P-ISSN 2355-2794
E-ISSN 2461-0275

Washback or Backwash? Revisiting the Status Quo of Washback and Test Impact in EFL Contexts

Mohammad Ali Salmani Nodoushan*

Humanities Department, Faculty of Encyclopedia Research, Institute for Humanities and Cultural Studies, Tehran 19977-43881, IRAN

Abstract

It has been argued in the literature on (language) testing that any act of testing/assessment can impact (a) educators' curriculum design, (b) teachers' teaching practices, and (c) students' learning behaviors. This quality of any given testing situation or act of assessment has been called washback, or backwash if you will. Washback falls into the two categories of positive or negative—that is, beneficial or harmful. After an overview of the existing scholarly knowledge on washback, this paper argues that washback is not necessarily a test quality. Drawing on the notion of test method facets, the paper lends support to claims that see washback as a main function of teaching, learning, and policy-making situations or conditions rather than a quality of any given test. The paper also argues that the concepts of facet design and analysis including formal research designs, structural hypothesis testing, and measurement are inevitable and inescapable in any comprehensive model of washback. A possible borderline between backwash and washback is also proposed.

Keywords: Assessment, backwash, coaching, test impact, testing, washback.

* Corresponding author, email: dr.nodoushan@gmail.com, m.nodoushan@ihcs.ac.ir

Citation in APA style: Salmani Nodoushan, M. A. (2021). Washback or backwash? Revisiting the status quo of washback and test impact in EFL contexts. *Studies in English Language and Education, 8*(3), 869-884.

Received May 19, 2021; Revised July 6, 2021; Accepted August 4, 2021; Published Online September 16, 2021

https://doi.org/10.24815/siele.v8i3.21406

1. INTRODUCTION

Any act of assessment or testing, be it diagnostic or prognostic in nature, is sure to impact how the assessed party (a) gets ready for it through coaching or (b) tries to make up for its failure through remedial instruction (Salmani Nodoushan, 2009, 2018a, 2018b, 2020a). This is much more important where and when the stakes are high (e.g., in high-stakes selection tests such as the Iranian national *Konkoor*, the *TOEFL*, the *IELTS*, etc.). This means that any act of assessment has some bearing and impact on the teaching that precedes or follows it. As such, testing is sure to influence both teachers' teaching and students' learning (Alderson & Wall, 1993). This influence of testing on both the students' learning processes and the teachers' teaching preferences and approaches has come to be known as 'washback', which may alternatively be called 'backwash' (Cheng & Curtis, 2004). Wall (2012) has defined washback as the influence which high-stakes tests have on classroom practices—in particular, on learning processes and instruction procedures. Any specifically devised test whose stakes are high is normally put to functions that impact the lives of not only students but also educators and schools/districts—mainly for the purpose of accountability (Phelps, 2019).

Much of our theoretical knowledge of washback and test impact has been produced in the 1990s and, to a lesser degree, in the noughties. The second decade of the 21st century has also witnessed a good number of seemingly empirical studies of washback mainly published in notorious standalone journals and/or by despicable publishers; due to certain conservative reservations, I am not going to give any examples here, but the reader is invited to see a bibliographical list of titles on washback and test impact[1] and juxtapose that list with the lists of standalone predatory journals, and vanity/predatory publishers to get the whole picture. Journals and publishers that do not follow academic codes of conduct (e.g., peer-review, academic scrutiny, etc.) have been called predatory by Jeffrey Beall (cf., Buschman, 2020). This has unfortunately created a lot of misinformation—e.g., pseudo studies claiming to be based on true experimental designs that have employed fake data and a naïve set of *t*-tests for data analysis; such journals and publishers have blatantly polluted our academia. Juniors who are less versed in the field have shown signs of this pollution in their understanding of washback as well as many other themes and topics in applied linguistics—and I would not be surprised if our colleagues in other disciplines step forward and speak up to reveal the same problem in their academic fields.

This observation makes it ever more crucial for us to revisit all of the topics that we naïvely assume our students know with precision. As embarrassing as it may seem, this is not the case. As for washback, the *IJLS* (*International Journal of Language Studies*) was lucky to have trustworthy peer reviewers who honor academic codes of conduct and integrity; to date, they have rejected around 15 pseudo-manuscripts on washback from many parts of the world. This is quite alarming in and by itself and tells us loudly and clearly that we need to revisit older taken-for-granted topics intermittently to combat the pollution that lies therein and to correct, reshape, and upcycle our students' knowledge. In fact, many of the works I have published in the past have been informed by this call of duty, and now this paper follows suit. It (a) revisits washback, (b) overviews what we already know, and (c) links washback to facets theory.

2. LITERATURE REVIEW

As Alderson and Wall (1993) have suggested, a good number of (language) teachers and educators (a) believe that tests influence students and teachers as well as classrooms, and (b) feel that the test impact is for the most part negative and harmful. They have also argued that few empirical studies have in fact addressed such test impacts. Their article created a lot of interest in the study of washback, and empirical research projects began to inspect the various sides of the phenomenon. In Sri Lanka, they conducted another study (Wall and Alderson, 1993) and noted that (a) the influence of tests on teaching content had to be differentiated from (b) the impact of teaching methodology; this observation led them to emphasize that teachers' and educators' feelings could not be taken for granted and that researching and re-examining one's beliefs may lead to unexpected results; they themselves were surprised to observe that their belief about washback was not that true in the context of Sri Lanka (cf., Wall, 1996, 1999). They concluded that the development or use of new tests was much less impactful than what they expected. They therefore suggested that testing specialists need to modify their own beliefs about test impact. Based on this, Alderson (2004) argues that testing specialists need to research their own hunches and beliefs and avoid accepting them as given truths.

At the heart of the notion of washback lies the claim that test developers are the party to be held responsible for test impact and washback. It seems as if a test is washback-prone because its developers have not been careful enough when they were busy constructing it. Although this assumption may be partly true, Alderson and Hamp-Lyons (1996) noted (a) that it is often a commonplace, and for the most part false, belief, and (b) that teachers' way of teaching to the test is constitutive of washback. Alderson (2004) argues that the 'teacher' factor is perhaps more responsible for washback than the test-developer factor. On this assumption, he rejects Morrow's (1986) coinage of 'washback validity' for the relationship he had seen between tests and the curricula to which the tests pertain (cf., Brown, 1997). Morrow's (1986) 'washback validity' is tantamount to what has been called 'consequences of test use' in Messick's perspective on test validation. Messick (1996) saw washback as an issue relating to the consequential side of the concept of construct validation. He argued that positive washback, direct assessment, and authentic testing are interrelated. He concluded that any attempt at minimizing construct-irrelevant factors and construct underrepresentation can and will turn up into positive washback.

Messick (1996) has argued that washback has to do with the extent to which a test—specifically its introduction and use—can influence teachers and learners so that they start behaving in ways they would not otherwise do. He therefore takes washback to be a validity issue; in fact, he takes washback to have a direct bearing on the concept of 'systemic validity'. Seen in this light, washback is part and parcel of the consequential aspect of construct validity. Therefore, an appraisal of washback is not possible unless we see it within the broader picture of the whole field of (language) testing.

Other people have also offered their definitions of washback. Shohamy (1992), for instance, defined washback as "the utilization of external language tests to affect and drive foreign language learning," a condition which she takes (a) to be "the result of the strong authority of external testing" and (b) to influence "the lives of test takers" (p. 513). Likewise, Gates (1995) defines washback as "the influence of testing on

teaching and learning" (p. 101). Shohamy et al. (1996) define washback very simply as "the connections between testing and learning" (p. 298). The various definitions of washback that are seen in the existing literature, as well as the confusion that exists as to which of the two terms (i.e., 'washback' or 'backwash') is precise, imply that we still do not have a complete picture. In this paper, an overview of what we already know on washback is presented, the need for a redefinition is noted, and a revised definition is offered.

3. WHAT DO WE KNOW ABOUT WASHBACK/BACKWASH?

Any instance of testing is said to have a tacit or overt impact on the coaching or teaching that precedes as well as the remedial instruction that follows it (Alderson & Wall, 1993); hence, washback. This has led testers to develop unusual beliefs and hypotheses about the influence of large-scale high-stakes tests of proficiency and specifically the *TOEFL* test. According to Alderson and Hamp-Lyons (1996), prior to the upsurge of interest in empirical research on washback, the *TOEFL* test was believed to lead to result in (a) unusual teaching, (b) coaching students to learn and use test-taking strategies that are aversive to language learning, (c) teaching *TOEFL*ese, (d) encouraging students to attend testwiseness classes rather than real language courses, and (e) boosting test scores rather than language competency.

Nevertheless, Alderson and Hamp-Lyons (1996) were the first scholars to have the foresight to notice that commonplace beliefs and ideas concerning washback could not and should not be trusted and that any belief and personal idea has to be studied empirically. In an empirical study conducted to inspect washback in *TOEFL* classes, they concluded that the 'teacher' factor was responsible for a sizeable share of washback. Alderson (2004) expresses his surprise that, in spite of all of the beliefs that exist about *TOEFL* washback, no one has sought to conduct similar empirical studies on the topic.

Although 'washback', as a professional term, is absent from the literature on educational testing, professionals do agree that it exists; different professionals have used different terms or phrases to refer to it. Baker (1991) used the term "test impact" to signal the existence of washback. Likewise, Messick (1989) used the term "consequential validity" to refer to what has come to be known as washback. Similarly, Frederiksen and Collins (1989) have used the phrase "systematic validity" to refer to the impact of testing on teaching. Our current knowledge of washback reveals four major points about the impact of testing on curricula. First, Madaus (1988) and Cooley (1991) have argued that washback results in the 'narrowing of curriculum'. Second, Smith et al. (1989) have emphasized that washback brings about what they call "lost" instructional time—or what Helen Abadzi (2009) calls instructional time wastage. Next, there is the problem, as claimed by Frederiksen (1984) and Darling-Hammond and Wise (1985), of inadequate emphasis on skills that need more complex processing and problem-solving abilities. Lastly, Haladyna et al. (1991) have indicated that washback results in test-score contamination or pollution; it results in higher test scores whereas the students' competence of the construct at hand is not equally developed.

In explaining what they mean by test-score pollution, Haladyna et al. (1991) have argued that the scores of standardized achievement tests have unfortunately been

taken as indicators achievement in educational settings. They have also argued that this short-sightedness has paved the way for attempts at coaching (or teaching to the test) the job of which is not to enhance and improve educational-achievement but to raise standardized achievement test scores. They argued that such practices contaminate the conclusions that we make on the basis of scores. They suggest two main sources for such contamination: (a) the way schools get their students prepared to take a high-stakes test, and (b) the nonstandard conditions and procedures that accompany the administration of such tests; (p. 2)—an example of 'b' would be performance-based assessment. Along the same lines, Haladyna (1991) argued that factors such as test preparation, the context and situation of testing, and the importance that policy makers, families, educators, teachers, and the like attach to test scores lead to coaching (or teaching to the test), and pave the ground for washback to emerge.

In connection to TEFL, Alderson and Wall (1993) proposed the 'washback hypothesis' which claims that tests can affect both teaching and learning. They posited a series of logically plausible, albeit monotonous, alternatives. According to Alderson and Wall (1993), a test will influence (1) teaching, (2) learning, (3) what and how teachers teach and learners learn, (4) the rate as well as the sequence of instruction and learning, (5) the degree, quality and depth of instruction and learning, (6) the approaches and attitudes of teachers towards the contents and methods of instruction as well as the attitudes of learners towards learning, and (7) the consequences of tests. They argued further that we should also expect washback when tests are expected to have important consequences.

Tsagari (2007) noted that, in their washback hypothesis, Alderson and Wall had aimed at laying out the territory for subsequent research studies of washback. All in all, Alderson and Wall (1993) had sought to propose that any test or gauge might be held accountable for why some teachers engage in coaching towards it.

Before Alderson and Wall's work, Buck (1988) had described washback and argued that teachers and learners have a natural inclination towards adjusting their classroom behaviors and procedures in such a way as to ensure high test scores. This is understandable because school authorities often judge teachers' success on the basis of their students' test scores. Buck (1988) argued that this can be harmful if attempts at boosting test scores do not result in higher learning (See also Bailey, 2006).

This quotation points to the existence of two types of washback: (1) beneficial or positive, and (2) harmful or negative. While beneficial washback is claimed to result in the enhancement of educational achievement, harmful washback misinterprets higher test scores as true educational achievement. As Pearson (1988) argued, harmful washback is the quality of a test that is not well-constructed enough to mirror (a) the aims, objectives and goals of learning, and (b) the tenets of curriculum. On the other hand, beneficial washback is part and parcel of a test that is conducive to the desired competence that a course or curriculum aims to instill into students.

What went before clearly shows that all tests and particularly the so-called high-stakes standardized language tests (e.g., *GRE*, *TOEFL*, *IELTS*, etc.) have been condemned for their potential and practical harmful impacts on teaching; the notion of negative washback is not new, and people like Vernon (1956) have spoken about it ever since the 1950s. Vernon (1956, p. 166), for instance, argued that tests may "distort the curriculum" in that teachers and coaches who teach 'to' the test oftentimes discard all the teaching materials and curriculum content that they assume would not contribute to the expected test scores; they simply welcome materials that will help

their 'customers' (i.e., would-be testees) to pass their tests (cf., Cheng & Curtis, 2004). Moreover, the students who attend such classes pay to be coached, and as such, cannot be called 'genuine' students with strong intrinsic motivation who aspire to learn; with all due respect, they are 'fake' students, would-be fraudsters—or with some pity, 'pathetic pragmatic would-be testees'—who see coaching as a tool that guarantees their desired outcome whereby enabling them to snatch away what is not their right in life.[3] Likewise, their teachers are not genuine teachers. Rather, they are criminal forgers who make forged copies of 'genuine' students with no shame, nor any respect for the future of mankind. As Davies (1968) rightly says, tests have been devised mainly for (a) evaluation, (b) prognosis, or (c) selection purposes. However, what has happened in reality is that paid coaching classes have turned them into what I would like to call *a posteriori* coaching materials for a priori purposes. Likewise, Wiseman (1961) argued that coaching is in essence corrosive to the principles and practices of real teaching and a waste of precious educational time. Nevertheless, Alderson and Wall (1993) did not see negative washback as bane of all tests, but oftentimes an inherent quality of "poor" tests that comprise materials that students do not intend to learn—nevertheless, good tests may also comprise such materials.

In an explanation of why negative washback happens, Fish (1988) noted that in the context of general education, (1) pressure from outside and (2) teachers' age, anxiety, inexperience, and accountability were constitutive of teaching towards the test (cf., Cheng & Curtis, 2004). Likewise, Noble and Smith (1994a) argued that high-stakes testing is conducive to coaching and washback. Elsewhere they have argued that coaching is not likely to turn up into achievement and general understanding (Noble & Smith, 1994b; cf., Cheng & Curtis, 2004). Along the same lines, Smith (1991) noted that harmful washback is inevitable when stakes are high. She argued that tests may (a) reduce teaching time, (b) limit curricula, (c) shrink teachers' affordance, and (d) limit modes of teaching. Similarly, Anderson et al. (1990) noted that using high-impact tests in the context of general education may (a) lead teachers to narrow their choices of teaching topics, methods, and materials and (b) motivate students to resort to a memorization approach in place of critical thinking (cf., Cheng & Curtis, 2004). Along the same lines, Widen et al. (1997) argued that tests have the potential to corner teachers between a rock and a hard place where they have to (a) give up their autonomy and their discretion in relation to the curriculum, and (b) teach what they are expected to teach, not what they want to teach (cf., Calder, 1990, 1997; Cheng & Curtis, 2004).

All in all, harmful washback is more of a bane than a boon for education and specifically for language teaching. It circumscribes teachers, students, curriculum developers, families, and so forth. Nevertheless, the picture is not always that gloomy, and the good news is that tests, albeit in theory, can also create positive washback—defined as the potential, in theory, of tests to promote genuine education, instruction, and achievement/learning.

The notion of 'positive washback' in the field of applied linguistics is the counterpart of 'measurement-driven instruction' in general education (Cheng & Curtis, 2004; cf., Turner & Purpura, 2015). It should be emphasized that theoretical claims about useful and positive washback have not been documented by many empirical findings. Nevertheless, there are people who take positive washback for granted. For instance, Heyneman (1987) suggested that a good number of academic achievement testing specialists sincerely believe that coachability is a virtue and a boon, not a bane,

for education (cf., Pearson, 1988). Likewise, Davies (1985) claimed that (1) innovative testing may turn up into a change in syllabus, and (2) a new syllabus can also affect testing. However, the question that remained unanswered by Davies is: Should a test serve education or lead it?

Anyway, Alderson and Wall (1993) recommended that where possible, teachers and educators should engage in activities that enhance positive washback. Likewise, Hughes (1989) suggested that beneficial washback could be achieved if the language tester (1) tests only the desired abilities which he/she expects the earners to master, (2) bases his/her test on a wide and direct sample, (3) engages in direct testing, (4) draws on criterion-referenced testing, (5) implements objectives-based achievement testing, (6) makes sure both students and teachers know and understand the test, and (7) provides teachers with support and assistance.

In this connection, one should be note that direct testing avoids too much abstraction and engages in a direct evaluation of the cognitive skill of interest (Frederiksen & Collins, 1989). A test item is direct when the learner's response to it involves the actual performance of the language recognition and/or production task or the communicative skill of interest. According to Frederiksen and Collins (1989), direct tests result in positive washback in that teaching to the test (a) boosts test scores, and (b) culminates in "improved performance on the extended task and on the expression of the cognitive skill within the context of the task (i.e., teaching to the task will be teaching to the domain)" (Gipps, 1994, p. 102).

It is also important to note that criterion-reference testing, like direct testing, can enhance positive, and defy negative, washback. A criterion-reference test does not compare test-takers with one another; rather, it involves a standard criterion—or a common metric, à la Bachman (1990)—to measure the level of the ability of a test taker to perform a cognitive task. Just like IQ tests that are administered in an individualized manner, criterion-reference language tests can also show test takers' ability levels without comparing them with each other. The higher the score on a criterion-referenced test, the higher the ability level. As such, teaching to the test that aims at boosting test scores concomitantly boosts ability levels; hence, positive washback.

Nevertheless, some of the suggestions by Hughes (1989) are both costly and run counter to the test evaluation criterion of 'practicality' which comprises—along with reliability and validity—the 'sine qua non' of language testing (Salmani Nodoushan, 2020b, 2021b). Another question that remains unanswered is whether we should at all be concerned about the 'quality' of any act of testing and evaluation when we talk about tests? Washback, if understood correctly, is a quality of any testing situation (be it in educational settings or, say, culinary settings where restaurants compete for Michelin stars); washback is not an inherent feature of any test. It does not need a high IQ to understand that tests are just what they are supposed to be—i.e., TESTS. They are expected (1) to have the qualities that any test is expected to have (i.e., reliability, validity, and practicality), and (2) to serve the function to which they have been tailored (be it diagnosis, prognosis, selection, and so forth). Perhaps Messick's (1996) conception of construct validation that envisages 'consequences of test use' to be constitutive of the construct validation process has been misunderstood. It seems as if his implicit plea for a criterion-referenced common metric for any construct at hand has been misinterpreted as an explicit plea for qualitative alternatives such as portfolio

assessment, or what Wolf et al. (1991) have called 'thick' descriptions of achievement or profiles of performance.

Nevertheless, testing may always want to strive for positive washback, and as Hughes (1989, p. 47) has rightly argued, "Before we decide that we cannot afford to test in a way that will promote beneficial backwash, we have to ask ourselves a question: what will be the cost of not achieving beneficial backwash?" It is in this context that Caine (2005) believes the "pursuit of positive washback should remain a primary objective in language test design" (p. 26). Along the same lines, Shohamy (1992) argued that washback is the act of putting to use what she called 'external language tests' with the aim of impacting how foreign languages are taught and learned in school settings. Seen in this light, testing bodies external to the school have the power and authority to disturb the lives of test takers; they are the main source of washback. It is the question of how they behave that gives teachers, learners, and school authorities the motivation and the stamina they need to behave in erratic ways that are conducive, through negative washback, to higher test scores at the cost of true and genuine educational achievement. Messick's (1996) perspective on this process is that no instance of learning and/or teaching effects could be called the washback of a given test unless we could link that instance to the implementation and/or introduction of that gauge.

4. WHERE ARE WE NOW?

In traditional educational systems, tests often served diagnostic functions (e.g., evaluation of course achievement or goal attainment), but in the brave new world in which we live today, it seems that tests oftentimes serve selection or admittance functions (e.g., *TOEFL*, *TOLIMO*, *IELTS*, etc. scores being required for immigration to Canada, USA, etc.). This has virtually reversed the traditional teaching-testing direction, and some tests, especially high-stakes selection tests, currently precede educational/social programs rather than being subsequent to them. It is not surprising, therefore, that so many coaching classes have popped up here and there, and teaching to the test has changed into a luxurious money-making business with a huge financial turnover. It is in this context that washback is both inevitable and inescapable.

Similarly, it was on this ground that Messick (1989, 1996) came forward with a plea for the inclusion of concerns about the unwanted consequences and effects of tests in discussions of construct validation. Shohamy (1993) also noted that the interplay between within-and-beyond school factors and forces did indeed impact how tests are developed, introduced, administered, and wielded to grant or deny individuals access to certain resources. She warned that these factors and forces have some bearing on test validity and argued that aspects of test use have to be included in attempts at construct validation in view of the fact that tests do not operate in isolation. Likewise, Linn (1992) entreated testing scholars and organizations to heed the desirable and undesirable consequences of new testing systems, especially where they are put to high-stakes gate-keeping functions. In this context, Messick (1989) argued that attempts at construct validation will have to rely on both (a) test score interpretation, and (b) variables external to the test that operate in the social context where the test is put to use for gate-keeping functions (cf., Bracey, 1989; Cooley, 1991; Cronbach, 1988; Gardner, 1992; Gifford & O'Connor, 1992; Linn et al., 1991; Messick, 1994).

After all, any means of evaluating educational effectiveness is naturally expected to display a comprehensive, reliable, and accurate picture of educational progress and goal attainment.

Table 1. The Trichotomy Backwash Model.

(1)	**Participants**	students, classroom teachers, administrators, materials developers, and publishers, whose perceptions and attitudes toward their work may be affected by a test
(2)	**Processes**	any actions taken by the participants which may contribute to the process of learning
(3)	**Products**	what is learned (facts, skills, etc.) and the quality of the learning

Note: Based on Hughes (1993, p. 2), cited in Cheng and Curtis (2004, p. 12).

Bailey (1996) referred to this as an assessment function and noted that "The Trichotomy Backwash Model" presented by Hughes (1993) could account for the plethora of factors that lie at the heart of the perplexing mechanisms that lead to washback in learning and teaching contexts. The model argues that (a) participants, (b) processes, and (c) products have their parts to play in the emergence of washback. Table 1 (above) displays the model.

The model holds (1) that the essence of any gauge inevitably impacts the attitudes and/or perceptions of the people involved, (2) that these attitudes and/or perceptions, in turn, affect the instructional and learning processes, and (3) that these processes, in turn, determine the outcome or product of instruction and/or education. As such, this model implies a chain reaction or a domino effect.

5. WHAT NEXT?

Based on what went before, it can safely be argued that any consideration of what assessment or testing—and specifically language assessment—should do will have to encompass the characteristics and qualities of not only the tests themselves but also the testing conditions and the larger social settings. Test method facets (cf., Bachman, 1990) cannot be ignored, and we also need to remember that the methods in which we apply tests are administration-specific in that they vary from one administration to the next. Test developers have potential professional control over the lion's share of test method facets, and this is where they can and should heed precision to guarantee that harmful washback will not ensue.

Needless to say, (language) test performance can vary as a function of (a) examinees' ability levels, (b) their construct-irrelevant personality attributes, and (b) the characteristics of the test method. Test developers need to bring an informed well-defined framework for the totality of observations in the area to be tested to bear on their test construction tasks. They simultaneously need (a) to base their tests on empirical designs and observations within that definitional framework and (b) bridge between the definitional framework and the empirical structures (Guttman & Greenbaum, 1998) that is, they need to clearly define the construct to be measured and understand its place within its relevant cognitive domain, and then develop an empirically-based criterion-referenced test to measure that construct. Only in this way can they be sure that their definitions for behavioral domains are mapped on to, and provide the rationale for, the structural relationships that they envisage among the

plethora of variables that threaten to affect the integrity of the tests they construct—washback and test method facets included. As such, formal research designs, structural hypothesis testing, and measurement are inevitable and inescapable (cf., Guttman, 1959).

It should be noted that the term "facet" was first used by Guttman (1954) in a discussion of facets design and analysis. As for (language) testing, there are so many test method facets that researchers have not been able to discover all of them yet—let alone their being clearly defined and studied. Some of the facets that we already know include (a) instructions describing test takers' task, (b) the range and extent of the tasks, the test stimuli, and the test situations covered by the gauge, (c) the type of response expected from the testees, and (d) the way responses are evaluated (cf., Luoma, 2001; Wigglesworth, 2008).

All in all, test method facets can be distilled to include (a) facets of the assessment environment, (b) facets of the test input, expected response, and test rubrics, and (c) facets that relate to input-response relationship. According to Bachman (1990), testing environment facets include familiarity with the location and tools of testing, proctors and other people involved in the act of testing, the testing time, and the physical properties of the test as well as the testing location. Likewise, facets that pertain to test rubrics include test organization, time allocation, and instructions. Test organization, in turn, includes considerations such as the importance of the different sections and parts of the test, the sequencing of the various parts of the test, and the relative significance of each of its parts. Facets of instructions, in turn, include consideration of the language of the test (i.e., whether the test in students' native language or a foreign language), the channel of the test (i.e., whether the test is aural and/or visual), the specification of test procedures and/or tasks, and the directness and vividness of the criteria based on which the correctness of test takers' responses is to be evaluated. The relationship between test-input and response could be reciprocal, nonreciprocal, or adaptive (Bachman, 1990). Facets of the input include considerations of the test format and the nature of language. According to Bachman (1990), a consideration of test format should take the following into account: (a) channel of presentation, (b) form, language, mode, and vehicle of presentation, (c) recognition and description of the problem, and (d) decision as to whether the test is supposed to be a power test or a speed test. The nature of the language has to do with a consideration of length, propositional content, features of the organization of the test, and pragmatic aspects of the test. Finally, test response facets include considerations of test format (i.e., test channel, its mode, and response forms, types, and language), nature of test language, and restrictions on test response (Bachman, 1990). Needless to say, all of these have the potential to lead to washback, and test developers need to ensure harmful washback is not what these factors and facets can create.

6. CONCLUSION

The line of argumentation followed in this paper brings us to more or less the same realization that Alderson and Wall (1993) have already stressed: washback is not necessarily a test quality. This entails the idea that washback is more often a function of teaching, learning, and policy making than the quality of any given test unless of course, the test suffers from poor construction. Any attempt at studying washback

should investigate the educational context in which testing takes place. Test irrelevant forces that exist in the society where a test is introduced may be responsible for almost all of the washback that is naively blamed on the test. After all, where money and power talk, construct-irrelevant forces are sure to play their parts to snatch away what they can. All in all, "Testing is a profession, but it is highly susceptible to political interference. To a large extent, the quality of tests relies on the ability of a test agency to pursue professional ends autonomous [*sic*]" (Heyneman, 1987, p. 262, as cited in Cheng & Curtis, 2004, p. 11). Fear of test, fear of the consequences of a test, desire to snatch away whatever one can, and so forth are just a few construct-irrelevant factors that are sure to cause washback through coaching. Where tests are seen as levers for change, washback is sure to loom on the horizon.

One final remark is that washback and backwash are interchangeable alternatives. People like Alderson, who consider Alan Davies as the doyen of British language testing, may want to use washback, but those who idolize Arthur Hughes of the University of Reading may want to use backwash (Alderson, 2004). Perhaps a distinction can be made between the two terms to show the 'direction' of test impact. Backwash might better be used for *a posteriori* situations where a test bounces back to impact remedial instruction, and washback might better be kept for *a priori* situations where coaching (or teaching to the test) is at stake; the reverse might also be envisaged.

Notes:

1. For a list of titles on test impact and washback, please see: www.tirfonline.org/wp-content/uploads/2021/06/WashbackAndTestImpact_SelectedReferences_26May2021.doc
2. Such journals do not follow the prerequisite genre (cf., Salmani Nodoushan, 2012), review, and publication practices that any scientific journal is expected to follow. For a complete list of such stand-alone predatory journals and predatory publishers, please see: www.beallslist.net
3. For a discussion of pragmatic topics, please see Salmani Nodoushan (2006, 2016a, 2016b, 2018c, 2019, 2021a).

ACKNOWLEDGMENTS

My special thanks go to Professor Dan Douglas (Iowa State University), who read, reacted to, and commented on the first draft of this manuscript.

REFERENCES

Abadzi, H. (2009). Instructional time loss in developing countries: Concepts, measurement, and implications. *The World Bank Research Observer, 24*(2), 267-290. https://doi.org/10.1093/wbro/lkp008

Alderson, J. C. (2004). Foreword. In L. Cheng, Y. J. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contents and methods* (pp. 3-17). Lawrence Erlbaum Associates, Inc.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*, 280-297. https://doi.org/10.1177/026553229601300304

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129. https://doi.org/10.1093/applin/14.2.115

Anderson, J. O., Muir, W., Bateson, D. J., Blackmore, D. & Rogers, W. T. (1990). *The impact of provincial examinations on education in British Columbia: General report*. British Columbia Ministry of Education.

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

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*, 257-279. https://doi.org/10.1177/026553229601300303

Bailey, K. M. (2006). Issues in teaching speaking skills to adult ESOL learners. In J. Comings, B. Garner & C. Smith (Eds.), *Review of adult learning and literacy: Connecting research, policy, and practice* (pp. 113-164). Lawrence Erlbaum Associates.

Baker, E. L. (1991, September 4-6). *Issues in policy, assessment, and equity* [Paper presented]. National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement, Washington DC, USA.

Bracey, G. W. (1989). The $150 million redundancy. *Phi Delta Kappa, 70*, 698-702.

Brown, J. D. (1997). The washback effect of language tests. *University of Hawaii Working Papers in ESL, 16*(1), 27-45.

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *Journal of the Japan Association of Language Teachers JALT, 10*(1), 15-42.

Buschman, J. (2020). A political sociology of the Beall's list affair. *The Library Quarterly, 90*(3), 298-313. https://doi.org/10.1086/708959

Caine, N. A. (2005). *EFL examination washback in Japan: Investigating the effects of oral assessment on teaching and learning* [Unpublished master's thesis]. The University of Manchester.

Calder, P. (1990). *Impact of diploma examinations on the teaching-learning process*. Alberta Teacher Association.

Calder, P. (1997). *Impact of Alberta achievement tests on the teaching-learning process*. Alberta Teacher Association.

Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. J. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contents and methods* (pp. 3-17). Lawrence Erlbaum Associates, Inc.

Cooley, W. W. (1991). State-wide student assessment. *Educational Measurement: Issues and Practice, 10*, 3-6. https://doi.org/10.1111/j.1745-3992.1991.tb00209.x

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

Darling-Hammond, L., & Wise, A. E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85*, 315-336. https://doi.org/10.1086/461408

Davies, A. (Ed.). (1968). *Language testing symposium: A psycholinguistic approach*. Oxford University Press.

Davies, A. (1985). Follow my leader: Is that what language tests do? In Y. P. Lee, C. Y. Y. Fok, R. Lord & G. Low (Eds.), *New directions in language testing* (pp. 1-12). Pergamon Press.

Fish, J. (1988). *Responses to mandated standardized testing* [Unpublished doctoral dissertation]. University of California.

Frederiksen, J. R., & Collins, A. (1989). A system approach to educational testing. *Educational Researcher, 18*(9), 27-32. https://doi.org/10.2307/1176716

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychology, 39*, 193-202. https://doi.org/10.1037/0003-066X.39.3.193

Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 77-119). Kluwer Academic.

Gates, S. (1995). Exploiting washback from standardized tests. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 101-106). Japanese Association for Language Teaching.

Gifford, B. R., & O'Connor, M. C. (Eds.). (1992). *Changing assessments: Alternative views of aptitude, achievement and instruction*. Kluwer Academic.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. The Falmer Press.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258-348). Free Press.

Guttman, L. (1959). Introduction to facet design and analysis. *Proceedings of the Fifteenth International Congress of Psychology* (pp. 130-132). North-Holland.

Guttman, R., & Greenbaum, C. W. (1998). Facet theory: Its development and current status. *European Psychologist, 3*(1), 13-36. https://doi.org/10.1027/1016-9040.3.1.13

Haladyna, T. M. (1991, September 4-6). *Test score pollution: Implications for limited English proficient students* [Paper presentation]. Proceedings of the National Research Symposium on Limited English Proficient Student. Washington DC, USA.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7. https://doi.org/10.2307/1176395

Heyneman, S. P. (1987). Uses of examinations in developing countries: Selection, research, and education sector management. *International Journal of Education Development, 7*, 251-263. https://doi.org/10.1016/0738-0593(87)90023-X

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

Hughes, A. (1993). *Backwash and TOEFL 2000* [Unpublished manuscript]. Department of Linguistics, University of Reading.

Linn, R. L. (1992). *Educational assessment: Expanded expectations and challenges* (Tech. Rep. 351). University of Colorado at Boulder, Center for the Study of Evaluation.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Luoma, S. (2001). *What does your language test measure?* [Unpublished doctoral dissertation]. University of Jyväskyla.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. 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.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan.

Messick, S. (1994). The interplay between evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. https://doi.org/10.2307/1176219

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256. https://doi.org/10.1177/026553229601300302

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing*. NFER/Nelson.

Noble, A. J., & Smith, M. L. (1994a). *Measurement-driven reform: Research on policy, practice, repercussion* [Tech. Rep. 381]. Arizona State University, Center for the Study of Evaluation.

Noble, A. J., & Smith, M. L. (1994b). *Old and new beliefs about measurement-driven reform: 'The more things change, the more they stay the same'* [Tech. Rep. 373]. Arizona State University, Center for the Study of Evaluation.

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

Phelps, R. P. (2019). Test frequency, stakes, and feedback in student achievement: A meta-analysis. *Evaluation Review. 43*(3-4), 111-151. https://doi.org/10.1177/0193841X19865628

Salmani Nodoushan, M. A. (2006). A comparative sociopragmatic study of ostensible invitations in English and Farsi. *Speech Communication, 48*(8), 903-912. https://doi.org/10.1016/j.specom.2005.12.001

Salmani Nodoushan, M. A. (2009). The Shaffer-Gee perspective: Can epistemic games serve education? *Teaching and Teacher Education, 25*(6), 897-901. https://doi.org/10.1016/j.tate.2009.01.013

Salmani Nodoushan, M. A. (2012). A structural move analysis of discussion sub-genre in applied linguistics. *DacoRomania, 17*(2), 199-212.

Salmani Nodoushan, M. A. (2016a). On the functions of swearing in Persian. *Journal of Language Aggression and Conflict, 4*(2), 234-254. https://doi.org/10.1075/jlac.4.2.04sal

Salmani Nodoushan, M. A. (2016b). Rituals of death as staged communicative acts and pragmemes. In A. Capone & J. L. Mey (Eds.), *Interdisciplinary studies in pragmatics, culture and society* (pp. 925-959). Springer.

Salmani Nodoushan, M. A. (2018a). Implementation of the Beghetto-Kaufman-Baer approach to creativity and the four-c developmental trajectory in common core foreign language classrooms. In L. Caudle (Ed.), *Teachers and teaching: Practices, challenges and prospects* (pp. 157-174). Nova Science Publishers, Inc.

Salmani Nodoushan, M. A. (2018b). Toward a taxonomy of errors in Iranian EFL learners' basic-level writing. *International Journal of Language Studies, 12*(1), 101-116.

Salmani Nodoushan, M. A. (2018c). Which view of indirect reports do Persian data corroborate? *International Review of Pragmatics, 10*(1), 76-100.

Salmani Nodoushan, M. A. (2019). Clearing the mist: The border between linguistic politeness and social etiquette. *International Journal of Language Studies, 13*(2), 109-120.

Salmani Nodoushan, M. A. (2020a). English for Specific Purposes: Traditions, trends, directions. *Studies in English Language and Education, 7*(1), 247-268. https://doi.org/10.24815/siele.v7i1.16342

Salmani Nodoushan, M. A. (2020b). Language assessment: Lessons learnt from the existing literature. *International Journal of Language Studies, 14*(2), 135-146.

Salmani Nodoushan, M. A. (2021a). Demanding versus asking in Persian: Requestives as acts of verbal harassment. *International Journal of Language Studies, 15*(1), 27-46.

Salmani Nodoushan, M. A. (2021b). Test affordances or test function? Did we get Messick's message right? *International Journal of Language Studies, 15*(3), 127-139.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal, 76*, 513-521. https://doi.org/10.2307/330053

Shohamy, E. (1993). The power of test: The impact of language testing on teaching and learning. *National Foreign Language Center Occasional Papers* (pp. 1-19). The National Foreign Language Center.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*, 298-317. https://doi.org/10.1177/026553229601300305

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11. https://doi.org/10.3102/0013189X020005008

Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1989). *The role of testing in elementary schools*. Center for Research on Educational Standards and Student Tests, Graduate School of Education, UCLA.

Tsagari, D. (2007). *Review of washback in language testing: How has been done? What more needs doing?* [Unpublished manuscript]. Department of Linguistics, Lancaster University.

Turner, C. E., & Purpura, J. E. (2015). Learning oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banjeree (Eds.), *Handbook of second language assessment* (pp. 255-272). DeGruyter Mouton.

Vernon, P. E. (1956). *The measurement of abilities* (2nd ed.). University of London Press.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*, 334-354. https://doi.org/10.1177/026553229601300307

Wall, D. (1999). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory* [Unpublished doctoral dissertation]. Lancaster University.

Wall, D. (2012). Washback. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 79-92). Routledge.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*, 41-69. https://doi.org/10.1177/026553229301000103

Widen, M. F., O'Shea, T., & Pye, I. (1997). High-stakes testing and the teaching of science. *Canadian Journal of Education, 22,* 428-444. https://doi.org/10.2307/1585793

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Language testing and assessment. Encyclopedia of language and education* (vol. 7, pp. 111-22). Springer.

Wiseman, S. (Ed.). (1961). *Examinations and English education*. Manchester University Press.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74. https://doi.org/10.3102/0091732X017001031