International Education Studies; Vol. 15, No. 4; 2022
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Consequences, Impact and Washback of CET Test Within Assessment for Use Argument to Validation

Chen Shijun[1]

[1] Shenzhen Bao'an No.1 Foreign Language School, Shenzhen, China

Correspondence: Chen Shijun, Shenzhen Bao'an No.1 Foreign Language School, Shenzhen, China.

Received: March 11, 2022 Accepted: April 15, 2022 Online Published: July 26, 2022

doi:10.5539/ies.v15n4p42 URL: https://doi.org/10.5539/ies.v15n4p42

Abstract

The high-stakes College English Test (CET), developed, administered, and reformed over the last 20 years, has received great attention in the aspect of washback on teaching and learning from previous research. Very few studies explored its consequences in the workplace domain—being used as a screening lever. This research aimed to 1) compare difference and similarities between skills measured in the test and performance required in the workplace, as well as the relevance between tasks in two domains, 2) investigate employers and employees' interpretation on the use of the test in the working environment, 3) explore the impacts of the test per se on both stakeholders and consequences of the test use. To reach this goal, the researcher adopted Bachman and Palmer's (2010) Assessment for Use Argument (AUA) framework and constructed three claims as research questions. This research employed qualitative method, carrying out in-depth interviews with eight participants consisted of employers and just-graduated students as employees. These participants' responses to the interviewing questions were fully transcribed and analyzed. The study found that though some task methods in two domains are different; there is a high level of similarity between skills measured in two areas. The test is proved in this study to be impartial, generalizable, and sufficient for employment. Therefore, the CET can be used for selection decision in commercial domain and beneficial to both groups of stakeholders.

Keywords: language assessment, washback, assessment for use argument, consequences

1. Introduction

The College English Test (CET) is a test serving for the educational attempt "to collect data connected with the learning of languages, and with assessing the degree of progress towards a learning goal" (Green, 2013). It is a test battery consisting of Band 4 (CET 4), Band 6 (CET 6), and the CET-Spoken English Test (CET-SET). It had experienced some significant changes in content and format within the last two decades to avoid "construct under-representation" and "construct irrelevant variance" (Messick, 1996), such as overloaded objective item format. In 2006, more focuses laid on integrated tasks and subject items were used alternatively in the reading section. Among all these changes, "the most significant development was the introduction of the CET Spoken English Test (CET-SET) (Jin & Yang, 2018).

Initially, the CET has been used for educational purposes intended by the Committee, however, there are some illegitimate uses of the CET and its test scores in recent years. For example, CET certificate at previous time is a graduation requirement for students to obtain their academic degree. Even if this policy has no longer been applied, many university administrations use certain stimulus to encourage students to sit for CET test as it is often regarded as one of the criteria to judge the prestige of a university. For most graduates, CET certificate is also an asset to stand a better chance in the job market (Cheng, 2008), or as a prerequisite for recruiting in the business domain.

Those expanded uses of the CET have become a spur for much research to investigate different aspects of the test, for example, perceptions of stakeholders (Li et al., 2012); problems of the test (Han, Dai, & Yang, 2004); washback on teaching and learning (Chen, 2007; Gu, Yang, & Liu, 2013), test reliability and validity (Li, 2019; Zhu, 2017). Among those studies, test validity is an essential issue to investigate as it correlates to the quality of assessment. Many scholars define validity as the degree to which a test measures what it claims to be measuring (Alderson et al., 1995; Brown, 1996) with lists of qualities, such as content validity, criterion validity and so on. However, the early definition relates these listed qualities without much consideration on the actual use in real life. McCall (1922, p. 196) argues that ensuing the validity of a test involves 'securing a close correspondence between

test results and practical life'. In addition, what needs to be valid are the inferences made about score meaning, namely, the score interpretation and its action implication for test use (Messick, 1996).

Test consequence is usually associated with the actual use or misuse of the test results. Messick (1996) stated that the consequences of tests are likely to be a function of factors both within the test itself and within the setting where the test is implemented. Similarly, Weir (2005) refers it to three aspects: washback in the classroom or workplace, differential validity and effects on the individual within society. As for its scope, test consequence is defined in this research as a broader sense referring to the effects of test and test use in both educational and social contexts. According to Wall (1997), test impact is "any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole". The distinction between impact and washback is the scope of effects. Washback or backwash is one dimension of impact which refers to the effect of teaching and learning in an educational context (Alderson & Wall, 1993; Shohamy, 1996; Alderson & Hamp-Lyons, 1996). Since washback has been widely used in applied linguistics nowadays, in this research washback and impact are used interchangeably as in many other studies (Andrews, Fullilove, & Wong, 2002; Li, Zhong, & Suen, 2012). Most studies (Wang, 2011; Xiao, 2014) conducted on washback were restrained to the educational context, yet very few targeted at the consequences in the social context, especially in the business domain.

This research aims to mainly explore consequences, impacts and washback of the CET test in the business domain by aligning it with the actual use of the test. Perceptions were adopted from both employer and employee parties. It first compared the similarities and differences between assessment tasks and performances that required in commercial activities (i.e. authentic task features, skills required and so on) to establish the intensity of the relevance. Next this study explored perceptions from both employers and students as employees on the test and their interpretation of test scores from three aspects: 1). The test used for screening impartial to all candidates, 2) The test is generalizable to competences needed in the workplace 3). The test is sufficient to demonstrate one's English proficiency. Another objective is to examine whether the consequences of test use and the decisions made are beneficial to the employers and the students in the workplace.

Consequential evidence derived from social context is of vital importance to validity study of the CET test based on three reasons. First, as the end-users of test scores, employers' perceptions and interpretation of scores provide information about actual use for test developers, which is beneficial from a validity perspective. Hamp-Lyons (1997) suggested that a variety of stakeholders must be investigated to generate a better understanding of how tests affect different strata of society. Second, it is not a given that employers' understanding of the test and test design are entirely the same with test developers. Depending on score use contexts, the meaning of test scores may be understood differently by stakeholders (Im, Shin, & Cheng, 2019). Views from employers and student-employees in the business domain may shed light on the future development of the CET or other large high-stake tests. Finally, this study helps enhance stakeholders' assessment literacy by equipping them with knowledge in testing, assessment methods, as well as the complexities of scoring criterion.

One main limitation of this research is the collection of qualitative data by using interviews only. Only eight informants involved in this study due to space and time restraints. It would be ideal if more employer participants from a variety of industries and student participants from different majors, grades, and universities being interviewed. Even though this study is devised to be an in-depth analysis, a mixed method with both quantitative methods and qualitative ones could be utilized to cancel out the sampling bias and establish a triangulation. Another limitation is that this research is based on participants' perceptions only with no other empirical evidence such as students test scores, the evaluation of employees' writing samples and so on. In addition, the inaccessibility of language use in certain business occasions preceded a more precise analysis of the characteristics of English language use.

2. Literature Review

2.1 Theoretical Frameworks

Based on Messick's progressive matrix and the argument-based structure proposed by Kane (2002, 2006, 2013), Bachman (2003, 2005), established assessment use argument (AUA) linking test performance with score meaning with two arguments: the assessment validity argument and the utilization argument. The framework initially specified four types of claims and warrants to be supported by evidence, focusing on the utilization argument. Claims are statements about measures, intended interpretations, decisions and intended consequences. The warrant refers to general statements that provide legitimacy of a particular step in the argument (Toulmin, 2003, p. 92) and is used in this case to justify the inference from data to claim.

Warrant 1: Relevance

Relevance warrant concerns with the correspondence between the characteristics of test tasks and tasks in Target Language Use (TLU) domain as well as the association between the ability assessed with the competence needed in the TLU domain. It justifies that the score- based interpretation is relevant to the decision being made.

Warrant 2: Utility

The utility is about the extent to which score-based interpretation provides information for making the right decisions and avoiding decision errors. Bachman (2005) explained this with multiple choices questions as an example, claiming that even though such test format is rarely utilized in real-life and only marginally relevant for a particular domain; it may be used to predict test-takers future performance.

Warrant 3: Intended Consequences:

Intended consequences have to do with the beneficial consequences of using the assessment and to what extent the decisions being made are of benefit to test takers and stakeholders. This warrant can incorporate some issues of test use such as test usefulness (Bachman & Palmer, 1996), fairness (Kunnan, 2003), Critical Language Testing (Shohamy, 2001) and ethics (Brown, 2012)

Warrant 4: Sufficiency

Sufficiency is about whether or not the assessment provides sufficient information for making decisions. These four types of warrants offer an argument structure which addresses the linkage between test score interpretation and utilization. However, Bachman and Palmer (2010) remodeled the framework by merging test score interpretation and utilization arguments with four articulating claims:

Claim 1: Assessment records are consistent

Claim 2: Interpretations are meaningful, impartial, generalizable, relevant and sufficient Claim 3: Decisions are valued sensitive and equitable

Claim 4: Consequences are beneficial

Each claim provides a list of warrants and rebuttals to be supported or refuted by evidence, which is useful for test users and researcher in test evaluation and test design. The AUA provides a useful, more thorough process for the evaluation of inferences concerning decisions and consequences with more emphasis on test use, by distinguishing the decisions and consequences (Im, Shin, & Cheng, 2019). The following figure illustrates a whole picture of all elements involved in the framework.

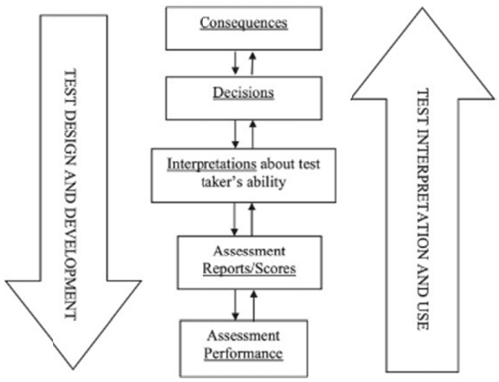

Figure 1. Inferential links from consequences to assessment performance or verse vise (Bachman & Palmer, 2010, p. 91)

To explore the consequences of the CET along with the test use and score interpretation, the current study stated three claims with a re-modelled Assessment Use Argument approach (Bachman, 2005; Bachman et al., 2010). These claims are about relevance, interpretation qualities including impartiality, generalizability and sufficiency and consequences. Relevance claim in this study differs with the one in AUA (2010), which is one quality under the interpretation claim. It is listed in this research as a separate claim because it is a fundamental function for test use. The three claims are listed as following with minor modifications for the context consideration.

Claim 1: Relevance

The characteristics of the CET tasks and performances they elicit correspond closely to the characteristics of the

tasks and the performances they required in the business domain.

Claim 2: Interpretation

The CET test scores demonstrate the test taker's performance are meaningful in the business context, and the use of scores is impartial to all test-takers, generalizable to job duties in the workplace and information derived from the CET scores are sufficient for the decisions to be made.

Claim 3: Consequences

The consequences of using CET and of the decisions that are made based on CET scores are generally beneficial to the test takers and the workplace.

2.2 Empirical Applications of AUA

Drawing on Bachman and Palmer's (2010) framework, Wang et al. (2012) reviewed the use of the Pearson Test of English Academic for making decisions on admission in the educational context. This research collected its data by documentation for evidence of claims about score interpretation, yet evidence had not been focused on the claims about decisions and consequences. Liu (2014) compared the scores of two tests to look for evidence of generalizability and meaningfulness interpretations. The findings showed that the use of the test score is not generalizable to the Target Language Domain (TLD). Bachman's (2005) assessment use argument has also been explicitly used in classroom assessment by Llosa (2008) who articulated claims, warrants, rebuttals and backing evidence to justify the link between scores on class assessment and the interpretations made about students' language ability. The findings of this research showed that this test measures English proficiency as defined by another test and support the rebuttals that teachers' scoring is inconsistent. Schmidgall (2017) adopted AUA to validate the use of TOIEC tests in his paper which highlighted a full implication of this approach. As for washback models, scholars (Alderson & Wall, 1993; Hughes, 1993; Green, 2013) had proposed different ones and had later been employed in empirical studies in a variety of regions, such as Sri Lanka (Alderson & Wall, 1993), Japan (Watanabe, 1996), Israel (Shohamy et al., 1996) and so on.

As for in Chinese context, Xu and Liu (2018) carried out a washback study on TEM and found that TEM brings positive impacts on teaching and learning from teachers' and students' perceptions. Xiao et al. (2011)'s study on washback of NMET showed a positive impact on students learning strategies and reading skills. With all these high-stakes tests, CET is the most widely investigated one. Researchers explored the impact and washback of CET from a variety of perspectives, such as teaching methods and teaching content (Gu, Yang, & Liu, 2013; Huang, 2002), students' learning outcomes (Li, Zhong, & Suen, 2012; Xiao, 2014), psychological effects on students such as test anxiety, self-efficiency suffer and so on (Jie, 2009), and changes of washback effects (Pang, 2017). Pan and Roever (2016) investigated the consequences of test use and their findings suggested that different stakeholders interpreted tests differently, and their perceptions on tests highlight one's attributes more than language abilities. Sun (2016) carried out a multi-phase, multi-method study to examine the intended and unintended consequences of the CET in both educational and social context. His research findings revealed the complexity of the test consequences and possible reasons causing this complexity. Some studies explored the use of test scores at the interface between study and work. For example, Knoch et al. (2016) explored the alignment between writing in the workplace and those of the IELTS writing. Their study shed lights on the development of workplace-specific writing skills, and a close connection between test tasks and workplace genres is possible to provide more meaningful standard setting. A similar test was conducted by Moore et al. (2015), who investigated both reading and writing tests by comparing them with literacy practices required in the workplace. This research raised the issue of having a test in effect two domains and the issue of adapting the test so that it has a strong relevance to tasks in the working area. These studies highlighted the significance of test use in social contexts.

2.3 Research Gap and Research Questions

As can be seen from the above-summarized studies, it can be concluded that most studies of Chinese large-scale and high-stakes tests limit the test effects in an educational context, which in short, are washback studies. Few studies address the impact of the CET in the social context, especially in the business domain. Moreover, what is needed in empirical research is evidence verifying which constructs are underrepresented and verification of the sources of irrelevant variance (Cheng & Deluca, 2011), yet very few addressing this issue. Previous studies generally refers to impact as positive or negative rather than considering these two factors as evidence for construct validity. Also, traditional washback models or their variants (Gu, 2007; Ren, 2011) instead of the validation approach were adopted for the consequences of large-scale high-stakes test.

To bridge the gap, this study is to investigate the consequences of test use to construct validity as evidence for the argument of interpretation and utilization in validation model proposed by Bachman and Palmer (2010). However,

it is worth noting that since the consequential study is comprehensive and systematic with a high number of factors to be taken into consideration, it is impractical and impossible to consider every aspect. Therefore, the justification of the test use and consequences of this use is the primary concern in this study.

Three research questions are proposed as followings:

Q1. To what extent do the CET tasks correspond to the performances required in workplaces that require CET for employment?

Q2. How are the CET and its test scores interpreted by students and employers? Q3. What are the consequences of the CET and its use in the business domain?

3. Methods

3.1 Participants

Participants in this study include four just graduated students who are employees or potential employees and four employers from different companies. These just-graduated students had learned English for 14 years on average (min=13, max=16), and were employed in fields ranging from education, Internet to real estate. All students have taken CET band 4, three of which have taken CET band six and two have taken CET-SET (speaking section). Employers were recruited from a range of companies and industries, from medium-sized workplace to large national or private companies. Their companies are based in Guangdong whose major international clients are from southeastern countries such as Singapore, India and Malaysia. English as a second language which is commonly used in email or reports writing, business negotiation at home and abroad and references reading. All employers are either in charge of recruiting or mentoring new graduates. Maximal variation sampling was used in this research. In recruiting student participants, the primary dimension concern was whether they have passed CET 4 and/or CET 6, the prestige of the university and whether they have taken.

CET-SEC (speaking) test. In recruiting employers, the nature of the companies they are from will be considered and to what extent English is used in their situation.

3.2 Instruments

The interviews were designed to be semi-structured in two rounds. The first round of interviewing covers a comprehensive area of questions and leaving with an interval which allows the researcher to supplement additional questions based on analysis of responses from the informants. The second round of interviews was to collect more focused data. Interview questions for the student participants focused on their background in English learning, English use in their context and CET test use, their perceptions on test design, while employers were asked about English language requirements in the workplace, their views on CET use, and test design and test-takers. These two sets of questions are aligned with each other for comparing the similarities and differences.

3.3 Data Collection

Data sets were collected in several processes. First, during the preparation stage, all interviewing questions were drafted by the researcher and later scrutinized by an expert in assessment filed for modification. And those questions were translated into the Chinese language. The researcher prepared a CET-4 and CET-6 test papers with oral and written tasks presented on. The whole process of interviews contains three procedures: the pilot study, first-round interview and second-round interview. A pilot study was first conducted with two participants, one student and one employer, for the revision of interview protocols before the first-round interviews being administrated to actual participants. Then the first-round interviews were taken individually, lasting around 50 minutes to one hour. Interviews with each participant were recorded along with field notes taken on papers. All interviews were conducted in the Chinese language to ensure smooth communication and to avoid any potential confusions. After data analysis, second-round interviews were carried out for further focused exploration. Only those whose responses needed an in-depth interpretation were involved in the second round.

3.4 Data Analysis

All interview recording data were transcribed and the statements are transformed into a formal and written style, repetitions, and meaningless expressions were removed. The transcripts were read carefully and repeatedly to ensure reliability. Inductive coding was utilized for data analysis. In terms of the approach to the analysis of interview texts, it was mainly focused on the meaning. Interview transcripts were subjected to meaning coding. Main categories and sub- categories were constructed by extracting the themes that emerged from the transcripts. Qualitative computer software programs were not employed due to the small size ofinterview data (with 48 pages only). Fragments with the same code was compared and referred to the transcripts to explore more details.

4. Results

4.1 Relevance Claim

To explore the first question of correspondence between the CET test tasks and performances that required in the workplace, relevance claim are stated as:

Claim 1: The characteristics of the CET tasks and performances they elicit correspond closely to the characteristics of the tasks and the performances they required in the business domain.

4.1.1 Test Characteristics

Test characteristics are analyzed from perspectives of task content and task format. In terms of reading, its passages cover a wide range of topics and background knowledge related to natural science, technology, economy, biography, culture and communication and common knowledge. With respects to text language, this research adopted Flesch Reading Ease[1] measurement tool to calculate the level of language difficulty. Even though the difficulty of each passage in CET reading cannot be merely measured by vocabulary and sentence complexity, it is useful to evaluate candidates' capability in comprehension to a certain level, especially to estimate whether they reach the level of reading academic materials. The formula may not be highly accurate, yet it can be "used as a rough guide, however, scores derived from readability formulas provide quick, easy help in the analysis and placement of educational material" (Klare et al., 1969). The average score of all passages which randomly selected from CET-4 and CET- 6 past test paper is 57.7 and 49.1 separately, according to Jin (2006)'s calculation. Readability of IELTS is 49.50 and of TOFEL is 44.85 (Li, 2018). The numbers are matched with numbers in scales of readability to evaluate its difficulty (See Table 1). The figure reflects readability of the CET test candidates who have acquired the CET report above the cut-score.

Table 1. Flesch Reading Ease (FRE) formulas

Score	Notes
90-100	Very easy to read, easily understood by an average 11-year-old student
80-90	Easy to read
70-80	Fairly easy to read
60-70	Easily understood by 13- to 15-year-old student
50-60	Fairly difficult to read
30-50	Difficult to read, best understood by college graduates
0-30	Very difficult to read, best understood by university graduates

The CET assesses a variety of reading skills and strategies in the deployment. This research applied Urquhart and Weir's (2014) model which further divided reading skills into careful reading and expeditious reading. What is required in the CET includes expeditious reading skills such as skimming and scanning, and careful reading skills like understanding contextual meaning, making proper inferences, deducing the meaning of unfamiliar lexical items and so on. In addition, the rate of reading is a request in the CET reading section. According to National College English Testing Committee (2006a, 2006b), the speed rate of reading should be as high as 70 wpm[2] for careful reading and 100 wpm for speed reading in CET-4; 90 wpm for careful reading and 120 wpm for speed reading in CET-6.

Writing section in a broader sense contains two tasks: composition writing and translation from Chinese to English. In composition writing assessment, test candidates are required to complete an essay task within the allotted time, usually 30 minutes. This research adopts Purves et al. (1984) and Hale et al. (1996)'s dimensions of tasks to clarify the measurement of CET writing. The subject matter of writing piece concerning personal experiences, affections, emotions and events. Test candidates are expected to deal with rhetorical tasks such as description or exposition and argument. The cognitive demands in the CET writing section include organizing information, reproducing clear ideas and facts, analyzing and evaluating. The length requirement is less than 120 words in CET-4 and less than 150 words in CET-6. Genres include but not limited to essay, letter and informal note etc.

CET-SET spoken test engage test takers "in several monologic and interactive tasks" (Jin, 2011) such as self-introduction, short essay reading, short answer, personal statement and interaction with another student (Band 4); and self-introduction, personal statement and discussion, and question and answer (Band 6) (See Table 2).

Table 2. Sections of CET-SET spoken test (adapted from https://wenr.wes.org/)

Sections	Tasks	Process	Time
Part 1	Self-introduction	Self-introduction	20 seconds
Part 2	Short Essay Reading	Short essay with about 120 words	1 minute +45 seconds preparation
Part 3	Short Answers	Two questions related to the short essay	40 seconds
Part 4	Personal Statement	Personal Statement	1 minute+45 seconds preparation
Part 5	Interaction with another student	Converse according to give prompts	3 minutes+1 minute preparation

4.1.2 Tasks in Business Domain

In the business context, employees are required to read a wide range of documents. Reading skills and strategies vary with the types of these documents. For example, in reading professional reports online, expeditious skills such as skimming and scanning skills, and rate of reading speed are required to identify specific information that may meet the needs of writing a summarized report within a limited time. On the contrary, contracts or legal documents needing to be read very carefully with little constrains on time issues, especially for identifying clients' needs and significant items, numbers and graphs etc. Professional dragons and terminology in these genres may contribute to the difficulty level of reading materials.

> *Our job relates to sales management and guidance; usually, the length of reports or contracts will not be too long. However, it has a high request for employees to be very cautious on the details of the items in contexts, like numbers and figures. (Employer 4)*

In addition, reading tasks are tied closely to productive skills. The most conducted written discourse in companies is emails either from internal or external sources. Internal emails are usually informal and low stakes correspondence, while external ones may lay more emphasis on the format and formal language usage. Employer informants explained that emails as a means of notification or formal discussion, serving a range of functions. Therefore, such kind of writing requires a particular tone, and it accounts for its audiences and purposes. Another type of genre which has been widely conducted is reports which include literature review (e.g. Employer 1), contract items (e.g. Employer 3), project plans and presentations (e.g. Employer 2 and Employer 3) in forms of either written papers or visual PowerPoint slides.

> *As for writing emails, usually the length is no more than 200 words for efficiency. Thus it requires graduate students to express relevant ideas clearly, even though minor grammatical errors or misusage of vocabulary occurs from time to time. What matters is the format of emails. Many just-graduate students have no idea of how to address salutations properly. (Employer 2)*

It can be concluded that, writing is expected to be more contextual-appropriate than linguistic-appropriate. Reading and responding to emails or reflecting by writing makes this combined activity both communicative and integrative. Because information decoded in reading will be transformed and reorganized into new ideas, thoughts and expressions, which requires higher-ordered cognitive strategies. Another skill of production is speaking, the tasks of which are available in several settings, such as job interview, presentation to co-workers and business partners or clients. These conversational activities imply either planned and rehearsed or even scripted monologues or impromptu casual dialogues. Speaking can be further categorized into talks of different types with regards to its functions. Speaking activities in business can be referred to description, narration, instruction and comparison and evaluative talk (Bygate, 1987), with its function of explanation, justification, prediction and decision. For example, introducing a product to clients may involve narration, description or comparison talks. Selling products to clients contains explanation and justification talks, while conferences are likely to have prediction and decision talks.

4.1.3 Warrant: Instrument Task(s) Being Similar to Task(s) in Business Domain

Before the review of test papers and constructs of assessment tasks presented and clarified by the researcher, all participants reached a consensus that there is no similarity between tasks in the CET and those in commercial jobs, considering the test format and content. However, after these informants were explained with information about test constructs, scoring criterions and so on later, they agreed that there are some similarities between what needed for job qualities and what assessed in the test.

First, skills, strategies and competences that candidates perform in assessment are similar to what they may utilize in future jobs. For example, reading skills measured in the test such as skimming and scanning and the emphasis on speed rate is of vital importance to achieve efficiency for reading tasks in the workplace. In CET-SET, some

speaking strategies such as justification and description are likely to occur to reach a logical and fluent speech. Similarly, those strategies are required in commercial talks like introducing/selling products to global clients (Employer 4), seeking for cooperative opportunities, or building close relationships with business partners (Employer 2) etc.

> *We need those just-graduate students to have a good command of communication skills when they converse with our clients. They are expected to learn to persuade clients to buy the products and justify themselves. (Employer 4)*
>
> *Careful reading is an important reading skill because we have to deal with contract, which requires employees to pay attention to details. (Employer 1)*

Second, some tasks in the CET-SET are authentic tasks in real world, such as self- introduction, dialogue, questions and answers, and presentation. Those tasks are often carried out in job interviews, business negotiations, and some meetings or conferences. One informant (Employer 3) mentioned that the length of CET-6 reading passages is similar to that of reading materials they need to deal with in translating.

> *Individual presentation is a big part either in internal meetings with staff members or external ones with clients. One needs to present their reports on a project or introducing our products to clients by sufficient details. (Employer 2)*

4.1.4 Instrument Task(s) Being Different to Task (s) in Business Domain

First and foremost, the major difference is that CET tasks stress its importance on receptive skills which account for 70% proportion of total scores. However, in daily jobs, speaking and writing skills as productive skills are more demanding for employees.

Second, the CET reading covers a wide range of topics and background information, while reading in the workplaces usually pointing to one specified field, such as financial, accounting, market analysis or sales strategies. Slim chances are there that employees have to deal with materials relating to different disciplines.

Third, some informants suggested that some writing tasks in assessment are not authentic tasks considering its design. The informants were presented with a writing task in 2018 test paper and one informant comment as following.

> *In the old days, the tone of letters to friends and families is quite formal, while emails and texts could be casual and informal. Even if we need to recommend Chinese-learning schools to our friends by letter. It is less likely that we would use formal languages because we are not experts or Chinese teachers. (Employer 2)*

Another aspect of difference in writing pointed out by the informant is the tone and formality of language use. In business context, there is always the intended audience for the writer to consider before writing. Because writing is 'an act that takes place within a context, that accomplishes a particular purpose, and that is appropriately shaped for its intended audience' (Hamp-Lyons & Kroll, 1997, p. 8). Yet 'writing to a particular audience' has not been included as one grading factor in the CET writing scoring criteria (See Table 3).

Table 3. Scoring criterion for writing

Score level	Description
14	Relevant to the subject. Clear and coherent; no linguistic mistakes in general; with one or two errors
11	Relevant to the subject, clear and coherent; with some linguistic errors
8	Almost relevant to the subject, some ideas are not well-expressed; almost coherent; some are major mistakes
5	Almost relevant to the subject, unclear ideas, not coherent, with many severe mistakes
2	Illogical, unconnected ideas, most sentences are wrongly expressed, with many severe mistakes

4.2 Interpretation Claim

Claim 2: The CET test scores demonstrate the use of scores is impartial to all test-takers, generalizable to job duties in the workplace and information derived from the CET scores is sufficient for the decisions to be made.

4.2.1 Impartiality

Warrant 1: The CET tasks include task formats or content that are impartial to all groups of test candidates.

Evidence: It has been clearly stated in the test speculation that CET test content covers a wide range of topics and

knowledge which is not in favour of any test-takers from different disciplines and cultural backgrounds. One employer (Employer 3) commented that the writing task and translation task is suitable for college students in either liberal arts or science majors, giving that topics are general and familiar to them.

Among the facets of test methods, there is a consistency in CET test considering test items, test rubric, item types, test environment and the nature of test input. Two informants referred to the multiple-choice items, the largest proportion of CET's item format in reading and listening sections, somehow guarantees a sense of impartiality. And for Chinese students who had gone through the National Matriculation Tests, the test format of multiple-choice questions is consistent with the cognitive strategy that test-takers used to apply. Multiple-choice items also ensure the objectivity of scoring and consistency of score interpretation.

One student participant stated that neither test content nor test format had changed much when he practiced on the CET past papers. Therefore, the consistency of test item types might eliminate the bias of one group of test-takers over another.

> *In IELTS reading section, there are different types of tasks, such as blanking filling, Yes/No/Not givenand so on; while in the CET, those items and rubrics do not change much. (Employer 2)*

> *Even though multiple choices are not what real-life tasks look like, it measures our ability to some extent. In addition, we do not have to worry about other factors that may affect our test performance, such as bad handwriting. (Student 1)*

Employer 4 believes that the policy of speaking test being non-obligatory is fair to students from impoverished areas of China, where English teaching and learning sources are scarce.

Rebuttal 1: Construct-irrelevant factors may reduce the impartiality to some extent

According to McNamara and Ryan (2011), fairness/impartiality is related to the psychometric qualities of the test. For the essential fairness of tests, there are two major concerns, one of which is construct-irrelevant factors, such as any particular tasks sets, raters' characteristics, noise distractions and so on. One student participant pointed out that score results of the CET-SET may be affected by the performance of the speaker's partner, which can be regarded as one of construction-irrelevant factor, leading to unfairness of the test.

> *My roommate told me that she could not perform well when her partner being too nervous to speak out anything during the CET-SET test. It is not fair to my roommate since it is not the fault of hers. (Student 3)*

Warrant 2: Employers perceive the CET report holders with the same cut score--above 425 equally.

Evidence: For employers in this research, the value of the CET use in the recruitment process is that it makes the employment objective and just for all applicants due to the authority, quality and high credibility of the test. Such value has also been affected by the enforcing policy of graduation in the past. Examinations continue to enjoy wide societal acceptance and recognition in China as a fair measurement for selection of the best talent into the social hierarchy (Cheng & Qi, 2006). Findings of CET in this research are consistent with this claim.

> *We have no other optional test in China to ensure fairness when screening out the candidates with regards to their language proficiency. (Employer 4)*

The specific score of the CET in each section will not be taken into consideration, as suggested by one informant. Employers will only consider the total score. Therefore, the standard for selecting employees guarantees a sense of equality. One informant mentioned, "We do not refer to the test scores of reading or listening sections, as long as the candidates pass the exam".

4.2.2 Generalizability

Warrant: Skills, strategies and competences are generalizable to those are needed in the business domain.

Evidence: Except skills measured in the CET being generalizable to what needed in the domain, student groups mentioned that some strategies that they employed during test preparation are also applicable in the workplaces.

> *I think the reading strategy of locating key items that I used in CET is generalizable to other reading comprehension tasks. That's the beneficial part of CET for me. (Student 4)*

Some employers reckoned the CET-6 report is not only a proof of one English language proficiency, but also an indicator of one's personality attributes such as diligence and problem- solving abilities.

4.2.3 Sufficiency

Warrant 1: The CET report provides sufficient information for screening function in the workplace.

Evidence: This warrant can be backed up with two pieces of evidence. First, the score of CET is sufficient to demonstrate one's English language proficiency. Most participants hold the belief that the CET report is proof of one's English language ability. The CET testing guideline contains a list of over 5,000 vocabulary words. "One has to grasp a certain amount of vocabulary words to pass the CET-6", as suggested by one employer (Employer 2). For some company divisions where jobs are more relevant with reading skills, CET certificate holders are thought to be qualified in handling their daily jobs. Some students argue that even though some test candidates may obtain scores by luck when dealing with multiple-choice questions, it is less likely that all the responses they give rely on that. This reflected that construct irrelevance variables are less likely to threat the general quality and validity of the test. Second, the information provided by the test is sufficient in making selection decisions based on the nature of CET, for example, it is a high-stake test with high acceptance in the job- hunting market. The fact that such a large-scale test has been developing steadily in the past 17 years is in itself substantial evidence to show that the CET has met social needs, won social recognition, produced beneficial effects on society and contributed significantly to the continual improvement of the quality of College English teaching in China (Wu, 2005).

Rebuttal 1: The score of CET is insufficient to demonstrate one's English language proficiency. Evidence: The fact that CET-SET is not a required test makes the whole CET test system insufficient for indicating one's language abilities. English is in primary for communication. One employer from the oversea division pointed out that some job applicants with CET high- score report are unable to produce utterances fluently and smoothly. It is important to note that CET-SET does not enjoy credibility as much high as the CET-4 and CET-6. Two employer participants reported that they did not even know this test and had no idea of its scoring criteria. Another insufficiency evidence has to do with writing tasks in the business domain. There is a variety of genre types in real-life work domain as discussed-above which are not measured in the CET. For example, language features for producing a market report could be complexing and involve many terminologies. But the writing tasks in the CET are general. In addition, what is required about writing strategies in jobs is the ability to synthesize and summarizing. Writing and reading are closely tied together, resulting in the output production mainly relying on the input. The CET separately measures reading and writing.

4.3 Consequential Claim

Claim 3: The consequences of using CET and of the decisions that are made on the basis of CET scores are generally beneficial or nonbeneficial to stakeholders.

4.3.1 Warrant: Beneficial Consequences for Stakeholders

This section deals with benefits that the CET may bring about to the employers and just- graduate groups. First, for the employers, the fact that the CET report being used as selection criterion for recruitment enhanced screening efficiency in English language proficiency to a large extent. Most companies need not to devise an extra test for language evaluation on candidates. Second, the CET report also affects employers' decisions in many ways. All informants hold the view that the CET-4 demonstrates not only candidates' English proficiency, but also their personality attributes. Two employers suggested that they would not be bothered to interview students whose CET-4 score is lower than 425. Because these students are considered not hard- working enough at college (Employer 1). Comparatively speaking, the CET-6 enjoys a higher standard than the CET-4 which refers to a general standard. Such difference affects the employers' perception on the students' ability. One employer claimed that they correlate the CET-6 with the level of Chinese universities.

> *If a student graduates from a top university in China, he is expected to pass CET-6. If not, he will be considered less hard-working than those students who passed the CET-6. (Employer 4)*

For the student participants, the CET is influential positively on their language knowledge and language skills during their test preparation stage. Three students reported that they are able to understand non-linguistic knowledge from a variety of disciplines when practicing on past test papers.

> *The various topics in CET reading is helpful for me because I have accumulated different background information about a variety of subjects, which I barely have any chances to read in my own major study. (Student 1)*

According to the schema theory (Ortony & Rumelhart, 1977; Rumelhart, 1980), comprehending a text is an interactive process between the reader's background knowledge and the text. Many empirical studies in Chinese context have demonstrated that background information helps to cultivate students interests and quicken their reading speed and fulfill the tasks more successfully (Che, 2014; Zhao & Zhu, 2012). Consequently, it may appear to be easier for these test candidates to deal with reading or listening materials of various background information. Furthermore, some instrumental tasks prepare them to be qualified for a particular position, such as the writing

patterns (Student 2) and speaking strategies (Student 4) they accumulated.

CET also brings some positive impacts on students in terms of motivation, confidence and self-fulfillment. One student reported that in order to get a higher score of the test so as to leave a good impression to future employers, she signed up CET-6 for a second time, even though she had passed the test at first time. All participants who obtained CET-6 report with score as high as 425 or above mentioned that they are more confident in job hunting and other test preparation. Though the CET has been criticized for inducing students to focus their English learning on the test (Cai, 2005; Han et al., 2004), nearly all participants in this research acknowledged that CET is a driven force for the improvement of their language proficiency in a short period of time and helped to develop certain learning strategies that applicable and generalizable in their future life.

4.3.2 Rebuttal: Nonbeneficial Impact on Stakeholders

The CET also proved to be nonbeneficial to employers and students in some aspect based on evidence provided by this research. Since CET-SET is an optional and separate test, many job applicants didn't take the test, making employers unable to predict their English oral ability. Even if some students hold the CET-SET certificate, neither their English-speaking abilities, nor will they are given any advantages during the interview stage. Consequently, overlapping assessments are arranged for English-related positions, such as oral tests, including questions and answers, self-introduction or presentation. The small proportion of writing section are regarded as insufficiency to test students writing ability for some employers (Employer 2 and Employer 3), they mentioned that sometimes, a writing task with higher requirements was given to the job applicants.

The CET may bring negative impacts on students' psychological aspects. Some participants mentioned that this test made them in the state of worrisome and nervousness, especially when their peers have passed the exam (Student 3) or when they take it for a second or third time. Much research findings have shown that test anxiety may interfere with a person's academic achievement (Cassady & Johnson, 2002; Jing, 2007; DordiNejad et al., 2011). Test anxiety also produces certain aversive patterns of motivation, coping and task strategies that interfere with learning and performance, resulting in competence and self-efficacy suffer, thus leading to further anxiety over time and generating a vicious circle of increasing anxiety and degrading competence (Zeidner, 2010). Those who did not hold CET-4 certificate reported that they are anxious and not confident enough while hunting for jobs.

> *One of my roommates in college spent most of her college time preparing for this test and under enormous pressure. She suffered from high anxiety, insomnia and lack of self- confidence from time to time. Even if she eventually passed CET-4, she had experienced so many psychological sufferings. (Employer 1)*

In a word, the CET test and its report bring both positive and negative impacts on employers and just-graduate students in the workplaces at the practical and psychological level. Even though, CET can still be used for screening in the business domain. The reasons are attributable to the close relevance between test tasks and performance required in the workplace, and the use of CET test is interpreted impartially to potential employees, generalizable to job requirements and sufficient for making selection decision to a certain. These results are used as the supporting evidence for the warrants and claims stated to clarify the appropriateness of screening use.

5. Discussion

The result showed in rebuttals that a few informants believed some tasks are authentic based on the test content while other tasks are not authentic due to the test methods. This finding correlates to Bachman and Palmer's (1996) notion that perceptions on test authenticity are varied with different stakeholders. Thus, to achieve authenticity, it requires ongoing validation and cooperation between different stakeholders. The reason that stakeholders perceive differently before and after they were explained with test constructs could be attributed to the lack of assessment literacy.

Authenticity in many studies is regarded as a core element for the content validity and reliability in the validation process of a test (Wu & Stansfield, 2001). While in this study, it is regarded as an argument to justify the screening use of the CET test. Such analysis is an essential step in this framework as the similarity somehow decided to what extent the CET and its score can predict one's ability to handle the demands of the target domain. Under the Bachman and Palmer's framework (2005) of validation, this argument justified that using the CET score are considered strictly relevant to the selection decisions being made by the employers. Informants reported in this research that some skills or competencies that they developed during the test preparation stage are applicable to the workplace. Moreover, the CET report is not only an indicator of one's English proficiency but also reveals implicitly test-takers personal attributes. This finding correspondence to Sun (2016)'s results of extended interpretation of the CET score, which represents students' efforts rather than their English proficiency.

Whether the CET is sufficient to demonstrate one's language ability vary with what is being required in a specific context. For example, when speaking is composed of most job duties, the CET is not sufficient to provide enough information as a speaking test is not an obligatory section. On the contrary, when reading skills are needed in the business domain, what the test presented is perceived to be sufficient for employers making appropriate decisions.

This finding concerned with the issue of fairness which is a particularly acute issue in language testing because of its high-stakes impact on test-takers is often made based language test scores (McNamara & Ryan, 2011). Fairness is taken as a facet of validity in previous studies (Kunnan, 2004; Xi, 2010), while in this study it is addressed in specific warrants of impartiality and sufficiency throughout the Assessment Use Argument. It is fair to use the test as a screening paradigm based on the analysis of impartiality and sufficiency.

To what extent the CET being used in the commercial area for decision making depends on the consequences it may bring to stakeholders. Comparatively speaking, the CET-4 enjoys higher credibility compared with the CET-6. Nearly every employer in this research refers to the CET-4 report as an indicator of one language ability as it represents a general standard which matches with tertiary English educational level. It is thought to prepare students well in topical knowledge literacy to a certain extent. This finding corresponds to previous studies about the positive impact of CET on students (Cheng et al., 2011).

6. Conclusion

In this study, the Assessment Use Argument (AUA) was employed for the purpose of reviewing the use of the CET and its consequences in the business domain where English as a second language is used for communicative purposes. Emphasis was laid on the actual use in social settings and obtains perceptions from employers as one party of stakeholders. Three claims associated with warrants are articulated with supporting evidence to justify the argument that CET can be used for employment selection. In a word, there is a high degree of correspondence between instrumental tasks and real-life tasks in the commercial domain in terms of the implicit skills and abilities that the candidates may apply. The finding is associated with the issue of authenticity, which is a fundamental element for making selection decisions with the CET. Responses to the second question illustrated that the use of CET scores for screening is impartial to all candidates, and what measured in the test are generalizable to the performance required in the workplace and scores are sufficient for making selection decisions. This interpretation concerns with the problem of test fairness as a quality of test validity. The third question relates to consequences of test use for selection decisions; it explores whether the consequences are in general beneficial or nonbeneficial to stakeholders. The results show that CET used as a selection criterion saves time and efforts for employers in evaluating job candidates' language proficiency and somehow personal attributes. And the CET well equips candidates with knowledge literacy to some extent and enhance certificate-holders' confidence during the job-hunting period. Admittedly, not all consequences are beneficial to employers and students. The lack of an obligatory spoken test in CET leads to overlapping assessments devised by companies. For those who fail to obtain CET certificate, they are likely to suffer from anxiety, worrisome and lack of confidence in the employment.

6.1 Implications for Test Design

To achieve authenticity in future test design, there seems to appear two options for meeting the needs of employers as test users in the business domain. One is to convert the CET-6 into several different tests pointing to different purposes such as CET-Academic, CET-Business, and CET-General, etc. This option is close to the notion of Test language for Specific Purpose, or LSP, which "refers to that branch of language testing in which the test content and test methods are derived from an analysis of a specific language use situation (Douglas, 2000, p. 1). Another option is to develop a new professional-oriented test which has a higher standard in English proficiency and more professional literacy than that of the CET-4 or CET-6. Both options are expected to be a useful complement to the current test system and are likely to bring positive impacts on test users. No matter which tests to opt, the authenticity of tasks is an important issue to address in the first place. As for the skills being measured, more focus should be placed on developing integrated tasks. The findings of this research show that the misuse of the CET-SET may attribute to the users' lack of assessment literacy. Therefore, test designer should shoulder the responsibility of better informing test users about "the dimensions and aspects of how the test results can be used most appropriately" (Yan, 2008).

6.2 Implications for Future Research

For future research, first, more employers as test stakeholders from a diverse range of industries (e.g., accounting, engineering, banking, retailing sales, hospitality) should be involved in the investigation. With more sampling, an in-depth analysis of interpretations and its underlying values may be possible to explore. Second, observation as a qualitative instrument should be adopted in the research if possible. For example, the researchers may observe actual business negotiating sessions and conferences, or audio record the use of English in different situations such

as one-on-one interview, business talks over meals, and so on. During the observation, more specific information could be provided for the research about the characteristics of the language they use and the tasks they must perform in English in different situations.

References

Abelson, R., & Schank, R. C. (1977). Scripts, plans, goals and understanding. *An inquiry into human knowledge structures New Jersey, 10*.

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

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Ernst Klett Sprachen.

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

Bachman, L. F. (2003). Constructing an assessment use argument and supporting claims about test taker-assessment task interactions in evidence-centered assessment design. *Measurement: Interdisciplinary Research and Perspective1*, 63-65.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press Oxford.

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

Brown, A. (2012). Ethics in language testing and assessment. *The Cambridge guide to second language assessment*, 113-121.

Brown, J. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Bygate, M. (1987). *Speaking*. Oxford University Press.

Cai, J. (2005). Validity, reliability and practicality of computer-based oral proficiency test. *Foreign Language World, 26*(4), 66-75.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary educational psychology, 27*(2), 270-295. https://doi.org/10.1037/t13906-000

Che, Y. (2014). A study on the application of schema theory to English newspaper reading. *Theory and Practice in Language Studies, 4*(2), 441-445. https://doi.org/10.4304/tpls.4.2.441-445

Cheng, L. (2008). Washback, impact and consequences. *Encyclopedia of language and education, 7*, 349-364. https://doi.org/10.1007/978-0-387-30424-3_186

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational assessment, 16*(2), 104-122. https://doi.org/10.1080/10627197.2011.584042

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational assessment, 16*(2), 104-122. https://doi.org/10.1080/10627197.2011.584042

Cheng, L., & Qi, L. (2006). Description and examination of the national matriculation English test. *Language Assessment Quarterly: An International Journal, 3*(1), 53-70. https://doi.org/10.1207/s15434311laq0301_4

Cronbach, L. J., & Thorndike, R. L. (1971). Educational measurement. *Test validation*, 443-507.

DordiNejad, F. G., Hakimi, H., Ashouri, M., Dehghani, M., Zeinali, Z., Daghighi, M. S., & Bahrami, N. (2011). On the relationship between test anxiety and academic performance. *Procedia-Social and Behavioral Sciences, 15*, 3774-3778. https://doi.org/10.1016/j.sbspro.2011.04.372

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.

https://doi.org/10.1017/CBO9780511732911

Green, A. (2013). *Exploring language assessment and testing: Language in action*. Routledge. https://doi.org/10.4324/9781315889627

Green, A. (2013). Washback in language assessment. *International Journal of English Studies, 13*(2), 39-51. https://doi.org/10.6018/ijes.13.2.185891

Gu, X. (2007). *Positive or negative: An empirical study of CET washback*. Chongqing: Chongqing University Press.

Gu, X. (2007). The empirical study of CET washback on college English teaching and learning in China. *Journal of Chong Qing University (Social Science Edition), 13*(4), 119-125.

Gu, X., Yang, Z., & Liu, X. (2013). A longitudinal study of the washback of CET on College English Classroom Teaching and Learning—Three College English Teachers' Classrooms Revisited. *Foreign Language Testing and Teaching, 1*, 18-29. https://doi.org/10.1515/cjal-2013-0021

Hale, S., & Myerson, J. (1996). Experimental evidence for differential slowing in the lexical and nonlexical domains. *Aging, Neuropsychology, and Cognition, 3*(2), 154-165. https://doi.org/10.1080/13825589608256621

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

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL (2000)–Writing: Composition, community, and assessment* (TOEFL Monograph Series Report N. 5). Princeton, NJ: Educational Testing Service. In: Weigle, SC (2002). Assessing writing. Cambridge: Cambridge University Press.

Han, B., Dai, M., & Yang, L. (2004). A survey on CET: Practices and problems. *Foreign Languages and Their Teaching (waiyu jiaoxue), 2*, 17-23.

Huang, D. (2002). *A preliminary investigation of CET-4 washback*. Paper presented at the International Conference on Language Testing and Language Teaching, Shanghai: Shanghai Jiao Tong University.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219. https://doi.org/10.1007/s11205-011-9843-4

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

Im, G.-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9*(1), 14. https://doi.org/10.1186/s40468-019-0089-4

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385-402. https://doi.org/10.1177/0265532208090158

Jin, Y. (2006). On the improvement of test validity and test washback—The CET washback study [J]. Foreign Lang. *World, 6*, 65-73.

Jin, Y. (2011). Fundamental concerns in high-stakes language testing: The case of the college English test. *Journal of Pan-Pacific Association of Applied Linguistics, 15*(2), 71-83.

Jin, Y., & Yang, H. (2018). Taking the road of language testing with Chinese characteristics: Enlightenment from the CET-4 and CET-6 exams for 30 years. *Foreign Language World, 2*, 29-39.

Kane, M. (2002). Validating high-stakes testing programs. *Educational measurement: Issues and practice, 21*(1), 31-41. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2006). Validation. *Educational measurement, 4*(2), 17-64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Klare, G. R., Rowe, P. P., St. John, M. G., & Stolurow, L. M. (1969). Automation of the Flesch reading ease readability formula, with various options. *Reading research quarterly*, 550-559. https://doi.org/10.2307/747070

Knoch, U., May, L., Macqueen, S. S., Pill, J., & Storch, N. (2016). *Transitioning from university to the workplace: Stakeholder perceptions of academic and professional writing demands*.

Kunnan, A. (2003). *Fairness and ethics in language assessment*. Course readings: TESL 567A. In Los Angeles,

California State University.

Kunnan, A. J. (2004). Test fairness. *European language testing in a global context*, 27-48.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. A Teacher's Book.

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

Li, J. (2019). *Analysis on Fairness of CET-4 Passage Reading Based on the Validity*. Paper presented at the 4th International Conference on Humanities Science, Management and Education Technology (HSMET 2019). https://doi.org/10.2991/hsmet-19.2019.142

Li, L. (2014). A Data-Based Investigation into Reliability and Validity of Computer-Assisted Oral English Test. *Applied mechanics and materials, 543-547*, 4494-4497. https://doi.org/10.4028/www.scientific.net/AMM.543-547.4494

Li, Y. (2018). A Comparison of TOEFL iBT and IELTS Reading Tests. *Open Journal of Social Sciences, 6*(08), 283. https://doi.org/10.4236/jss.2018.68023

Liu, F., & Stapleton, P. (2014). Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System, 45*, 117-128. https://doi.org/10.1016/j.system.2014.05.005

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational measurement: Issues and practice, 27*(3), 32-42. https://doi.org/10.1111/j.1745-3992.2008.00126.x

McCall, W. A. (1922). *How to measure in education*. Macmillan. https://doi.org/10.1037/13551-000

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*(2), 161-178. https://doi.org/10.1080/15434303.2011.565438

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives, 1*(1), 3-62. https://doi.org/10.1207/S15366359MEA0101_02

Moore, T., Morton, J., Hall, D., & Wallis, C. (2015). Literacy practices in the professional workplace: Implications for the IELTS reading and writing tests. *IELTS Research Reports Online Series*, 46.

National Academies of Sciences, Engineering, and Medicine 2020. *A Principled Approach to Language Assessment: Considerations for the U.S. Foreign Service Institute*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25748.

National College English Testing Committee. (2006a). *CET-4 Test Syllabus and Sample Test*s (Revised edition). Shanghai: Shanghai Foreign Language Education Press.

National College English Testing Committee. (2006b). *CET-6 Test Syllabus and Sample Tests* (Revised edition). Shanghai: Shanghai Foreign Language Education Press.

Ortony, A., & Rumelhart, D. E. (1977). The representation of knowledge in memory. *Schooling and the acquisition of knowledge*, 99-135. https://doi.org/10.4324/9781315271644-10

Pan, Y.-C., & Roever, C. (2016). Consequences of test use: A case study of employers' voice on the social impact of English certification exit requirements in Taiwan. *Language Testing in Asia, 6*(1), 1-21. https://doi.org/10.1186/s40468-016-0029-5

Pang, Y. (2017). An Investigation into Longitudinal CET Washback from Teachers' Perspective. Paper presented at the *2017 3rd International Conference on Social Science and Technology Education*. https://doi.org/10.12783/dtssehs/icsste2017/9368

Purves, A. C., Söter, A., Takala, S., & Vähäpassi, A. (1984). Towards a domain-referenced system for classifying composition assignments. *Research in the Teaching of English*, 385-416.

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

https://doi.org/10.5172/ijpl.2011.6.3.243

Rumelhart, D. E. (1980). *On evaluating story grammars*. https://doi.org/10.1207/s15516709cog0403_5

Schmidgall, J. E. (2017). Articulating and evaluating validity arguments for the TOEIC® tests. *ETS Research Report Series, 2017*(1), 1-9. https://doi.org/10.1002/ets2.12182

Shohamy, E. (2001). Democratic assessment as an alternative. *Language testing, 18*(4), 373-391. https://doi.org/10.1177/026553220101800404

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

Sun, Y. (2016). *Context, construct, and consequences: Washback of the College English Test in China*.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press. https://doi.org/10.1017/CBO9780511840005

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language testing, 29*(4), 603-619. https://doi.org/10.1177/0265532212448619

Wang, J. (2011). A Study of the Role of the'teacher Factor'in Washback.

Watanabe, Y. (1996). Investigating washback in Japanese EFL classrooms: Problems of methodology. *Australian Review of Applied Linguistics. Supplement Series, 13*(1), 208-239. https://doi.org/10.1075/aralss.13.09wat

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan, 10, 9780230514577. https://doi.org/10.1057/9780230514577

Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language testing, 18*(2), 187-206. https://doi.org/10.1177/026553220101800205

Wu, W. S. (2005). Using blogs in an EFL writing class. Paper presented at the *international conference on TEFL and applied linguistics*.

Xi, X. (2010). How do we go about investigating test fairness? *Language testing, 27*(2), 147-170. https://doi.org/10.1177/0265532209349465

Xiao, W. (2014). The Intensity and Direction of CET Washback on Chinese College Students' Test-taking Strategy Use. *Theory & practice in language studies, 4*(6). https://doi.org/10.4304/tpls.4.6.1171-1177

Xiao, Y., Sharpling, G., & Liu, H. (2011). Washback of national matriculation English test on students' learning in the Chinese secondary school context. *Asian EFL Journal, 13*(3), 103-129.

Xu, Q., & Liu, J. (2018). *A study on the washback effects of the Test for English Majors (TEM)*. Springer. https://doi.org/10.1007/978-981-13-1963-1

Zeidner, M. (2010). *Test Anxiety* (Vol. 3). https://doi.org/10.1002/9780470479216.corpsy0984

Zhao, X., & Zhu, L. (2012). Schema Theory and College English Reading Teaching. *English Language Teaching, 5*(11), 111-117. https://doi.org/10.5539/elt.v5n11p111

Zhu, D. (2017). Analysis of the application of artificial intelligence in college English teaching. Paper presented at the *2017 2nd International Conference on Control, Automation and Artificial Intelligence* (CAAI 2017). https://doi.org/10.2991/caai-17.2017.52