Language Testing in Asia

RESEARCH Open Access

Standardized proficiency tests in a campus-wide English curriculum: a washback study

Shao-Ting Alan Hung[1] and Heng-Tsung Danny Huang[2*]

* Correspondence:
dannyhuang123@ntu.edu.tw
[2]Department of Foreign Languages and Literatures, National Taiwan University, Taipei, Taiwan
Full list of author information is available at the end of the article

Abstract

Washback refers to the influence of tests on learning and teaching. To date, a number of studies have revealed that tests affect teaching content, course design, and classroom practices. However, in Asian higher education contexts, little research has examined the washback of proficiency tests on English learning in comparison with the efforts on teaching. Thus, the current study bridged this research gap by exploring the washback effects of a proficiency test on student learning in a campus-wide English curriculum, uncovering relationships between washback and learner characteristics such as major, gender, and proficiency level. A total of 694 students from engineering-, business-, and foreign language-related disciplines at a national university in Taiwan were surveyed. The results revealed washback effects on such aspects as personal image, learning motivation, emotion, and future job opportunities were especially salient. In addition, the relationship between washback and proficiency level was found to be statistically significant. However, male and female students did not differ statistically in washback nor was there a statistically significant difference in washback among different majors. With detailed information and consideration of different aspects of washback, stakeholders, including instructors, school administrators, and language policymakers, can make informed decisions when formulating language-related policies.

Keywords: Washback, Standardized testing, Test impact, Test-taker characteristics

Background

Nowadays, a great deal of higher education institutions around the world rely on standardized English proficiency tests not only in evaluating teaching effectiveness and learning outcomes, but also in achieving desirable pedagogical changes (Cheng, Andrews &Yu, 2010; Qi, 2007). Policymakers and administrators at schools of higher education generally perceive these proficiency tests as possessing a certain degree of credibility in determining learning performance. Hence, believing that tests have much power, they use them to implement educational policies (Shohamy, 2001).

Given the prevalence of standardized English proficiency tests around the world, a number of researchers have claimed that these tests affect not merely individual learning styles and future careers but also teaching approaches, syllabi, curriculum planning, and educational systems (Alderson & Wall, 1993; Wall, 1998). Additionally, to date, a plethora of studies have revealed that tests influence teaching content, course design, and classroom practices (Cheng, 2005; Cheng, 2007; Green, 2007; Pan, 2009; Shih, 2010; Wall,

2005). However, studies done in Taiwan's higher education contexts are limited in number and scope. Thus far, only a few studies, e.g., Shih (2007, 2010) and Pan (2012), have put forth efforts in researching how tests influence teaching and learning within classes or departments in Taiwan. Additionally, most if not all of them were small-scale case studies with large-scale investigations that focus on campus-wide English curricula being largely absent. Therefore, the present study was undertaken to bridge gaps in the relevant research.

Literature review

Washback

The influence of tests on learning and teaching is called washback or backwash (Alderson & Wall, 1993; Biggs, 1996). More specifically, washback is generally known to be the positive or negative impact tests have on teachers' instruction and students' learning (Cheng, 2004). Alderson and Wall (1993) argued that washback must not be assumed, but examined for specific aspects of influences (e.g., teaching methodology, assessment techniques, and materials), for the direction of washback (i.e., positive, negative), and for the extent of impact (i.e., strong, weak). Furthermore, Bachman and Palmer (1996) defined washback in broader notions of test impact and consequences. In the micro level, tests influence individual learners and teachers while in the macro level, tests affect educational systems and society. They further proposed three ways in which learners can be affected—through their experiences of taking and preparing for the exam, the feedback they receive about their performance, and the decisions made about them. Finally, Green (2007) posited that washback effects are complex and variable for different individuals in terms of the washback intent, direction, and intensity. To elaborate, Green (2007) discussed that washback intensity varies in relation to the test-takers' perceptions of the test stakes and the test difficulty. He hypothesized that washback will be most intense where test-takers (1) value success on the test above developing skills for the target language use domain, (2) consider success on the test to be challenging, and (3) work in a context where these perceptions are shared by other participants (as cited in Purpura, 2009). Drawing on Alderson and Wall's (1993) and Hughes' (1989) framework on washback, Bailey (1996) proposed a model that consisted of two aspects: washback to the learner and washback to the program. Washback to the learner refers to "the effects of test-derived information provided to the test-takers and having a direct impact on them" (p. 263) while washback to the program means the results of test-derived information given to such stakeholders as teachers, administrators, and curriculum developers. She further suggested a number of strategies to promote positive washback; for example, the integration of language learning goals, authenticity, learner autonomy, and detailed score reporting.

To date, a number of studies have investigated washback effects that large-scale proficiency tests have on language learning and teaching (Alderson and Hamp-Lyons, 1996; Allen, 2016; Cheng, 2004; Cheng, Klinger & Zheng, 2007; Green, 2007; Qi, 2004; Read & Hayes, 2003; Watanabe, 1996). For instance, researching courses offered in the private extracurricular institutions called *yobiko* (preparatory school), a type of school that prepared secondary school students for the English section of university entrance examinations, Watanabe (1996) examined the effect of the university entrance examination on the use of the grammar-translation method in Japan and found that the

entrance examination failed to play a significant role in the choice of teaching methodology. Rather, it was teacher factors that dictated how the course was taught. Next, in comparing Test of English as a Foreign Language (TOEFL) preparation course and non-TOEFL preparation course taught by the same instructor, Alderson and Hamp-Lyons (1996) discovered that the TOEFL affected language teachers on the content of instruction and teaching strategies. In addition, Cheng, Klinger, and Zheng (2007) also conducted a washback study to investigate the impact of Ontario Secondary School Literacy Test (OSSLT), a large-scale literacy test, on second language (L2) students in Canada. Test performances of English as a Second Language (ESL) and English Literacy Development (ELD) students were analyzed to identify the factors why ESL/ELD students had lower scores than their non-ESL/ELD counterparts. The results indicated that the reading test formats, text types, skills, strategies, and writing tasks impacted ESL/ELD and non-ESL/ELD learners differently and significantly. Qi (2004) probed the intended washback of the National Matriculation English Test in China and found a great inconsistency between test developers' intentions and school practice. Qi concluded that the exam achieved little of the intended washback and failed to produce instructional changes in schools in China. Similarly, situated in the Hong Kong Certificate of Education Examination, Cheng (2004) explored washback effects by examining how such examination reform influenced ESL teachers and their classroom practices in Hong Kong's secondary curriculum. The study surveyed teachers' perspectives and teaching methods around an examination change. Cheng found certain washback effects on instructors' perceptions toward the new examination. However, teachers' classroom practices were not influenced by the new examination.

Next, targeting the International English Language Testing System (IELTS) Writing Test, Green (2007) examined how preparation classes impacted score gains. The results found no significant improvement in test scores, suggesting that test-driven instruction did not necessarily raise students' scores. Hence, Green proposed that a more beneficial way of improving students' scores was to integrate materials covered on the test with regular teaching. Choi (2008) investigated the overwhelming washback effects of EFL tests on EFL teaching in Korea and found that most stakeholders do not think favorably of the EFL tests because of negative washback effects in terms of the mismatch between test scores and English proficiency. Additionally, another cause of negative washback pertained to the failure of multiple-choice items in promoting EFL learners' communicative skills. Next, Pan (2013) examined whether exit requirements have produced washback on teaching in the tertiary institutions in Taiwan. Results revealed that the exit requirements have elicited a minor degree of changes in teaching, suggesting that teachers considered test factors and test-related activities the lowest priority in their selection of materials and pedagogy. Nevertheless, the study found that teachers at schools with exit requirements have a significantly higher consideration of test factors and employ more related activities than their counterparts at school without exit requirements. Last but not least, Pan (2014) tested Alderson and Wall's hypotheses (1993) regarding learner washback variability and examined the washback of a standardized exit test. The findings uncovered the differential effects of exit requirements for different groups of learners depending on their years of study, proficiency levels, and viewpoints concerning the tests. To be specific, the intermediate- and high-proficiency students had more favorable views of the tests than low-proficiency students.

Test-taker characteristics

Bachman and Palmer (2010) have proposed that test-taker characteristics are one of the key factors that affect performance on educational measurements in general and language tests in particular. In the washback contexts, Shih (2007) has proposed a washback model and argued that extrinsic, intrinsic, and test factors determine the washback of a test on students' learning and personal psychology. In discussing intrinsic factors, Shih posited that individual differences, personal characteristics, and personal perceptions of the test could play vital roles in students' learning. Gender, as one of the individual differences, has received considerable attention in a great number of research studies (Karami, 2011; Karami, 2013; O'Loughlin, 2002; O'Sullivan, 2000; Lumley & O'Sullivan, 2005). For instance, Karami (2013) used generalizability theory to examine the presence of gender differential performance on a language proficiency test—the University of Tehran English Proficiency Test (UTEPT). The results from his study showed that there was little evidence of gender bias in the scores. The patterns of score variance were roughly the same across the groups and the dependability indices were also comparable. Hence, Karami (2013) concluded that the UTEPT had a high level of dependability and was free of gender bias. This finding was consistent with O'Loughlin's (2002) study in which he concluded that little gender effect existed in the IELTS scores. However, Takala and Kaftandjieva (2000) examined the differential item functioning (DIF) in the vocabulary section of a foreign language proficiency test and found that over 25% of the items showed DIF in favor of either male or female test-takers. Although the gender effects are not consistent across studies, their relationship with performance has been well researched.

In addition to gender, the test-taker characteristic of academic background constitutes another factor that affects performance on tests. A number of studies have examined the effects of academic backgrounds on test performance (Kreleler, 2006, Karami, 2012a, b; Pae, 2004). For instance, targeting test-takers in the humanities and sciences, Pae (2004) found that test-takers in the humanities performed better on items related to human relationships while sciences majors performed better on items related to number counting. Similarly, Hale (1988) indicated that test-takers in the humanities and social science majors outperformed test-takers in the biological and physical sciences on the items related to the humanities or social sciences. Nevertheless, Karami (2012a, b) explored the impact of academic background on the dependability of the scores from a high-stakes language proficiency test, UTEPT, and found that there was no significant interaction between items and academic fields. Thus, he concluded that academic background did not exert a large influence on the performance of the UTEPT test-takers.

In brief, the area of test-taker characteristics has been researched. Their relationship with test performance has been supported by empirical studies. Nevertheless, how test-taker characteristics interact with washback effects still remained underexplored.

The proficiency test GEPT and its washback studies

Developed in 1999 and launched in 2000 by the Language Training and Testing Center (LTTC) in Taiwan, the General English Proficiency Test (GEPT) aims to establish a fair, valid, and reliable test for all local English language learners and promotes lifelong

learning in English (LTTC, n.d.a); it is a five-level criterion-referenced English proficiency test that includes elementary, intermediate, high intermediate, advanced, and superior. It has been widely used by universities, government offices, and private sectors for various purposes. For universities, GEPT has been used as a graduation benchmark; for government offices, it has been used as a criterion for promotion; for private sectors, it has served as a reference for recruitment, pay raises, and promotion. The fact that GEPT was partially funded by the Ministry of Education increases its credibility in the public opinion (Vongpumivitch, 2010). In fact, according to the LTTC, approximately 40 government offices and private sectors, 126 universities and colleges, and 315 junior and senior high schools are using GEPT scores to make a variety of decisions. To date, approximately 6.5 million people have registered for the GEPT (LTTC newsletter, 2016). It is the largest standardized English proficiency test in Taiwan with approximately 500,000 test-takers each year at over 100 test sites around the country (Wu & Lee, 2017).

Due to its high popularity, a number of researchers have invested efforts in studying the GEPT in a variety of areas. For instance, studies have been conducted on corpus analysis of written performance (Kuo, 2005), parallel form reliability (Weir & Wu, 2006), validation (Chin & Kuo, 2004), comparisons of GEPT scores with other tests (Chin & Wu, 2001; Wu, Wu & Dunlea, 2015), item analysis (Tung, Bai, & Shi, 2005), and motivation and anxiety in test performance (Cheng, Klinger, Fox, Doe, Jin, & Wu, 2014; [name deleted to maintain the integrity of the review process, 2016]. These studies have enriched the testing literature on the GEPT in the past decade.

According to LTTC's research agenda for new directions, the impact of the GEPT will be one of the key future developments (Kunnan & Wu, 2010). However, only a few studies have been done on the relationships between test impact and test-taker characteristics on tertiary education in Taiwan (Shih, 2010; Wu & Chin, 2006). Specifically, Wu and Chin's (2006) washback study focused on students and teachers in high schools rather than in colleges. Next, Shih (2007) conducted a qualitative study to uncover the washback of the GEPT on students' English learning in two institutions of higher education in Taiwan. He interviewed department chairs, English teachers, students, and family members; observed teachers' instruction and activities; and reviewed documents related to the GEPT policy. The findings revealed that the GEPT generated little impact on learners at both schools, indicating that existing theories failed to fully explicate the washback of the GEPT on higher education. Therefore, a washback model that contains extrinsic, intrinsic, and test factors was proposed to bring a more insightful view of the impact on student learning. Subsequently, employing the same instruments, interviews, and document analysis, Shih (2010) investigated the washback of the GEPT on school policies. The results revealed that when making policies on English requirements, a number of factors needed to be considered, such as social and educational factors, school factors, and parental and student factors. Although Shih's studies (2007, 2010) targeted Taiwanese tertiary education, they did not investigate the relationships between test impact and test-taker characteristics. Therefore, the current project aims to bridge that gap by examining the test impact, also known as washback, of the GEPT and its relationships with test-taker characteristics, such as gender, learner backgrounds, and proficiency levels.

Statement of purpose

The purpose of the current research project is threefold. First and foremost, according to the LTTC's research agenda on new directions for language testing research, the impact of the proficiency tests will be one of the key future developments (Kunnan & Wu, 2010). Nevertheless, there is not enough research on the washback effects of proficiency tests in the Taiwanese context. In fact, Shih's (2007, 2010) and Pan's (2012) studies were the only published journal articles that examined the washback effects of the GEPT in Taiwan's higher educational context. Due to the limited number of participants, however, the results from Shih's small-scale case studies may not be generalized to larger populations. Hence, surveying a larger sample size of learners becomes necessary for a more holistic picture of washback effects on English learning and teaching in the Taiwanese context.

Second, according to some researchers (e.g., Cheng, 2008; Shih, 2010; Wall, 2000; Watanabe, 2004), little research has examined the washback of tests on learning in comparison with the efforts put forth to examine the washback on teaching. As stated by Cheng (2010), students are the ultimate stakeholders in any assessment practices, but they have been researched less compared with other stakeholders in previous washback studies. Therefore, it becomes urgent to find out students' attitudes on how tests affect their language learning in various aspects.

Third, thus far, few washback studies have investigated the relationships between test impact and learner characteristics. Cheng and Curtis (2012) pointed out that some areas needed to be strengthened in washback research. One of the areas pertains to the exploration of the relationships between test-taker characteristics and test performance. That is to say, future washback research could center on such characteristics as gender, age, learning strategies, motivation, and test anxiety. In addition, little research has investigated the interplay between washback effects and proficiency levels. Whether more proficient test-takers experience a greater strength of washback remains unknown. Similarly, further exploration is needed to ascertain whether gender and academic background play a role in washback.

Therefore, the present study was proposed to bridge these three gaps by examining the washback effects of a proficiency test in a larger context—a campus-wide English curriculum—and their relationships with learner characteristics.

Research questions

Proceeding from the foregoing rationale, this study posed four research questions as follows:

1. Are there identifiable washback effects of a proficiency test on student learning in tertiary institutions in Taiwan?
2. What is the relationship between washback effects and proficiency levels?
3. Is there a significant difference in the strength of washback experienced by male and female students?
4. Is there a significant difference in the strength of washback experienced by learners of different academic majors (i.e., Engineering, Management, and Foreign Languages)?

Methods

Context of the study

The data for the current study were collected from a national university in Taiwan. The university has undergraduate and graduate programs in five colleges: Engineering, Electrical Engineering and Computer Science, Management, Finance and Banking, and Foreign Languages. In this university, the graduation benchmark for undergraduate students is to reach the B1 level of the Common European Framework of Reference for Languages (CEFR), which corresponds to the intermediate level of the GEPT, 550 on the Test of English for International Communication (TOEIC) and 4.0 on the International English Language Testing System (IELTS).

Participants

The current study recruited a total of 753 participants from different academic disciplines. For the questionnaire development, nine students were invited to participate in focus group interviews for item generation, and 50 students were invited to fill out the questionnaire for piloting purposes. For the official study, 694 students were invited to respond to the finalized, official questionnaire. Among these 694 participants, 390 were recruited from engineering-related departments (hereafter Engineering cohort), 231 were from business-related majors (hereafter Business cohort), and 73 were from the foreign language department (hereafter Foreign language cohort). The percentages of learner distribution in Engineering, Business, and Foreign Language cohorts are 56%, 33%, and 11%, respectively, also shown in Fig. 1.

Proficiency tests at the research site

The proficiency test administered at the research site was a simulation test of the GEPT. Among the five levels of the GEPT (elementary, intermediate, high-intermediate, advanced, and superior), GEPT-intermediate was selected by the school since the graduate benchmark in that school was to reach a B1 level on the CEFR, equivalent to GEPT-intermediate. A general level description of GEPT-intermediate is to be able to use basic English to communicate on topics in daily life. To be specific, test-takers who pass the listening section of GEPT-intermediate can (1) understand general English conversation in daily life situations and (2) grasp the general meaning of announcements, advertisements, and broadcasts while

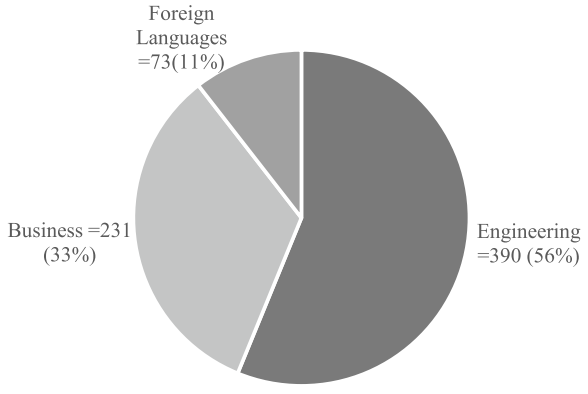

Fig. 1 Learner distribution

test-takers who pass the reading section can (1) read stories and news articles of familiar or concrete topics and (2) read personal letters. As for the test type, the listening section includes picture description, answering questions, and conversations while the reading section includes sentence completion, cloze, and reading comprehension (see Table 1).

The graduation requirement at the research cite is to pass reading and listening sections of GEPT-intermediate or equivalent (see Appendix A). Speaking and writing were not tested because of the logistical and practicality reasons. To better prepare students for meeting the English graduation benchmark, the school has incorporated GEPT-intermediate into its English curriculum by providing test preparation materials as self-study resources. Students were required to study those materials and take practice tests. At the end of each semester, all students were required to take the mock GEPT-intermediate. Their scores on the tests accounted for 10% of their final grade for the English class they were taking. Hence, students at the research site all had experience with standardized proficiency tests, GEPT in particular, and were aware of the role of proficiency tests in the curriculum.

Instruments

The research data for the current study came from questionnaires and students' test scores.

Questionnaire development

To investigate the washback effects of the GEPT on student learning, a questionnaire was adapted from Cheng (2005). The questionnaire examined students' reactions toward the impact of proficiency tests on various aspects of learning. Cheng's (2005) questionnaire was designed to explore students' attitudinal and behavioral changes in relation to classroom in the context of the Hong Kong College Entrance Exam (HKCEE). It also explored the role of the public examination in student learning. Hence, the aims of Cheng's questionnaire well served the purposes of the current study, which also aimed to explore how students' learning was influenced by a proficiency test.

In Cheng' questionnaire, a number of themes were covered, such as classroom teaching and learning activities, perceptions of aspects of learning, preferred learning strategies, aspects of examination influences on students, and students' attitudes towards the public exam. Since some themes were not so pertinent to the current study, only some items under certain themes were adapted. For example, items under the themes, aspects of examination influences on students, and students' attitudes towards the public exam were adapted to better suit the present study. After item modification, the finalized questionnaire contains 9 items that explicitly and implicitly examined the impact of tests and probed changes after the tests were used. The alpha coefficient of the revised questionnaire reached 0.89 in the current study, suggesting that the questionnaire was a reliable scale. The

Table 1 GEPT-intermediate: skills, test types, items, time, and scores

Test	Skill	Test type	Number of items	Time	Full score
GEPT-intermediate	Listening	• Picture description • Answering questions • Conversations	45	30 mins	120
	Reading	• Sentence completion • Cloze and reading • Comprehension	40	45 min	120

correlation between each individual item and the total of scores on the other items was provided in Table 2. Corrected item-total correlations refer to "the correlations between each item and the total score from the questionnaire" and an item that comes with a value smaller than .30 does not correlate well with the scale (Field, 2009, p. 678). As shown in Table 2, the corrected item-total correlations of the nine items all emerged to be larger than .30, indicating their adequate correlation with the entire scale and justifying the decision of the retaining them on the scale. The questionnaire was then piloted in a group of 50 students before officially administered.

In an attempt to further establish the validity of this custom-designed questionnaire, the current researchers performed a confirmatory factor analysis (CFA) on the gathered responses via taking several steps. To begin, they ensured that the assumptions underlying the conduct of CFA have been met, namely they confirmed the absence of cases with missing values, univariate and multivariate outliers, pairwise nonlinearity, and multicollinearity. Next, they developed a hypothesized model with all of the 9 items postulated to measure one single latent construct, i.e., washback. Then, the researchers drew on the maximum likelihood estimation method of the AMOS 22.0 software package to estimate this hypothesized model with reference to the fit indices

Table 2 Items used to determine washback

Items	*M*	Std. Dev.	VNI[a]	NI[a]	*F*	PI[a]	VPI[a]	Corrected item-total correlation
1. My performance on standardized tests affects my personal image.	3.45	0.98	26 (4%)	63 (9%)	282 (41%)	220 (32%)	103 (15%)	.62
2. My performance on standardized tests affects my learning motivation.	3.35	0.99	37 (5%)	61 (9%)	310 (45%)	197 (28%)	89 (13%)	.58
3. My performance on standardized tests affects my relationship with teachers.	2.85	1.02	82 (12%)	123 (18%)	350 (50%)	92 (13%)	47 (7%)	.65
4. My performance on standardized tests affects my relationship with my peers.	2.80	1.06	101 (15%)	122 (18%)	328 (47%)	98 (14%)	45 (6%)	.64
5. My performance on standardized tests affects my emotion.	3.20	1.02	49 (7%)	79 (11%)	322 (46%)	169 (24%)	75 (11%)	.73
6. My performance on standardized tests affects my future job opportunities.	3.50	1.04	34 (5%)	47 (7%)	294 (42%)	178 (26%)	141 (20%)	.69
7. Participating in mock proficiency tests helps me improve my English ability.	3.32	0.97	40 (6%)	62 (9%)	302 (44%)	218 (31%)	72 (10%)	.63
8. Tests push me to study harder.	3.50	1.01	35 (5%)	46 (7%)	266 (38%)	231 (33%)	116 (17%)	.64
9. I perform better on tests than on regular learning.	3.09	0.91	36 (5%)	97 (14%)	380 (55%)	129 (19%)	52 (7%)	.64

[a]*VNI* very negative influence, *NI* negative influence, *F* fair, *PI* positive influence, *VPI* very positive influence. *N* = 694

recommended by Kline (2005): the χ^2 test statistic, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). After performing a series of model estimations and modifications based upon substantive rationales, they derived the finalized model that fitted the collected data reasonably well: χ^2 (20) = 126.11, $p < 0.01$; CFI = 0.96; TLI = 0.95; RMSEA = 0.08; SRMR = 0.09, further the 9 items constituted significant and strong indicators of the latent construct, as they all came with a factor loading larger than 0.60 (Garson, 2013) (Fig. 2). Taken together, the results of the CFA verified the hypothesized relationship that the questionnaire items collectively assess a univariate construct, namely washback, and thus lent additional support for the validity for this questionnaire.

The scores

Students' scores on GEPT-intermediate, which are composed of two sub-scores on listening comprehension and reading comprehension, were obtained to be compared with their responses from the questionnaires. After the university administered one simulated GEPT-intermediate at the end of the semester, the researcher requested the scores and ensured confidentiality by replacing students' numbers and names with codes.

Results

Results to research question 1: in what aspects do washback exist in student learning?

The data from the questionnaire (Table 2) showed that among its 9 items, the items that received the highest mean scores were item 6: *My performance on standardized tests affects*

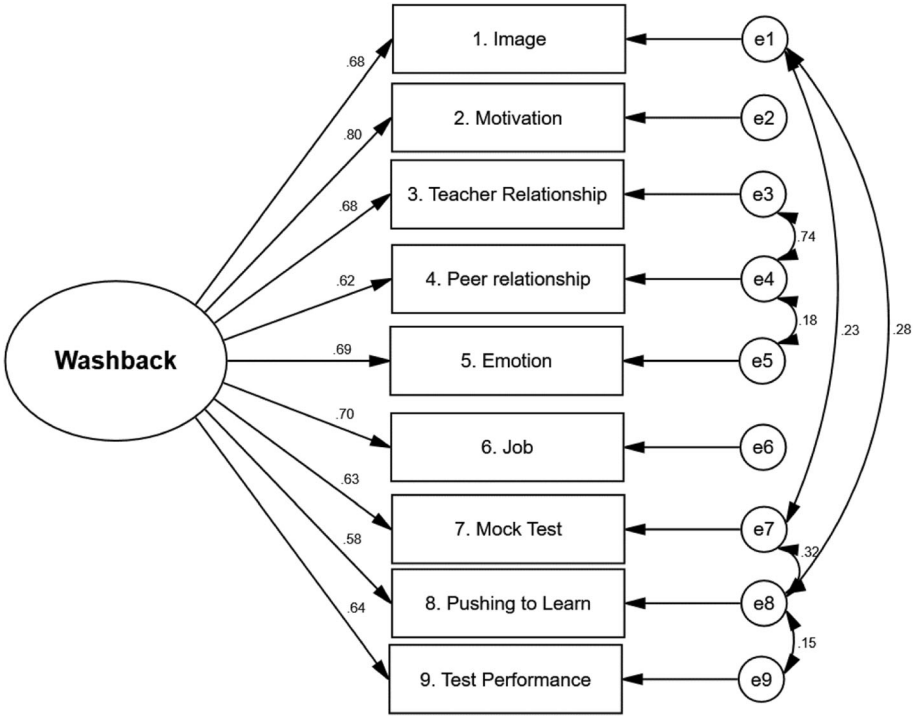

Fig. 2 The finalized confirmatory model with estimated parameters

my future job opportunities (*M* = 3.50) and item 8: *Tests push me to study harder* (*M* = 3.50), followed by item 1: *My performance on standardized tests affects my personal image* (*M* = 3.45). Next, the items that received mean scores above 3 were item 2: *My performance on standardized tests affects my learning motivation* (*M* = 3.35), item 7: *Participating in mock proficiency tests helps me improve my English ability* (*M* = 3.32), item 5: *My performance on standardized tests affects my emotion* (*M* = 3.2), and item 9: *I perform better on tests than on regular learning* (*M* = 3.09). The results not only indicated that washback existed in various aspects of students' English learning, such as personal image, learning motivation, emotion, and future job opportunities, but also revealed that the proficiency test induced such changes on student learning as improving English ability, studying harder, and performing better on tests.

However, compared with the abovementioned washback effects on various learning aspects, the ones on relationship with teachers and peers tended to be weaker. For instance, item 3: *My performance on standardized tests affects my relationship with teachers* (*M* = 2.85) and item 4: *My performance on standardized tests affects my relationship with my peers* (*M* = 2.80) received mean scores lower than 3, suggesting that external factors seemed less salient as compared with internal factors, such as personal image and learning motivation.

Results to research question 2: is the strength of washback related to proficiency levels?

The second research question investigated whether the strength of washback is related to learners' English proficiency. The grand mean of 9 items on washback was used because it represents all aspects of washback being examined in the current study. Furthermore, after ranking all of the participants' scores from highest to lowest, the top quartile (high-proficiency group) and bottom quartile (low-proficiency group) scores were used. These two sets of scores were then compared using the *t* test for independent samples. The relationship between the strength of washback and proficiency levels was found to be statistically significant. Specifically, the high-proficiency group experienced a significantly stronger washback effect than the low-proficiency group ($t(344) = -2.97$, $p < 0.01$) (see Table 3).

Results to research question 3: is there a significant difference in the strength of washback experienced between male and female students?

The third research question explored the differences in washback experienced between male and female students. The independent *t* test performed on the collected data revealed that male and female students did not differ statistically in the strength of the washback experienced ($t(692) = 0.50$, $p = 0$s.88). (See Table 4).

Table 3 The *t* test for independent samples comparing the mean difference in the strength of washback experienced between high-proficiency and low-proficiency students

Group	*N*	*M*	SD	Skewness	Kurtosis
High proficiency	173	29.97	5.73	0.05	1.21
Low proficiency	173	28.00	6.57	− 0.13	0.93
Independent comparisons					
Comparison	M		SE	*t*(344)	Sig.
Washback	− 1.97		.66	− 2.97	0.003**

**$p < .01$

Table 4 The *t* test for independent samples comparing the mean difference in the strength of washback experienced between male and female students

Gender	*N*	*M*	SD	Skewness	Kurtosis
Male	390	29.10	7.20	− 0.02	0.74
Female	304	29.02	5.65	− 0.12	0.98
Independent comparisons					
Comparison	*M*		SE	*t*(692)	Sig.
Washback	0.08		0.50	0.16	0.88

Results to research question 4: is there a significant difference in the strength of washback experienced among different academic majors (Engineering, Management, and Foreign Languages)?

Research question four intended to uncover any differences in washback experienced by students of different majors. A one-way ANOVA disclosed no statistically significant difference in the strength of washback experienced among Engineering, Management, and Foreign Languages majors ($F(2, 691) = 2.75$, $p = 0.06$). (See Table 5).

Discussion and conclusion

The results of this study confirm to language testing researchers and classroom teachers that washback effects on learning do exist. To be more specific, washback effects on such aspects as personal image, learning motivation, emotion, and future job opportunities were especially salient. In addition, such changes as improving English ability, studying harder, and performing better on tests were also induced. Among these, the most influenced aspect relates to future careers. Students become aware that higher scores on English proficiency tests increase their job opportunities as job markets become more competitive. The findings were found to be consistent with Hayes (2014). Investigating Thai university students' perceptions of the spread of English, Hayes (2014) found that they subscribed to the belief of English as a tool for personal economic advancement. Thus, it can be argued that the implementation of campus-wide standardized tests helps students realize the importance of English language proficiency in the future workplace. In addition to job opportunities, washback related to motivation was also notable. Examining students' attitudes toward a public exam, Cheng (2005) revealed that students' motivation was the aspect that the exam scores impacted the most ($M = 3.73$) compared with other aspects of learning. Similarly, in the present study, motivation was also found to be one of the aspects influenced ($M =$

Table 5 Results of the one-way ANOVA comparing the mean difference in the strength of washback experienced among different majors

Major	*N*	*M*	SD	Skewness	Kurtosis
Engineering	390	28.67	6.74	0.04	1.01
Management	231	29.89	6.10	0.02	0.92
Foreign Languages	73	28.56	6.83	− 0.41	1.03
Source of variation	SS	df	MS	*F*	Sig.
Between groups	235.83	2	117.92	2.75	0.06
Within groups	29605.38	691	42.84		
Total	29841.21	693			

3.35) by the campus-wide standardized proficiency exam, more so than other aspects like relationships with peers and teachers. Hence, it can be concluded that the implementation of a campus-wide standardized proficiency exam can highly motivate students. These findings correspond to Cheng (2005) in that the most influential factor that motivated students to learn English was the examination, with 30% of students agreeing with this, followed by the impact on their future jobs (27%).

Particularly noteworthy is that less washback was found on the interaction with the teachers and peers. It was believed that tests would affect test-takers' interactions with teachers and peers (Cheng, 2004; Cheng, Klinger & Zheng, 2007). However, in the current study, such aspects of washback were not salient. This finding may probably be explained by the fact that the standardized proficiency tests were not an integral part of pedagogy. Although test scores accounted for 10% of students' final grades, they were neither greatly emphasized by the classroom instructors nor discussed by the peers. Hence, tests produced little washback between teachers and students and little interaction among peers. In brief, washback on intrinsic factors in Shih's (2007) model such as motivation and personal image appeared to be stronger than on extrinsic factors such as interactions with peers and teachers.

The findings also revealed that the relationship between the strength of washback and proficiency levels was statistically significant: the high-proficiency group experienced a significantly stronger washback effect than the low-proficiency group. This finding can be explained by the possibility that more proficient learners take greater responsibility for their learning and have higher levels of extrinsic and intrinsic motivation. Thus, they care more about what tests mean to their learning. The finding was in line with Gan, Humphreys, and Hamp-Lyons (2004) and Pan (2014) who argued that more proficient students appeared to have more favorable views toward and be more willing to study for the English proficiency tests because they saw the value of proficiency tests in evaluating their English ability. On the other hand, less proficient students tended to develop negative and even resentful attitudes toward taking the test and experience a greater state of stress (Pan, 2012). In the current study, low-achieving students tended to care less about the test and thus failed to see how the test might influence their future, personal image, or job opportunities. Another reason why low-proficient students were less influenced by tests in the present study pertains to perceived test difficulty, as noted by Watanabe (2001) and Pan (2014). Given that the learners' perceptions of test difficulty can partially determine their level of motivation and amount of effort they devote to preparing for the test (Watanabe, 2001), low-achieving learners in this study may have become frustrated, discouraged, and unconcerned about the test impact.

However, the study revealed no significant difference between washback and test-taker characteristics. This finding suggests that the influences generated from standardized tests have little relationship with gender and academic background. This could be explained by the fact that these participants came from the same education system and shared similar learning experiences. Thus, their distinct characteristics may not be heterogeneous enough to play a role in washback effect. Hence, the findings that the GEPT induces similar washback effects on male and female test-takers can be compared with those from previous studies that found no gender differential performance on tests (Karami, 2013; O'Loughlin, 2002). Although their studies researched different proficiency tests, the former on UTEPT and latter on IELTS, they confirmed that these

proficiency tests possessed a high level of dependability, and thus, they argued, the tests have little gender effect. In a similar vein, with the results that showed no significant differences in the current study, it can be argued the GEPT has little gender bias in terms of test impact. Moreover, no significant difference in washback was found among different academic backgrounds, suggesting that the GEPT induces similar washback effects on students from different disciplines. While Karami (2012a, b) confirmed that academic background did not exert an influence on test performance, the current study goes a further step and concludes that academic background did not exert an impact on washback intensity, either.

Finally, the results from the study carry some implications for language teachers, policymakers, and school administrators. Derived from the current study, the significant relationship between the strength of washback and proficiency levels suggests that low-proficient students tend to be less influenced by and devote fewer efforts to standardized tests. Hence, a mechanism should be established to encourage low-proficient students to take greater responsibility for their learning as well as better prepare for standardized tests. As Allen (2016) revealed, students who reported preparing more intensely for the test made a significant improvement on their scores. Moreover, since standardized proficiency tests such as the GEPT influences students of different majors and genders no differently, these variables need not be considered when formulating foreign language-related policies, such as graduation benchmarks.

Limitation and future research

The current study has two limitations that suggest areas for future research. First, aiming for a larger size of participants (N = 694), the study employed a questionnaire to elicit washback effects on learning. Although the results from the questionnaire may be generalized to the larger population, they may fail to reflect in-depth influences on learning, as they only relied on learners' self-reported data in 9 items, and the researcher had no opportunities to conduct follow-up interviews (Munoz & Alvarez, 2010). Hence, future research should include such instruments as interviews, observation, and reflection to obtain qualitative data to corroborate the findings from the questionnaire. Next, due to the scope, the present washback study only focused on two test-taker characteristics: gender and academic background. However, other characteristics may also play a role in washback. For instance, as discovered by Scott (2007), the degree of the test impact varies in different grades in the primary school context. Future research should look into grade levels as well as other characteristics such as learning style, age, anxiety, among others.

In light of the findings that showed fewer washback effects on external factors, it would be beneficial for classroom teachers and curriculum developers to align classroom instruction with content standards in higher education and tests to verify achievement of exit-level benchmarks. Constant guidance and support over time should be provided to learners in order to induce positive washback. In other words, learners in the curriculum should be informed of the stakes and the objectives of the standardized test in the campus-wide curriculum. As noted by Munoz and Alvarez (2010), when students are well-informed of assessment practices, they can focus their learning on specific goals and thus achieve their learning outcomes.

Appendix

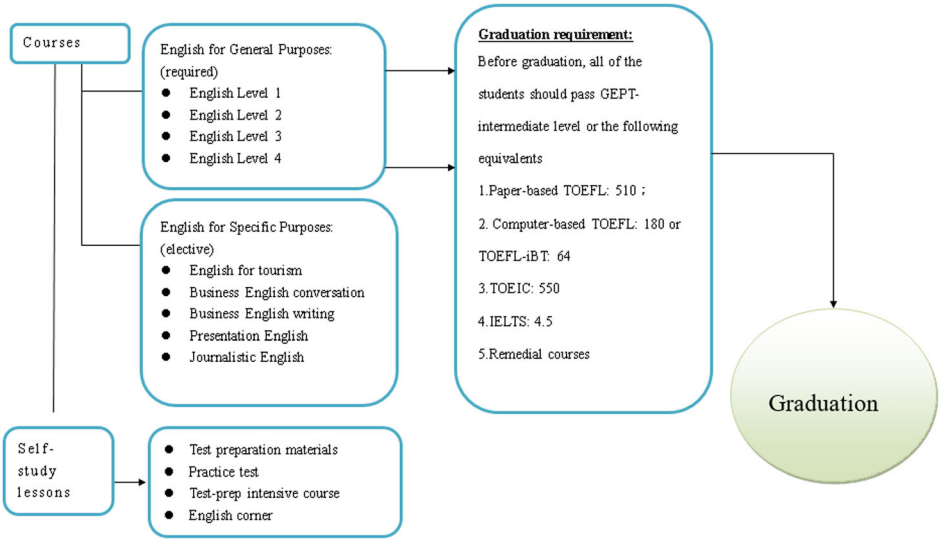

Fig. 3 English curriculum and graduation requirement

Abbreviations

ANOVA: Analysis of variance; CEFR: Common European Framework of Reference for Languages; DIF: Differential Item Functioning; ELD: English Literacy Development; ESL: English as a Second Language; GEPT: General English Proficiency Test; HKCEE: Hong Kong College Entrance Exam; IELTS: International English Language Testing System; L2: Second language; LTTC: Language Training and Testing Center; OSSLT: Ontario Secondary School Literacy Test; TOEFL: Test of English as a Foreign Language; TOEIC: Test of English for International Communication; UTEPT: University of Tehran English Proficiency Test

Acknowledgements

We would like to thank the editorial board and anonymous reviewers for their insightful comments on this paper.

Authors' contributions

Hung and Huang worked collaboratively on the research project. They collected, analyzed, and interpreted the data. They read and approved the final manuscript.

Authors' information

Dr. Shao-Ting Alan Hung (Ph.D. Indiana University, Bloomington) is a Professor and Chairperson at the Department of Applied Foreign Languages, National Taiwan University of Science and Technology. With research interests in language curriculum design, second language assessment, and educational technology, he has published in *Computer Assisted Language Learning*, *TESOL Quarterly*, *Computers and Education*, and *The Encyclopedia of Applied Linguistics*. He is also the Editor-in-Chief of *Taiwan International ESP Journal*.

Dr. Heng-Tsung Danny Huang (Ph.D. University of Texas at Austin) is an Associate Professor at National Taiwan University and Editor-in-Chief of *Language Education and Assessment*. His research interests include language testing and individual differences in Second Language Acquisition. His publications appear in *TESOL Quarterly*, *Language Testing*, *Language Assessment Quarterly*, *System*, *Learning and individual differences*, and so on.

Funding

The study was funded by the Ministry of Science and Technology (MOST), Taiwan. (Grant Number: NSC101-2410-H-011-032)

Availability of data and materials

The datasets used for the present study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Applied Foreign Languages, National Taiwan University of Science and Technology, Taipei, Taiwan.
[2]Department of Foreign Languages and Literatures, National Taiwan University, Taipei, Taiwan.

Received: 19 July 2019 Accepted: 31 October 2019
Published online: 30 December 2019

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A case study. *Language Testing, 13*, 280–297.
Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.
Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*, 1–20.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*, 257–279.
Biggs, J. B. (Ed.). (1996). *Testing: To educate or to select? Education in Hong Kong at the cross-roads*. Hong Kong: Hong Kong Educational Publishing.
Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 147–170). Mahwah, New Jersey: Lawrence Erlbaum Associations, Inc..
Cheng, L. (2005). Changing language teaching through language testing: A washback study. In M. Milanovic & C. Weir (Eds), Studies in language testing. Cambridge, UK: Cambridge University Press.
Cheng, L. (2007). What does washback look like? *Selected Papers of the 16th International Symposium on English Teaching*. November 9-11, 2007 (pp. 1-6). Taiwan: EnglishTeachers' Association/Republic of China (ETA-ROC).
Vongpumivitch, V.* (2010) "The General English Proficiency Test" In L. Cheng & A. Curtis (Eds.) English language assessment and the Chinese learner (pp.158-172). New York: Routledge.
Cheng, L., & Curtis, A. (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, Davidson, P. O'Sullivan, & Stoynoff, S (Eds), The Cambridge guide to second language assessment. New York, NY: Cambridge University Press.
Cheng, L., Klinger, D., Fox, J., Doe, C., Jin, Y., & Wu, J. (2014). *Motivation and Test Anxiety in Test Performance Across Three Testing Contexts: The CAEL, CET, and GEPT. TESOL Quarterly, 48*(2), 300–330.
Cheng, L., Klinger, D., & Zheng, Y. (2007). The challenges of the Ontario Secondary School Literacy Test for second language students. *Language Testing, 24*(2), 185–208.
Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.
Gan, Z., Humpherys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese university. *The Modern Language Journal, 88*(2), 227–244.
Garson, G. D. (2013). *Factor analysis*. Asheboro: Statistical Associates Publishers.
Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge: Cambridge University Press.
Hayes, D. (2014). The value of learning English in Thailand and its impact on Thai: Perspectives from university students. *Asia Pacific Journal of Education, 36*(1), 73–91.
Hughes, A. (1989). *Testing for language teachers*. New York: Cambridge University Press.
Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5*(2), 167–178.
Karami, H. (2012a). The relative impact of persons, items, subtests, and academic background on performance on a language proficiency test. *Psychological Test and Assessment Modeling, 54*(3), 211–226.
Karami, H. (2012b). The quest for fairness in language testing. *Educational Research and Evaluation, 19*(2&3), 158–169.
Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.
Kunnan, A. J., & Wu, R. W. (2010). The language training and testing center, Taiwan: Past, present, and future. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 77–92). New York: Routledge.
LTTC newsletter (2016). *LTTC newsletter*. Retrieved September 11th, 2017, from https://www.lttc.ntu.edu.tw/enewsletter.htm. Accessed 15 Dec 2019.
Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape- mediated assessment of speaking. *Language Testing, 22*(4), 415–437.
Munoz, A., & Alvarez, M. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing, 27*(1), 33–49.
O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169–192.
O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373–386.
Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing, 21*(1), 53–73.
Pan, Y. C. (2009). The impact of test design on teaching. *The international journal of educational and psychological assessment, 3*, 94–103.
Pan, Y. C. (2012). Tertiary EFL proficiency graduation requirements in Taiwan: A study of washback on learning. *Electronic Journal of Foreign Language Teaching, 9*(1), 108–122.
Purpura, J. (2009). The impact of large-scale and classroom-based language assessments on the individual. In L. Taylor & C. Weir (Eds.), *Language testing matter* (pp. 301–325). Cambridge: Cambridge University Press.
Qi, L. (2004). Has a high-stakes test produced the intended changes? In L. Cheng, Y. Watanabe & A. Curtis (Eds), Washback in language testing: Research contexts and methods (pp. 171-190). Mahwah, New Jersey: Lawrence Erlbaum Associations, Inc.
Read, J. & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. In R. Tulloh (ed.), *International English Language Testing System Research Reports, 4,* Canberra, Australia: ITLTS.
Shih, C. M. (2007). A new washback model of students' learning. *Canadian Modern Language Review, 64*, 135–162.
Shih, C. M. (2010). The washback of the General English Proficiency Test on university policies: A Taiwan case study. *Language Assessment Quarterly, 7*(3), 234–254.
Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*(3), 323–340.
Wall, D. (1998). Impact and washback in language testing. In C. M. Clapham & D. Corson (Eds.), *Language testing and assessment. Encyclopedia of language and education, Vol. 7* (pp. 291–302). Dordrecht: Kluwer Academic.

Wall, D. (2005). *The impact of high-stakes testing on classroom teaching: A case study using insights from testing and innovation theory. In Studies in Language Testing* (Vol. 22). Cambridge: Cambridge University Press.
Watanabe, Y. (1996). Does grammar translation come from the entrance examination? *Preliminary findings from classroom-based research, Language Testing, 13*, 318–333.
Watanabe, Y. (2001). Does the university entrance examination motivate learners? A case study of learner interview. In A. Murakami (Ed.), *Trans-equator exchanges: A collection of academic papers in honor of Professor David E. Ingram* (pp. 100–110). Akita: Akita University.
Wu, J., & Lee, C. L. (2017). The relationships between test performance and students' perceptions of learning motivation, test value, and test anxiety in the context of the English benchmark requirement for graduation in Taiwan's universities. *Language Testing in Asia, 7*, 9.
Wu, R. W., & Chin, J. (2006). An impact study of the Intermediate level GEPT. In *Proceedings of the Ninth International Conference on English Language Testing in Asia, Taipei* (pp. 41–65).
Wu, R. Y. F., Wu, J., & Dunlea, J. (2015). *Glocalization in English language examinations: Aptis and GEPT comparison study*. Paper presented at 2015 LTTC International Conference, Taipei.

Publisher's Note