An Investigation into the Consequential Validity of a Diagnostic College English Speaking Test

Zhongbao Zhao[1]

[1] Foreign Languages College, Zhejiang Gongshang University, Hangzhou, China

Correspondence: Zhongbao Zhao, No.18, Xuezheng Str., Foreign Languages College, Zhejiang Gongshang University, Xiasha University Town, Hangzhou, China. E-mail: Michaelzhao998@hotmail.com

The research is financed by National Research Center for Foreign Language Education in China (Project ZGWYJYJJ2012A25)

Received: November 26, 2012 Accepted: December 20, 2012 Online Published: January 9, 2013

doi:10.5539/elt.v6n2p171 URL: http://dx.doi.org/10.5539/elt.v6n2p171

Abstract

This paper reports the verification of the consequential validity of a Diagnostic College English Speaking Test. A case study was conducted with 28 sophomore students from a national key university in China engaged in seven sets of DCEST tests. The analysis of the DCEST scores of the students in the experiment group indicates that progress has been made in their oral English proficiency over the two-month period. The survey data analysis reveals that the provision of diagnostic feedback is welcomed by a great majority of students, and they think that the diagnostic feedback of the DCEST can reflect the strengths and weaknesses of their oral English ability. Results of both quantitative and qualitative data analyses provide supportive evidence to the consequential validity of DCEST. The limitations and future research directions are finally discussed.

Keywords: diagnostic testing, oral English assessment, consequential validity, validation study

1. Introduction

Validity is defined by Messick (1989: 13) as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". And the concept of consequential validity was put forward by Messick (1996) as one aspect of the construct validity. Messick (1996: 249) argues that "the consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness and distributive justice (Messick 1980; 1989), as well as to washback".

Validation is considered as an essential component of language test development, for it can examine whether the test has achieved its intended purposes. Messick (1996) suggests that evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with positive or negative washback effects on teaching and learning should be collected to support the consequential aspect of construct validity.

Furthermore, Weir (2005: 210-215) suggests that consequential validity can be considered from three perspectives: differential validity, washback, and effect on society. Differential validity deals with the construct under-representation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers (American Educational Research Association *et al.* 1999). Washback examines the impact of tests on teaching and learning in a variety of settings. Effect on society refers to the effect of high-stakes tests on a wider community. Consequential validity has attracted more and more attention from designers of high-stakes tests and English teaching in China in recent years (Gong, 2012; Jin 2000; 2004; Yang and Gui 2007; Zhao, 2010; Zhao and Fan, 2012). However, fewer studies have been conducted to investigate the consequential validity of formative assessments in China, whose impact on English teaching and learning should never be neglected. The present study will focus on exploring the consequential validity of a Diagnostic College English Speaking Test (DCEST) in terms of its impact on oral English teaching and learning in real educational settings, as the DCEST is a formative assessment designed to diagnose students' oral English proficiency.

The specific questions addressed by the study are as follows:

1) What is the impact of DCEST on students' oral English proficiency?

2) What do students think of the test and the usefulness of the test feedback?

3) What kind of impact will the feedback exert on students' oral English learning?

2. Method

The Diagnostic College English Speaking Test (DCEST) is designed as a 15-minute face-to-face interview test for the purpose of identifying the strengths and weaknesses of students' English speaking ability at the tertiary level in China, which involves three task types: reading aloud, individual presentation and information gap (Zhao 2011). A checklist is designed for the examiner to record each test-takers' performance with respect to pronunciation, intonation, grammatical accuracy, grammatical complexity, vocabulary accuracy, vocabulary range, fluency, communicative strategy, coherence, discourse size. In addition to the report of a five-level composite grade to each test-taker, individualized feedback is provided detailing students' strengths and weaknesses.

As part of the a posteriori validation of the DCEST, a case study was conducted with 28 sophomore students from a national key university in China engaged in seven sets of DCEST tests from April to June in 2008. Apart from the student participants, one college English teacher and two doctoral students of applied linguistics were invited to help with data collection and analysis. A variety of instruments were employed for the purpose of obtaining various types of information to validate the consequential validity of the DCEST (see Table 1). To explore the impact of diagnostic feedback on students' oral English learning, both the control group and the experiment group took the same tests (DCEST 1 and DCEST 7) at the beginning and the end of the main study, and the experiment group took another five tests (DCEST 2 to DCEST 6) during the two-month experiment period. Students' evaluation of the usefulness of the DCEST feedback was gathered through the student questionnaire of feedback evaluation (SQFE) survey and the student and the teacher interviews at the end of the main study. The SQFE was composed of 14 questions which were divided into two sections: an evaluation of the overall usefulness of the feedback and an evaluation of each parameter used to report the profile scores. All the questions were designed using the five-point Likert scale. Following the SQFE survey, the researcher conducted face-to-face interviews with eight students in the experiment group and their college English teacher. Altogether the raw data collected in the present study were 7 sets of test scores from DCEST, 28 questionnaires and 9 interviews.

Table 1. Research instruments for the validation study

Research instruments	Targeted user	Purpose
Pre-test (DCEST 1)	EG and CG	For measuring both the EG and CG's oral English proficiency at the beginning of the main study
Scoring Sheet	Research assistant	For raters to assess test takers' performance in DCEST
Feedback Descriptors	Researcher	For the researcher to provide feedback reports to EG
DCEST 2-6	EG	For diagnosing the oral English proficiency of EG
Post-test (DCEST 1)	EG and CG	For measuring the EG and CG's oral English proficiency at the end of the main study
SQFE	EG	For collecting EG's evaluation of the tests' feedback
SI	EG	For collecting EG's comments on the usefulness of the tests' feedback
TI	English teacher	For collecting English teacher's comments on the usefulness of the feedback

Note: CG= control group, EG= experiment group, SQFE= student questionnaire of feedback evaluation, SI= student interview, TI= teacher interview.

3. Results

3.1 What is the Impact of DCEST on Students' Oral English Proficiency?

To explore the impact of diagnostic feedback on students' oral English learning, both the control group and the experiment group took the same tests (DCEST 1 and DCEST 7) at the beginning and the end of the main study, and the experiment group took another five tests (DCEST 2 to DCEST 6) during the two-month experiment period.

DCEST 1 scores of the control group (CG1) and those of the experiment group (EG1) were compared using the Independent Samples t-test. The result indicated that there was no significant mean difference between the two groups (p=0.203) (see Tables 2 and 3).

Table 2. Means and SDs of CG1 and EG1, CG 2 and EG2

	Mean	**N**	**Std. Deviation**
CG1	35.00	14	5.738
EG1	37.43	14	3.936
CG2	34.14	14	4.167
EG2	40.86	14	3.900

Table 3. Independent Samples t-test for CG1 and EG1, CG2 and EG2

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
CG1-EG1	Equal variances assumed	2.497	.126	-1.306	26	.203	-2.429	1.860	-6.251	1.394
CG2-EG2	Equal variances assumed	.411	.527	-4.402	26	.000	-6.714	1.525	-9.850	-3.579

However, the Independent Samples t-test of DCEST 7 scores of the two groups (CG2 and EG2) showed that there was a significant mean difference between them (p=0.000) (also see Tables 2 and 3). Further comparison of DCEST 1 and DCEST 7 scores of the experiment group (EG1 and EG2) also confirmed a significant mean difference between the two scores (p=0.006) (see Table 4).

In contrast, the Paired Samples t-test of the control group's DCEST 1 and DCEST 7 scores did not show significant mean difference (p=.325) (also see Table 4), indicating that the control group made little progress in their oral English proficiency over the two-month period of the main study, during which the control group didn't take any oral English test except the pre- and post-tests of the DCEST, nor did they receive any feedback from their English teacher about their strengths and weaknesses in oral English communication.

Table 4. Paired Samples t-test for EG1 and EG2, CG1 and CG2

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	EG1-EG2	-3.429	3.917	1.047	-5.690	-1.167	-3.275	13	.006
Pair 2	CG1-CG2	.857	3.134	.838	-.953	2.667	1.023	13	.325

Furthermore, the 10 analytic scores of the experiment group on DCEST 1 to DCEST 7 were compared to see on which aspects students had made more progress. Table 5 indicated that the experiment group students made progress in all the aspects of oral English concerned in the study, with the progress in the following three aspects being most substantial: pronunciation (improved by 0.86), intonation (improved by 0.57), coherence (improved by 0.43).

Table 5. Analytic mean scores of the experiment group in DCEST 1 to 7

	Test1	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**
PR	3.93	4.36	4.29	4.21	4.36	4.29	4.79
IN	3.79	3.86	4.29	4.21	4.29	4.14	4.36
GA	3.71	3.79	3.86	3.93	3.71	3.93	3.93
GC	3.29	3.50	3.21	3.43	3.50	3.71	3.50
VA	3.71	3.79	3.79	3.86	3.86	3.79	3.93
VR	3.36	3.29	3.43	3.50	3.71	3.64	3.50
FL	3.93	3.43	3.86	4.00	4.00	3.93	4.21
CS	3.79	3.64	3.86	3.79	3.93	3.79	4.07
CO	3.93	3.93	3.93	3.86	4.29	4.21	4.36
DS	4.00	4.07	4.21	4.21	4.43	4.36	4.21

Note: PR=pronunciation, IN=intonation, GA=grammatical accuracy, GC=grammatical complexity, VA=vocabulary accuracy, VR=vocabulary range, FL=fluency, CS=communicative strategy, CO= coherence, DS=discourse size.

In addition to the comparison between the control and the experiment groups, a close look at the descriptive statistics of the seven total scores of the experiment group revealed that the mean score of one test was always slightly higher than that of the previous one with the only exception of DCEST 6 (see Table 6).

In other words, students in the experiment group were making steady and consistent progress in their oral English performance on the DCEST tests. DCEST 6 had a mean 0.28 points lower than that of DCEST 5. This could be explained by the fact that the students were somewhat distracted by their final exams which were administered in late June when they took DCEST 6.

Table 6. Means and SDs of seven DCEST test scores of the experiment group

	Mean	**Std. Deviation**
DCEST 1	37.43	3.936
DCEST 2	37.64	3.734
DCEST 3	38.71	3.024
DCEST 4	39.00	3.187
DCEST 5	40.07	3.339
DCEST 6	39.79	3.786
DCEST 7	40.86	3.900

To facilitate the analysis of experiment group's improvement in oral English proficiency, the researcher calculated the mean of the seven DCEST total scores for the experiment group, and categorized the students with a mean score below 35 as the lower-intermediate subgroup, and those with a mean score between 35 and 40 as intermediate, and those with a mean score above 40 as advanced. Each group's mean score of the seven tests was 34, 38 and 42.4 respectively (see Table 7).

Table 7. Three proficiency subgroups in the experiment group

Level of Oral Proficiency	**Student**	**Mean**
Lower-intermediate	Student C, M	34
Intermediate	Student A, B, G, H, K, L, N	38
Advanced	Student D, E, F, I, J	42.4

A closer examination of the changes in students' test scores in the main study revealed that the three students of the advanced proficiency group exhibited steady progress over time. Two students from the intermediate proficiency group (Student A and B), however, made the most dramatic progress on their performances. Students in the lower-intermediate group showed changes in their scores in both upward and downward directions.

In sum, the analyses of the test scores suggested that the experiment group students on average improved their oral English proficiency, whereas the control group students showed little progress. This progress could be attributed to the constant provision of diagnostic feedback after each test session, which guided the students to improve their oral English in the right direction. However, it seemed too early to claim that such improvement was the strongest evidence of positive effects of the diagnostic feedback on students' learning. It could be due to students' self-learning during the eight-week period. Therefore, the SQFE survey and the interview data would be analyzed for more supportive evidence to prove the usefulness of the diagnostic feedback provided.

3.2 What do Students Think of the Test and the Usefulness of the Test Feedback?

The assumption was that students from the experiment group would have a good understanding of the usefulness of the feedback and make positive comments on the DCEST tests and the accompanying detailed feedback reports. The descriptive data of the experiment group students' evaluation of the feedback were summarized in Table 8.

Table 8. Descriptive statistics of responses to SQFE

Question	**Responses (frequency)(valid percentage)**					**Mean**	**Std**
	a	b	C	D	e		
1	0 (0%)	12 (85.7%)	2 (14.3%)	0 (0%)	0 (0%)	3.86	.363
2	2 (14.3%)	11 (78.6%)	1 (7.1%)	0 (0%)	0 (0%)	4.07	.475
3	2 (14.3%)	10 (71.4%)	2 (14.3%)	0 (0%)	0 (0%)	4.00	.555
4	2 (14.3%)	9 (64.3%)	2 (14.3%)	1 (7.1%)	0 (0%)	3.86	.770
5	3 (21.4%)	9 (64.3%)	1 (7.1%)	1 (7.1%)	0 (0%)	4.00	.784
6	2 (14.3%)	8 (57.1%)	3 (21.4%)	1 (7.1%)	0 (0%)	3.79	.802
7	3 (21.4%)	9 (64.3%)	2 (14.3%)	0 (0%)	0 (0%)	4.07	.616
8	3 (21.4%)	8 (57.1%)	3 (21.4%)	0 (0%)	0 (0%)	4.00	.679
9	7 (50.0%)	5 (35.7%)	2 (14.3%)	0 (0%)	0 (0%)	4.36	.745
10	3 (21.4%)	8 (57.1%)	2 (14.3%)	1 (7.1%)	0 (0%)	3.93	.829
11	5 (35.7%)	7 (50.0%)	2 (14.3%)	0 (0%)	0 (0%)	4.21	.699
12	3 (21.4%)	10 (71.4%)	1 (7.1%)	0 (0%)	0 (0%)	4.14	.535
13	3 (21.4%)	7 (50.0%)	4 (28.6%)	0 (0%)	0 (0%)	3.93	.730
14	3 (21.4%)	7 (50.0%)	4 (28.6%)	0 (0%)	0 (0%)	3.93	.730

Note: a=very helpful, b=quite helpful, c=so-so, d=not quite helpful, e=no help at all, a=5, b=4, c=3, d=2, e=1.

Questions 1-4 inquired about the overall usefulness of the feedback from several aspects. Question 1 was about the extent to which the feedback report can reflect students' general oral English proficiency. The majority of the students (85.7%) agreed that the feedback report on the whole can be a valid indicator of their oral English proficiency. Responses to Question 2 indicated that a great majority of the students (92.9%) thought that the feedback report can accurately describe the strengths of their oral English ability. With regard to Question 3, 85.7% of the students thought that the feedback report can provide useful diagnostic information on their weaknesses in oral English communication. Answers to Question 4 showed that the majority of the students (78.6%) agreed to the usefulness of the diagnostic feedback for their oral English learning.

Questions 5-14 focused on the evaluation of each feedback parameter. The means of the questions revealed that feedback on vocabulary accuracy was considered the most useful, followed by fluency and use of communicative strategies. Feedback on intonation was perceived as the least useful by the experiment group (see Table 9). However, the analysis of the experiment group students' seven DCEST scores showed a different picture (see Table 5). The test results showed that experiment group students made great progress in their test scores in these aspects: pronunciation, intonation, and coherence, which were quite different from those aspects evaluated by the experiment group students as the most useful. The discrepancy may be explained by the fact that it might take longer time for students to make improvements in those aspects they considered most useful.

Table 9. Means and SDs of the usefulness of analytic feedback parameters

Feedback parameter	**Mean**	**Std. Deviation**
Vocabulary accuracy	4.36	.745
Fluency	4.21	.699
Use of Communicative strategy	4.14	.535
Grammar accuracy	4.07	.616
Grammatical complexity	4.00	.679
Pronunciation	4.00	.784
Vocabulary range	3.93	.829
Discourse size	3.93	.730
Coherence	3.93	.730
Intonation	3.79	.802

Note: 1= not useful at all, 2= not quite useful, 3= so-so, 4= quite useful, 5= very useful.

Furthermore, the comparison of the three proficiency groups' evaluation of the usefulness of the 10 feedback parameters indicated that the intermediate group commented most favorably on the usefulness of the feedback, whereas the lower intermediate group least favorably (see Table 10).

Table 10. Descriptive statistics of the three proficiency subgroups' overall evaluation of the usefulness of feedback

Oral English proficiency group	**N**	**Mean**	**Std. Deviation**
1 Lower-intermediate level	2	35.50	6.364
2 Intermediate level	7	42.29	4.889
3 Advanced level	5	39.60	5.177

The evaluations of the usefulness of each feedback parameter by the three proficiency subgroups were also investigated. Table 11 indicated that students of the lower intermediate group regarded the feedback on using communicative strategies as the most useful. Students at the intermediate level also reported that the feedback on using communicative strategies as the most useful, and considered the feedback on vocabulary range as the least useful, whereas students at the advanced level thought that the feedback on vocabulary accuracy was the most useful and the feedback on intonation was least useful.

Table 11. The three proficiency subgroups' evaluation of the usefulness of each feedback parameter

Feedback	**Lower intermediate**		**Intermediate**		**Advanced**	
	Mean	SD	Mean	SD	Mean	SD
PR	3.50	.707	4.29	.488	3.80	1.095
IN	3.50	.707	4.14	.690	3.40	.894
GA	3.50	.707	4.29	.488	4.00	.707
GC	3.50	.707	4.14	.690	4.00	.707
VA	3.50	.707	4.29	.756	4.80	.447
VR	3.50	.707	3.86	1.069	4.20	.447
FL	3.50	.707	4.29	.756	4.40	.548
CS	4.00	.000	4.43	.535	3.80	.447
CO	3.50	.707	4.29	.756	3.60	.548
DS	3.50	.707	4.29	.756	3.60	.548

This section explored the effect of the feedback on students' test performance and students' evaluation of the usefulness of the feedback. For a better understanding of how the feedback would be used by students in oral English learning, the next section analyzed the Student Interview and Teacher Interview data.

3.3 What Kind of Impact will the Feedback Exert on Students' Oral English Learning?

The interview data were transcribed and then subjected to a qualitative analysis through a hermeneutic process of reading, analyzing and re-reading. When asked about the usefulness and the impact of feedback on their oral English learning, some students commented that the feedback could raise their awareness of the linguistic problems in oral English communication. The following is one illustrative piece of interview excerpt:

I think the feedback is quite useful; at least it makes me aware of my weaknesses in oral English ability… After knowing my problems, I pay special attention to them in learning. (Student B)

Some students thought that the feedback enabled them to make an overall evaluation of their oral English proficiency:

The feedback enables me to make a systematic and all-round evaluation of my oral English proficiency. (Student D)

The feedback is quite helpful. The score profile and feedback descriptors show me an objective and accurate picture of my oral English proficiency on both macro and micro levels. Then in my daily oral English learning, I will pay special attention to these problems, and in this way, I think I can make greater progress. (Student H)

Some students considered the feedback on some aspects as more useful to their learning. The following interview excerpt illustrated this point:

I was able to improve my grammatical and vocabulary accuracy, and I also paid more attention to fluency of my speech. I used to use a lot of fillers such as 'er', 'um', etc., now I am using them less frequently. (Student A)

I paid a lot of attention to the criterion of accuracy of pronunciation, then I made fewer mistakes, and I thought I made some progress in this aspect. (Student D)

However, comments on intonation were not as positive as those on accuracy of vocabulary and grammatical structure. The following excerpt might give us some idea:

I don't think that I can make progress in intonation within a short period of time, and I spent little time practicing it. (Student B)

Furthermore, some students thought that the feedback not only diagnosed their oral English proficiency but also showed them the way forward in their oral English learning:

The feedback provides macro-level diagnostic information; it shows us the directions to move forward. For example, if the feedback tells you that your major problems lie in grammar and vocabulary, then you will have a *clear learning objective. (Student H)*

The feedback has raised my awareness of the importance of oral English learning and influenced positively my oral English learning methods. If I had no chance to communicate in English as I did in the DCEST, then I would not have known my problems. Feedback from the DCEST tests enables me to realize my weaknesses and then I know how to improve my oral English. (Student G)

In addition to these positive comments, concerns raised by several students were also worth mentioning and discussing. One student pointed out that the distinctions between some levels of the rating scale were too subtle:

Generally speaking, I think the feedback can reflect the strengths and weaknesses of my oral English proficiency. But I think distinctions between Level 3 and Level 4 and between Level 4 and Level 5 are too subtle. I hope that more information could be provided to distinguish these adjacent levels. (Student E)

Though efforts have been made to make the feedback descriptors as accurate as possible, there is room for improvement. One of the possible methods is to refine the descriptors on the basis of students' actual test performances. This was pointed out as a recommendation in the final chapter.

Some students also expressed their hope to be provided with specific guidance on the appropriate types of actions they need to take in addition to the diagnosis of their difficulties and problems. One student commented that:

The feedback reveals the problems of my oral English proficiency. But more importantly, I would like to have more information on how to overcome the problems and make improvements in these aspects. (Student F)

In sum, the above quantitative and qualitative data analyses indicated that a great majority of the students welcomed the diagnostic feedback report and agreed that the diagnostic feedback report can provide accurate information on their weaknesses and strengths in oral English proficiency. Students' perceptions of each criterion and the accompanying feedback descriptor indicated that diagnostic information focusing on the lexical-grammatical knowledge such as vocabulary and grammatical accuracy, use of communicative strategies and coherence was considered more useful than feedback on other aspects.

Since the main study was conducted in the middle of a term, the college English teacher for the experiment and control group was not able to participate in the study to evaluate the impact of the feedback on oral English teaching due to the conflict of teaching plans. The researcher therefore invited the teacher to observe the test-taking process of the experiment group, and had a brief interview with the teacher for his comments and suggestions.

The teacher agreed that the diagnostic feedback would be useful for students to know about their strengths and weaknesses in oral English, but he pointed out that the effectiveness is largely dependent on how students would make use of it. The following interview excerpt illustrated this point.

The usefulness of the feedback depends on how students will use it in their oral English learning. (English teacher)

As for the impact of feedback on teaching, the teacher argued that the usefulness of the feedback on oral English teaching would depend on the teacher's pedagogical approach, the purpose of learning and the context of learning. Just as the following interview excerpt indicated:

The feedback may exert little impact on oral English teaching, because the speaking activities in classroom are very limited, and it is impossible to take all students' needs into account in oral English teaching. (English teacher)

The teacher's interview data implied that the effectiveness of the feedback may depend largely on the degree to which it was compatible with the teachers' teaching plan and students' learning attitude.

4. Discussion

The results of both quantitative and qualitative data analyses indicated that students who took the DCEST tests over a period of eight weeks and received feedback regularly made great progress in their oral English proficiency. The survey data analysis revealed that the provision of diagnostic feedback was welcomed by a great majority of students, and they thought that the diagnostic feedback of the DCEST could reflect the strengths and weaknesses of their oral English ability. In all, the results of DCEST scores, SQFE, SI and TI analyses provided supportive evidence to the consequential validity of the DCEST.

As with any scholarly investigation, this study has its share of limitations. First, due to the limitation of human resources and time constraints, the study was administered to a small sample size (N=28) over a two-month period. Considering the relatively small sample size, further studies with larger sample sizes are necessary to generalize the findings beyond the participants in this study. Another limitation in the research is that the teacher's participation was restrained due to the conflict of teaching plans. Since the study began in the middle of a term, the college English teacher of the student participants was not able to participate in the study and use the test feedback in his oral English teaching. Hence, this study only focused on investigating students' evaluation of the usefulness of the feedback and the impact of feedback on students' oral English learning, without giving much attention to the impact of feedback on oral English teaching.

It is hoped that future research should be conducted on a larger sample with students from a variety of majors and universities that could better represent Chinese undergraduates for the purpose of confirming the generalizability of the results of the present study. In addition to student participants, college English teachers should also be invited to participate in future research to investigate the impact of feedback on oral English teaching over a period of time.

References

AERA, APA, & NCME. (1999). *Standards in Educational and Psychological Testing*. Washington D.C.: APA.

Gong, Z. (2012). Research on Foreign Language Classroom: Status quo and Implications—An Empirical Study on the Certain Papers Published from 1996 to 2011. *Journal of Zhejiang Gongshang University*, (3), 82-88.

Jin, Y. (2000). The Washback of CET-SET on Teaching. *Foreign Languages World*, (4), 57-62.

Jin, Y. (2004). The Reform of CET. *Foreign Languages in China*, (1), 27-29.

Messick, S. (1989). Meaning and Values in Test Validation: the Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing*, *13*(3), 241-256. http://dx.doi.org/10.1177/026553229601300302

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Yang, H. Z., & Gui, S. C. (2007). The Sociological View of Language Testing, *Modern Foreign Languages*, (4), 368-74.

Zhao, Z. B. (2010). Review of Diagnostic Foreign Language Proficiency: The Interface between Learning and Assessment. *Foreign Languages World*, (4), 91-94.

Zhao, Z. B. (2011). Development and Validation of the Diagnostic College English Speaking Test (Unpublished doctoral dissertation, Shanghai Jiao Tong University).

Zhao, Z. B., & Fan, J. S. (2012). Review of Language Assessment in Practice: Developing Language Assessment and Justifying their use in the Real World. *Modern Foreign Languages,* (1), 105-107.