integrated ways by examining activity systems (here, various stakeholders) as a whole. For instance, when examining stakeholders' perceptions and practices (i.e., mediating artifacts) other components (i.e., rules, community, division of labor) will also be considered, since all components of the activity system (re)construct teachers' perceptions, practices, and objects.

DISCUSSION AND CONCLUSION

This review of washback research highlights the power attributed to the impact of testing on teaching and learning. As seen, a test, especially a high-stakes examination, has increasingly become the primary curriculum in a language classroom, imposing new textbooks, teaching methods, and learning strategies geared toward the test; thus, the test, rather than the curriculum, dictates how teaching and learning take place. Beyond Alderson and Wall (1993) and Wall and Alderson (1993), washback studies have identified only limited intended positive washback effects on teaching and learning of second and foreign language.

The absence of positive washback is found to be a direct result of restricting teaching and learning to skills that are tested in high-stakes examinations. This conclusion is consistent with O's (2001) washback research findings on CSAT. As Noble and Smith (1994) asserted, stakeholders of language tests should make a special extra effort toward achieving test-curriculum alignment. In addition, as other washback researchers have recommended, test designers must ensure that the design of the test, not only its content, is clear to students and teachers (Cheng, 1999; Wall & Alderson, 1993); they must sample test questions widely and unpredictably (Shih, 2007); and use direct testing (Saif, 2006; O, K.-M. 2001). In the case of CSAT, test designers may consider including speaking and writing sections to encourage teachers and learners to begin focusing on these skills when preparing for CSAT.

Prior research has also shown that washback is interwoven with the varying macro- and micro-level variables. Existing models of washback, however, investigate only teachers or only students. Therefore, they could not explain conflicting findings, such as individual differences, or offer a comprehensive understanding of washback on teaching and learning. The present study illustrates theoretical (AT) and empirical strategies appropriate for understanding and facilitating the impact of testing on teaching and learning. It uses activity theory as a unifying piece to explore washback on teachers and learners.

References

Alderson, J. C. (2004). Forward. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research context and methods* (pp. ix-xii). Lawrence Erlbaum Associates.

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.

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37-50). Lawrence Erlbaum Associates.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback - A case study. *System*, *30*(2), 207-223.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice* (Vol. 1). Oxford University Press.

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

Baker, E. (1991). *Alternative assessment and national policy*. Paper presented at the National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement.

Buck, G. (1988). Testing listening comprehension in Japanese university entrance exams. *JALT Journal*, *10*(1), 15-42.

Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 135-150). Lawrence Erlbaum Associates.

Center for Activity Theory and Developmental Work Research. (2020, March 14). An Activity System Model. https://www.helsinki.fi/en/researchgroups/center-for-research-on-activity-development-and-learning

Cheng, L. (1999). Changing assessment: Washback on teacher perspectives and actions. *Teaching and Teacher Education*, *15*(3), 253-271.

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

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopaedia of language and education: Language testing and assessment* (pp. 349-364). 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). Lawrence Erlbaum Associates.

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

Cole, M., & Engeström, Y. (1995). Mind, culture, person: Elements in a cultural psychology - Commentary. *Human Development*, *38* (1), 19-24.

Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Orienta-Konsultit Oy.

Engeström, Y. (1999). Activity theory and individual and social transformation. In Y. Engeström, R. Miettinen & R.-M. Punamáki (Eds.), *Perspectives on activity theory* (pp. 19-38). Cambridge University Press.

Estaji, M. (2013). Demystifying the complexity of washback effect on learners in the IELTS academic writing test. *Studies in English Language Teaching*, *1*(1), 211-226.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2011). *Educational research: Competencies for analysis and applications* (10th ed.). Pearson.

Gosa, C. M. C. (2009). *Investigating washback: A case study using student diaries*. VDM Verlag Dr. Müller.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Vol. 25). Cambridge University Press.

Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, *13*(2), 39-51.

Grossman, P. L., Smagorinsky, P., & Valencia, S. (1999). Appropriating tools for teaching English: A theoretical framework for research on learning to teach. *American Journal of Education*, *108*(1), 1-29.

Hughes, A. (1993). *Backwash and TOEFL 2000*. [Unpublished manuscript]. Reading: University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Hughes, A. (2010). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Johnson, K. E. (2009). *Second language teacher education: A sociocultural perspective*. Routledge.

Kim, D. J., & O, K.-M. (2002). Washback on 12th grade English classes of the English tests within Korean university entrance exams. *English Teaching*, *57*(3), 303-331.

Kim, S. H. (2014). Examining the students' perceptions on the English section of the new 2015 college scholastic ability test. *Modern English Education*, *15*(2), 157-176.

Lantolf, J. P. (2000). *Sociocultural theory and second language learning*. Oxford University Press.

Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural theory and L2. *Studies in Second Language Acquisition*, *28*(1), 67-109.

Lee, H. (2019). What are the effects of the change in the assessment systems on the alignment between curriculum, teaching and assessment? A case in Korean middle schools. *Journal of Asia TEFL*, *16*(1), 91-106.

Leont'ev, A. N. (1978). *Activity, consciousness, and personality*. Prentice Hall.

Madaus, G. F. (1988). The influence of testing on the curriculum. In N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the national society for the study of education* (pp. 83-121). University of Chicago Press.

McEwen, N. (1995). Educational accountability in Alberta. *Canadian Journal of Education*, *20*(1), 27-44.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). ACE and Macmillan.

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

Noble, A. J., & Smith, M. L. (1994). Old and new beliefs about measurement driven reform: "The more things change, the more they stay the same." National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://cresst.org/wp-content/uploads/TECH373.pdf

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology*, *8*(5), 375-387.

O, K.-M. (2001). *Washback on high school classrooms of the English tests within Korean university entrance exams* (Publication No.

3014794) [Doctoral dissertation, Columbia University]. ProQuest Dissertations and Theses Global.
Pan, Y. C. (2008). A critical review of five language washback studies from 1995-2007: Methodological considerations. *JALT Testing & Evaluation SIG Newsletter*, *12*(2), 2-16.
Pearson, I. (1988). Tests as levers of change. In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (pp. 98-107). Modern English.
Popham, W. J. (1987). The merits of measurement-driven instruction. *The Phi Delta Kappa*, *68*(9), 679-682.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high- stakes test. *Language Testing*, *22*(2), 142-173.
Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, *23*(1), 1-34.
Saville, N. (2000). Investigating the impact of international language examinations. *Research Notes, 2,* 4-7.
Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test. *Educational Measurement: Issues and Practice*, *9*(3), 15-22.
Shih, C. M. (2007). A new washback model of students' learning. *The Canadian Modern Language Review*, *64*(1), 135-162.
Shohamy, E. (1993). *The power of test: The impact of language testing on teaching and learning.* National Foreign Language Center.
Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, *13*(3), 298-317.
Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, *9*(1), 5-29.
Stobart, G. (2003). The impact of assessment: Intended and unintended consequences. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 139-140.
Tasker, T. (2011). Teacher Learning through Lesson Study: An activity theoretical approach toward professional development in the Czech Republic. In Johnson, K. E. & Golombek, P. R. (Eds.). *Research on second language teacher education: A sociocultural perspective on professional development* (pp. 204-223). Routledge.
Thorne, S. L. (2004). Cultural historical activity theory and the object of Innovation. In O. S. John, K. van Esch & E. Schalkwijk (Eds.). *New insights into foreign language learning and teaching* (pp. 51-70). Peter Lang Verlag.
Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revisited. In Shohamy, E., Or I. & May, S. (Eds.) *Encyclopedia of Language and Education: Language Testing and Assessment* (pp. 359-372). Springer.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.
Vygotsky, L. S. (1997). *The collected works of L.S. Vygotsky: The history of the development of higher mental functions* (W. Rieber, Vol. 4 Ed. & M. J. Hall, Trans.). Plenum Press.
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*(4), 499-509.
Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory* (Vol. 22). Cambridge University Press.
Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing*, *10*(1), 41-69.
Wall, D. (2012). Washback. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 79-92). Routledge.
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). Methodology in washback studies. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 19-36). Lawrence Erlbaum Associates.
Yoon, T-J. (2017). Measuring washback effect on learning English using student response systems. *Lingua Humanitatis*, *19*(1), 223-239.

Appendix A

Proposed View of Washback: A Curriculum Innovation Model (Burrow, 2004, p. 126)

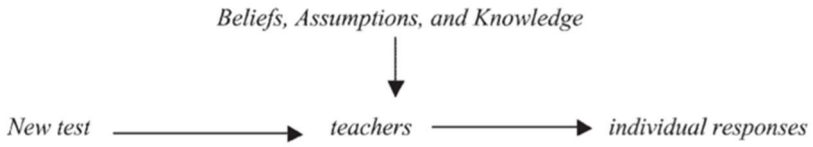
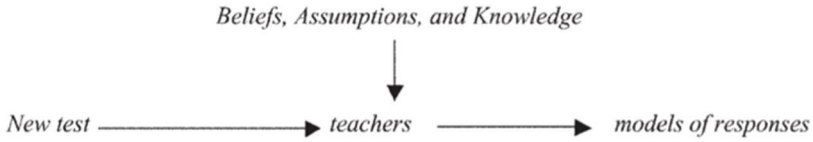
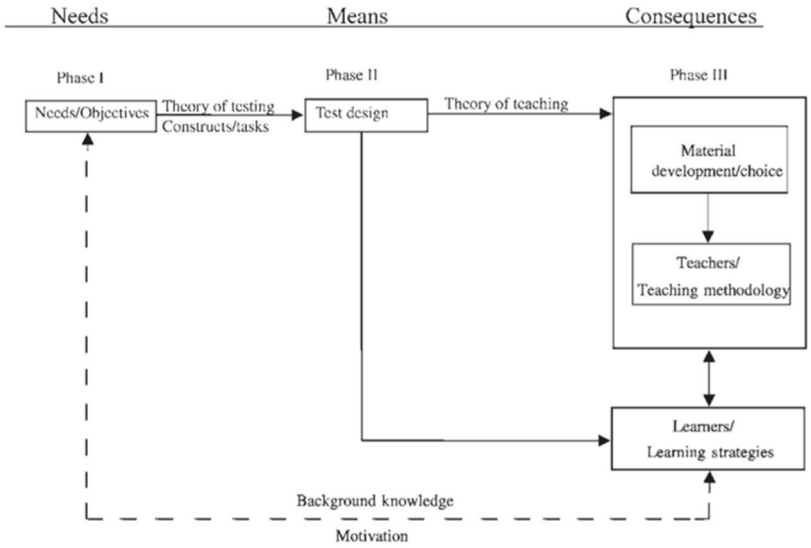


FIG. 7.5. 1990s view of washback: A "black box" model.



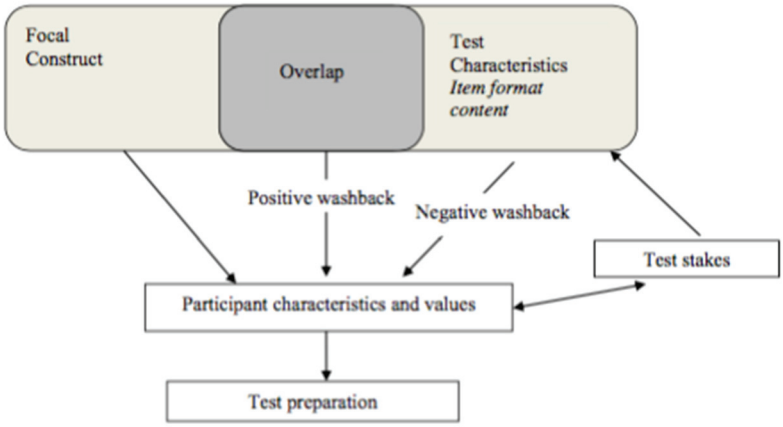
Appendix B

A Conceptual Framework for Washback (Saif, 2006, p. 5)



Appendix C

Model of Washback, Incorporating Intensity and Direction (Green, 2007, p. 24)



Appendix D

A Washback Model of Students' Learning (Shih, 2007, p. 151)