THE JOURNAL OF ASIA TEFL
Vol. 17, No. 4, Winter 2020, 1214-1235
http://dx.doi.org/10.18823/asiatefl.2020.17.4.4.1214

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Washback of the College English Test – Band Four on English Teaching and Learning in China

Linlin Wang
Temple University, United States

Elvis Wagner
Temple University, United States

This study examined the possible washback effects of the latest version of the College English Test–Band Four (CET-4) test from the test-takers' perspective, focusing primarily on the listening section. Using survey methodology, 305 Chinese EFL learners from both a first- and a second-tier city in China rated the difficulty, usefulness, and level of authenticity of the CET-4 listening, the non-listening test sections, the degree to which they adapted their learning behaviors to the test, and how much their teachers adapted their instruction to address the CET-4. The results indicated learners from the second-tier Chinese city considered the CET-4 more difficult, authentic, and useful, and that it had more of an impact on teaching and learning. Despite these differences, both groups rated the listening section as more difficult, less useful, and less authentic than the other sections of the test. Finally, the listening section was perceived to have more of an impact on student's learning than the perceived impact on teaching. Based on these results, we argue that the possible washback effects of a test should be a fundamental consideration in the development and validation of a test, not just an afterthought (Bachman, 2005; Chalhoub-Deville & O'Sullivan, 2020; Messick, 1996).

Keywords: test washback, CET-4, L2 listening assessment

Introduction

Large-scale, standardized tests of L2 English proficiency are becoming more and more high-stakes. The results of these tests have a huge impact on the educational opportunities and careers of millions of test-takers every year (Andrews, 1995; Cheng, 2008; Gu, 2005). However, the impact of these language tests extends far beyond the individual test takers, as these high-stakes tests can impact teachers, materials developers, curriculum designers, and even entire educational systems (e.g., Alderson & Wall, 1993; Hughes, 1993). Messick (1996) considered test washback a critical part of construct validity, in that washback was "… the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning" (p. 241). Understanding the impact of a particular test on specific learning and teaching behaviors is therefore of great importance.

The College English Test–Band Four (CET-4) is one of the most influential English tests in China, as it is a required test for graduation from most Chinese universities. For literally millions of Chinese students, failing this test would mean not graduating from university. Analyzing the washback of this test could be helpful in understanding (and hopefully improving) current college English teaching and learning in

China. There are surprisingly few studies that have focused on the washback of the CET-4, and a number of them were written and published in Chinese, leading to difficulty for non-Chinese researchers that hope to build on this research. In addition, the CET-4 underwent a major revision in 2016, and very little research has been conducted examining the new version's washback effects, and virtually no washback research has closely examined its listening component. Furthermore, no research has focused on the potential differential effects of washback on different learning contexts. Therefore, the current study aims to address these research gaps and investigate the possible washback effects of the CET-4, especially the listening section, on English teaching and learning in China from the students' perspectives. The study will be used as the basis for subsequent studies that focus on other CET-4 stakeholders, including teachers, curriculum developers, materials developers, and even the CET-4 test developers.

Conceptualizing Test Washback

A number of researchers have conceptualized test washback from multiple perspectives. In this section we briefly review five influential interpretations of test washback (i.e., Alderson & Wall, 1993; Bailey, 1996; Hughes, 1993; Watanabe, 1997, 2004), and propose a model of test washback based on a synthesis of these interpretations.

Hughes (1993) proposed a washback trichotomy that includes how participants, process, and production can be affected by a language test:

> The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work. (p. 2)

This trichotomy is useful because it outlines the mechanism of test washback, highlighting elements that may be affected by a test.

Pointing out that the existing literature had not unpacked the complexity of test washback, Alderson and Wall (1993) specified 15 Washback Hypotheses about possible testing impact:

1) A test will influence teaching.
2) A test will influence learning.
3) A test will influence what teachers teach.
4) A test will influence how teachers teach.
5) A test will influence what learners learn.
6) A test will influence how learners learn.
7) A test will influence the rate and sequence of teaching.
8) A test will influence the rate and sequence of learning.
9) A test will influence the degree and depth of teaching.
10) A test will influence the degree and depth of learning.
11) A test will influence attitudes to the content, method, etc., of teaching and learning.
12) Tests that have important consequences will have washback; and conversely
13) Tests that do not have important consequences will have no washback.
14) Tests will have washback on all learners and teachers.
15) Tests will have washback effects for some learners and some teachers, but not for others. (pp. 120-21)

Alderson and Wall proposed these hypotheses to further delineate the possible impact a test may have. They suggested that a variety of factors need to be taken into consideration when examining possible test washback, and they called for empirical research to substantiate washback.

Bailey (1996) reviewed these two conceptualizations and argued that the idea of impact can be categorized into two types: washback to the learner (including Alderson and Wall's Hypotheses 2, 5, 6, 8, and 10), and washback to the program (including Alderson and Wall's Hypotheses 1, 3, 4, 7, 9, and 11).

Watanabe (1997) further classified the dimensions of washback into five continua: 1) specificity (whether the washback can be generated by all tests or specific tests and specific aspects of tests); 2) intensity (whether the test has a strong or weak impact); 3) length (whether the washback lasts for a long or short period of time); 4) intentionality (whether the washback is intended or unintended); and 5) value (whether the washback is positive or negative). Watanabe (2004) also summarized how five factors can mediate how washback could work:

> … test factors (e.g., test methods, test contents, skills tested, purpose of the test, decisions that will be made on the basis of test results, etc.); prestige factors (e.g., stakes of the test, status of the test within the entire educational system, etc.); personal factors (e.g., teachers' educational backgrounds, their beliefs about the best methods of teaching/learning, etc.); micro-context factors (e.g., the school setting in which the test preparation is being carried out); and macro-context factors, that is, the society where the test is used. (p. 25)

Based on a synthesis of the five washback interpretations described above, we have conceptualized a model of test washback that incorporates the different components of these models. In our model (shown in Figure 1), a test may exert an influence on learners and program in a broad sense. Multiple agents or participants are involved including students, teachers, material and curriculum developers, and researchers. Their learning, teaching, material and curriculum design, and research processes may be affected at various levels, which may lead to different learning, teaching, material and curriculum design, and research results respectively. In addition, these impacts can vary according to the different contexts of the test user. A test's influence on learners and programs may be mediated by factors such as test-related, prestige-related, personal, micro-context, and macro-context ones. As a result, test washback may vary regarding its specificity, intensity, length, intentionality, and value (Alderson & Wall, 1993; Bailey, 1996; Hughes, 1993; Watanabe, 1997, 2004).

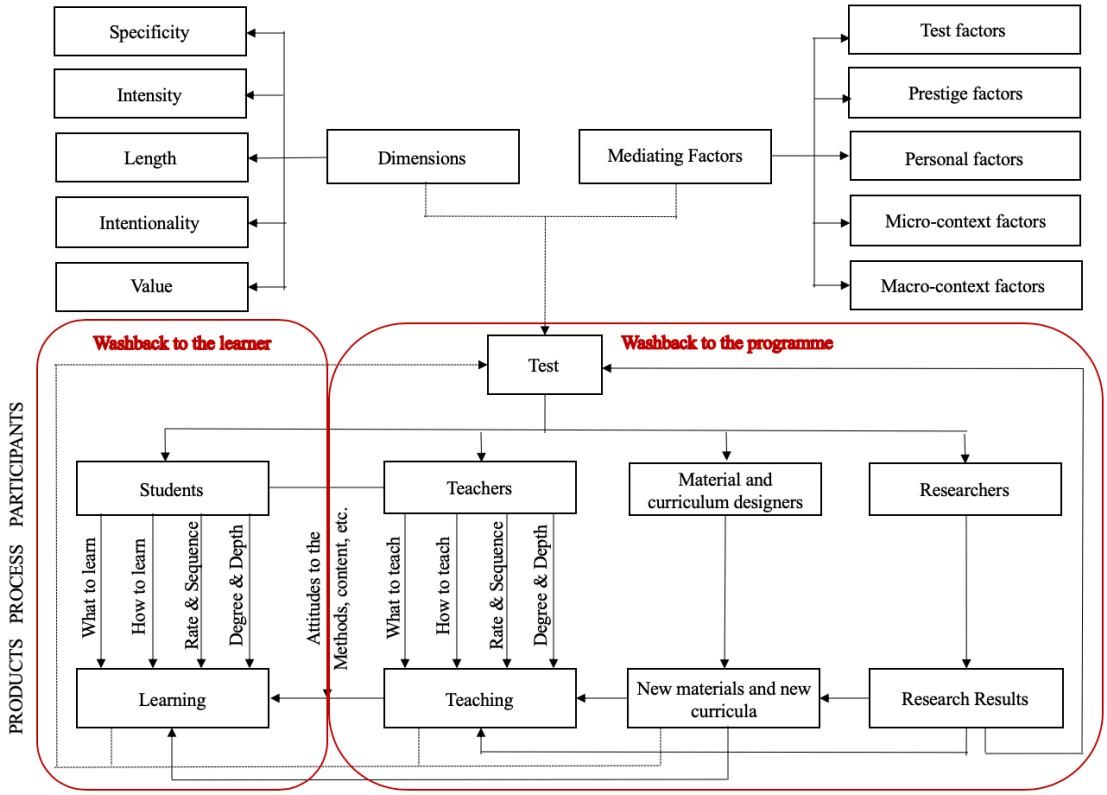

Figure 1. Conceptualizing test washback (adapted from Alderson & Wall, 1993; Bailey, 1996; Hughes, 1993; Watanabe, 1997, 2004).

Literature Review

As shown in Figure 1, test washback is a complex notion. One of the biggest challenges that researchers face in investigating washback is to choose a starting point of the research. Previous studies have had a myriad of foci understanding certain parts of test washback. Synthesizing them reveals four threads that may shed light on the key considerations of a research design on test washback.

First, researchers have chosen specific contexts all over the world to examine test washback effects, such as the United States, Japan, Australia, China, and South Korea (e.g., Allen, 2016; Cheng, 1997; Choi, 2008). Second, different test stakeholders have been examined, including students and learning (e.g., Gosa, 2004; Hawkey, 2006; Purpura, 2009), teachers and teaching, (e.g., Han, Dai, & Yang, 2004), and institutions (e.g., Mickan, P., & Motteram, 2009; Rozzi, Pinto, González, & Crimi, 2014; Zou & Xu, 2017). Third, high-stakes tests such as TOEFL and IELTS (e.g., Bailey, 1999; Rea-Dickins, Kiely, & Yu, 2007; Yu et al., 2017) have been examined in different washback studies. Lastly, different research methodologies have been utilized, including ethnography, participant and nonparticipant observation, narrative inquiries, and other research techniques such as interviews and questionnaires (Watanabe, 2004).

The current study focuses on English tests in China, mainly because Chinese students make up the largest population of English learners in the world. This population has been the focus of a number of washback studies (e.g., Cheng, 2001; Cheng & Falvey, 2000; Crusan & Cornett, 2002; Li, 2002; Li, Zhong, & Suen, 2012; Qi, 2006, 2007; Yan, Gu, & Khalifa, 2014). Among the tests that are widely used in the Chinese EFL context, the College English Test Band Four (CET-4) is one of the most influential English tests, as it is a required test for graduation for most Chinese universities. The CET-4 is designed as an English proficiency test that is composed of four sections (writing, listening, reading, and

translation), and takes about 125 minutes to complete. The test is administered twice a year throughout China. Millions of undergraduate students in China take the CET-4 every year, and numerous institutions utilize its score as a measure of students' English abilities. Table 1 shows the format of the test.

TABLE 1
CET-4 Test Format

Item	Form	Number of questions	Proportion of scoring (%)	Time (minutes)
Writing	Writing	1	15	30
Listening	Short news	7	35	25
	Long dialogues	8		
	Passages	10		
Reading Comprehension	Vocabulary: cloze	10	35	40
	Long passages: matching	10		
	Detailed reading: multiple-choice	10		
Translation	Chinese to English passage	1	15	30
Total		57	100	125

Even though it is such a high-stakes test, there are surprisingly few empirical studies that have focused on the washback of the CET. Some studies have examined the test washback effect on teachers and their teaching. For instance, Han et al. (2004) surveyed 1,194 English teachers in 40 colleges and universities about their attitudes toward the CET tests. Even though the teachers reported that their teaching was largely evaluated by their students' CET test scores, only 40% of the teachers reported that their teaching was greatly influenced by the tests. While about 70% of the teachers reported that they hoped that the CET would continue to be administered due to the complexity in designing a new test system if the CET were to be abandoned, over 70% of the teachers expressed they did not believe the CET actually succeeded in improving English teaching and learning in China. About 25% of the teachers cautioned that rather than measuring students' English abilities, the CET tests are susceptible to test strategy use and the test-wiseness of the test-takers. Around 40% of the teachers pointed out the inauthentic nature of the tests, and around 30% of the teachers believed the scores were not accurate indicators of students' English abilities.

Zhang (2003) also investigated the washback effect of the CET on teaching, and reported a stronger washback effect on how teachers chose teaching materials, while there was little influence on their actual teaching methods. Shao (2006) provided additional insights that the CET-4 did not exert a great impact on teaching until the fourth semester when students were actually required to take the test. The latter half of the fourth semester was specifically allocated to hold CET-4 preparation classes, in which teachers' teaching strategies, materials, and activities were all adapted.

With the same focus of the CET's impact on teachers, Gu (2003) observed and video recorded three college teachers' classrooms. The data showed that the greatest impact of the CET on teaching was manifested in the last month of the fourth semester with an increased teaching pace, similar to Shao's (2006) conclusion, and a use of more complex materials. Gu (2003) also suggested that the English classes were mostly characterized as material- and teacher-led cramming, and the CET overall did not lead to many changes on classroom instruction. Instead, Gu found that a teacher's training and personal beliefs were much more influential than the impact of the CET on the way the teachers taught in their classrooms.

In a subsequent CET washback study, Gu (2007) observed 38 college English teachers' classes in three Chinese colleges, surveyed 100 English teachers and 1,200 students, and interviewed several school administrators, teachers, and students. Similar to Zhao (2003), Gu's (2007) findings revealed that the CET had a greater impact on the teaching content, pace, and attitude than the teaching method. Gu identified both positive and negative washback. Positive washback included meeting the requirements in the national College English Teaching Syllabus, increasing schools' attention to English teaching and

learning, motivating teachers and students, and enhancing students' reading abilities. Negative washback was mainly associated with the fast pace observed in the second year. The teachers had to rush through the materials required by the school curriculum to spare time for CET-related instruction. The study further suggested that the washback effects varied according to schools and grade levels. Contradicting her previous findings, Gu (2007) identified greater impacts in the first two semesters compared to the third and fourth ones. Teachers' traits and whether they could creatively utilize learning materials contributed to better CET performance. Based on Gu's (2003, 2007) studies, Yang, Gu, and Liu (2013) revisited 13 teachers in the previous studies and an additional 16 teachers, and completed a follow-up study regarding the newer version of the CET at that time. The test was again found to have had a greater impact on the teaching content than method. Specifically, because the listening test was allocated with more weight in this version, the teachers increased their instruction time on listening. Nevertheless, the teacher-dominant mode stayed the same.

With a focus on both teachers and students, Huang and Yang (2002) aimed to investigate the washback effect of the CET-4 on English teaching and learning in order to further improve the test (cited in Li, 2009 and Li et al., 2012). English instructors and undergraduate students from 11 universities in Chongqing participated in interviews, a survey, and observations. The results indicated that the CET-4 mainly influenced materials selection, instruction methods, and students' learning goals and strategies. The majority of the participants suggested more positive than negative washback. Interestingly, compared to the washback effect of the CET on teaching and learning in key and non-key universities, a greater impact was found in the non-key universities because the teachers and students felt more challenged. Additionally, although the participants held a positive attitude toward new adaptations in the latest reform (e.g., adding a speaking section), they also reported doing little to actually prepare for these changes.

Li (2009) investigated whether the CET writing section led to "teaching to the test". Employing both questionnaires and interviews, Li surveyed 20 college English teachers and 128 students. Li found that the CET exerted a stronger influence on students than on teachers. Teachers did not report much teaching to the test, perhaps due to the easy nature of the requirement and the fact that the writing component of the test is weighted relatively lowly. Similar to what Gu (2003) and Yang et al. (2013) found, teacher-related factors such as the training the teachers received outweighed the impact of the test itself.

Focusing on learners, Zhao (2006) used mixed methods to examine Chinese university students' attitudes toward the CET-4, as well as the relationship between their attitudes and their test performance. Qualitative data reflected students' high motivation yet self-doubt about their performance. The results of a stepwise regression analysis revealed that participants' affective factors best predicted their test performance. Specifically, test-taking motivation and test-taking anxiety/lack of concentration accounted for about 12.4% of the variance in their test performance, and belief in the CET-4 (as to how well they believed the CET-4 score indicated their real English ability) accounted for an additional 3.0%. Test-taking anxiety/lack of concentration also differentiated female and male students, as female students tended to get higher scores if they experienced lower anxiety, but the anxiety level did not significantly account for male students' test performance. In addition, three variables (test-taking anxiety/lack of concentration, test-taking motivation, and belief in CET-4) differentiated high-achieving students from low-achieving students.

Li et al. (2012) took a slightly different approach and solicited 150 students to complete a questionnaire before they took the CET-4. The questionnaire explored students' perceptions of the test's impact on their learning and affection. A greater influence on the learning content was found compared to the learning method. A majority of the participants experiencing increased motivation, investment, and self-efficacy, but many also reported greater pressure and anxiety.

One study that has focused specifically on the listening section of the CET-4 is Yan (2016). Yan surveyed 120 undergraduate students who took part in the 2014 version of the CET-4 and CET-6. She also interviewed several school administers, teachers, and students. She found some evidence suggesting that students started to pay more attention to English listening and their confidence was boosted due to the impact of the CET-4. She also found that schools devoted more time and attention to listening teaching,

and most teachers and students reported that they were not satisfied with the materials in use. A majority of the students reported that the impact of the CET only lasted for a short time. However, the study did not report the specific instruments used, and there was no data analysis section, which make the results less interpretable.

These studies provide a wealth of information about the washback effect of the CET, and informed the current study, especially regarding research design and methodology. However, most of the studies reviewed above have focused on teachers and their teaching, while there has been much less research focusing on how test washback affects students and learning (Hamp-Lyons, 1997; Li et al., 2012). The few CET washback studies that focused on learners are also somewhat outdated because the CET-4 underwent a major revision in 2016. Since the test content and format have changed extensively, it is likely the impact of the new test differs from the impact of the old test. Additionally, the washback of the CET-4 listening section has only been superficially examined. To conclude, reliable research is scarce pertaining to the washback of the CET-4 listening after its 2016 reform, especially research that focuses on its impact on students.

Moreover, a test's washback may vary according to the social context in which the test is utilized (e.g., Alderson & Hamp-Lyons, 1996; Brown, 1997; Shohamy et al., 1996; Wall, 1997; Watanabe, 2004). The very impacts a test has on test users in one social context may not be found on another group of test users. Therefore, it is necessary to examine the CET-4 washback in different contexts and populations in China. The Chinese City Tier System is widely used to indicate the cities' overall economic development, which is closely associated with the cities' educational resources (Zhang & Li, 2014). The washback effects of the CET-4 may differ for students from first- or second-tier cities. However, no research has focused on comparing and contrasting the CET-4 washback in different-tier cities.

Current Study

Research Questions

The following research questions are addressed in the study to examine students' perceptions of the CET-4 and the specific washback effect of its listening section:

1a. How do test-takers perceive the difficulty, usefulness, and authenticity of the CET-4 listening? Do the test-takers' perceptions differ if they are from first- or second-tier Chinese cities?
1b. How do test-takers perceive the difficulty, usefulness, and authenticity of the other sections (non-listening) of the CET-4? Do the test-takers' perceptions differ if they are from first- or second-tier Chinese cities?
1c. Do the test-takers in each city perceive the difficulty, usefulness, and authenticity of the CET-4 listening differently from the other sections of the test?
2. What impact does the CET-4 listening have on test-takers' English learning behaviors and on students' perceptions of how L2 listening is taught? Do the test-takers' perceptions differ if they are from first- or second-tier Chinese cities?

Methods

This study uses survey methodology to examine the potential washback effect of the CET-4 on English learners in China, specifically focusing on the "Washback to the Learner" section shown in Figure 1. English learners in both first- and second-tier cities in China completed a 40-item survey that examined their perceptions and attitudes towards the CET-4.

Participants

Convenience sampling was utilized. Considering the potential impact of social context on students' learning experiences, one first-tier city, Shanghai, and one second-tier city, Kunming, were selected as the sites for the research. Kunming was listed as the "new first-tier city" in a later ranking in 2019, but this tier is still one level below the first-tier cities including Shanghai. We contacted a number of English lecturers from a university in Kunming and several PhD students and professors at two universities in Shanghai, and they agreed to act as liaisons for the research, introduce the study to their students, and invite their students to participate in it. A total of 148 undergraduates from Shanghai (n1 = 148) and 157 undergraduates from Kunming (n2 = 157) participated in the study (n = n1 + n2 = 305). These students had all taken the CET-4 after the 2016 revision. Table 2 summarizes the participants' demographic information.

TABLE 2
Demographic Information of the Participants

			Gender		Grade level					
City	*n*	*M* age	F	M	1	2	3	4	*M* years of English learning	*M* CET-4 score
Shanghai	148	21.00	93	55	7	87	23	30	9.39	535.97
Kunming	157	19.57	100	57	2	147	1	7	9.96	505.70

Instrument

We utilized survey methodology to investigate test-takers' perceptions of the CET-4 overall, but with a special focus on the listening section of the test. Created based on the synthesized framework in Figure 1, the survey was composed of 30 likert items and 10 open-ended questions that constituted eight sub-scales examining participants' perceptions, learning behaviors, and perceived in-class instruction related to the CET-4 (the survey is provided in Appendix A). For the likert items, participants had to rate the degree to which they perceived the test on each dimension. The open-ended questions were designed to allow participants to add additional comments on the construct.

The first sub-scale asked the participants about how difficult they found the listening section of the CET-4. The second sub-scale asked about how difficult they found the non-listening sections of the test. In these two sub-scales, test difficulty was further broken down to how challenging the participants perceived the test content, task format, and test preparation. The third sub-scale asked them to rate how useful they perceived the listening section. The fourth sub-scale asked about the perceived usefulness of the non-listening sections. Test usefulness in these two sub-scales included the extent to which the test scores indicated participants' real listening abilities, and to which the test was helpful in improving their English learning, pursuing future education, and applying for jobs. The fifth sub-scale asked about how authentic they found the listening section. The sixth sub-scale asked about the perceived authenticity level of the non-listening sections. Authenticity was evaluated based on how closely the types of test materials, task type, and test content resembled those in real-life situations. The seventh and eighth sub-scales were designed to focus on the washback effects of the listening section only. The seventh sub-scale asked about the perceived extent to which the participants' English teachers had adapted their teaching to the CET-4 listening section. The eighth sub-scale asked the participants to report the extent to which their English learning behaviors were impacted by the CET-4 listening. Each of these two constructs were divided into five categories: the content and topic, approach and strategies, pace and speed, sequence, and length of time (Watanabe, 1997). Table 3 summarizes the different components of the questionnaire, and the number of items in each sub-scale.

TABLE 3
Questionnaire

Sub-scale	Focus	*n* of likert items	*n* of open-ended questions
1	Difficulty of CET-4 listening	3	1
2	Difficulty of CET-4 other than the listening section	3	1
3	Usefulness of CET-4 listening	4	1
4	Usefulness of CET-4 other than the listening section	4	1
5	Authenticity of CET-4 listening	3	1
6	Authenticity of CET-4 other than the listening section	3	1
7	How CET-4 listening influences instruction	5	2
8	How CET-4 listening influences learning	5	2

Data collection

The study was approved by the university IRB, including the survey instrument and consent forms. The electronic version of the survey was created on Wenjuanxing, a Chinese survey website. The consent form and survey link were sent to individual participants by the instructors at the different universities who were acting as liaisons for this study. The participants were instructed to read the consent form and contact the research team if they had any questions. Once they had no further questions, they could proceed to open the survey link, confirm that they had read the consent form with a checkbox, and then complete the survey. They first answered seven background demographic questions. They then completed the eight-part survey that was written in both Chinese and English; participants had the option to choose in which language to answer the open-ended questions. It took approximately 20 minutes to complete the survey. Each participant was automatically compensated with 5 to 20 RMB.

Data analysis

The reliability for each sub-scale was first computed using Cronbach's alpha, and the average rating of each sub-scale reflected by the three to five items was computed.

Research Question 1a. A MANOVA was conducted in order to compare students' perceptions of the difficulty, usefulness and authenticity levels of the CET-4 listening between cities, with city (first-tier or second-tier) as the independent variable and the ratings of the difficulty, usefulness, and authenticity levels of the listening section as the dependent variables.

Research Question 1b. A MANOVA was conducted in order to compare students' perceptions between cities regarding the difficulty, usefulness and authenticity levels of the non-listening sections of the CET-4, with city (first tier or second tier) as the independent variable and the ratings of the difficulty, usefulness, and authenticity levels of the non-listening sections as the dependent variables.

Research Question 1c. Six paired-sample *t*-tests were conducted to compare students' perceptions of the listening section to their perceptions of the other sections within each city pertaining to the difficulty, usefulness, and authenticity levels.

Research Question 2. A MANOVA was conducted with city (first-tier or second-tier) as the independent variable and the ratings of the teaching and learning impact of the listening section as the dependent variables. In addition, two paired-sample *t*-tests were conducted to compare students' perceptions of how the learning and teaching behaviors were impacted by the test within each city.

The quantitative data were examined for normality by checking the skewness and kurtosis of each variable. If the absolute values are within 2, the data can be considered as normally distributed (Bachman, 2004). For the MANOVAs, homoscedasticity was further checked by the Box's test, and the assumption is met if the *p* value is greater than .005 (Huberty & Petoskey, 2000). Linearity was checked by plotting bivariate scatterplots between each pair of dependent variables in each group, and the assumption is met if the scatterplots are elliptical (Mertler & Reinhart, 2016). For the paired-sample *t*-tests, there were a total of eight comparisons conducted. The alpha level was set as $\alpha = .1$, and a Bonferroni adjustment was made due to multiple comparisons ($\alpha = .1/8 = .013$) (Larsen-Hall, 2010).

Due to space constraints, the analyses of the qualitative data collected from the open-ended questions is not reported in this paper.

Results

Reliability

The results indicate a very high reliability for the different sub-scales in the survey, ranging from a low of $\alpha = .845$ for the listening authenticity sub-scale, to high of $\alpha = .967$ for the impact on teaching subscale, as seen in Tables 4.

TABLE 4
Reliability of the Survey

	Difficulty ($k = 3$)	Usefulness ($k = 4$)	Authenticity ($k = 3$)	Impact on teaching ($k = 5$)	Impact on learning ($k = 5$)
Listening	.883	.856	.845	.967	.959
Other sections	.891	.892	.894		

Research Question 1a. Difficulty, Usefulness, and Authenticity Levels of the CET-4 Listening in First- versus Second-Tier Cities

Participants' ratings for each variable were shown to be normally distributed, with absolute values of skewness and kurtosis smaller than 2 (Bachman, 2004). The Box's test indicated that the assumption of homoscedasticity was violated ($F = 6.382$, $p < .005$), probably due to the unequal sample sizes for the two cities. The linearity between each pair of dependent variables in each city was ensured indicated by the bivariate scatterplots. Therefore, we determined to use Pillai's Trace, which is robust to violation of homoscedasticity, in interpreting the MANOVA result (Mertler & Reinhart, 2016).

Overall, there was no statistically significant difference in participants' overall perceived difficulty, usefulness, and authenticity levels of the CET-4 listening based on city, $F(3, 301) = 2.528$, $p > .05$; Pillai's $V(s) = 0.025$, multivariate $\eta^2 = .025$. However, as Table 5 indicates, students from Kunming consistently rated the CET-4 listening as more difficult, useful, and authentic, than the students from Shanghai reported, although the differences were only statistically significant for the difficulty and usefulness variables. This indicates that the CET-4 listening test was perceived differently by the students from different-tier cities, similar to the results comparing students from key and non-key universities in China (Huang & Yang, 2002).

TABLE 5
Difficulty, Usefulness, and Authenticity Levels of the CET-4 Listening

	Kunming (tier 2)				Shanghai (tier 1)						
	M	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis	*p*	*F*	Partial η^2
D	6.04	1.34	-.140	.043	5.63	2.10	-.667	-.462	.041*	4.203	.014
U	5.30	1.57	-.086	.028	4.81	1.93	-.204	-1.025	.016*	5.834	.019
A	5.55	1.34	-.072	.056	5.23	1.57	-.210	-.262	.056	3.683	.012

Note. D: Difficulty, U: Usefulness, A: Authenticity.
*$p < .05$

Research Question 1b. Difficulty, Usefulness, and Authenticity Levels of the Non-Listening Sections of the CET-4 for First- versus Second-Tier Cities

The participants' ratings on each of the variables were shown to be normally distributed. The Box's test indicated that the assumption of homoscedasticity was violated ($F = 8.974$, $p < .005$). The assumption of linearity was also met checked by the bivariate scatterplots. Therefore, we again decided to use Pillai's Trace (Mertler & Reinhart, 2016).

Overall, there was no statistically significant difference in participants' perceived difficulty, usefulness, and authenticity levels of the CET-4 non-listening sections based on city, $F(3, 301) = 1.520$, $p > .05$; Pillai's $V(s) = 0.013$, multivariate $\eta^2 = .015$. More specifically, as Table 6 indicates, students from Kunming consistently rated the non-listening sections of the CET-4 as slightly more difficult, useful, and authentic than students from Shanghai, although these differences were small and not statistically significant. This suggests that the non-listening sections of the CET-4 test were also perceived slightly differently by the students from different-tier cities as well, similar to Huang and Yang's (2002) study.

TABLE 6
Difficulty, Usefulness, and Authenticity Levels of the CET-4 Non-Listening Sections

	Kunming (tier 2)				Shanghai (tier 1)						
	M	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis	*p*	*F*	Partial η^2
D	5.20	1.28	-.538	1.009	4.98	1.77	-.591	-.255	.224	1.490	.005
U	5.54	1.34	-.284	.314	5.23	1.69	-.283	-.627	.074	3.224	.011
A	5.71	1.27	-.279	.444	5.66	1.59	-.381	-.322	.725	.124	.000

Note. D: Difficulty, U: Usefulness, A: Authenticity.
*$p < .05$

Research Question 1c. Difficulty, Usefulness, and Authenticity Levels of the CET-4 Listening versus Non-listening Sections

Students from both cities consistently rated the listening section more difficult, less useful, and less authentic compared with the non-listening sections of the test. These differences on the three sub-scales were statistically significant for four of the six comparisons as seen in Table 7.

TABLE 7
Within Cities

		Listening		Other sections				
		M	*SD*	*M*	*SD*	*p*	*t*	*d*
Kunming (tier 2)	Difficulty	6.04	1.34	5.20	1.28	.000*	7.62	.61
	Usefulness	5.30	1.57	5.54	1.34	.022	-2.32	.18
	Authenticity	5.55	1.34	5.71	1.27	.071	-1.82	.15
Shanghai (tier 1)	Difficulty	5.63	2.10	4.98	1.77	.000*	5.22	.43
	Usefulness	4.81	1.93	5.23	1.69	.001*	-3.29	.27
	Authenticity	5.23	1.57	5.66	1.59	.000*	-3.88	.31

* $p < .013$

Research Question 2. The Impact of the CET-4 Listening on Students' English Learning Behaviors and Their Instructors' Teaching

Participants' ratings on each variable were shown to be normally distributed, with absolute values of skewness and kurtosis smaller than 2 (Bachman, 2004). The Box's test indicated that the assumption of homoscedasticity was violated (F = 7.850, p < .005). The assumption of linearity was again met, as suggested by the bivariate scatterplots. Therefore, Pillai's Trace was chosen for use (Mertler & Reinhart, 2016).

In general, the participants from the two different cities perceived the impact of the CET-4 listening on their learning behaviors and their instructors' teaching differently, $F(2, 302) = 5.244$, $p < .05$; Pillai's $V(s) = 0.034$, multivariate $\eta^2 = .034$. As Table 8 shows, students from Kunming reported a greater impact of the CET-4 listening on their learning behaviors compared with students from Shanghai, although the difference was not statistically significant. In addition, students from Kunming reported that their instructors adapted their teaching in preparing for the CET-4 listening to a greater extent compared to students from Shanghai, and this difference was statistically significant.

TABLE 8
Impact of CET-4 Listening on Learning and Teaching

	Kunming (tier 2)				Shanghai (tier 1)						
	M	*SD*	Skewness	Kurtosis	*M*	*SD*	Skewness	Kurtosis	*p*	*F*	Partial η^2
L	4.99	1.73	-.318	-.045	4.61	2.27	-.337	-1.188	.106	2.636	.009
T	4.82	1.75	-.443	-.288	4.05	2.40	.131	-1.304	.001*	10.419	.033

Note. L: Learning, T: Teaching.
*p < .05

In addition, as Table 9 shows, students from both cities reported that their learning behaviors were impacted more by the CET-4 listening than their teachers' instruction was impacted by the CET-4 listening, although the difference was only statistically significant for the Shanghai participants.

TABLE 9
Washback on Students' Learning versus Perceived Teaching Behaviors

	Learning		Teaching				
	M	*SD*	*M*	*SD*	*p*	*t*	*d*
Kunming (tier 2)	4.99	1.73	4.82	1.75	.287	-1.07	.09
Shanghai (tier 1)	4.61	2.27	4.05	2.40	.000*	-3.75	.31

*p < .013

Discussion

Previous research on test washback demonstrated a need to examine "students' views and their accounts of the effects on their lives of test preparation, test taking and the scores they have received on tests" (Hamp-Lyons, 1997, p. 299). By addressing this stated need, the current study yielded some surprising results.

Firstly, the test's impact on students was less than expected. The means of students' ratings on the 9-point impact scales for both learning and perceived teaching behaviors were around the middle of the scale (4.5). High-stakes tests are associated with crucial decisions that immediately and directly affect students' lives (Andrews, 1995; Cheng, 1998, 2008; Gu, 2005). Typically, the higher the stake of a test, the stronger washback effects would be expected (Cheng, 1998). Considering the important role the CET-4 plays in terms of students' graduation and job opportunities, and the fact that the listening section constitutes 35% of the overall score, it was expected that students would report much higher levels of test impact (Li et al., 2012). The participants' ratings of the listening difficulty, usefulness, and authenticity

may shed some light on this finding. As described above, students' ratings of the difficulty level of the listening section were higher than their ratings of its usefulness and authenticity. This suggests that for a test that students perceive to be difficult but not particularly useful or authentic, they would not feel the need to put in tremendous efforts in preparing for something that they think is not very helpful for them personally, and for a test that they feel lacks many of the aspects of real-life communication.

Secondly, students from both cities rated the listening section as more difficult and less useful and authentic than the non-listening sections of the test. It has been shown that L2 listening ability is relatively difficult to improve compared to reading and writing ability, especially within a short time (Fan, 1993). In addition, students may be more used to learning grammar and vocabulary intensively, and it is probably easier for learners to see and recognize the improvement they make in grammatical accuracy and vocabulary knowledge compared to the more amorphous improvements in listening ability. Therefore, they may not pay as much attention to the listening section as to other sections in the test preparation process. A similar mindset may be adopted by the teachers as well. Because of the difficulty in making measurable and salient improvements in listening ability, the teachers may not focus on improving students' listening ability but rather teach more to the other sections of the test.

The results were also somewhat surprising that a stronger impact was found on students' learning behaviors than perceived instruction, although statistical significance was only found with students from Shanghai. This is in line with Li's (2009) findings, which focused on the impact of the CET writing section. Pan (2013) found similar results suggesting that teachers' instruction was only affected to a minor degree by the exit requirements of universities. It seems logical to believe that teachers' instruction directs students' learning, and this seems especially true since teachers are more knowledgeable and experienced in language learning and testing. Numerous studies have demonstrated the extent to which English classrooms were teacher-centered (e.g., Gu, 2003; Yang et al. 2013). Research has also indicated East Asian learners are able to learn English outside of the classroom when directed by their teachers, but this outside learning is much less effective without their teachers' help (Sakai, Chu, Takagi, and Lee, 2008). However, in the current study, the results indicated that students reported taking more initiative than their teachers to prepare for the listening section of the CET-4, perhaps because teachers were more concerned about alternative or more difficult tests that their students need to pass. This interpretation of the research findings can also be seen from students' current English levels. In the first demographic section of the survey, students' CET-4 scores were also reported. These reported scores were relatively high, indicating that these students would probably have little reason to worry about their "passing" the listening section. Their teachers may have realized this and therefore did not adapt their teaching much to prepare for the test, leaving the student to take a more active role in their own test preparation. Another possible reason for this result can be related to teachers' lack of professional support from colleges (Li, 2010). The teachers perhaps had not received enough training on how to best prepare their students for this test, and as a result, they were not able to make effective instructional adjustments. This finding is contradictory to some of the previous research findings indicating that teaching to the test was a severe problem in China because CET certificates were required for graduation for the majority of Chinese universities (e.g., Gu & Liu, 2005; Zheng & Cheng, 2008). As evident in the case of the universities in the current study, teachers' instruction was not heavily directed or constrained by the test from the students' perspective.

One important contribution of the present study is comparing and contrasting the washback on students from a first- and second-tier city, because it is important to acknowledge and research how a test's impact can differ according to the different contexts in which a test is used. Although no causal relationships can definitively be established, these results suggest that students from the second-tier city are more influenced by the test regarding both their learning behaviors and the perceived instruction they received from their teachers. This result is similar to the findings in Huang and Yang's (2002) research looking at students' and teachers' perceptions in key and non-key universities in China. Teachers from the latter institutes indicated more washback effects because their students and their instruction were more challenged by the test. In the current study, although the participants were all enrolled in key-universities,

those students from Kunming, a second-tier city, considered the CET-4, especially the listening section, more challenging than students from Shanghai, and therefore more washback effects were identified for the Kunming students.

Additionally, the present study adds crucial insights to the relationship between test validity and washback. Messick (1989, 1996) suggested that avoiding construct under-representation is indispensable to ensure test validity. In this case, for example, evaluating students' English ability in real-life situations (authenticity), which is listed as one of the constructs of CET-4, is not clearly tested or represented in the test. Such a mismatch under-represents the construct of communicative language ability, and may in turn negatively affect students' learning and teachers' instruction, which was reflected in students' ratings and comments. This association aligns with Messick's (1996) argument that places "washback within the consequential aspect of construct validity" (p. 242), and supports the argument that test developers should critically evaluate potential intended and unintended impact a test may exert on teaching and learning.

Conclusion

This study was designed to understand the possible washback effects of the latest version of CET-4 test from the students' perspective. Using a survey, students in different-tier cities in China were asked about the difficulty, usefulness, and authenticity levels of CET-4 listening and the other (non-listening) sections of the test. They also specifically reported the degree to which their own learning and their teachers' instruction were impacted by the listening section. The results indicated that, comparing participants from the first- and second-tier city, the latter group considered CET-4 more difficult, authentic, and useful across sections, and the listening section had more of an impact on their teachers' instruction and their own learning. Despite these differences, for both groups, the listening section had a stronger influence on learning than perceived teaching. The listening section was also rated as more difficult, less useful, and less authentic compared with the non-listening sections. These results reiterate the belief that test washback should be considered as a fundamental component of test validity and should be taken into account during test development (Bachman, 2005; Messick, 1996).

This study also has a number of limitations. First, the only instrument utilized was the survey. Although both quantitative and qualitative data were collected, space constraints did not allow us to report on the qualitative data in this paper. Incorporating additional methods to triangulate the data would be useful. Class observations, interviews, or ethnographic research could be done to achieve this purpose. In addition, future studies should include other test stakeholders, beyond students. After all, research on students' perspectives is only the tip of the iceberg, and other participants of the whole teaching and learning process deserve equal attention. Lastly, the effect size for most comparisons was quite small. Future studies could consider further increasing the sample size and collecting various other types of data in addition to surveys.

Even with these limitations, this study expands the current understanding of washback effects of the CET-4 in that, first, it is the first study to focus on the listening section of the latest version of the test. Additionally, this study examines how test impact (on both teaching and learning) might vary according to the different tiers of cities in China. Lastly, this paper addresses the call commonly found in the literature (e.g., Hamp-Lyons, 1997; Li et al., 2012) that studies need to include and focus on learners when researching test washback.

Acknowledgment

We would like to thank Ms. Xuanchen Ye and Ms. Xueye Yan for assisting with our data collection. We appreciate the anonymous reviewers who provided insightful feedback to improve the paper.

The Authors

Linlin Wang is a PhD candidate in the Dept. of Teaching and Learning at Temple University. Her research interests include L2 assessment, L2 pedagogy, and multicultural educational issues. Her recent research focus investigates the effect of visual cues on test-takers' L2 listening behaviors.
Department of Teaching and Learning
Temple University, College of Education and Human Development
1301 Cecil B. Moore Ave, Philadelphia, PA 19122
Mobile: +1 2672648892
Email: linlin.wang@temple.edu

Elvis Wagner is an associate professor in the Dept. of Teaching and Learning at Temple University. He is interested in the teaching and testing of second language oral communicative competence. His primary research focus examines how L2 listeners process and comprehend unscripted, spontaneous spoken language, and how this type of language differs from the scripted spoken texts learners often encounter in the L2 classroom. He recently co-authored (with Gary Ockey) the book *Assessing L2 listening: Moving towards authenticity*.

Department of Teaching and Learning
Temple University, College of Education and Human Development
1301 Cecil B. Moore Ave, Philadelphia, PA 19122
Email: elviswag@temple.edu

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, *13*(3), 280-297. https://doi.org/10.1177/026553229601300304

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

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

Andrews, S. (1995). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, V. Berry, & R. Berry (Eds.), *Bringing about change in language education* (pp. 67-81). Hong Kong, China: Department of Curriculum Studies, University of Hong Kong.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1-34. https://doi.org/10.1207/s15434311laq0201_1

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, *13*(3), 257-279. https://doi.org/10.1177/026553229601300303

Bailey, K. M. (1999). *Washback in language testing*. Princeton, NJ: Educational Testing Service.

Brown, J. D. (1997). Do tests washback on the language classroom? *The TESOLANZ Journal, 5*(5), 63-80. https://www.tesolanz.org.nz/publications/tesolanz-journal/volume-5-1997/

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. Sheffield, UK: Equinox.

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

Cheng, L. (1998). *The washback effect of public examination change on classroom teaching: An impact study of the 1996 Hong Kong certificate of education in English on the classroom teaching of*

English in Hong Kong secondary schools (Unpublished doctoral dissertation). University of Hong Kong, China.

Cheng, L. (2001). Washback studies: Methodological considerations. *Curriculum Forum, 10*(2), 17-32.

Cheng L. (2008) Washback, impact and consequences. In N. H. Hornberger (Eds.), *Encyclopedia of language and education* (pp. 349-364). Boston, MA: Springer. https://doi-org.libproxy.temple.edu/10.1007/978-0-387-30424-3_186

Cheng, L., & Falvey, P. (2000). What works? The washback effect of a new public examination on teachers' perspectives and behaviors in classroom teaching. *Curriculum Forum, 9*(2), 1-33.

Choi, I. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*(1), 39-62. https://doi.org/10.1177/0265532207083744

Crusan, D., & Cornett, C. (2002). The cart before the horse: Teaching assessment criteria before writing. *The International Journal for Teachers of English Writing Skills, 9*, 20-33.

Fan, Y. (1993). Listening: Problems and solutions. *English Teaching Forum, 31*(2), 16-19. http://www.valrc.org/courses/esolbasics/lesson5/docs/Listening_a.pdf

Gosa, C. M. C. (2004). *Investigating washback: A case study using student diaries* (Unpublished doctoral dissertation). Lancaster University, UK.

Gu, X. (2003). Case studies of college English teacher' lessons. *Research in Foreign Language and Literature, 4,* 45-51. http://www.cnki.com.cn/Article/CJFDTOTAL-WGYW200304007.htm

Gu, X. (2005). Positive or negative? An empirical study of CET washback on college English teaching and learning in China. *ILTA Online Newsletter*, *2*. http://www.iltaonline.com/newsletter/02-2005oct/

Gu, X. (2007). Characteristics of college English classroom teaching and learning: The impact of CET on classroom teaching and learning. *Journal of Xi'an International Studies University, 4*, 39-45. https://doi.org/10.16362/j.cnki.cn61-1457/h.2007.04.020

Gu, W., & Liu, J. (2005). Test analysis of college students' communicative competence in English. *Asian EFL Journal, 7*(2), 118-33. http://asian-efl-journal.com/June_05_wg.pdf

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

Han, B., Dai, M., & Yang, L. (2004). On the problems of CET. *Foreign Languages and Their Teaching, 179*(2), 17-23. http://en.cnki.com.cn/Article_en/CJFDTotal-WYWJ200402005.htm

Hawkey, R. (2006). *Studies in language testing: Vol. 24. Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge, UK: Cambridge University Press.

Huang, D., & Yang, L. (2002). Analyzing the problems of the College English Test based on a survey. *Foreign Language Testing and Research, 4*, 288-293.

Huberty, C. J., & Petoskey, M. D. (2000). Multivariate analysis of variance and covariance. In H. Tinsley., & S. Brown (Ed.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 183-208). Cambridge, MA: Academic Press.

Hughes, A. (1993). *Washback and TOEFL 2000* (Unpublished manuscript). University of Reading, UK.

Larsen-Hall, J. (2010). *A guide to doing statistics in second language research.* New York, NY: Rutledge.

Li, J. (2002). The current College English Test in China: Problems and thoughts. *Foreign Language Education, 23*(5), 33-38.

Li, H. (2009). Are teachers teaching to the test? A case study of the College English Test (CET) in China. *International Journal of Pedagogies and Learning*, *5*(1), 25-36. https://doi.org/10.5172/ijpl.5.1.25

Li, J. (2010). An explorative study of Chinese college English teachers' professional development. *The Journal of Asia TEFL*, *7*(4), 1-48. https://search.proquest.com/docview/2266429358?pq-origsite=gscholar&fromopenview=true

Li, H., Zhong, Q., & Suen, H. K. (2012). Students' perceptions of the impact of the college English test. *Language Testing in Asia*, *2*(3), 77. https://doi.org/10.1186/2229-0443-2-3-77

Mertler, C. A., & Reinhart, R. V. (2016). *Advanced and multivariate statistical methods: Practical application and interpretation*. Glendale, CA: Pyrczak Publishing.

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

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

Mickan, P., & Motteram, J. (2009). The preparation practices of IELTS candidates: Case studies. In *International English Language Testing System (IELTS) Research Reports 2009: Volume 10.* https://search.informit.com.au/documentSummary;dn=103243737967283;res=IELHSS

Pan, Y. (2013). Does teaching to the test exist? A case study of teacher washback in Taiwan. *The Journal of Asia TEFL*, *10*(4), 185-213. https://search.proquest.com/docview/2266419800?pq-origsite=gscholar&fromopenview=true

Purpura, J. (2009). The impact of large-scale and classroom-based language assessments on the individual. In L. Taylor & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 301-325). Cambridge, UK: Cambridge University Press.

Qi, L. (2006). Some reflections on washback. *Foreign Languages and their Teaching, 8*, 29-32. http://en.cnki.com.cn/Article_en/CJFDTotal-WYWJ200608007.htm

Qi, L. (2007). Examining the intended and actual washback of the proofreading subtest in the National Matriculation English Test. *Curriculum, Teaching Material and Method, 10*, 43-46. http://en.cnki.com.cn/Article_en/CJFDTotal-KJJF200710010.htm

Rea-Dickins, P., Kiely, R., & Yu, G. (2007). Student identity, learning and progression: The affective and academic impact of IELTS on successful candidates. In *International English Language Testing System (IELTS) Research Reports 2007: Volume 7.* https://search.informit.com.au/documentSummary;dn=078890444532816;res=IELHSS

Rozzi, A. M., Pinto, V., González, M., & Crimi, Y. (2014). The impact of Cambridge English examinations on institutional change. *Research Notes, 57*, 31-40. https://www.cambridgeenglish.org/images/177881-research-notes-57-document.pdf#page=33

Sakai, S., Chu, M. P., Takagi, A., & Lee, S. (2008). Teachers' roles in developing learner autonomy in the East Asian region. *The Journal of Asia TEFL*, *5*(1), 97-121. https://www.researchgate.net/profile/Shien_Sakai/publication/279979157_Teachers'_Roles_in_Developing_Learner_Autonomy_in_the_East_Asian_Region/links/55a15a6b08ae1c0e0464091c.pdf

Shao, H. (2006). An empirical study of washback from CET-4 on college English teaching and learning. *CELEA Journal*, *29*(1), 54-59. http://www.celea.org.cn/teic/65/65-54.pdf

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, *13*(3), 298-317. https://doi.org/10.1177/026553229601300305

Wall, D. (1997). Impact and washback in language testing. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education* (pp. 291-302). Dordrecht, Netherlands: Kluwer Academic Publishers.

Watanabe, Y. (1997). *The washback effects of the Japanese university entrance examinations of English: classroom-based research* (Unpublished doctoral dissertation). Lancaster University, UK.

Watanabe, Y. (2004). Methodology in washback studies. In L. E. Cheng, Y. E. Watanabe, & A. E. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 41-58). Mahwah, NJ: Lawrence Erlbaum Associates.

Yan, S. (2016). Washback of the CET on English teaching. *Literature Education, 12*, 110-111. https://doi.org/10.16692/j.cnki.wxjyx.2016.12.055

Yan, Q., Gu, X., & Khalifa, H. (2014). Impact of Cambridge English: Key for schools on young learners' English learning: Voices from students and parents in Beijing, China. *Research Notes, 58*, 44-50. https://www.cambridgeenglish.org/Images/182921-research-notes-58-document.pdf

Yang, Z., Gu, X., & Liu, X. (2013). A longitudinal study of the CET washback on college English classroom teaching and learning in China: Revisiting college English classes of a

university. *Chinese Journal of Applied Linguistics*, *36*(3), 304-325. https://doi.org/ 10.1515/cjal-2013-0021

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., Zhang, Y., Xu, S., & Fang, L. (2017). Preparing for the speaking tasks of the TOEFL iBT Test: An investigation of the journeys of Chinese test takers. *ETS Research Report Series*, *1*, 1-59. https://doi.org/10.1002/ets2.12145

Zhang, L. (2003). *An investigation into the issue of washback in language testing, with reference to the college English testing in China* (Unpublished master's thesis). University of Warwick, UK.

Zhang, H., & Li, Z. F. (2014). Residential properties, resources of basic education and willingness price of buyers. *China Finance Review International, 4*(3). https://doi.org/10.1108/CFRI-09-2013-0117

Zhao, J. (2006). *Exploring the relationship between Chinese Students' attitudes toward College English Test and their test performance* (Unpublished master's thesis). Queen's University, Canada.

Zheng, Y., & Cheng, L. (2008). Test review: College English Test (CET) in China. *Language Testing*, *25*(3), 408-417. https://doi.org/10.1177/0265532208092433

Zou, S., & Xu, Q. (2017). A washback study of the test for English Majors for Grade Eight (TEM8) in China: From the perspective of university program administrators. *Language Assessment Quarterly*, *14*(2), 140-159. https://doi.org/10.1080/15434303.2016.1235170

Appendix

Washback study survey
大学生英语四级影响的研究

1. How difficult do you think the listening section of the CET-4 is?
您怎样评价四级听力考试的难度？

(1) Please rate the difficulty level regarding the content of the spoken texts. 听力材料的内容令四级听力考试（1=非常容易，9=非常困难）：

1	2	3	4	5	6	7	8	9

1 = very easy very difficult = 9

(2) Please rate the difficulty level regarding the format of listening test (listening to each text once only, cannot preview question stems, time limit, etc.).听力考试的形式（比如只能听一次材料，不能看到听力问题，时间限制等）令四级听力考试（1=非常容易，9=非常困难）：

1	2	3	4	5	6	7	8	9

1 = very easy very difficult = 9

(3) Please rate the difficulty level regarding the listening test preparation. 四级听力考试的备考令四级考试（1=非常容易，9=非常困难）：

1	2	3	4	5	6	7	8	9

1 = very easy very difficult = 9

(4) Do you have any comments on the difficulty of the CET-4 listening?您对四级听力考试的难度想要发表任何评论吗？

_____

2. How difficult do you think the sections other than listening of the CET-4 are?
您怎样评价四级考试除听力以外其他部分的难度？

(1) Please rate the difficulty level regarding the content of the reading passages, writing and translation prompts, and speaking topics if applicable. 四级考试材料的内容（包括阅读篇章的内容，写作和翻译的考试内容，和口语话题）令四级考试（1=非常容易，9=非常困难）：

1	2	3	4	5	6	7	8	9

1 = very easy very difficult = 9

(2) Please rate the difficulty level regarding the test format (time limit, question type, etc.). 四级考试的形式（考试时间限制、题型设定等）令四级听力以外的考试部分（1=非常容易，9=非常困难）：

1	2	3	4	5	6	7	8	9

1 = very easy very difficult = 9

(3) Please rate the difficulty level regarding the test preparation. 四级考试的备考令四级考试（1=非常容易，9=非常困难）：

1	2	3	4	5	6	7	8	9

1 = very easy very difficult = 9

(4) Do you have any comments on the difficulty of the CET-4 sections other than listening?您对四级考试除听力以外部分的难度想要发表任何评论吗？

_____

3. How useful do you think the listening section of the CET-4 is?
您怎样评价四级听力考试的有用程度？

(1) How well do you think the listening section indicates your listening ability? 四级听力考试对您真实听力水平的评价（1=没有用或非常不准确，9=非常有用或非常准确）：

1	2	3	4	5	6	7	8	9

1 = not useful at all very useful = 9

(2) How helpful do you think the listening section is in improving your English learning?四级听力考试对提高您的英语听力水平（1=没有任何帮助，9=帮助非常大）

1	2	3	4	5	6	7	8	9

1 = not helpful at all very helpful = 9

(3) How helpful do you think the listening section is in pursuing further education? 四级听力考试对您日后求学深造 （1=没有任何帮助，9=帮助非常大）

1	2	3	4	5	6	7	8	9

1 = not helpful at all very helpful = 9

(4) How helpful do you think the listening section is in preparing for job applications? 四级听力考试对您日后求职（1=没有任何帮助，9=帮助非常大）

1	2	3	4	5	6	7	8	9

1 = not helpful at all very helpful = 9

(5) Do you have any comments on the usefulness of the CET-4 listening?您对四级听力考试的有用程度想要发表任何评论吗？

_____

4. How useful do you think the sections other than listening of the CET-4 are?
您怎样评价四级考试除听力以外的其他部分的有用程度？

(1) How well do you think the sections other than listening indicate your English ability? 四级除听力以外的其他部分对您真实英语水平的评价（1=没有什么用或非常不准确，9=非常有用或非常准确）：

1	2	3	4	5	6	7	8	9

1 = not useful at all very useful = 9

(2) How helpful do you think the sections other than listening are in improving your English learning? 四级除听力以外的其他部分对提高您的英语水平（1=没有任何帮助，9=帮助非常大）

1	2	3	4	5	6	7	8	9

1 = not helpful at all very helpful = 9

(3) How helpful do you think the sections other than listening are in pursuing further education? 四级除听力以外的其他部分对您日后求学深造（1=没有任何帮助，9=帮助非常大）

1	2	3	4	5	6	7	8	9

1 = not helpful at all very helpful = 9

(4) How helpful do you think the sections other than listening are in preparing for job applications? 四级除听力以外的其他部分对您日后求职（1=没有任何帮助，9=帮助非常大）

1	2	3	4	5	6	7	8	9

1 = not helpful at all very helpful = 9

(5) Do you have any comments on the usefulness of the CET-4 sections other than listening?您对四级考试除听力以外部分的有用程度想要发表任何评论吗？

_____

5. How authentic do you think the listening section of the CET-4 is?
您怎样评价四级听力考试的真实性（贴近真实英语使用）？

(1) How authentic do you think the spoken texts are? 四级听力材料和现实生活中的英文听力相比（1=完全脱离，9=非常贴近）

1	2	3	4	5	6	7	8	9

1 = not authentic at all very authentic = 9

(2) How often do you think you may encounter the types of tasks in CET-4 listening? 四级听力考试问题形式（单项选择题）和现实生活中能遇到的听力问题形式相比（1=完全脱离，9=非常贴近）

1	2	3	4	5	6	7	8	9

1 = very little very often = 9

(3) How often do you think you may encounter the content of CET-4 listening? 四级听力考试材料设计的话题和内容和现实生活中能遇到的听力内容相比（1=完全脱离，9=非常贴近）

1	2	3	4	5	6	7	8	9

1 = very little very often = 9

(4) Do you have any comments on the authenticity of the CET-4 listening?您对四级听力考试贴近真实英语使用的程度想要发表任何评论吗？

_____

6. How authentic do you think the sections other than listening of the CET-4 are?
您怎样评价四级听力考试的真实性（贴近真实英语使用）？

(1) How authentic do you think the materials in the prompts other than the listening section are? 四级考试材料和现实生活中的英文材料相比（1=完全脱离，9=非常贴近）

1	2	3	4	5	6	7	8	9

1 = not authentic at all very authentic = 9

(2) How often do you think you may encounter the types of tasks in the CET-4 other than the listening section? 四级除听力以外的考试问题形式（单项选择题）和现实生活中能遇到的听力问题形式相比（1=完全脱离，9=非常贴近）

1	2	3	4	5	6	7	8	9

1 = very little very often = 9

(3) How often do you think you may encounter the content of CET-4 other than the listening section? 四级除听力以外的考试材料设计的话题和内容和现实生活中能遇到的听力内容相比（1=完全脱离，9=非常贴近）

1	2	3	4	5	6	7	8	9

1 = very little very often = 9

(4) Do you have any comments on the usefulness of the CET-4 sections other than listening?您对四级考试除听力以外部分贴近真实英语使用的程度想要发表任何评论吗？

_____

7. To what extent does the listening section of the CET-4 influence the instructions you received at school? 您怎样评价四级听力考试对您学校英语教学的影响？

(1) My English teacher adjusts the content/topic of learning according to the CET-4 listening: 我的老师根据四级听力考试调整教学内容或主题（1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(2) My English teacher teaches us the approach and strategies of learning according to the CET-4 listening: 我的老师根据四级听力考试教授听力学习的方法和策略（1=并不教授，9=教授很多）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(3) My English teacher adjusts the pace or speed of learning each topic/content (or the time spent on each topic/content) according to the CET-4 listening: 我的老师根据四级听力考试调整每个教学内容的教学速度（1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(4) My English teacher adjusts the sequence of learning each topic/content according to the CET-4 listening: 我的老师根据四级听力考试调整教学内容的顺序 （1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(5) My English teacher adjusts the time spent on each topic/content and how deeply we should learn each topic according to the CET-4 listening: 我的老师根据四级听力考试调整每个话题的教学时间长短和对每个内容了解的深度（1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(6) Are there any other influences that CET-4 listening has on your teacher's English instruction at school? 四级听力考试对您学校老师的英语教学还产生了其他影响吗？

_____

(7) Do you like the above influences that CET-4 listening has on English instructions at school? Why or Why not? 您喜欢上述四级听力对学校英语听力教学的影响吗？为什么？

_____

8. To what extent does the listening section of the CET-4 influence your own learning behavior? 您怎样评价四级听力考试对您自身英语学习的影响？

(1) I adjust the content/topic of learning according to the CET-4 listening: 我根据四级听力考试调整学习的内容或主题（1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(2) I specifically adopt the approach and strategies of learning according to the CET-4 listening: 我采用与四级听力考试相适应的听力学习的方法和策略（1=并不采用，9=采用很多）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(3) I adjust the pace or speed of learning each topic/content (or the time spent on each topic/content) according to the CET-4 listening: 我根据四级听力考试调整每个学习内容的学习速度（1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(4) I adjust the sequence of learning each topic/content according to the CET-4 listening: 我根据四级听力考试调整学习内容的顺序 （1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(5) I adjust the time spent on each topic/content and how deeply I should learn each topic according to the CET-4 listening: 我根据四级听力考试调整每个话题的学习时间长短和对每个内容了解的深度（1=并不调整，9=调整很大）

1	2	3	4	5	6	7	8	9

1 = not very much a very large extent = 9

(6) Are there any other influences that CET-4 listening has on your own English learning at school? 四级听力考试对您的英语学习还产生了其他影响吗？

_____

(7) Do you like the above influences that CET-4 listening has on your own English learning at school? Why or Why not? 您喜欢上述四级听力对您英语学习的影响吗？为什么？

_____