The Asian Journal of Applied Linguistics
Vol. 1 No. 2, 2014, pp. 178-91

AJAL

Washback of university-based English language tests on students' learning: A case study

Jinsong Fan
College of Foreign Languages and Literatures, Fudan University, China

Peiying Ji
College of Foreign Languages and Literatures, Fudan University, China

Xiaomei Song
Georgia Southern University, United States of America

Many universities in China have developed or are in the process of developing local English tests, claiming that such tests can bring about more beneficial washback on English teaching and learning. However, empirical evidence is scanty regarding the washback of these university-based English tests. This study investigated the washback of the Fudan English Test (FET) on students' English learning, and explored the roles of gender and English language ability in shaping students' reported washback. Research data were collected from 335 students through a questionnaire and semi-structured follow-up interviews. This study showed that though students considered the FET as giving them motivation or pressure in learning English, the test had limited washback on students' learning practices. Though students' perceptions of the test became significantly more positive with the increase of their English ability levels, their gender and English ability were found to play insignificant roles in shaping test washback. The findings of this study suggest that to foster positive washback, it is vital for university-based English test developers to not only focus on minimizing construct-irrelevant variance in test development and administration but also provide resources and support which are deemed essential to students' English learning and test preparation.

Keywords: English language testing; test washback; language learning; test preparation; Fudan English Test

Introduction

Though China boasts a long history of testing and examinations, modern testing theory and practice did not develop and thrive until the late 1970s, following the resumption of the National Matriculation Examination (known in China as *gaokao*) and the open-door policy. A large number of English language tests have been developed in China by educational and examinations authorities at different levels to meet the demand of the huge and ever-growing English learner population (Cheng & Curtis, 2010). Among the many English language tests developed and used in China, the College English Test (CET) has been recognized as a reliable and valid instrument to assess university students' English achievement and proficiency levels (Jin, 2010). In recent years, however, the CET has come under heavy criticism from language educators and researchers for its test format, lack of alignment between the CET and teaching curriculums developed within particular universities, and its negative washback effects on English teaching and learning (for example, Han, Dai, & Yang, 2004). Though many of these criticisms are politically motivated or emotionally charged rather than empirically grounded, some universities in China have developed or are attempting to develop their own English language tests in the hopes of addressing the deficiencies of the CET and exercising more beneficial washback effects on English teaching and learning (for example, TOPE Project Team, 2013; Tsinghua University Testing Team, 2012). Despite the claims from these universities that locally developed tests can bring about positive washback effects, currently there is almost no empirical evidence in support of these claims. Numerous studies have shown that language tests may not be effective agents in triggering pedagogical reforms or promoting the desired changes in students' learning practices (for example, Cheng, 2005; Cheng, Watanabe, & Curtis, 2004; Qi, 2005). Consequently, empirical investigations are much needed to ascertain whether positive washback of these university-based English tests indeed occurs in practice.

In the field of language testing, the number of empirical washback studies has been growing rapidly since Alderson and Wall (1993) posited the 15 washback hypotheses (for example, Andrews, Fullilove, & Wong, 2002; Cheng, 2005; Cheng et al., 2004; Ferman, 2004; Shohamy, Donitsa-Schmidt, & Ferman, 1996). These studies have indicated that washback is a highly complex rather than monolithic phenomenon. Also, these studies have demonstrated that the influence of language tests is on various aspects of learning and teaching, and the process of washback being generated is mediated by numerous factors (Cheng et al., 2004). However, a review of previous washback studies reveals that the vast majority of them have been focused on large-scale high-stakes language tests such as TOEFL, IELTS, and CET with few tapping into the washback of university-level language tests. Furthermore, though students are the most important stakeholders in any assessment situation, they tend to have been less investigated than other stakeholders in previous washback studies (Cheng, Andrews, & Yu, 2011). Compared with the many studies which examine teachers' perceptions and practices (for example, Alderson & Hamp-Lyons, 1996; Qi, 2005), investigations into the washback on learners and learning processes are only piecemeal. Of the 15 washback hypotheses proposed by Alderson and Wall (1993), eight were related to learners and learning, yet few of these have been empirically tested. Finally, few studies have investigated whether a language test exercises differential washback effects on different groups of test takers. Gender has long since been included as a variable in second language acquisition research (for example, Ellis, 1994), but its roles in shaping test washback are largely unknown. Although test takers' English ability has been included as a variable in some previous washback studies (for example, Cheng et al., 2011; Ferman, 2004), most of them are qualitative. Also, the criteria adopted to divide test takers into different ability groups are often vague (for example, students' self-rated English competence, see Cheng et al., 2011). Due to the close relationship between university-based English tests and English teaching and learning, as claimed by the test developers, it is particularly necessary to understand the roles of gender and English ability level in the washback mechanisms.

The current study was designed to investigate the washback of the Fudan English Test (hereafter the FET) on students' English learning, and to explore whether test takers' gender and English ability affect the test washback. This study, therefore, seeks to answer the following four research questions:

RQ1: What are students' perceptions and views of the FET?
RQ2: Are there any differences in students' perceptions and views of the FET that are related to their gender and English ability level?
RQ3: What is the washback of the FET on students' English learning practices?
RQ4: Are there any differences in the washback of the FET on students' learning practices that are related to their gender and English ability level?

The Fudan English Test

The target test of this study was the Fudan English Test, developed by the College English Center of Fudan University. According to the FET Test Syllabus (Fudan University Testing Team, 2014, forthcoming), the FET serves two purposes: 1) to accurately measure students' abilities and skills as reflected in the English teaching curriculum at Fudan University, and 2) to exercise more positive washback effect on English teaching and learning at Fudan University. The FET was initially launched in 2011, following a series of pilot studies and trials (Fan & Ji, 2013). By the end of 2013 over 6,000 students had taken the FET.

Designed on the basis of recent models of communicative language ability (for example, Bachman, 1990; Bachman & Palmer, 1996), the FET assesses students' English language abilities in the four modalities of listening, writing, reading, and speaking, each accounting for 25% of the test score (Fudan University Testing Team, 2014, forthcoming). Unlike the CET, spoken English is placed on a par in the FET with the other three language modalities. For practical and logistical reasons, the speaking subtest in the FET adopts the semi-direct method. In this type of English speaking test, computers are used to present tasks and to capture spoken responses which are later evaluated by human raters (Shohamy, 1994). In the FET speaking test, students are required to complete two tasks. In the first task, students are required to answer a question after listening to a short passage, and to give brief comment on a topic mentioned in the passage; in the second task, students are required to describe a picture before presenting their views and opinions about a social phenomenon shown in the picture. Another difference between the CET and the FET lies in the fact that the FET adopts a larger proportion of constructed-response items as well as some multi-modality tasks (for example, writing a summary based on an academic lecture that students have heard). The test results that students receive include a composite score and four profile scores on listening, writing, reading and speaking, all reported on a scale using five levels of A, B, C, D, and F. The design principles and score-reporting policy have explicitly demonstrated the intention of the FET developers to exercise more beneficial washback effects through test innovation. Currently, the FET Test Syllabus and practice test papers are available on the learning website of the College English Center[1], the provider of the FET. Though there is a wealth of learning resources on the website, no resources are found to be explicitly directed at the preparation for the FET. Taking into account the widespread phenomenon of teaching and learning for the test in China's language education (for example, Cheng & Curtis, 2010), it seems that the FET developers have tried to shift the trend from "assessment of learning" to "assessment for learning" (Cheng et al., 2011, p. 211).

Existing research shows that the FET is on the whole a reliable test, with internal consistency reliability coefficient reported at 0.83 (Fan & Ji, 2013). Confirmatory factor analyses suggest that there is a higher-order general language competence factor and four first-order factors representing listening, reading, writing, and speaking, lending support to the test's construct validity and score-reporting policy (Fan & Ji, 2013). A recent study (Fan & Ji, 2014) demonstrated that students generally held positive attitudes to the FET, particularly in terms of test administration and the multi-modality tasks. An important purpose of implementing the FET, according to the FET Test Syllabus, is to exercise more positive washback on English teaching and learning. No studies, however, are currently available about whether the FET test developers' intentions have indeed materialized in practice.

Method

Participants

The participants in this study were 335 students who took the FET in December, 2013. Of the 335 participants, 182 (54.3%) were female and 153 (45.7%) were male. At the time of investigation, all participants were Bachelor of Arts students studying with Fudan University with 96 (28.7%) majoring in humanities and social sciences, 140 (41.8%) in science and engineering, 46 (13.7%) in medical science, and 53 (15.8%) in business management. Based on students' test scores on the FET, students were divided into three ability groups: high-ability group (those scoring above the 75th percentile, n=63, 25.2%), intermediate-ability group (those scoring between 25th and 75th percentile, n=129, 51.6%), and low-ability group (those scoring below the 25th percentile, n=58, 23.2%). 85 scores were not retrievable and were therefore treated as missing data. After the questionnaire survey, 13 students participated in the follow-up interviews on a voluntary basis.

Instruments

Two instruments were developed for this study both of which were conducted in Chinese. They consisted of a structured questionnaire and an interview guide (see Appendices 1 and 2 respectively for English translations). The design of the questionnaire was based on a review of the conceptual frameworks of test washback on language learning (Alderson & Wall, 1993; Bailey, 1996) and the empirical washback studies on learning practices (Cheng et al., 2011; Ferman, 2004). In this study, learning practices were operationalized as three dimensions which were closely related to students' learning processes: learning attitudes, learning content, and test preparation strategies (for example, Bailey, 1996; Ellis, 1994). The questionnaire included a total of 29 items, all using a six-point Likert-scale. The 29 items were divided into two parts: Part 1: Students' perceptions of the FET (items 1-11), and Part 2: Washback of the FET on students' English learning (items 12-29). The questionnaire was piloted on 75 Year Two undergraduates at Fudan University. The pilot study demonstrated that the internal consistency reliability of the questionnaire was very satisfactory (Cronbach's α=0.94). The interview guide, which was developed on the basis of the questionnaire, also consisted of two parts with the same foci as in the questionnaire. .

Data collection

Due to the exploratory nature of this study as well as practical difficulties, convenience sampling, rather than strictly stratified sampling, was adopted. A total of 400 questionnaires were administered to the students with the aid of their English teachers in December, 2013. As a result, 335 questionnaires were returned and considered valid, achieving a response rate of 83.8%. Following the administration of the questionnaire, the interviews were conducted by one researcher and his assistant. With the participants' consent, every interview was audio-recorded. The recordings were later transcribed verbatim for analysis.

Data analysis

This study adopted the mixed-method sequential explanatory design which consisted of two consecutive phases, first quantitative, then qualitative. According to Ivankova, Creswell, and Stick (2006), priority is typically given to the quantitative approach in the sequential explanatory design "because the quantitative data collection comes first in the sequence and often represents the major aspect of the mixed-method data collection process" (p. 9). The smaller qualitative component follows in the second phase of the research.

In the first stage of quantitative analyses, exploratory factor analysis (EFA) was performed to investigate the construct structure of the questionnaire; Cronbach's Alphas at both the factor and scale levels were computed to investigate the reliability of the questionnaire. To address *RQ1* and *RQ3*, descriptive statistics at both factor and item levels were computed. To address *RQ2*, a 2×3 Analysis of Variance (ANOVA) was performed. To address *RQ4*, a 2×3 Multivariate Analysis of Variance (MANOVA) was performed. All quantitative analyses in this study were performed in IBM SPSS 21.0 (IBM, 2012), and the level of all tests of significance was set at 0.05.

Following quantitative analyses, the qualitative data in this study were analysed by means of analytic induction (Given, 2008). A coding scheme was developed on the basis of several preliminary readings and analyses to identify the salient and recurring themes and patterns in the data. Based on the coding scheme, the transcriptions of the interview data were coded by two researchers. Inter-coder reliability was confirmed by using the Cohen's kappa statistic (*k*=0.88).

Results[2]

Factor analysis and reliability estimates

Principal axis factoring with oblimin rotation was performed on the two parts of the questionnaire separately. Oblimin rotation was used to enhance the interpretability of the factor solutions because the two parts of the questionnaire were each basically measuring one overriding construct, and the dimensions under these two constructs should be correlated. Prior to the analysis, statistical assumptions for the EFAs were checked. Firstly, univariate normality was checked through examining the skewness and kurtosis values of the data at the item level (see Appendix 1), which were all within the acceptable range. Secondly, for the two parts of the questionnaire, the Kaiser-Mayer-Olkin (KMO) measure of sampling adequacy was 0.91 and 0.94 respectively. The Bartlett's spherical tests were both significant ($p<0.01$, two-tailed). These analyses demonstrated that the datasets were suitable for factor analysis. The Kaiser Criterion was adopted which meant only factors with eigenvalues over one would be extracted (Field, 2009). Factor loadings lower than 0.3 were deleted and were not counted toward any factors. Factor analysis revealed that for the first part of the questionnaire, only one factor was extracted, explaining 45.42% of the common variance. This factor was interpreted as students' perceptions of the FET (hereafter students' perceptions). For the second part of the questionnaire, three factors were extracted, explaining 63.04% of the common variance. Based on the item loadings, these three factors were interpreted respectively as washback on students' learning attitudes (hereafter learning attitudes), students' test preparation strategies (hereafter test preparation strategies), and washback on students' learning content (hereafter learning content).

Furthermore, internal consistency reliability estimates were computed at both the factor and scale levels. Results of the EFAs and reliability estimates are summarized in Table 1. As shown in this table, reliability estimates were consistently high across the four factors in this study (0.84-0.92). Cronbach's α for the whole questionnaire was 0.95 (n=29). These analyses demonstrated that the questionnaire was a reliable and valid instrument for the intended purpose.

Table 1. Summary of factor analysis and reliability estimates

Factor	Item Number	Cronbach's α
Students' perceptions of the FET	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	0.90
Learning attitudes	12, 13, 14, 15, 16, 17	0.92
Test preparation strategies	22, 24, 25, 26, 27, 28, 29	0.92
Learning content	18, 19, 20, 23	0.84

RQ1: Students' perceptions of the FET

Descriptive statistics at the factor level were first of all computed, demonstrating that students' perceptions of the FET were on the whole moderately positive (M=3.94, SD=0.84). Descriptive statistics of all items in this factor (see Appendix 1) indicated that students commented most positively on the materials used in the listening subtest (Item 9, M=4.25), and the design of the listening (Item 5, M=4.18), reading (Item 7, M=4.13), and writing subtests (Item 6, M=4.10), believing that the design reflected the language abilities that they needed in real life situations. Meanwhile, students tended to believe that the FET as a whole reflected their English language ability (Item 2, M=4.13). On the other hand, students commented least positively on the speaking subtest, raising doubts over its validity (Item 11, M=3.30) and believing that the subtest was not authentic (Item 8, M=3.87). Also, students seemed not satisfied with the level of difficulty of the test (Item 3, M=3.40).

The qualitative data showed that on the whole students spoke positively of the FET, particularly in terms of the overall design of the test, the listening and writing subtests, and the score-reporting policy. In terms of test design, students believed that the test was a comprehensive assessment of their English language abilities because "listening, speaking, reading, and writing were all assessed in the FET, and the design gave me a feeling that it was like another version of the TOEFL or IELTS developed in China" (Participant ZY). Echoing what we found from quantitative analyses, students believed that the materials used in the listening subtest were authentic because "most of them were probably clipped directly from English-speaking radio broadcasts or TV programs" (Participant BS). In particular, students commented very positively on the essay writing task which required students to write an essay based on an academic lecture they had heard because "we really need such abilities in academic studies" (Participant YL). Furthermore, students believed that the current score reporting policy could help them identify the strengths and weaknesses and take appropriate remedial actions in their English learning. Almost all students, however, complained about the difficulty of the test, believing that the test was too difficult for them. Also, many students commented rather negatively of the speaking subtest, believing that such a speaking test could not accurately assess their spoken English abilities. The reasons, as reported by students, were threefold. First, they didn't have the opportunity to familiarize themselves of the test format and procedures prior to the test. Second, they believed that the two tasks in the speaking subtest were not adequate proof of their spoken English abilities. Third, the testing environment was quite noisy. In addition to students' rather negative comments on the speaking subtest, they also hoped that the university could better promote the test within and beyond the university so that the test scores could be recognized as reliable proof of their English proficiency. As Participant WT remarked, "if the test enjoyed better recognition, undoubtedly I would have taken it more seriously and put more effort in the preparation for the test".

RQ2: Gender, English ability level, and students' perceptions

To explore whether gender and English ability level affected students' perceptions of the FET, a two-way ANOVA was performed with students' gender and English ability level as the independent variables and the factor score of students' perceptions as the dependent variable. Prior to the analysis, statistical assumptions necessary for the ANOVA procedure were checked. Two key assumptions to ANOVA are that the data in the samples are each normally distributed, and have variances that are equivalent to each other (Field, 2009). Based on skewness and kurtosis values, all samples were shown to be normally distributed. Levene's test demonstrated that the error variance of the dependent variable was equivalent across all samples under investigation ($F(5, 239)=0.33$, $p=0.89$). These analyses indicated that the dataset was suitable for the ANOVA procedure. The results of two-way ANOVA are presented in Table 2 which shows that no significant differences were found to exist between the perceptions of male and female students ($F(1, 239)=0.19$, $p=0.66$).The interaction effects between gender and English ability level on the dependent variable were also shown to be statistically not significant ($F(2, 239)=0.17$, $p=0.84$). However, significant differences were found between the perceptions of students from the three English ability groups ($F(2, 239)=11.18$, $p=0.00$). Post hoc tests demonstrated that with the increase of students' English ability level, their perceptions of the test also became significantly more positive ($p<0.05$). The factor score means of students' perceptions as a function of gender and English ability level are presented schematically in Figure 1.

RQ3: Washback on students' English learning

To portray a general picture of the washback of the FET on students' learning practices, descriptive statistics at the factor level were computed (Table 3). The mean values in this table indicated that the FET did not have much impact on students' learning across the three dimensions of learning attitudes, test preparation strategies, and learning content. Comparatively speaking, the FET had the most impact on test preparation strategies (M=3.82), and least impact on learning attitudes (M=3.67). In addition, descriptive statistics at the item level demonstrated that students seemed to attach more importance to their English learning because of the test (item 12, M=4.16), and they tended to spend more time on English learning (item 16, M=4.11). However, students tended not to agree that preparing and taking the FET had improved their confidence in

English learning (item 14, M=3.15), and they were engaged in more English speaking activities because of the test (item 23, M=3.19).

Table 2. Results of two-way ANOVA

Source	Sum of Squares	df	Mean Square	F-value	p-value
Corrected Model	17.57	5	3.51	4.69	0.00
Intercept	2.60	1	2.60	3.47	0.06
Gender	0.14	1	0.14	0.19	0.66
Ability	16.75	2	8.37	11.18	0.00
Gender × Ability	0.26	2	0.13	0.17	0.84
Error	179.05	239	0.75		
Total	200.08	245			
Corrected Total	196.61	244			

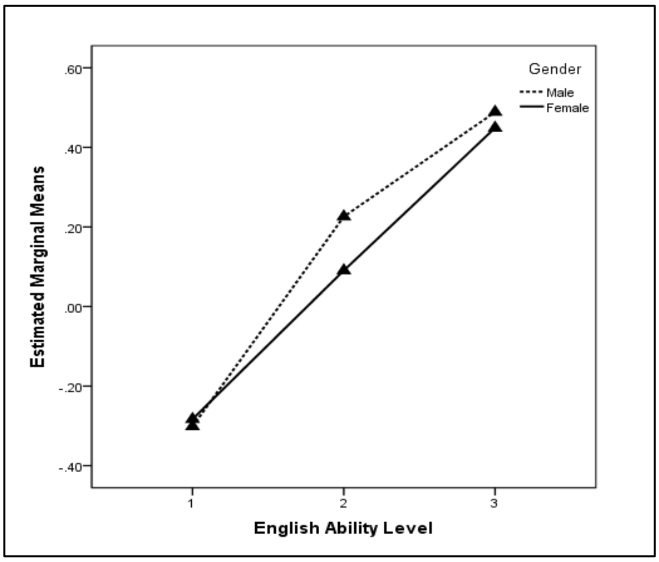

Figure 1. Factor score means of students' perceptions of the FET

Table 3. Descriptive statistics at the factor level

Factor	Mean	SD	Skewness	Kurtosis
Learning attitudes	3.67	1.08	-0.27	0.01
Test preparation strategies	3.82	1.00	-0.45	0.23
Learning content	3.73	1.08	-0.32	-0.12

In most cases, the qualitative data corroborated the findings derived from the quantitative analyses. For example, most students agreed that the FET had given them some external motivation or pressure to learn English, and they would spend more time on their English learning. A typical comment was: “The good thing about the FET lies in that it can motivate me to learn English, and I have really paid more attention to English learning. This is a difficult test, and we have to study harder in order to get a good score on the test” (Participant WT). However, most students also commented that the impact was limited in scope and intensity. Students mentioned that when the test was remote, they would normally just focus on their routine learning activities but when the test was drawing closer (for example, one month or three weeks before the test), they would spend more time preparing and studying for the test such as doing simulated exercise. Students agreed that getting a good score on the FET would certainly boost their confidence in learning English because “it demonstrates that my English is indeed really good” (Participant YR). Students did not think that the test had much impact on their learning motivation, and the reason was largely attributed to the fact that the FET was a newly developed test used only within the university. Contrary to what we had expected, the test did not motivate students to study the coaching materials, and the reason, as Participant ZY explained, was “currently, we couldn’t find any coaching materials in the market or on the website. I am at a loss about what to do even though I want to study the materials”. As far as test preparation strategies were concerned, some students mentioned that as the test was approaching, they would formulate and implement study plans and try to better manage their time. Participant BS, for example, remarked, “one or two months before the test, I would start to formulate study plans in order to be more efficient in preparing for the test. I would set aside time specifically for vocabulary, reading, and listening”. Some students also reported that an important strategy they deployed was preparing the CET or TOEFL together with the FET because “preparing one test means you are also preparing another” (Participant MJ).

RQ4: Gender, English ability, and test washback

To determine whether gender and English ability level affected the washback on students’ learning practices, a two-way MANOVA was adopted with students’ gender and English ability level as the independent variables and the factor scores of learning attitudes, test preparation strategies, and learning content as the dependent variables. The advantage of MANOVA lies in that it takes into account correlations among dependent variables. However, the power of MANOVA decreases as the correlation between dependent variables increases. MANOVA works acceptably well with moderately correlated dependent variables in both directions (Field, 2009). In this analysis, Pearson’s product moment correlation coefficients were computed to investigate the correlations between the three dependent variables. The results demonstrated that the three factor scores were moderately correlated (r from -0.64 to 0.72, p<0.01). Therefore, one of the basic reasons for using MANOVA was justified. However, Box’s Test of Equality of Covariance Matrices showed that the assumption of homogeneity of covariances across groups was violated (Box’s M=70.69, F=2.27, $df1$=30, $df2$=57120.53, p=0.00). Therefore, Pillai’s trace, the multivariate statistic which is most robust to violations of statistical assumptions, was adopted in this study. The results of two-way MANOVA are summarized in Table 4. The multivariate tests of significance indicated that the main effects of gender and English ability level were statistically not significant (gender: Pillai’s trace=0.00, F(3, 236)=0.15, p=0.93; English ability level: Pillai’s trace=0.05, F(6, 474)=1.88, *p*=0.08). Also, the analysis indicated that the interaction effects between gender and English ability level were statistically not significant, either (Pillai’s trace=0.01, F(6, 474)=0.38, *p*=0.89). The MANOVA results indicated that gender and English ability level did not affect the FET washback on the three dimensions of students’ learning practices, including learning attitudes, test preparation strategies, and learning content.

Table 4. Results of two-way MANOVA

IV	Pillai’s trace	F-value	Hypo df	Error df	p-value	η2
Gender	0.00	0.15	3	236	0.93	0.00
Ability	0.05	1.88	6	474	0.08	0.02
Gender×Ability	0.01	0.38	6	474	0.89	0.00

Discussion and conclusions

In the Chinese examination culture, the power of testing has often been described figuratively as “the assessment tail wagging the educational dog” (Biggs, 1992, p. 11). In this study, we investigated the washback of the FET and explored the factors which might have contributed to positive washback or prevented positive washback from occurring in students’ learning practices. Since students’ views of a test may affect their learning and test preparation process (Hughes, 1994), we first investigated how students perceived the FET before exploring the washback of the FET on students’ learning. We also investigated the roles of gender and English ability level in shaping the FET washback.

The findings of this study suggest that most of the test developer’s intentions to engineer positive washback through testing innovation were endorsed by students. For example, students commented positively on the inclusion of all four language modalities in the assessment of their English abilities, the multi-modality tasks, and the score-reporting policy. Students’ positive views in these regards, as suggested by Hughes (1994), are likely to bring about beneficial washback on their learning and test preparation. On the other hand, students commented rather negatively on the speaking subtest, raising doubts over its authenticity, testing environment, and face validity. These findings concur closely with a recent investigation into test takers’ attitudes to the FET which identified similar factors contributing to test takers’ negative attitudes towards the speaking subtest (Fan & Ji, 2014). Messick (1989) related test washback to test validity, arguing that to foster positive washback, one should concentrate on minimizing construct underrepresentation and construct-irrelevant variance. The issues reported by students (for example, noisy testing environment and unfamiliarity with test format) were very likely to give rise to measurement error, resulting in construct-irrelevant variance. This might have in turn prevented positive washback from occurring in students’ learning. Therefore, to pursue positive washback, it is vital for the FET provider to apply more rigorous quality control procedures to test development and administration so that construct-irrelevant variance can be minimized.

In addition to the findings about students’ perceptions, this study also revealed that the FET, as intended by the test provider, had given students some motivation and pressure in learning English. Taking into account the examination culture in China, it is not surprising that students consider a newly developed English test as the source of motivation and pressure in English learning (Cheng, 2008; Cheng & Curtis, 2010). Also, consistent with many previous washback studies (for example, Cheng et al., 2004; Qi, 2005), this study indicated that as the test drew closer, test washback became more overt and intense. However, this study also demonstrated that the FET washback on students' learning was limited in intensity and scope. Two reasons may explain the limited washback of the FET. The first is related to the recognition of the test. Unlike the big tests such as the CET whose results enjoy extensive recognition from society, the FET currently does not enjoy any recognition from outside the university. In consequence, students may not be motivated to achieve a high score on the test since they believe that their performance is not recognized as adequate proof of their English language ability. The second reason is the lack of study materials and support, as suggested by the qualitative data. Hughes (1994) rightfully pointed out that the availability of resources is vital for positive washback to occur. Though a well-designed test is much more likely to bring about positive washback than a test featuring poor design, positive washback seldom, if ever, occurs simply because of the introduction of a test (for example, Cheng et al., 2004; Qi, 2005; Shohamy et al., 1996). Very often in language education, much attention is focused on the test itself, but woefully insufficient attention has been paid to the resources and support which are vital for the intended washback to occur.

In this study, gender and English ability level were included as two variables in our analyses. Despite the increasing concerns over possible gender effects in language assessment (McNamara & Roever, 2006), the role of gender in shaping test washback on learning is largely unknown. English ability level was included as a variable because whether a university-based test has the same amount and direction of washback on students at different ability levels is of particular interest to the test provider and policymakers. Results of this study indicate that gender and English ability level did not affect students' reported washback on their learning practices. This finding does not resonate with some previous washback studies (for example, Cheng et al., 2011; Ferman, 2004) which demonstrated that higher- and lower-ability students deployed different learning strategies in test preparation. The discrepancy may be related to the nature of the FET. As a university-based English test, the lack of the recognition as well as the lack of resources and support, as discussed above, might have prevented the intended positive washback from occurring in practice. Therefore, the test had limited washback on students' learning practices, irrespective of their gender and English ability levels. Nonetheless, this study found that as students' English ability level increased, their perceptions of the test became correspondingly more positive. The finding is not surprising since students with higher ability levels are normally more motivated to learn English and also tend to view the test more positively (for example, Cheng et al., 2004).

A few limitations need to be addressed to support the accurate interpretations of the research findings. First of all, all quantitative and qualitative data collected in this study were self-reported by students. Therefore, the research findings should be interpreted and accepted with caution. Secondly, though EFA extracted three factors representing students' learning and qualitative interview data provided additional evidence relating to their learning practices, we did not incorporate such data as students' study journals or diaries in our investigation which could portray a fine-grained picture of their learning processes. Given the priority of the quantitative approach in the sequential mixed-method design, the qualitative component was given less weight in this current research. To enhance the depth of qualitative analysis, a multiple perspective study involving different test stakeholders (for example, students, teachers, and test developers) might be conducted in the future. Thirdly, a stratified sampling method was not adopted in this study. It is likely that the voluntary interview participants were from the intermediate- or high-ability groups, and were therefore more active and motivated in learning English. A more rigorous and systematic sampling approach can therefore be adopted in future studies to further investigate the washback of the FET.

In light of the findings of this study, we believe that to engineer positive washback, it is important for language test designers and educators to heed the advice of Messick (1989), Hughes (1994), and Bailey (1996) to not only focus on test design but pay more attention to reducing construct-irrelevant variance and providing resources and support which are vital to students' English learning and test preparation. We believe the findings of this study warrant attention from other universities which have developed or are currently in the process of developing local English tests with a view to engineering positive washback. The intended washback effects may fail to occur if factors other than the test itself (for example, resources, support, and information) are not taken into full consideration at the very beginning of a local testing project.

Acknowledgements

The preparation of this manuscript was sponsored by the National Social Science Fund, or *Guojia Sheke Jijin*, under the project title *The Development and Validation of Standards in Language Testing* (Grant Number: 13CYY032). An earlier draft of this article was presented at the Second National Language Testing Symposium at Changzhou University in Changzhou, China. We would like to express our appreciation to Miss Wang Qian, an MA student studying at Fudan University, for her assistance in data collection and input. Also, we would like to thank Dr David Gardner, the editor of AJAL, and the two anonymous reviewers for their insightful feedback on the previous draft of this article.

Notes

1. The FET Test Syllabus will be published soon in 2014. The test syllabus and practice test papers are available at: http://elearning.fudan.edu.cn (in Chinese).
2. All participant quotations in this paper are taken from interview transcripts. Pseudonyms are used to ensure anonymity.

About the authors

Jinsong Fan is a lecturer at the College of Foreign Languages and Literatures and Associate Director of the Language Testing Center, Fudan University, China. His research interests are language testing and assessment.

Peiying Ji is a professor of Applied Linguistics and Vice Dean of the College of Foreign Languages and Literatures, Fudan University, China. Her research interests are in second language acquisition, foreign language teaching and learning, language testing, and EFL materials development.

Xiaomei Song is a senior research associate and instructor at Georgia Southern University in the United States of America. Her research interests are language testing and assessment.

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280-297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

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

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

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

Biggs, J. (1992). The psychology of educational assessment and the Hong Kong scene. *Bulletin of the Hong Kong Psychological Society, 28/29*(4), 5-26.

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

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.

Cheng, L., Andrews, S., & 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.

Cheng, L., & Curtis, A. (2010). *English language assessment and the Chinese learner*. New York and London: Routledge, Taylor and Francis Group.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing*. London: Lawrence Erlbaum.

Ellis, R. (1994). *Understanding second language acquisition*. Oxford: Oxford University Press.

Fan, J., & Ji, P. (2013). Examining the validity of the Fudan English Test: Test data analysis. *Foreign Language Testing and Teaching, 2*, 45-53.

Fan, J., & Ji, P. (2014). Test candidates' attitudes and their test performance: The case of the Fudan English Test. *University of Sydney Papers in TESOL, 9*, 1-35.

Ferman, I. (2004). The washback of an EFL national oral matriculation test on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *In Washback in language testing: Research methods and contexts* (pp. 190-210). London: Lawrence Erlbaum.

Field, A. (2009). *Discovering statistics using SPSS (3rd ed.)*. London: SAGE publications.

Fudan University Testing Team. (2014, forthcoming). *The FET Test syllabus*. Shanghai: Fudan University Press.

Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods (Volume 1 & 2)*. London: Sage Publications.

Han, B., Dai, M., & Yang, L. (2004). Problems with College English Test as emerged from a survey. *Foreign languages and their teaching, 179*(2), 17-23.

Hughes, A. (1994). *Backwash and TOEFL 2000*. Unpublished manuscript, University of Reading. Educational Testing Service (ETS).

IBM. (2012). *IBM SPSS statistics 21 core system user's guide.* New York: IMB Corp.

Ivankova, N., Creswell, J., & Stick, S. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods, 18*(1), 3-20.

Jin, Y. (2010). The National College English Testing Committee. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 44-59). London: Routledge, Taylor & Francis Group.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99-124.

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

TOPE Project Team. (2013). *Syllabus for Test of Oral Proficiency in English (TOPE)*. Beijing: China Renming University Press.

Tsinghua University Testing Team. (2012). *Syllabus for Tsinghua English Proficiency Test (TEPT).* Beijing: Tsinghua University Press.

Appendix 1: Questionnaire Items and Descriptive Statistics

Item		Mean	SD	Skewness	Kurtosis
1)	Test design is on the whole satisfactory.	4.10	1.12	-0.74	0.51
2)	Test results truly reflect my English proficiency.	4.13	1.12	-0.84	0.78
3)	The level of difficulty is appropriate.	3.40	1.26	-0.28	-0.62
4)	Objective and subjective items are well balanced.	3.97	1.10	-0.61	0.38
5)	Language abilities tested in listening are what I need.	4.18	1.25	-0.56	-0.18
6)	Language abilities tested in writing are what I need.	4.10	1.16	-0.55	-0.01
7)	Language abilities tested in reading are what I need	4.13	1.11	-0.59	0.55
8)	Language abilities tested in speaking are what I need.	3.87	1.27	-0.38	-0.41
9)	Materials used in listening are authentic.	4.25	1.14	-0.67	0.34
10)	Writing tasks reflect language use in real life.	3.94	1.23	-0.47	-0.13
11)	Speaking test can reflect my oral English ability.	3.30	1.38	-0.06	-0.81
12)	I have paid more attention to English by taking the FET.	4.16	1.25	-0.58	-0.04
13)	Taking the FET has improved my interest in English.	3.26	1.30	0.10	-0.49
14)	Taking the FET has improved my confidence in English.	3.15	1.33	0.11	-0.54
15)	By taking the FET, I become more active in English.	3.79	1.27	-0.36	-0.32
16)	I have spent more time on English by taking the FET.	4.11	1.32	-0.55	-0.21
17)	Taking the FET has improved my learning efficiency.	3.54	1.23	-0.07	-0.27
18)	Because of the FET, I listened to more English programs.	3.85	1.38	-0.31	-0.72
19)	Because of the FET, I watched more English videos.	4.05	1.36	-0.42	-0.55
20)	Because of the FET, I read more English newspapers.	3.83	1.28	-0.23	-0.41
23)	Because of the FET, I attended more speaking activities.	3.19	1.21	0.12	-0.31
22)	FET has motivated me to study coaching materials.	3.60	1.33	-0.23	-0.51
24)	To prepare the FET, I have set English learning goals.	3.86	1.19	-0.39	-0.14
25)	To prepare the FET, I have made study plans.	3.90	1.20	-0.42	-0.20
26)	To prepare the FET, I have often reviewed lessons.	3.71	1.21	-0.15	-0.37
27)	To prepare the FET, I have managed my time better.	3.80	1.19	-0.36	-0.23
28)	I have tried to find a good learning method for the FET.	3.93	1.20	-0.41	-0.21
29)	I have tried to change my study habits for the FET.	3.91	1.25	-0.29	-0.40

Note. 1) Due to space limitations, items reported in tables are not exactly the same as they appeared in the original questionnaire; 2)The factor of students' perceptions of the FET is represented by Item 1-11 in this table; the factor of impact on learning attitudes is by Item 12-17; the factor of impact on test preparation strategies by Item 22, and 24-29; the factor of impact on learning content by Item 18, 19, 20, 23; 3) Item 21 was deleted from further analyses because it had double loadings on two factors and each loading was over 0.3.

Appendix 2: Interview Guide

Part 1: Students' perceptions of the FET

Q1. Please comment on the FET in terms of test design, administration, and score reporting, etc.

Q2. Please comment on the four components of the FET, i.e. listening, writing, reading, and speaking.

Part 2: Washback of the FET on students' English learning

Q3 Please comment on the impact of the FET, if any, on your English learning practices.

Q4. Please comment in more detail on the impact of the FET, if any, on your English learning in terms of learning attitudes, test preparation strategies, and learning content.