A literature review of washback effects of assessment on language learning

Nguyen Thi Thanh Ha[1*]

[1]Ho Chi Minh City University of Education, Vietnam
[*]Corresponding author: hantt@hcmue.edu.vn

ARTICLE INFO

DOI:10.46223/HCMCOUJS.soci.en.9.2.257.2019

Received: October 23rd, 2019
Revised: December 9th, 2019
Accepted: December 13th, 2019

Keywords:

mechanism of washback, test impact, washback effect, washback

ABSTRACT

This paper reviews the empirical studies on washback effects of assessment on language learning. The study begins with the definitions of washback, its equivalent terms, and dimensions of washback. Then it summarizes the empirical studies of washback on three most frequently investigated areas, namely learners' motivation, behaviours, and achievement. Finally, it examines the mechanism by which washback on learning is generated. The findings show how complex and context-dependent test washback is and, based on these findings, the authors provide some recommendations for future research.

1. Introduction

Test washback on teaching and learning has long been of wide interest in general education. However, in language education, empirical evidence of the phenomenon only started to flourish in the 1990s, especially after Alderson and Wall (1993) posed their famous question "Does washback exist?" and proposed an agenda for future research. Since then, a considerable number of studies in language education have been done to seek empirical evidence for the widespread belief that tests have an impact on teaching and learning.

When referring to the effects of tests, language testers usually use two different terms: *washback* and *impact*. Washback is commonly understood as the influence of testing on teaching and learning (Alderson & Wall, 1993; Bailey, 1996; Cheng, Watanabe, & Curtis, 2004; Hamp-Lyons, 1997; Messick, 1996; Tsagari, 2007). However, there is not an unanimous understanding of impact. For some language testers, the impact is a broader construct referring to "any of the effects that tests may have on individuals, policies or practices, within the classroom, the school, the educational system, or society as a whole" (Tsagari, 2007, p. 4), and thus washback is only one of its dimensions. Other language testers distinguish washback and impact as micro and macro effects of testing within society with washback as the effects of tests on the teaching/learning context and impact as the effects on the teaching and learning context (Taylor, 2005). This paper used these terms as defined in previous studies.

The nature of washback can be described in five aspects or dimensions: specificity, intensity, lengths, intentionality, and value (Watanabe, 2004). In terms of *specificity*, washback can be general or specific. "General washback means a type of effect that may be produced by

any test" (Watanabe, 2004, p. 20). In the same way, Alderson and Wall (1993) hypothesize that a test impacting the content taught by teachers can be considered general washback. On the other hand, specific washback "refers to a type of washback that relates to only one specific aspect of a test or one specific test type" (Watanabe, 2004, p. 20). Thus, a finding that a multiple-choice test does not encourage learners to learn productive language skills, for instance, relates to specific washback.

Intensity, which is synonymous with the term *extent* used by Bachman and Palmer (1996), describes the strength of washback. It was first coined by Cheng (1997) "to refer to the degree of washback effect in an area or a number of areas of teaching and learning affected by an examination" (p. 43).

Length is used to describe how long the washback of a test lasts (Watanabe, 2004). This aspect of washback can be perfectly illustrated by Shohamy, Donitsa-Schmidt, and Ferman's study (1996). The effect of the ASL existed only before it was first administered, so this effect is short-term. However, the washback of the English as a Foreign Language (EFL) exam still persisted many years after its introduction, so this is a long-term washback.

Washback can be intended or unintended. To the best of our knowledge, language testers, including Watanabe, have not given an official definition for the term *intentionality* yet. However, a number of them have acknowledged and discussed this aspect of washback in their papers (Andrews, 2004; Messick, 1996; Tsagari, 2007).

Watanabe's term *value* is equivalent to *direction*, which is used by many other authors (Alderson & Wall, 1993; Bailey & Masuhara, 2013; Green, 2007a; Tsagari, 2007). Washback may be positive or negative. According to Bailey and Masuhara (2013), the judgment of washback as positive or negative depends "on our view of the desirable outcomes of language learning" (p. 304).

To date, there have been very few empirical studies of washback on language learning compared to those on teaching (Cheng, Sun, & Ma, 2015; Damankesh & Babaii, 2015; Shih, 2007). This paper reviews these studies to give an overview of the current state of knowledge on the research topic and make some recommendations for future research. It examines washback on three most frequently investigated areas, namely learners' motivation, behaviours, and achievement as well as the mechanism of washback. Washback is discussed in terms of its intensity and direction. Many studies investigated more than one area simultaneously so the same study may be mentioned in different sections of the paper.

2. Washback on learning

2.1. Washback on learners' motivation intensity and direction of washback

Research findings of test washback on learners' motivation are mixed. Some studies (Li, 1990; Shohamy, 1993) show positive effects of language tests on the motivation of most students. For example, Li (1990) found positive changes in students' learning after adding the language use component to the Matriculation English Test in China. These changes most clearly manifested through "a new enthusiasm for after-class learning of English" with increased extra learning materials and activities, i.e., reading readers and journals, listening to the radio, and watching TV (p. 401). The researcher attributed the positive changes to the change in the test design - the weight put on the use of language component of the test.

Hirai and Koizumi (2009) also reported positive washback effects on students' motivation, but unlike Li (1990), who investigated the effects of a high-stakes test, they examined a classroom test called the Story Retelling Speaking Test. A questionnaire was used to find out test-takers' perceptions of the test qualities. The results showed that students were motivated to learn English by the test with a relatively-high mean score for the motivation question. A few students even commented that the test was very interesting and beneficial. Unfortunately, no explanation for this positive effect of the test was given in the paper.

In contrast, some other studies found only limited or even detrimental washback effects of testing on learners' motivation. Shih (2007), for example, examined the washback of the General English Proficiency Test (GEPT) on English learning at two departments of two Taiwanese universities. One department had no GEPT requirement, while the other required students to pass the first stage of the GEPT's intermediate level or their own in-house exam. Qualitative data were collected by interviewing different stakeholders; observing teachers, activities in the self-study center, and course meeting; and reviewing the universities' documents and records. The researcher found that GEPT had detrimental effects on some students' motivation for learning English. One of the students, who failed the test three times, said that "the GEPT had gradually eroded his self-confidence and eventually snuffed out his learning motivation" (p. 144). His failure might be due to the mismatch between the content of the test and the school's teaching.

Cheng (1998) and Pan and Newfields (2012) found only minimal test impact on students' motivation. However, it is noticeable through their research that motivation was not measured in terms of level but type, or more specifically, reasons for learning English. Examining the impact of the new revised Hong Kong Certificate of Education Examination in English (HKCEE) on students' English learning, Cheng (1998) administered the same questionnaire to two different cohorts of students over a two-year period. The first cohort took the old version of the HKCEE, and the second took the new integrated and task-based version of the test. Comparing the data, she found that "the changes in students' motivation and learning strategies remained minimal" (p. 280). Only three out of nine reasons for learning English showed significant change after two years: *meeting the requirements of the society and watching English movies and listening to English programmes* became more motivating, while *fulfilling parents' expectations* was less motivating. Cheng considered the changes in students' motivation, to some extent, the washback effect of the new test on student learning because "these types of motivation were also related to the requirements of the new 1996 HKCEE" (p. 295).

Using a research method similar to Cheng's (1998), Pan and Newfields (2012) investigated the washback of the EFL proficiency graduation requirements (EGR) on university students' learning in Taiwan. They carried out a survey with two groups of students - one at schools with such requirements and the other at schools without them. To determine the impact of EGR on motivation, they also compared the two groups' responses to questions about reasons for learning English, many of which were borrowed from Cheng's (1998) questionnaire. They found that only three out of twelve reasons (*to earn certificates, to pass the test to graduate,* and *to improve their English for further education*) had statistically significant differences, and the effect sizes for these differences were only small. The researchers associated these changes with the pressure of the EGR on students. They also found that EGR appeared to somewhat

motivate some EGR students, but impede low ability students.

Washback variability

Washback may vary significantly depending on a variety of factors. Tests of different statuses can produce different effects on motivation. Stoneman (2006) compared the effects of the Graduating Students' Language Proficiency Assessment (GSLPA) and the IELTS used as Hong Kong's territory-wide Common English Proficiency Assessment (IELTS-CEPAS) on students at Hong Kong Polytechnic University. She found that the percentage of the respondents who sat the IELTS-CEPAS (74.9%) were more highly motivated to prepare for it than that of the respondents who sat the GSLPA (18.8%). The reason was that the IELTS had a higher status than the GSLPA did. Also, 81.8% of the IELTS-CEPAS students did not put as much effort into preparing for the test as they did into other public exams they had taken. This was due to the fact that the IELTS-CEPAS was essentially low stakes. It was only optional and students could choose not to include the test results in their transcripts.

The same test may have different impacts on different groups of students (Ferman, 2004; Gan, Humphreys, & Hamp-Lyons, 2004; Shohamy, 1993). The most important factor that caused the variability in washback on students' motivations was probably their ability. The EFL National Oral Matriculation Test in Israel, for example, produced different effects on the three different ability levels: "The lower the students' ability level, the higher the intensity of learning" to the test (Ferman, 2004, p. 199). Gan et al. (2004), on the other hand, found that successful students were more willing to prepare for and take tests.

Tests of different nature can generate different effects on learners' motivation. Huang (2011) compared the effects of two different types of assessment, convergent assessments (CA) and divergent assessments (DA) on grammar learning. CA aims to determine whether learners have mastered the knowledge and skills predetermined by the assessor, while DA focuses on what learners know, understand, and can do. Motivation as measured in this research included five subcomponents: perceived task characteristics, perceived self-efficacy, amount of invested mental effort, mastery goal orientation, and performance goal orientation. In speaking classes, the students were more motivated by the DA than the CA, while, surprisingly, in listening class, the CA had stronger motivational effects than the DA. Huang attributed higher motivation for the DA in the speaking classes to the higher-order thinking and depth of engagement required by the DA. As for listening class, the students' unfamiliarity with the type of task used in the DA might be an influential factor as students' self-efficacy about a particular classroom assessment depending on their previous experiences with similar kinds of assessments. Their lack of experience with the DA listening test might have resulted in lower perceived task characteristics, lower self-efficacy, a lower amount of invested mental effort, and lower performance goals, i.e., lower motivation.

2.2. Washback on learners' behaviours

In this paper, the effects of tests on learners' behaviours are understood as the changes in what and how they learn as a result of testing requirements. Some of the studies reviewed in this section have already been summarized in Section 2.1, so only their findings related to learners' behaviours are mentioned here. Due to a small number of studies on this aspect of test washback, the empirical evidence is still far from conclusive.

Minimal washback

Some studies found an only superficial influence of tests on learners' behaviours (Andrews, Fullilove, & Wong, 2002; Cheng, 1998; Pan & Newfields, 2011, 2012; Shih, 2007). Cheng (1998) found that only four out of eleven preferred learning strategies were significantly changed during her two years of research. In particular, note-taking, a skill required by the new HKCEE, was still the least preferred, and the slight increase in the mean score of this item was not statistically significant. Similarly, Pan and Newfields (2012) also discovered only minimal washback of the EGR on university students' learning in Taiwan. Their data showed that the test requirements "did not lead to a noteworthy amount of 'studying for the test'" (p. 118). To a small extent, EGR students might have adopted communicatively oriented and test-preparation approaches, which resulted in a slight increase in their productive skills. However, most of the students in both groups still embraced traditional methods of learning, i.e., text reading, rote memorization, and practicing grammar exercises. In line with other researchers, Pan and Newfields (ibid) attributed students' traditional learning methods to their teachers' traditional teaching methods. However, their conclusion about the impact of EGR based on the comparison of the two EGR and non-EGR groups of students should be taken cautiously because EGR was maybe not the only factor responsible for the differences between them. The compositions of the two groups of students were very different in terms of majors. The majority of the non-EGR group (67%) were Business Management students, while more than half of the EGR group (52.2%) were Engineering students. Engineering students could be very different from Business Management students in their motivation and ability to learn English.

More positive washback

A few studies (Allen, 2016; Hung, 2012; Xiao, 2014) found positive washback of testing on learners' behaviours. Allen (2016) investigated the effects of the IELTS test on students' test preparation strategies and score gain. Two hundred and four undergraduates, who were all high academic achievers in a Japanese university, voluntarily took the IELTS test twice during one year. In order to find out their test preparation strategies, the researcher carried out a survey after the students' second test. The results showed that overall the washback intensity was still relatively weak, but the washback direction was positive. The students studied the productive skills (speaking and writing) more, practiced more spontaneous speaking, and engaged more in speaking activities involving both daily and abstract topics. These changes were considered positive because they were important for the target language use domain. A closer look at the group of students who most intensively prepared for the test showed that they also practiced listening more. The interviews with the students showed that various factors were involved in shaping the test washback. These factors will be discussed in Section 2.3.

In a different context, Xiao (2014) used a questionnaire to examine the washback effects of the CET test on Chinese non-English majors' test-taking strategies as well as the intensity and direction of washback. The 284 participants of the study came from two universities in Southeast China, a proportion of them having sat the CET. Generally, the participants used test-management and test-wiseness strategies more frequently than cognitive and metacognitive strategies. However, the comparison between those joining and not joining the test revealed that the former used all four types of strategies more often than the latter, with the difference in cognitive strategies being statistically significant and the effect size approaching the medium

level. The researcher suggested that the washback of the CET test on strategy use was positive because cognitive strategies involve the use of language ability. Although the differences in test management and test-wiseness strategies were not statistically significant with the significance level being slightly higher than .05, the researcher believed that the test had weak washback on the use of these strategies, and this washback was both negative and positive in nature because test-wiseness strategies "involve using abilities to exclusively rely on test facets of the environment to answer test items" and test management strategies involve "both the language ability and the exploitation of test characteristics."

While Allen (2016) and Xiao (2014) examined the washback effects of high-stakes tests, Hung (2012) explored the influence of an alternative assessment technique, e-portfolios, on 18 student teachers in a Master's program in Teaching English to Speakers of Other Languages. The researcher corroborated data obtained by multiple qualitative methods including interviews, observations, document analysis, and reflective journals to collect data. The findings showed that e-portfolio assessments produced positive washback on learning, which included "building a community of practice, facilitating peer learning, enhancing the learning of content knowledge, promoting professional development, and cultivating critical thinking" (Hung, 2012, p. 33).

More negative washback

A few studies (Damankesh & Babaii, 2015; Ren, 2011; Zhan & Andrews, 2014) show more negative washback on learners' behaviours than a positive one. Ren (2011) reported findings from a pilot study of the washback of the CET-4 on English teaching and learning at five universities in China. Data were collected by means of questionnaires and semi-structured interviews. Findings suggested that the test affected both learning content and methods. Although the test drove students to learn English, it generally exerted negative effects on what and how students learned. Students tended to study for the test by memorizing CET-4 word lists and doing past and practice tests, which, according to the researcher, was not useful for non-test contexts. Additionally, they learned writing by memorizing model answers because the test task was highly predictable. Some students admitted ignoring the textbook tasks that were not included in the CET-4 test. The researcher blamed CET-4 for "students' low proficiency of using English for real-life purposes" (Ren, 2011, p. 258), and he recommended that the test should include a speaking component to promote teaching and learning for real-life communication. It is unclear why CET effects on learning strategies are negative in this study but positive in Xiao's (2014).

Damankesh and Babai (2015) studied the washback effects of high school final exams on students' test-taking and test preparation strategy use. The tests included three sections: vocabulary, grammar, and language function. Eighty Iranian male high school students joining the high-stakes national English test at the end of their school year participated in the study. Information about the students' psychological processes and strategies were collected through a think-aloud methodology. The results showed that the tests negatively affected students' use of strategies because they drove them toward "a measurement-driven approach to learning" (Damankesh & Babaii, 2015, p. 67), i.e., learning to the test including rote memorization, practicing grammar exercises, doing past tests, reviewing teachers' notes, and studying exam-related parts of the book. Students also had some strategic behaviors while taking the test such

as cheating, random guessing, translating words, phrases or sentences directly into Persian, eliminating bad alternatives, and copying words or phrases from previous items and alternatives. However, the tests also had a slightly positive effect on learning by generating some desirable behaviours during the test such as guessing meaning based on lexical cohesion, employing reasoning, applying world knowledge, and considering semantic and grammatical clues. All of these may foster students' cognition and attention.

2.3. Washback on learners' achievement

Most studies on learners' achievement focus on the relationship between direct test preparation and students' test scores. These studies often used the same method of comparing test scores of groups of involved students to different extents in direct test preparation. The findings showed various influences of direct test preparation on score gains.

No clear washback

Green (2007b) and Read and Hayes (2003) both studied the effect of direct IELTS preparation on score gains of students who were preparing for their academic/tertiary studies in an English speaking country and found no clear evidence of test impact. Green (2007b) examined the influence of direct test preparation on writing scores. His study involved three different groups of international students in the UK. The first group (85 students) participated in IELTS preparation courses, the second one (331 students) - in English for Academic Purposes courses (EAP) (with no IELTS component), and the third one (60 students) - in combination courses (EAP courses with IELTS preparation strands). All participants completed IELTS writing tests at the course entry and exit. They also responded to a questionnaire about factors that might account for the differences in the mean score gains. The results showed that the groups with IELTS preparation did not improve their scores compared to the EAP group. They also suggested that, in the context of the study, the individual learner's response to the demand of the test affected their outcomes more than the content of their classes.

Read and Hayes (2003) compared score gains on the IELTS listening, reading, and writing tests of international students participating in two IELTS courses at different language schools in New Zealand. Course A mainly aimed at test preparation. Course B used a topic-based approach and focused not only on test tasks, but also on the development of language knowledge and academic skills. Retired versions of IELTS were used for pre-and post-testing. Paired-samples t-tests showed no statistically significant differences in the overall pre- and post-test scores of both courses. There was only a statistically significant difference in the listening scores in course A, which might have resulted from a large amount of class time spent on listening tests and exercises.

Gan (2009) also looked at the influence of IELTS preparation on score gains. However, in his study, IELTS was used as an exit language test for university students in Hong Kong. Comparing the exit IELTS test scores of students who had taken an IELTS preparation course before the test and those who had not, Gan found no statistically significant difference between the groups' scores. However, the findings also showed that the two groups were significantly different in their university English entrance exam scores, and that students with lower entrance scores were more likely to take a test preparation course before the exit test. Gan suggested that university English learning, IELTS preparation in particular, made an important contribution to

narrowing down the gap between the two groups in the entrance exam scores. Thus, the relationship between test preparation and score gains in Gan's study may be more positive than those conducted by Green (2007b) and Read and Hayes (2003).

Weak positive washback

Some other researchers found only a small influence of test preparation on score gains. Robb and Ercranbrack (1999) investigated the effect of TOEIC preparation on score gains. They used two samples of students - English major and non-major students. They assigned each sample to three different treatments: TOEIC preparation, Business English, and General (four-skills) English. The results showed only statistically significant gains for non-majors' reading component. However, the researchers were cautious about the results, saying that they were "by no means conclusive" because the non-major students had much lower pre-test scores than the English majors did, and so they could benefit from test preparation more than higher ability students.

Stronger positive washback

A few studies (Allen, 2016; Muñoz & Álvarez, 2010; Safa & Goodarzi, 2014; Saif, 2006), however, reported strong positive effects of tests on learning outcomes. Like Green (2007b), Read and Hayes (2003), and Gan (2009), Allen (2016) investigated the washback effects of the IELTS test on score gain. However, the students in his study did not take a test preparation course. Instead, they prepared for the test independently and took the test twice within a one-year period. Their test scores were analysed using a paired-samples t-test. The results showed an improvement in the speaking scores for all participants and also listening scores for those who prepared most intensively for the test. Students with lower initial proficiency also gained more than those with higher initial proficiency.

Another example of studies showing positive washback on learners' achievement is one done by Saif (2006). The researcher investigated the washback of a needs-based test of spoken language proficiency. The test was developed for international teaching assistants (ITAs) based on an analysis of their needs. The subjects of the study were then divided into two groups. The control group continued with the old orientation programme for ITAs. The experimental group took a new course, which was taught by an English as a Second Language teacher who was involved in the preliminary administration and scoring of the test. The results showed that the experimental group "performed significantly better than the control group". (Saif, 2006, p. 26) attributed the superior performance of the experimental group to the improved teaching which was aligned with the needs-based test in terms of teaching objectives, content and activities, and, therefore, suggested a cause-effect relationship between the test and the learning outcomes.

Another experiment carried out by Muñoz and Alvarez (2010) to assess the impact of the oral assessment system (OAS) on teaching and learning outcomes in EFL classrooms in Columbia also showed positive effects of the test on learning outcomes. The experimental group improved their scores in some areas, including communicative effectiveness, grammar, and pronunciation. This improvement may be because the teachers of the experimental group emphasized more on assessing "not only linguistic competence, but also communicative competence as a whole." Other factors might have also played a role. Those are the ongoing training on assessment practices provided to the teachers, students' awareness of the test objectives, criteria, procedures, and techniques. In both studies by Muñoz and Álvarez (2010)

and Saif (2006), the positive washback of the test was a result of stake-holders' clear understanding of the test they were aiming for and the alignment of teaching/learning with the test. This supported other language testers' opinions that only introducing a new test alone is not enough to make positive changes to teaching and learning.

All the above studies examined the effects of tests on students' achievements quantitatively, i.e., through their test score gains. Andrews et al. (2002), however, investigated this aspect of test impact not only "quantitatively," but also "qualitatively." The researchers looked at the influence of the Use of English (UE) oral exam on students' performance in spoken English. A spoken test was administered to three different cohorts of Secondary 7 students in three consecutive years in 1993 (before the introduction of the exam), 1994, and 1995. A comparison of the mean scores showed an apparent, although not statistically significant, improvement in students' test scores between 1993 and 1995, suggesting a positive effect of the EU exam on students' performance. The analysis of students' recorded performances also revealed the changes in their organizational and language features. For example, students stopped introducing themselves when starting group discussion, probably, because they had become familiar with the exam format. Another example is students' use of a wider range of language to introduce their presentation topic. Sometimes the changes in the content of students' performance reflected the influence of published materials, which acted as a mediating factor in the relationship between the test and students' performance. Andrews et al. (2002) also noted that the test impact was delayed, with more obvious differences in the second year of the test.

2.4. Mechanism of washback

Many studies above revealed several factors influencing washback. The intensity and direction of washback may depend on the test design (Li, 1990; Muñoz & Álvarez, 2010), the alignment between teaching and test characteristics (Saif, 2006), published materials (Andrews et al., 2002), etc. However, no explanation of the mechanism of washback is given in many cases (Hirai & Koizumi, 2009). It is still unclear, for example, why IELTS test preparation had no effect on score gain in one case (Green, 2007b; Read & Hayes, 2003) but had a positive effect in another (Allen, 2016). There is a need to find out what bring about washback and how it occurs. Below is a review of some studies that attempt to explore the mechanism on which washback on learning operates.

Some researchers explore the mechanism of washback using a qualitative approach. In her study, Shih (2007) found that the English exit exam (GEPT) for university students in Taiwan generated little washback on their learning although it was supposed to have important consequences for them. She argued that the low-stakes and high-stakes dichotomy could not explain this minimal impact. Her interviews with participants also revealed various contextual and personal factors involved in shaping the washback. These included the absence of immediate importance of the test; students' major; their proficiency, which was higher than required for the test; their learning attitudes; their laziness, their unavailability prior to the test; their resistance to learning for the test, and the university's use of a make-up examination. Shih grouped these into three sets of influencing factors called extrinsic factors, intrinsic factors, and test factors, and proposed a tentative model of washback of students' learning which involved the interaction of these factors.

In line with Shih (2007), Allen (2016) employed interviews to explore the factors that contributed to the positive effects of the IELTS test on students' test preparation strategies and score gains. The factors were found strongly associated with students' sociocultural and educational contexts. They included students' perceptions of the test difficulty, their beliefs about efficiency and effectiveness, their knowledge of how to study and improve, assistance from others, their perceptions of the importance of the test, their interest, time available to study, or concurrent English classes.

Unlike Shih (2007) and Allen (2016), Zhan and Andrews (2014) conducted a longitudinal case study to investigate the washback of the revised CET-4 on Chinese non-English major undergraduates' out-of-class learning and identify factors mediating washback. They collected data from three cases in a university by diary and post-diary interview. The results revealed that the test influenced the students' learning content more than methods. The mechanism of washback was explained by possible self-theories. The washback of the CET-4 on the participants was mediated by several factors including their beliefs about the revised CET-4, their self - knowledge, their past learning and test-taking experience, others' experience in taking CET-4, and the learning environment (English curriculum, English teachers, commercial CET-4 preparation books and CET-4-related websites). All of these factors seemed to interact with one another and contribute to the participants' possible CET-4 selves (the "individualised image of what and how a possible CET-4 taker would like to/ought to act" (p. 84). The learner with an ideal self "experienced more of the washback intended by the CET-4 designers" than those with ought to selves. However, the researcher indicated that the relationship between the possible CET self and the types of CET-4 washback as found in their study was only tentative.

Some other studies attempt to explain the mechanism of washback using quantitative methods. One of these studies was reported in Xie and Andrews (2013) and Xie (2015). Xie (2015) used Structural Equation Modeling to investigate the washback mechanism. The study examined the relationship between students' perceptions of two design aspects of the CET test, component weighting and testing methods, and their test preparation activities. One thousand test-takers from one university in South China took part in the study. The participants answered two questionnaires before taking the revised College English Test 4 (R-CET4), one concerning their perceptions of the R-CET4 and one regarding the test preparation activities they were involved in. Their R-CET4 scores were also collected. The findings revealed that changes in component weighting had a small effect on test-takers' time management, i.e., test-takers spent more time on paper with higher weight than one with a lower weight. Favourable perception of test validity increased both desirable language learning activities and undesirable test preparation activities (drilling and cramming).

Xie and Andrews (2013) had a slightly different focus from Xie (2015). They used Expectancy-value motivation theory to explain the influence of test design and test uses on test preparation. Results showed that perceptions of test design and test uses affected test preparation simultaneously, but perceptions of test design appeared to have a stronger influence on test preparation. The two factors also followed different paths to test preparation. Endorsed test uses influenced test preparation mainly via Test-value, while the perception of test design affected test preparation via both Expectation and Test-value. Therefore, to

successfully use tests to drive education, learners should endorse and perceive the test design as valuable and doable.

Green's (2006, 2007a) model of washback may be the most comprehensive. It explains the direction, variability, and intensity of washback by showing the interaction of the factors involved. Regarding the direction of washback, the focal construct (the target skills of a curriculum or a target domain as understood by course providers and learners) and the test characteristics are the responsible factors. The more they overlap, the more positive washback is likely to occur and vice versa. The variability of washback is explained by the differences in the participant characteristics, their knowledge and acceptance of test demands, and their resources to meet test demands. The intensity of washback is shaped by the interaction between the test stake and its difficulty. For example, a test perceived as important and challenging will have intense washback while a test considered important but easy will have no washback. This model, however, ignores context as a factor responsible for washback variability.

The studies on the mechanism of washback revealed that washback is very complex. It results from the interplay of myriad factors, many of which depend on specific sociocultural and educational contexts. Therefore, more studies in new research contexts are needed to fully understand this phenomenon.

3. Conclusion

This paper reviews studies of washback on learning in much research with multiple research methods including surveys, interviews, case studies, and experiments. The review shows that the number of research is still small and the research findings are still inconclusive. Washback is so unpredictable and complex. The nature of the test alone does not determine what and how learners learn. Having got the positive answer to his question "Does washback exit?" after approximately ten years of empirical research on the topic, Alderson (2004) noted that the existence of washback "raises more questions than it answers" (p. xii). As washback is context dependent, it is necessary for future research to be conducted in new research contexts to fully understand how it operates and to generate positive effects of testing on learning if we want to use testing to drive education.

References

Alderson, J. C. (2004). Foreword. In L. Cheng, Y. J. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.

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

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*(1), 1-20. doi:10.1186/s40468-016-0030-z

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. J. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Erlbaum.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback - A case-study. *System, 30*(2), 207-223. doi:10.1016/S0346-251X(02)00005-2

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

Bailey, K. M., & Masuhara, H. (2013). Language testing washback: The role of materials. In B. Tomlinson (Ed.), *Applied linguistics and materials development* (1st ed.). London and New York: Bloomsbury Academic.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education*, *11*(1), 38-54. doi:10.1080/09500789708666717

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation*, *24*(3), 279-301. doi:10.1016/S0191-491X(98)00018-2

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching*, *48*(4), 436-470. doi:10.1017/S0261444815000233

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

Damankesh, M., & Babaii, E. (2015). The washback effect of Iranian high school final examinations on students' test-taking and test-preparation strategies. *Studies in Educational Evaluation*, *45*, 62-69. doi:10.1016/j.stueduc.2015.03.009

Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L. Cheng, Y. J. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.

Gan, Z. (2009). IELTS preparation course and student IELTS performance: A case study in Hong Kong. *RELC Journal*, *40*(1), 23-41. doi:10.1177/0033688208101449

Gan, Z., Humphreys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese universities. *The Modern Language Journal*, *88*(2), 229-244. doi:10.1111/j.0026-7902.2004.00227.x

Green, A. (2006). Watching for washback: Observing the influence of the international English language testing system academic writing test in the classroom. *Language Assessment Quarterly*, *3*(4), 333-368. doi:10.1080/15434300701333152

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

Green, A. (2007b). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education*, *14*(1), 75-97.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, *14*(3), 295-303. doi:10.1177/026553229701400306

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly*, *6*(2), 151-167.

Huang, S. (2011). Convergent vs. divergent assessment: Impact on college EFL students' motivation and self-regulated learning strategies. *Language Testing*, *28*(2), 251-271.

Hung, S. T. A. (2012). A washback study on e-portfolio assessment in an English as a foreign language teacher preparation program. *Computer Assisted Language Learning*, *25*(1), 21-36. doi:10.1080/09588221.2010.551756

Li, X. (1990). How powerful can a language test be? The met in China. *Journal of Multilingual and Multicultural Development*, *11*(5), 393-404. doi:10.1080/01434632.1990.9994425

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

Muñoz, A. P., & Álvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing*, *27*(1), 33-49.

Pan, Y. C., & Newfields, T. (2011). Teacher and student washback on test preparation evidenced from Taiwan's English certification exit requirements. *International Journal of Pedagogies and Learning*, *6*(3), 260-272.

Pan, Y. C., & Newfields, T. (2012). Tertiary EFL proficiency graduation requirements in Taiwan: A study of washback on learning. *Electronic Journal of Foreign Language Teaching*, *9*(1), 108-122.

Read, J., & Hayes, B. (2003). The Impact of IELTS on preparation for academic study in New Zealand. *IELTS International English Language Testing System Research Reports*, *4*, 153-206.

Ren, Y. (2011). A study of the washback effects of the College English Test (band 4) on teaching and learning English at tertiary level in China. *International Journal of Pedagogies & Learning*, *6*(3), 243-259.

Robb, T. N., & Ercanbrack, J. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese University students. *TESL-EJ*, *3*(4).

Safa, M. A., & Goodarzi, S. (2014). The washback effects of task-based assessment on the Iranian EFL learners' grammar development. *Procedia - Social and Behavioral sciences*, *98*, 90-99. doi:10.1016/j.sbspro.2014.03.393

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, *23*(1), 1-34. doi:10.1191/0265532206lt322oa

Shih, C. M. (2007). A new washback model of students' learning. *The Canadian Modern Language Review*, *64*(1), 135-161.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning. NFLC occasional papers.* Retrieved September 20, 2019, from National Foreign Language Center website: http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED362040

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

Stoneman, B. (2006). *The impact of an exit English test on Hong Kong undergraduates: A study investigating the effects of test status on students' test preparation behaviours* (Unpublished master's thesis). Hong Kong Polytechnic University, Hong Kong.

Taylor, L. (2005). Washback and impact. *ELT Journal*, *59*(2), 154-155. doi:10.1093/eltj/cci030

Tsagari, D. (2007). *Review of washback in language testing: How has been done? What more needs doing?* Retrieved September 21, 2019, from https://eric.ed.gov/?q=review+of+washback+in+language+testing%3a+how+has+been+done%3f+what+more+needs+doing%3f&ft=on&id=ED497709

Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.

Xiao, W. (2014). The intensity and direction of CET washback on Chinese College students' test-taking strategy use. *Theory & Practice in Language Studies*, *4*(6), 1171-1177. doi:10.4304/tpls.4.6.1171-1177

Xie, Q. (2015). Do component weighting and testing method affect time management and approaches to test preparation? A study on the washback mechanism. *System*, *50*, 56-68. doi:10.1016/j.system.2015.03.002

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with Structural equation modeling. *Language Testing*, *30*(1), 49-70. doi:10.1177/0265532212442634

Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: Insights from possible self theories. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 71-89. doi:10.1080/0969594X.2012.757546