Allen *Language Testing in Asia* (2016) 6:7
DOI 10.1186/s40468-016-0030-z

Language Testing in Asia

RESEARCH Open Access

Investigating washback to the learner from the IELTS test in the Japanese tertiary context

David Allen

Correspondence:
allen.david@ocha.ac.jp
Foreign Language Education Centre, Ochanomizu University, 2-1-1, Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Abstract

Background: This study investigated the consequential validity of the International English Language Testing System (IELTS) Academic exam, specifically focusing on washback upon learners' test preparation strategies and score gain, and the mediating factors influencing washback when learners in an EFL context are not enrolled in test preparation courses.

Methods: Two IELTS Tests were administered to 190 undergraduates at a Japanese university over a 1-year period. A survey instrument was used to collect data about test preparation strategies for both tests. Test scores were compared to assess score gain. Interviews were conducted with 19 participants to investigate the factors mediating washback.

Results: Test results revealed a significant increase in speaking ability, with more significant increases in speaking and listening for participants who reported preparing more intensely for the test. Survey results showed that students focused significantly more on speaking and writing, and significantly less on reading, when preparing for the second test, and those who prepared most intensely also focused significantly more on listening. A qualitative analysis of the interview data revealed a complex array of factors related to learner perceptions and their access to resources, which are highly dependent on learners' sociocultural and educational context, and which shape washback to the learner.

Conclusions: The IELTS Test created positive washback on learners' language ability and test preparation strategies, specifically regarding productive skills, which learners in the study context had neglected in their previous language study. However, a range of mediating factors must be addressed in order to ensure positive washback in EFL contexts and in the absence of instruction.

Keywords: Consequential validity, Validation, Washback, IELTS, Japan

Background

As O'Sullivan and Weir (2011, p.13) note, 'there is a growing awareness in the stakeholder community of the need for a sound theoretical model that underlies a test (validity) and the generation of evidence concerning the operationalisation and interpretation of the model in practice (validation)'. The socio-cognitive model of test validation (Weir 2005) is one such model and provides a theoretical framework that allows test designers and researchers to evaluate and modify tests through a cyclical process

of development. The model has served as the basis for validating the Cambridge examinations (Weir et al. 2013) and other tests, such as the Test of English for Academic Purposes (TEAP) designed by the EIKEN Foundation of Japan (e.g., Taylor 2014).

The model includes context, cognitive (theory-based), scoring, criterion-related, and consequential validity components. To offer an acceptable validity argument to stakeholders, evidence must be gathered in support of each of these components (Weir 2005). Consequential validity, the focus of this study, requires evidence in support of the following question: 'What effects does the test have on its various stakeholders?' (O'Sullivan & Weir 2011). When focusing on the effects upon learning and teaching, evidence of washback must be sought and evaluated to make a claim about the consequential validity of the test.

According to O'Sullivan and Weir (2011, pp.21–22) the type of consequential validity is essentially derived from successful realization of construct validity. Thus, whether a test creates positive or negative washback is determined initially by the overlap of test items with the Target Language Use (TLU) domain (Green 2007; Messick 1996). However, washback is manifested through the interaction between the test and the various participants, processes and products (Bailey 1996; Hughes 2003), and thus researchers must provide evidence of washback in contexts in which the tests are actually being used.

Researchers have noted that although washback concerns teaching and learning equally (Alderson & Wall 1993), research has tended to focus more on the former (Cheng 2014; Watanabe 2004). In response to this, there has been a recent growth in the number of studies focusing on washback upon learning (e.g., Cheng et al. 2011; Gosa 2004; Green 2006a, 2007; Mickan & Motteram 2009; Pan 2014; Shih 2007; Smyth & Banks 2012; Stoneman 2006; Tsagari 2007; Tsai & Tsou 2009; Xie 2013; Xie & Andrews 2012; Zhan & Andrews 2014; Zhan & Wan 2016). In almost all of these studies, however, research on learning has taken place in contexts where learners were, or had been, enrolled in test preparation courses. Therefore, it is difficult to separate out the influence of teaching upon learning, particularly because the effect of teaching has been noted to be one of the most important factors influencing students' test preparation practices (Gosa 2004; Zhan & Wan 2016). Another factor that has hampered washback studies into learning is that participants were often already familiar with the test prior to the study, many having taken the exam prior to the study (e.g., Green 2007). Both instructed contexts and prior familiarity should ideally be controlled in order to see how tests affect learners directly, i.e. in non-instructed contexts and when the test is really 'new' for the test takers. The present study sought to address the aforementioned issues by focusing on the washback to learners from a newly introduced test, the International English Language Testing System (IELTS) Academic exam, when learners were not enrolled in preparation courses.

Literature review

The literature review provides an overview of relevant IELTS washback research, followed by a review of other key findings from the literature regarding the factors that mediate the washback process.

IELTS washback studies

The IELTS Test (www.IELTS.org) is an established and widely used international English language proficiency exam that comes in two formats, each for a different

purpose: Academic and General Training. The test has four sections, one for each of the language skills, which are equally weighted to give an overall averaged band of proficiency measured from 0 (lowest) to 9 (highest).

Research investigating the consequential validity of the IELTS Test has considered washback on teaching practices (Green 2006b, 2007; Mickan & Motteram 2008), teaching materials (Saville & Hawkey 2004), learners' approaches to test preparation (Brown 1998; Elder & O'Loughlin 2003; Green 2007; Read & Hayes 2003; Mickan & Motteram 2009), learner perspectives on IELTS preparation course expectations and outcomes (Green 2006a) and score gain (Elder & O'Loughlin 2003; Green 2007; Humphreys et al. 2012; O'Loughlin & Arkoudis 2009).

Washback on learning, that is, whether preparing for a test leads to measurable increases in language ability, has been investigated using test takers' score gain on the IELTS Test. Studies have shown that there is considerable individual variation in score gain (Elder & O'Loughlin 2003; Green 2005; Humphreys et al. 2012; Read & Hayes 2003) and a greater likelihood of score increase for those at lower levels of initial proficiency (Elder & O'Loughlin 2003; Green 2005; Humphreys et al. 2012). To observe score gain on tests such as IELTS, it has been suggested that a considerable amount of time and intensive preparation is typically required, especially at higher levels of initial proficiency (Green 2005; Read & Hayes 2003).

Washback on learners' test preparation strategies has been investigated within instructed ESL contexts (e.g., Brown 1998; Elder & O'Loughlin 2003; Green 2007; Read & Hayes 2003). In IELTS preparation courses, students tend to focus on test-related tasks and materials when preparing for the test (Green 2007; Mickan & Motteram 2008). Mickan and Motteram (2008) found that 'the dominant activities were test practice, skills-focused activities, and explanations of the format and content of the IELTS modules and test-taking procedures' (p.8). Thus, in instructed contexts, learners are 'apprenticed into the semiotic activities connected with the IELTS examination' (p.23). A similar narrowing of the curriculum has been observed in preparation courses for TOEFL (Wall & Horak 2011) as well as other regional English language exams (Gosa 2004; Shih 2007; Stoneman 2006; Xie 2013; Zhan and Andrews 2014; Zhan & Wan 2016).

Only one IELTS study has considered washback on learning in a non-instructed context. Mickan and Motteram (2009) surveyed 78 test takers' preparation strategies for the IELTS General Training Exam in Australia. Participants mainly took the exam for immigration or university entrance purposes and many had taken it repeatedly, indicating not only that it was important for their future, but also that they were familiar with the exam. In addition, most had not enrolled in test preparation programs and thus reported preparing by themselves.

Similar to instructed contexts, test takers relied mainly on published IELTS practice materials and practice tests, with a minority reporting other general activities, such as watching TV and reading newspapers. A key finding was that while most test takers studied alone, more than half reported gaining advice from friends or teachers about test taking. Furthermore, Mickan and Motteram's (2009) case study data revealed test takers' apparent dependency on others: Test takers viewed success as dependent on expert help and reflection and self-analysis as dependent on feedback (2009, p. 20). They reported not knowing how to improve their scores and some did not prepare at all because they felt there was little that they could do without a mentor. In other words,

participants lacked personal agency and strategic action in preparation for the exam. Seeking assistance has been referred to as a socio-affective test preparation strategy (Xie & Andrews 2012), and may be especially important when learners are preparing for tests without input from the classroom. Based on the findings of one study, however, it is unclear whether these strategies and beliefs are generalizable to other non-instructed test preparation contexts.

Importantly, the participants in Mickan and Motteram's (2009) study were already familiar with IELTS, many having taken it a number of times, and some of the case studies reporting that they had studied for the test prior to arriving in Australia. Therefore, the washback effect being investigated was delayed relative to their initial experience of the test. Also, the study, like other IELTS washback studies (e.g., Brown 1998; Elder & O'Loughlin 2003; Green 2007; Mickan & Motteram 2008, 2009; Read & Hayes 2003), was conducted in an ESL environment. However, the IELTS Test is also prepared for and taken in many EFL contexts. Thus, to better understand how washback to the learner is generated in these EFL contexts, and to provide more comprehensive evidence for the consequential validity of the test, research into test preparation in such contexts is necessary.

Mediating factors in washback

Washback studies have revealed that beliefs, educational experience and contextual circumstances mediate washback to learners and learning (Gosa 2004; Xie & Andrews 2012; Zhan & Andrews 2014; Zhan & Wan 2016). Participants' characteristics and values, including their knowledge and understanding of the test, resources to meet the test demands and their acceptance of these demands, as well as their perceptions of test importance and test difficulty, may all mediate washback (Green 2007).

The context in which tests are used is also crucial for understanding washback variability. In one study, Shih (2007) identified a number of learner factors that were inherently related to the educational context and that appeared to limit washback from the General English Proficiency Test in Taiwanese technical colleges. Notably, learners had little opportunity to practice speaking, which appeared to be peripheral to the Taiwanese students' language learning experience. Consequently, students' appeared to lack the ability (the resources) to prepare for the speaking component, as indicated by a greater, and somewhat haphazard, variety of test preparation strategies. The importance of context and how it can influence learners' choices of test preparation strategies has also been noted in other studies (Andrews et al. 2002; Gosa 2004).

Another factor that that may determine test takers' preparation strategies is their previous experience of tests. Test takers may hold negative attitudes towards tests (Cheng 1998) and experience varying levels of anxiety (Shohamy 1993; Smyth & Banks 2012; Tsagari 2007). However, tests often function as important motivators for learning, and many learners respond positively, even if their scores are lower than necessary or expected. In Tsagari's (2007) study, for instance, one student's experience 'made him aware of his level in relation to the requirements of the exam and strengthened his determination to increase his efforts in the future' (p.265). In situations where learners take a test multiple times, as is common for high-stakes tests such as IELTS and TOEFL, it could be assumed that learners modify their strategies on the basis of their experience of and initial performance on the test (i.e., test difficulty). Such modifications of behaviour would

provide evidence of direct washback upon test preparation strategies, thereby providing evidence for the consequential validity of the test.

The present study

Recently, as a result of a ministry-led drive for internationalization of higher education in Japan (MEXT 2016), many universities have been utilizing international four skills tests to promote English language learning, to prepare students to study abroad and to evaluate their language proficiency development. As part of this initiative, undergraduates at one Japanese university were invited to take the IELTS Academic Exam twice over the period of 1 year, free of charge, during their normal course of study at the university.

This project provided a unique opportunity to investigate washback from the IELTS Test from several points. Firstly, most test takers were unfamiliar with the IELTS Test prior to the study. The initial test thus served as the baseline to evaluate change in behaviour and scores following the introduction of a test. Secondly, washback could be investigated for students who were not participating in test preparation courses. Thus, washback directly from the test, without the influence of teachers and teaching, could be investigated. Finally, few IELTS-related studies have been conducted in EFL contexts, which differ considerably from ESL contexts. In these ways, the present study sought to assess the consequential validity of the IELTS Test. To this end, the following research questions were proposed:

1. Do learners in a non-instructed context adopt different test preparation strategies for two consecutive IELTS tests? And if so, do the changes evidence positive washback from the test?
2. Do learners' scores increase in any of the four skills? And if so, is this related to any observed changes in test preparation strategies?
3. What factors influence the choices of test preparation strategies for the two tests, thereby mediating washback to the learner?

Method

Participants

Three hundred first-year undergraduate students at a Japanese university in the Tokyo metropolitan area were recruited for the study. Participants were selected on a first come, first served basis, and they agreed to take two fully funded IELTS Tests and complete the survey that followed the second test. Of these, 204 participants completed both IELTS Tests (a completion rate of 68 %). One hundred and ninety of these participants also completed the survey (a completion rate of 93 % for the survey). At the time of the first test, the test takers (127 male, 62 female, 1 no response; mean age = 20 years) had just completed their first, or in a smaller number of cases, were just completing their second semester. They were in their second year of academic study at the time of the second test, survey and interview. Nineteen students (12 male, seven female) were recruited randomly for interviews via the survey and were paid 1500-yen for participation. All participants gave their consent to participate in the study.

The participants were all high academic achievers, having succeeded in gaining entry to a prestigious and extremely competitive national university. Most of the learners attended cram schools during their preparation for the entrance exams (Allen 2016) and consequently students entering this university certainly understand 'the rules of the game' (Bourdieu 1990); in other words, they had a good understanding of what is required of them in exam-oriented educational systems, an attribute often associated with aspiring middle-class students (e.g., Smyth & Banks 2012).

The students were enrolled in the liberal arts program, in which one or two English language courses are compulsory and a number of elective English courses are offered each semester. Given that English is only one of many required subjects, it was necessary for students to distribute their study efforts. Consequently, while motivation to study English appeared to be reasonably high, learners had to juggle their study efforts according to the demands of the wider curriculum.

Materials and procedure

The two tests were taken at one of four officially designated test centers in the Tokyo area at times convenient to the applicants. The gap between the two tests was on average 11 months, with the shortest gap being 7 months and the longest being 13 months.

Participants prepared for the tests independently; however, there were two half-day workshops given about 6 months apart, which focused on the productive skills components as participants were expected to be less familiar with these skills. Free access to online IELTS Test preparation materials was also provided for 30 weeks (http://www.britishcouncil.jp/exam/ielts/resources/free-practice).

The survey instrument was developed through a lengthy process of expert reviews and trials with focus groups. Students took the final online survey in Japanese immediately after taking the second IELTS Test. The survey contained nine sections with 122 questions, of which only those relevant to IELTS Test preparation are considered here (23 questions; Table 6 in Appendix 1). Seven categorical questions targeted use of preparation resources (website and workshops), IELTS-related tuition, previous experience of IELTS, and reasons for taking the tests. Sixteen Likert scale questions were on a 6-point scale of agreement and targeted three main aspects:

- Preparation for the four skills
- Types of activities prepared for (based on the tasks in IELTS Exam)
- Focus on form, fluency and test taking techniques

These 16 items were repeated for the first and the second tests. Cronbach's α reliability for the 32 items was 0.95.

As the researcher was teaching at the institution in which the study was carried out, and might have held a position of authority over some of the interviewees as their teacher, semi-structured interviews were conducted in Japanese by trained postgraduate research assistants (see Appendix 2), which enabled students to speak freely about their preparation strategies. Interview questions were similar to

those in the survey, but interviewees were encouraged to discuss the reasons for their strategies. Interviewers referred to the participants' survey responses as necessary.

Analyses

For the quantitative analyses, statistical comparisons were conducted. First, Likert scale responses to the survey questions were compared for the two tests. Data were checked to see whether the residuals (errors) were normally distributed by observing plots and checking skewness and kurtosis. Residuals for Likert scale data were not normally distributed and therefore Wilcoxon Signed-rank tests were performed with r as the effect size (.1 = small, .3 = medium, .5 = large). A Bonferroni correction for multiple tests was applied to allow for a more conservative estimation statistical significance (p at .05/16 = .0031).

Second, participants' test scores were compared for the two tests. Residuals for test scores were normally distributed and so t-tests were performed with Cohen's D as the effect size (.2 = small, .5 = medium, .8 = large). A Bonferroni correction for multiple tests was applied to allow for a more conservative estimation statistical significance (p at .05/4 = .0125).

For the qualitative research questions, interview data regarding the approaches to preparation and the reasons for these approaches were analyzed. The methodology for analyzing the interview data followed recommendations in Kvale and Brinkmann (2009) and Mason (2002). The interview transcripts were read for each participant individually and also across participants when identifying recurrent patterns and themes. Information regarding approaches to study as well as the reasons given for these approaches was highlighted and coded according to categories that evolved during the process of analysis. The approach was both inductive and deductive; in other words, themes emerged from the data without *a priori* hypotheses but the researcher also had some expectations of what the data may reveal based on knowledge of the test, the context, and previous research. All English translations provided in this report were checked for accuracy by a native speaker of Japanese who was proficient in English.

Results and discussion

Survey data

Categorical responses

The categorical response data revealed that most students prepared for less than 20 h for the first and second tests while smaller proportions studied for longer durations (Table 1). Those who reported studying for 0 h for both tests (n = 27, 14 % of participants) were removed on the basis that washback on preparation strategies cannot exist when participants do not prepare. This left 163 participants (31 % females, 69 % males, mean age 20.2 years) in the following analyses.

Table 1 Test takers' preparation for the two tests

Preparation (h)	0	<20	>20<40	>40 <60	>60 <80	>80 <100	>100
Test 1	44	123	18	4	1	0	0
Test 2	47	103	28	7	5	0	0

Categorical responses showed that only 8.5 % had previously taken the IELTS Test. In preparation for the tests, 32 % of participants attended one or more of the half-day workshops, 42 % used the online resources, 4 % attended a conversation school during the period, and 3 % prepared with the assistance of a personal acquaintance. The reasons for taking the test were because it was free (93 % agreed), for study abroad (55 %), for the qualification (53 %), to find out about IELTS (43 %), and finally, for work (10 %). In sum, most participants were unfamiliar with the test, prepared alone and were either motivated by study abroad and/or qualification prospects, but particularly because the test was being provided without charge.

Preparation for the two tests

Mean scores for all items showed that learners tended to 'disagree' at least somewhat with the statements (responses between 1 and 3), indicating limited preparation overall. In other words, washback intensity was relatively weak. Nevertheless, significant differences were observed in the Likert scale response data. For the second test, learners studied more speaking and writing (Table 2); did more speaking activities involving both everyday and abstract topics; and practiced more spontaneous speaking in response to prompts/questions. In contrast, they did fewer reading activities for the second test. All effect sizes were small-to-medium. These changes in learning processes show positive washback, because learning was re-focused towards aspects of language use (speaking/writing) that are important for the target language use domain (i.e., using English for academic purposes) but which had hitherto been lacking in the learners' language educational experience.

The above analysis was repeated for test takers who reported studying more than 20 h in preparation for the second test ($n = 40$; Table 3). The higher means compared to those

Table 2 Two-way non-parametric comparisons (test takers who studied >0 h, $n = 163$)

Category	Sub-category	Mean T1	Mean T2	*Z*-value	*p*-value	Sig	Effect size (*r*)
Skills	Speaking	2.1 (1.21)	2.7 (1.5)	−4.54	.0000	**	.25
	Writing	2.3 (1.3)	2.8 (1.7)	−2.97	.0028	*	.16
	Listening	2.7 (1.5)	2.5 (1.4)	1.78	.0757		.10
	Reading	2.7 (1.5)	2.5 (1.4)	2.52	.0114		.14
Activities	Reading	2.6 (1.5)	2.3 (1.4)	3.48	.0004	*	.19
	Listening: 2 people	2.3 (1.2)	2.2 (1.3)	0.29	.7722		.02
	Listening: 3 (+) people	2.2 (1.2)	2.2 (1.3)	0.06	.9515		.00
	Speaking: Everyday topics	1.9 (1.2)	2.4 (1.3)	−3.15	.0015	*	.17
	Speaking: Abstract topics	1.8 (1.1)	2.2 (1.4)	−3.44	.0005	*	.19
	Writing: Graphs	2.3 (1.4)	2.6 (1.7)	−1.70	.0893		.09
	Writing: Essays	2.4 (1.4)	2.6 (1.7)	−1.05	.2959		.06
Form	Fluency	2.2 (1.4)	2.6 (1.6)	−3.15	.0014	*	.17
	Pronunciation	1.6 (0.9)	1.7 (1.1)	−1.52	.1257		.08
	Grammar	2.1 (1.2)	1.9 (1.0)	2.39	.0165		.13
	Vocabulary	2.6 (1.3)	2.6 (1.4)	0.78	.4392		.04
	Test techniques	2.5 (1.4)	2.6 (1.4)	−1.08	.2807		.06

Note: Standard deviation in parentheses; Asterisks denote Bonferroni corrected *p*-values: * <0.0031 ** <0.0001

Table 3 Two-way non-parametric comparisons (test takers who studied >20 h, $n = 40$)

Category	Sub-category	Mean T1	Mean T2	*z*-value	*p*-value	Sig	Effect size (*r*)
Skills	Speaking	2.4 (1.3)	3.6 (1.2)	−4.15	.0000	**	.46
	Writing	2.7 (1.2)	4.0 (1.6)	−3.76	.0000	**	.42
	Listening	3.2 (1.5)	3.6 (1.3)	−4.15	.0000	**	.46
	Reading	3.3 (1.5)	3.6 (1.5)	−1.07	.1198		.12
Activities	Reading	3.1 (1.5)	3.2 (1.5)	−0.19	.8696		.02
	Listening: 2 people	2.7 (1.2)	3.1 (1.4)	−1.84	.0687		.21
	Listening: 3 (+) people	2.7 (1.4)	3.0 (1.4)	−1.40	.1734		.16
	Speaking: Everyday topics	1.8 (0.9)	3.1 (1.2)	−4.57	.0000	**	.51
	Speaking: Abstract topics	1.8 (0.9)	3.2 (1.3)	−5.02	.0000	**	.56
	Writing: Graphs	2.8 (1.5)	3.8 (1.7)	−3.08	.0017	**	.34
	Writing: Essays	2.8 (1.4)	4.0 (1.5)	−3.75	.0000	**	.42
Form	Fluency	2.6 (1.5)	3.3 (1.5)	−2.71	.0059		.30
	Pronunciation	1.6 (0.7)	2.1 (1.1)	−2.03	.0434		.23
	Grammar	2.4 (1.1)	2.6 (1.1)	−1.18	.2360		.13
	Vocabulary	3.1 (1.4)	3.5 (1.4)	−1.52	.1310		.17
	Test techniques	2.9 (1.5)	3.7 (1.4)	−3.36	.0006	*	.38

Note: Asterisks denote Bonferroni corrected *p*-values: * <0.0031, ** <0.001

in Table 2 show that these learners reported preparing more for all aspects of the test. For the second test, learners reported studying more speaking, writing and listening skills (medium-to-large effects); more speaking about both everyday and abstract topics (large effects); and more spontaneous speaking, though this difference was only close to significant ($p = .0059$). They also focused more on writing about both graphs and writing essays and studied more test techniques (medium-to-large effects). In sum, for test takers who gave a greater priority to the test (based on the number of hours studied), the changes were quite similar to the whole group (i.e., more productive skills on the second test) but the extent of the changes was greater, indicating more intense washback.

To test whether there were any differences in test preparation for high and low proficiency learners, two groups were formed using the overall IELTS score for the first test (high group, bands 6.0 to 8.0, $n = 74$; low group, bands 4.0 to 5.5, $n = 89$). Wilcoxon tests were conducted comparing mean ratings for all 32 items for the two groups using a conservative alpha (p at $.05/32 = .0015$). Of all comparisons, the only significant difference was in the amount of fluency practice for the first test (high group mean = 2.7, low group mean = 1.9; W = 4246, $p = .00083$). Therefore, in contrast to previous studies that have observed differences according to learners' levels (e.g., Cheng et al. 2011; Shohamy et al. 1996), washback on test preparation strategies did not vary greatly for high and low proficiency learners. The discrepancy between the findings here and those of previous studies may be due to differences in the way that learners were grouped into high and low categories (i.e., test scores or self-ratings), differences in the way that test preparation strategies were defined, and differences in the homogeneity of learners (all high academic achievers or various levels of academic achievement).

In sum, the survey data showed that test takers did change their preparation strategies for the second test, changing their focus from receptive to productive skills. Those who prepared the most focused significantly more on all skills except reading in

preparation for the second test. These findings indicate that the IELTS Test generated a positive washback effect on the study of productive skills.

Test data

Paired *t*-tests were performed on the IELTS Test scores for the 163 test takers in the study (Table 4). A significant increase in speaking scores was observed with a medium-sized effect. This finding is similar to Humphreys et al. (2012) who also found a significant gain in speaking on the IELTS Test for international students over the course of one semester at an Australian university. Interestingly, although the survey data suggested a greater focus on both productive skills, writing scores did not increase. In two other IELTS studies, score gain for IELTS writing was the least (Craven 2012), or second to least (Humphreys et al. 2012) of all components. This may be due to the relative difficulty level of the IELTS Writing Test, which consistently has the lowest mean score of all components (www.IELTS.org).

The scores were also compared for those who studied the most ($n = 40$). Table 5 shows that for this subset of test takers, both their speaking and listening scores significantly increased (by 0.3 bands) with medium-sized effects.

To investigate whether the increase in speaking scores was more likely to occur for participants at a lower initial speaking proficiency (i.e., speaking score on Test 1), mean scores for those whose initial score was between 3.0 and 5.5 ($n = 103$), and those whose score was between 6.0 and 8.5 ($n = 60$) were compared using an independent *t*-test. There was a significant difference between the two groups ($t = -3.39$, df = 161, $p < .001$), such that those at lower initial proficiency made greater gains (0.40) compared to those at higher initial proficiency (0.02). This finding suggests that score gain over a shorter period is most likely to be observed for test takers at lower bands, and is in line with the findings of other studies investigating score gain on the IELTS exam (Elder & O'Loughlin 2003; Green 2005; Humphreys et al. 2012).

In sum, the test data indicate positive washback on speaking skills for all participants and also on listening skills for those who prepared the most for the test. Taken together, the survey and test data suggest that studying for the IELTS Test resulted in a greater focus on productive skills, with evidence of an increase in proficiency in speaking, particularly at lower initial levels of proficiency.

Interview data

Test preparation strategies

The most frequently referred to materials were official IELTS materials, including textbooks, reference books, workbooks, past exam collections and the online materials

Table 4 Two-way parametric comparisons (test takers who studied >0 h, $n = 163$)

Category	Sub-category	Mean T1 (SD)	Mean T2 (SD)	*t*-value (df = 162)	*p*-value	Sig.	Cohen's *D*
Scores	Speaking	5.4 (1.0)	5.6 (1.0)	−4.579	.0000	**	.36
(0-9)	Writing	5.5 (0.6)	5.6 (0.6)	−1.576	.1171		.12
	Listening	6.5 (1.0)	6.6 (1.1)	−2.299	.0228		.18
	Reading	7.2 (0.9)	7.3 (0.9)	−1.880	.0619		.15

Note: Asterisks denote Bonferroni corrected *p*-values: * <0.0125 ** <0.001

Table 5 Two-way parametric comparisons (test takers who studied >20 h, $n = 40$)

Category	Sub-category	Mean T1 (SD)	Mean T2 (SD)	*t*-value (df = 39)	*p*-value	Sig.	Cohen's D
Scores	Speaking	5.4 (1.0)	5.7 (1.0)	−2.893	.0062	*	.46
	Writing	5.5 (0.5)	5.6 (0.5)	−2.481	.0175		.39
	Listening	6.4 (0.9)	6.7 (0.9)	−3.313	.0020	*	.52
	Reading	7.2 (0.9)	7.4 (0.9)	−1.617	.1140		.26

Note: Asterisks denote Bonferroni corrected *p*-values: * <0.0125 ** <0.001

provided by the British Council. The interviewees adopted a test-focused approach, that is, they focused primarily on tasks found in the exam and aimed to improve their ability to successfully complete these tasks within specified time limits. This was particularly the case for the second test.

For reading and listening, the interviewees typically described how they worked through IELTS test tasks. Four participants additionally used materials that were not directly related to the IELTS test, including a variety of materials for reading (university reading textbooks, and English stories) and listening (CNN news magazine, TED talks, entrance exam materials, and TOEFL materials).

For writing, ten interviewees mentioned explicitly using test-related materials and writing responses to test-like questions. They read through textbooks, practiced writing and then checked their responses and the model answers. Two interviewees also practiced writing using other materials from their high school/entrance exam preparation, and two wrote a diary. Conversely, two participants reported only reading about the writing component without actually writing themselves.

For speaking, eight interviewees utilized test preparation materials and practiced speaking using the test format. Interestingly, only two stated that they practiced speaking aloud. Four interviewees practiced by speaking 'in their heads' only and two read or watched a video on the IELTS preparation website about the speaking exam but did not actually practice formulating answers themselves.

All in all, students tended to focus on test-related tasks and materials when preparing for the test though there was evidence of other language learning activities that were not directly related to the test as well.

Changes in approaches to test preparation

Sixteen participants said that their test preparation strategies changed after taking the first test. Among those, five reported significant changes (P2, P4, P6, P8, P15). The most obvious and widely observed change was in the focus from receptive skills to productive skills. This was accompanied by an increase in the amount of time spent on productive skills. However, not all interviewees reported changing their preparation in this way, highlighting the complex array of factors influencing participants' choices. These factors, in other words the reasons for changing or not changing their approaches, are outlined below.

Perceived difficulty

Twelve interviewees described explicitly how their experience of taking the first test made them aware of their language proficiency deficiencies and led them to reevaluate

their study approaches. In other words, taking the test had a washback effect on their preparation for the subsequent test.

The most difficult parts of the test for these test takers were the productive skills sections as shown in the comments below:

> The speaking exam was the hardest. For instance, even if I knew what I wanted to say in response to the question, I often couldn't find the words to reply well and I realized that I couldn't convey myself very well (P1)

> Basically, I hadn't done speaking before like that on the exam so I wasn't used to it […] I hadn't gotten used to speaking with a partner, so it was difficult to do it without getting nervous (P2)

> I thought the writing part was difficult. Compared to the writing tasks I've done before, in English classes and for the entrance exams, I felt it was much harder […] my writing score was not very good so I concentrated on it (P17)

Four participants (P1, P2, P9, P18) reflected that they could not formulate their responses within the allotted time during the speaking and writing tests (see P1 and P2 above). Most importantly, the participants felt the difficult parts of the test were those they were most unfamiliar with, specifically the following sections: Oral interviews, writing about visually presented data and academic writing in general. In other words, the difficulty of the test was partly due to the novelty of the test tasks in relation to previous learning and test-taking experience.

The first test was also instrumental in highlighting perceived strengths in language ability; in other words, some interviewees did not perceive certain aspects of the test to be difficult. Seven interviewees reported being satisfied with their reading and/or listening scores on the first test, or that they were confident in these skills, and therefore they did not spend time studying these skills for the second test. They felt that their high school English classes and their preparation for university entrance exams, which focused primarily on reading and listening, had enabled them to sufficiently develop these skills.

> I was forced to read and listen a lot while preparing for the entrance exams, so I thought I don't need to work on those skills (P15)

In sum, the experience of the initial test and its results was the primary driving force behind learners' strategies employed for the second test. These findings align closely with previous observations about how tests can raise learners' awareness of their strengths and weaknesses and motivate learning (e.g., Tsagari 2007, Zhan & Andrews 2014). Washback can thus be observed on behaviour and attitudes both before and after taking the test (Bachman & Palmer 1996; Bailey 1999). Moreover, the findings illustrate how washback effects are mediated by learners' experience of the test in combination with their prior learning experiences, which are in turn partly determined by the social and cultural context in which they live. In Japan, learners almost inevitably focus more attention on developing receptive skills as they are the focus of the high-stakes university entrance exams (Allen 2016).

Taking the IELTS Test highlighted these strengths as well as the participants' relative weakness in productive skills and thus motivated them to change their focus of test preparation.

Perceived efficiency and effectiveness

Related to perceptions of difficulty were beliefs about efficiency. Three test takers (P8, P9, P19) focused on receptive skills for the first test because they believed it was efficient to do so within the limited time available. These interviewees focused on their strengths of reading and/or listening, a strategy that was intended to maximize their overall score.

> Until now my entrance exam preparations had focused on reading and listening so I thought I could get a good score on those and thought it'd be efficient to focus on them (P9)

Two other participants (P3, P5) focused on developing their receptive skills for the second test, explicitly stating that they had taken advice from other students who had said that concentrating on reading skills would allow them to get a high score overall. Thus, their strategies were aimed at maximizing their score, rather than improving their overall English skills.

> I was told that Japanese should first do well in reading and listening, then if you score well in those two, move on to writing and speaking, you know, extending the practice in the first two skills. Also, they said that if you get a good score on reading and listening, it will be easy (to get a high score), so I thought I needed to really get solid scores on those skills (P5)

The above examples illustrate the importance of peers as advice-givers, and it highlights how beliefs are shared within the test takers' social environment. Such 'folk-knowledge' about the test (Bailey 1999), along with official test information, can influence test takers' preparation strategies and as a result mediate the potential for washback. In the cases above, this folk-knowledge can be seen to disrupt the flow of washback: Instead of focusing on their weaker, productive skills, which have equal weighting in the test, they instead focused on receptive skills, which they were already much stronger at.

Knowledge of how to study and improve

Related to the above issues of difficulty and efficiency was knowledge about how to study and how to improve one's skills (and scores). Four interviewees focused on reading for the first test because they could develop this skill by themselves or because reading was 'easiest' for them to prepare for. In contrast, two participants said that they did not know how to practice speaking (P4, P14) and another (P11) said that he did not know how to improve speaking. As a result, these test takers did not prepare for the speaking component. Another (P2) did not know how the writing task would be assessed, so he practiced only writing fluency (i.e., writing sufficient amount within the time limit).

> I didn't know what kind of writing would score high so I at least I thought I need to write the required number of words within the time limit, and I checked my grammar by myself by reviewing my work (P2)

Assistance from others

Although three learners sought and gained assistance from others (an instructor, P10; peers, P16; and a parent, P18) who acted as interlocutors for speaking practice, most participants said that they had no one to assist them in test preparation.

> For speaking and writing if there isn't anyone to check then there's nothing you can do […] There wasn't anyone around who seemed to be able to correct my writing (P3)

> It was difficult to prepare for the speaking section by myself, so I didn't […] I've had little experience or opportunity to speak, but I did think I really needed to practice (P17)

The belief that assistance was necessary directed the test takers' preparation strategies: Eleven interviewees stated that they did not study, only studied a little, or only focused on certain aspects of speaking and/or writing because their was no opportunity to practice or no-one to provide feedback on their work. This implicates people (i.e., interlocutors, peers, teachers and other speakers of English) as an essential resource that mediates the washback effect of exams that include productive skills. It resonates with the comments of IELTS test takers in the non-instructed ESL context reported by Mickan and Motteram (2009), where test takers were dependent on the assistance and feedback of others, without which they felt incapable of preparing for the test. The finding also confirms the concern of learners in EFL contexts that there is little opportunity to practice speaking, which leads them to ignore practice of this skill (Shih 2007).

It is important to note that the lack of practice of productive skills reported here contrasts with the results of the survey, which indicated that speaking and writing both received greater attention when test takers prepared for the second test. The contradiction is explained by the fact that most interviewees did not report exclusively not studying the productive skills but felt it was difficult to do so and thus limited or modified how they practiced them, as exemplified by P10, below:

> Finishing writing within the time limit was my priority because there was no one to check my writing (P10)

Other factors

A number of other factors were also influential in directing test preparation behavior. Participants who reported that the test was not important for them tended to study little for it (P2, P5, P12, P15), demonstrating how washback is impeded through lack of perceived importance (Cheng 1997, 2005). Test takers' interests also played a role in determining their choices of preparation strategies, especially towards learning activities that were not directly related to the test. For example, two participants (P11, P14) said that they listened to English not for the test but because they liked listening to particular programs (e.g., TED Talks). Similar

findings were reported in Zhan and Andrews (2014) who noted that content of test preparation is partially directed by learners' interests. Also, factors related to the learning environment, such as the time available and concurrent English classes, played a role. Eight test takers reported having very little time to study, thus impeding potential washback from the test. Finally, participants were concurrently taking classes as part of their undergraduate study within the liberal arts program, which apparently influenced their test preparation strategies as well. For instance, three participants had recently taken courses on reading and writing at university (P11, P16, P6) and two more (P1, P8) were taking such courses at the time of the exam so they did not practice those skills for the test. Therefore, the potential for washback was mediated by participants' wider learning environment.

Conclusions

The present study found that the IELTS Test generated positive washback on productive skills in the Japanese tertiary context. Moreover, this appears to have led to an increase in test takers' language proficiency, particularly speaking proficiency. In addition, a range of mediating factors were identified that shaped washback to the learner in this context.

All in all, students tended to focus on test-related tasks and materials when preparing for the test as observed in instructed contexts (Green 2007; Mickan & Motteram 2008; Shih 2007; Stoneman 2006; Zhan & Andrews 2014). In non-instructed contexts, learners are apprenticed through the use of textbooks and practice tests, though there is considerable variance in the preparation strategies adopted. The facilitatory role of learning resources is clear, which underscores the value of studies that evaluate the impact of published material on stakeholders (e.g., Saville & Hawkey 2004).

When confronted with a test that is radically different from previous experience (i.e., a four skills, equally weighted test), many of the learners in this study reported changing their preparation strategies to accommodate the novel features of the test. However, past learning and test taking experience led some learners to maintain strategies for the new test (e.g., focusing on receptive skills) and thus these strategies could not be linked to washback from the new test. The fact that learners had little experience of studying productive skills for previous exams was reflected in the strategies that they adopted for studying those skills. There was evidence of uncertainty, futility, and dejection, leading some participants to avoid studying these skills or adopt dubious approaches (i.e., not actually practicing speaking aloud). Previous researchers have suggested that learners are more willing to change what they learn for tests rather than how they learn (Zhan & Andrews 2014), a washback effect that is thought to be superficial (Cheng 1998). In this study, learners showed a willingness to practice productive skills and thus to adopt new ways of studying English, but experienced difficulty in achieving this.

The prevalence of folk-knowledge and gaining advice from others, as well as an apparent dependency on assistance from those who can offer feedback was observed. In non-instructed contexts, test takers are perhaps more likely to rely on advice from peers. They may also feel even more dependent on the feedback of others for improving their productive skills, and feel that without others to speak with or write to, there

is little they can do by themselves to develop their abilities. As in Mickan and Motteram's (2009) study, many learners in the present study lacked personal agency in test preparation, at least when faced with preparing for productive skills. Interestingly, this lack of agency has now been observed in two very different contexts (ESL vs. EFL) and with different test takers (a variety of mainly Asian nationals taking the test for immigration purposes vs. Japanese high achieving undergraduates). Thus it appears that when faced with preparation for the IELTS exam (either General Training or Academic), test takers often feel incapable of studying by themselves for the test. This is clearly an important issue that can hinder positive washback from the test and thus must be subject to further scrutiny in future research.

In the Japanese tertiary context, four skills tests might be one way to generate washback upon productive skills that have hitherto been sidelined in favour of testing receptive skills. The IELTS Test in this study provided a stimulus, which oriented students towards the study of productive skills, which is a positive step forward for the test takers' development of a rounded language proficiency in the four skills. While the increase in spoken language proficiency cannot be attributed solely to IELTS test preparation, it does provide an indication of the potential for positive washback on learning from the IELTS Test in the Japanese context. However, the outcome of using the IELTS Test, or any other four skills test, as a way to generate positive washback to the learner is dependent on numerous factors which are intrinsically related to the sociocultural and educational context. Test designers and test users therefore must consider these factors when planning to introduce tests that are intended to promote positive washback to the learner.

First and foremost, learners must understand the test demands, content, format, and weighting of the test sections; they must know how to prepare and how to improve; and they must know how to interpret their scores. Such information must be provided by test developers for stakeholders in order to promote positive washback from the test. In the case of IELTS, the developers and administrative institutions (i.e., British Council, Cambridge English Language Assessment, and IDP: IELTS-Australia) do provide considerable online support for learners. In addition, learners in this study had access to free online resources ('Road to IELTS: IELTS Preparation and Practice'). However, even still, many did not fully understand the test demands, how to prepare for the test and how to improve their scores.

Importantly, test takers need guidance on how to study for the test, specifically the parts of the tests that may be novel to test takers from different contexts. Moreover, guidance must be appropriate for the local context in which there may be few opportunities to productively use English in daily life. In this study, although it is possible to practice aspects of speaking and writing individually, this was not known, or not accepted by many of the interviewees. Guidance provided by test developers about how to prepare independently for the speaking and writing exams, both with and without others to act as interlocutors or providers of feedback, is therefore crucial to generating positive washback.

Provision of such guidance is the responsibility of the test developers as part of their commitment to promoting positive washback and thus ensuring the consequential validity of the test. Others who seek to assist learners in their preparation for the test (i.e., instructors in universities, schools, cram schools, and other institutions) must be able to access to this information in order to facilitate the promotion of positive washback.

Although the extent to which test developers can control the social consequences of tests is disputed (Alderson 2004), striving to achieve positive washback is necessarily one aspect of an ethical approach to language test development (O'Sullivan & Weir 2011). Given the increased interest in four skills tests for university entrance purposes in Japan (In'nami et al. 2016), test-related guidance will need to address the issues highlighted here. If positive washback is to be generated, learners must have the ability, as well as the inclination, to study productive skills in their everyday environment.

Limitations

As with all washback studies, a number of limitations must be taken into consideration when considering the findings of this study. Firstly, the findings are derived from self-report data. Participants may not have accurately recalled all of their preparation behaviour, which is particularly important regarding preparation for the first test, which was taken up to a year before the survey/interview data was collected. However, the interview data did appear to support those from the survey, which increases the reliability of the findings. Secondly, the study focused on high academic achievers at one university in Japan, and the generalizability of the findings is potentially limited by this constraint. However, there is no reason to suggest that the findings are not generalizable to other high academic achievers in universities across the country and elsewhere in the region. In fact, the mediating factors observed here overlapped greatly with those reported in Zhan and Andrews' (2014) small-scale study with three Chinese test takers, indicating some generalizability across contexts. The findings may also apply similarly to other well-designed, balanced four skills tests of academic English language proficiency that are used in the Japanese context (e.g., TEAP). Such four skills tests have the potential to impact learners' test preparation and proficiency development in the near future in Japan, though the mediating factors outlined here must be accounted for to ensure such positive washback can be generated.

Future directions

Following the validation framework proposed by Weir (2005) and O'Sullivan and Weir (2011), demonstrating the consequential validity of a test requires, *inter alia*, evidence of washback upon teaching and learning. Washback was clearly evident in the present study, and this is in part due to the fact that when a test contains features that are novel in the context, it is more likely to elicit changes in behaviour. In contrast, when a test does not include features that are innovative in the context, it is less likely to do so. Consequently, demonstrating consequential validity for a less innovative test will be more challenging as washback effects may be difficult to discern. Similarly, when teachers and learners are already familiar with a test, washback effects will be more difficult to establish. Future validation studies must be aware of these issues when attempting to demonstrate consequential validity. Moreover, even when contextually innovative tests are introduced, different learner populations may well behave differently. To develop a stronger and broader validity argument, washback research with a variety of learner populations within a particular context is required. For example, washback studies with test takers who are not self-selecting, and who are not already high achievers, would strengthen the present argument put forward for the IELTS Test in EFL contexts such as Japan.

Appendix 1

Table 6 Survey questions regarding IELTS Test preparation

Survey question* Items repeated for Test 2	Question sub-categories	Responses
Have you taken IELTS prior to the first test? (If yes, what was your score?)		Yes/No (Free response)
Why did you decide to take the IELTS Test?	o For study abroad o To find out about IELTS o For work o For qualification o Because it was free	Select all that apply
Did you attend either of the IELTS preparation courses?		Yes/No
Did you use the online materials?		Yes/No
Did you receive additional tuition for your tests?		Yes/No
If yes (40), where?	o Cram school o Conversation school o Personal contact o Other	Select all that apply
*How many hours did you study for the test?		0, <20, 20–40, 40–60, 60–80, 80–100, 100+
Likert scale responses (1 = Strongly disagree, 2 = Disagree, 3 = Somewhat disagree, 4 = Somewhat agree, 5 = Agree, 6 = Strongly agree)		
*In preparation for the first/second test I studied *mainly* _____.	o Reading o Listening o Writing o Speaking	
*In preparation for the first/second test I spent a lot of time _____.	o Reading texts then answering questions o Listening to monologues/conversations between two people then answering questions o Listening to conversations between more than two people then answering questions o Speaking about familiar topics spontaneously with a partner or partners o Speaking about abstract topics spontaneously with a partner or partners o Writing a paragraph to summarize information from a chart or table o Writing an essay	
*I practiced speaking immediately with little or no preparation time.		
*My preparation activities focused a lot on _____.	o Pronunciation o Grammar o Lexis	
*Overall, I studied test-taking techniques a lot.		

Appendix 2

Semi-structured interview questions

Interviews were semi-structured and included some or all of the following prompts:

- Please tell me how you studied for the first (second) test?
- How long did you prepare for the first (second) test?
- How did you study reading/writing/speaking/listening for first (second) test?

- Why did (or didn't) you focus on reading/writing/speaking/listening for first (or second) test?
- What kind of activities did you do in preparation for the first (second) test?
- What materials did you use for the first (second) test?
- Did you study grammar/vocabulary/pronunciation for the first (second) test? How?
- Did you practice speaking spontaneously by yourself or with someone for the first (second) test? How?
- Did you study any techniques for the first (second) test? How?

Acknowledgements
I am grateful to Sayaka Meguro, Masaaki Ogura, Shoko Tanaka and Kimie Yamamura for their assistance in the data collection stages of this project, Dr. Akiko Katayama for providing training for the assistants, Prof. Yoshinori Watanabe for helpful discussion, and to Prof. Barry O'Sullivan and Chie Yasuda at the British Council for their assistance.

Funding
This research is part of a study that was funded by the British Council.

Author's information
David Allen gained his PhD in Psycholinguistics from the University of Nottingham. His research concerns a variety of areas in psycholinguistics and applied linguistics, specifically bilingual lexical processing and representation, peer feedback in second language writing, corpus studies, and most recently, test washback in the Japanese context. He is Associate Professor at Ochanomizu University in Tokyo, where he teaches applied linguistics and English language related courses.

Competing interests
The author declares that he has no competing interests.

Received: 23 August 2016 Accepted: 19 September 2016
Published online: 03 October 2016

References

Alderson, J. C. (2004). Foreword. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. ix–xii). Mahwah: Lawrence Erlbaum Associates.
Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.
Allen, D. (2016). Japanese cram schools and entrance exam washback. *Asian Journal of Applied Linguistics, 3*(1), 54–67.
Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting Washback: A Case-Study. *System, 30*(2), 207–223.
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: 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. (1999). *Washback in Language Testing* (TOEFL Monograph Series. Report Number: RM-99-04, TOEFL-MS-15). Princeton: Educational Testing Service.
Bourdieu, P. (1990). *In other words: Essays towards a reflexive sociology*. Cambridge: Polity Press.
Brown, J. D. (1998). An investigation into approaches to IELTS preparation, with particular focus on the Academic Writing component of the test. In S. Wood (Ed.), *IELTS Research Reports* (Vol. 1, pp. 20–37). Sydney: ELICOS/IELTS Australia.
Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38–54.
Cheng, L. (1998). Impact of a public examination change on students' perceptions and attitudes towards their English learning. *Studies in Educational Evaluation, 24*(3), 279–301.
Cheng, L. (2005). *Changing Language Teaching through Language Testing: A washback study* (Studies in Language Testing 21). Cambridge: Cambridge ESOL/Cambridge University Press.
Cheng, L. (2014). Consequences, Impact and Washback. In A. J. Kunnan (Ed.), *The companion to language assessment: Evaluation, Methodology and Interdisciplinary themes* (pp. 1130–1145). Singapore: Wiley Blackwell.
Cheng, L., Andrews, A., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing, 28*(2), 221–249.
Craven, E. (2012). The quest for IELTS Band 7.0: Investigating English language proficiency development of international students at an Australian university. In J. Osborne (Ed.), *IELTS Research Reports* (Vol. 13, pp. 1-61). Canberra: IELTS Australia and Manchester: British Council.
Elder, C. & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. In R. Tulloh (Ed.), *IELTS Research Reports* (Vol. 4, pp. 207-254). Canberra: IELTS Australia.
Gosa, C. M. C. (2004). *Investigating Washback: A Case Study Using Student Diaries*. Lancaster: Unpublished PhD thesis, Department of Linguistics and Modern English Language, Lancaster University.
Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test. *Assessing Writing, 10*, 44–60.
Green, A. (2006a). Washback to the learner: Learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing Writing, 11*, 113–134.

Green, A. (2006b). 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.
Green, A. (2007). *IELTS Washback in Context: Preparation for academic writing in higher education. Studies in Language Testing 25*. Cambridge: Cambridge University Press.
Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.
Humphreys, P., Haugh, M., Fenton-Smith, M., Lobo, A., Michael, R. & Walkinshaw, I. (2012). Tracking international students' English proficiency over the first semester of undergraduate study. In J. Osborne and G. Lim (Eds.), *IELTS Research Report Series* (Vol. 1, pp. 1-41). Canberra: IDP: IELTS Australia.
In'nami, Y., Koizumi, R. & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. Language testing in Asia, 6(3), doi:10.1186/s40468-016-0025-9
Kvale, S., & Brinkmann, S. (2009). *InterViews: Learning the Craft of Qualitative Research Interviewing* (2nd ed.). London: Sage.
Mason, J. (2002). *Qualitative Researching* (2nd ed.). New Delhi: Sage.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.
Mickan, P., & Motteram, J. (2008). An ethnographic study of classroom instruction in an IELTS preparation program. In J. Osborne (Ed.), *IELTS Research Reports* (Vol. 8, pp.17-43). Canberra: IELTS Australia.
Mickan, P., & Motteram, J. (2009). The preparation practices of IELTS candidates: Case studies. In J. Osborne (Ed.), *IELTS Research Reports* (Vol. 10, pp. 223-262). Canberra: IELTS Australia and Manchester: British Council.
Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2016). Support for internationalization of universities. http://www.mext.go.jp/english/highered/1326670.htm. Accessed 29 Mar 2016.
O'Loughlin, K., & Arkoudis, S. (2009). Investigating IELTS exit score gains in higher education. In J. Osborne (Ed.), *IELTS Research Reports* (Vol. 10, pp. 95-180). Canberra: IELTS Australia and Manchester: British Council
O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language Testing: Theories and practices* (pp. 13–32). Basingstoke: Palgrave Macmillan.
Pan, Y. C. (2014). Learner Washback Variability in Standardized Exit Tests. *TESL-EJ, 18*(2), 1–30.
Read, J., & Hayes, B. (2003). The Impact of IELTS on Preparation for Academic Study in New Zealand. In R. Tulloh (Ed.), *IELTS Research Reports* (Vol. 4, pp. 153-206). Canberra: IELTS Australia.
Saville, N., & Hawkey, R. (2004). The IELTS impact study: Investigating washback on teaching materials. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 73–96). Mahwah: Lawrence Erlbaum Associates.
Shih, C. M. (2007). A new washback model of students' learning. *Canadian Modern Language Review, 64*(1), 135–162.
Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington: National Foreign Language Center.
Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317.
Smyth, E., & Banks, J. (2012). High stakes testing and student perspectives on teaching and learning in the Republic of Ireland. *Educational Assessment, Evaluation and Accountability, 24*(4), 283–306.
Stoneman, B. W. (2006). *The impact of an exit English test on Hong Kong undergraduates: A study investigating the effects of test status on students' test preparation behaviors (Unpublished PhD thesis)*. Hong Kong: The Hong Kong Polytechnic University.
Taylor, L. (2014). *A Report on the Review of Test Specifications for the Reading and Listening Papers of the Test of English for Academic Purposes (TEAP) for Japanese University Entrants*. Tokyo: Eiken Foundation of Japan.
Tsai, Y., & Tsou, C. (2009). A standardized English Language Proficiency test as the graduation benchmark: Student perspectives on its application in higher education. *Assessment in Education: Principles, Policy & Practice, 16*(3), 319–330.
Tsagari, D. (2007). *Investigating the Washback Effect of a High-Stakes EFL Exam in the Greek Context: Participants' Perceptions, Material Design and Classroom Applications*. UK: PhD Thesis, Submitted in Department of Linguistics and English Language, Lancaster University.
Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Context and method in washback research: The influence of language testing on teaching and learning* (pp. 129–146). Hilsdale: Lawrence Erlbaum.
Wall, D., & Horak, T. (2011). The Impact of Changes in the TOEFL® Exam on Teaching in a Sample of Countries in Europe: Phase 3, The Role of the Coursebook Phase 4, Describing Change. TOEFL iBT® Research Report, TOEFL iBT-17.
Weir, C. J. (2005). *Language Testing And Validation: An Evidence-based Approach*. London: Palgrave-Macmillan.
Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English language examinations 1913–2012* (Studies in Language Testing 37). Cambridge: Cambridge ESOL/Cambridge University Press.
Xie, Q. (2013). Does Test Preparation Work? Implications for Score Validity. *Language Assessment Quarterly, 10*(2), 196–218.
Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing, 30*(1), 49–70.
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.
Zhan, Y. & Wan, Z. H. (2016). Test Takers' Beliefs and Experiences of a High-stakes Computer-based English Listening and Speaking Test. RELC Journal (pre-print online version). doi:10.1177/0033688216631174