English Language Teaching; Vol. 15, No. 7; 2022
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Validation of the Results of Linking Speaking Test of IELTS to China's Standards of English Language Ability

Yuyue Chen[1] & Xuefeng Wu[1]

[1] School of Foreign Studies, Nanjing Forestry University, Nanjing, China

Correspondence: Yuyue Chen, School of Foreign Studies, Nanjing Forestry University, Nanjing, China.

Received: May 16, 2022 Accepted: June 13, 2022 Online Published: June 15, 2022

doi: 10.5539/elt.v15n7p75 URL: https://doi.org/10.5539/elt.v15n7p75

Abstract

This study, taking into consideration of teachers' as well as students' judgments, explored the generalizability and consistency of the results of linking speaking test of IELTS to China's Standards of English Language Ability (CSE). Nine college English teachers and 81 undergraduate students judged the degree of congruence between the IELTS speaking test and 72 relevant CSE descriptors to generate evidence of generalizability; 2 teachers judged the CSE levels of 11 videos of IELTS speaking test and 113 students assessed themselves based on the self-assessment scale, to provide data for exploring consistency from three perspectives: consistency of teachers' judgments, consistency between teachers' judgments and empirical scores, and consistency between students' actual ability and the ability demonstrated in the self-assessment scale. The evidence showed that the linking results performed well both in generalizability and consistency from the perspective of both teachers and students, but students rated comparatively low than teachers both in the two facets. However, considering that certain descriptors are relatively abstract and separate from daily life situations in IELTS speaking test, it is understandable that students rate low in generalizability; and considering students' insufficient self-assessment ability, their relative low recognition in consistency also get explained. In general, therefore, the linking results show good generalizability and consistency.

Keywords: consistency, CSE, generalizability, IELTS, speaking test

1. Introduction

For years, there has been no unified English language ability scale in China. In June 2018, the Ministry of education and the State Language Commission officially issued China's Standards of English Language Ability (CSE) to fill this gap. On January 15, 2019, the results of linking IELTS and Aptis to China's Standards of English Language Ability were first published. The linking results show cut scores of language tests that linked to corresponding levels of CSE, covering both overall language ability and itemized language skills such as listening, reading, writing, speaking, etc.

Linking language tests to language standards is equivalent to building a bridge between test scores and ability descriptors (Pan & Wang, 2021), and providing unified language ability descriptors for different English tests helps to give concrete interpretations to abstract test scores (Tannenbaum & Cho, 2014) and make them comparable. However, it should be noticed that the language ability scale is a description of language level. It does not describe any test attributes or item requirements, nor is it based on any item difficulty theory. In this case, the task of linking tests to the language standards must be carried out through expert judgment (Harsch & Hartig, 2015), but expert judgment inevitably has a certain degree of subjectivity and will bring errors. Therefore, in order to ensure the fairness of linking-based decisions, the validation of the linking results is vital. Taking oral skills as an example, this study examined the validity of the results of linking IELTS test to China's Standards of English Language Ability from two perspectives: generalizability and consistency.

2. Literature Review

2.1 Validation of Linking Language Tests to Language Standards

Previous studies mostly examined validity issues through three traditional sources of evidence, namely procedural validity, internal validity and external validity (Martyniuk, 2010; Dunlea, 2015). Procedural validity focuses on whether each step of the linking project is effectively implemented; Internal validity answers questions about the accuracy and consistency of standard setting results; External validity verifies the linking

results by collecting external evidences. This method has strong operability and helps researchers to clarify the sources of relevant validity evidence. However, the connection between the three validity evidences is loose, and the simple superposition of them can not fully represent the linking validity, which poses a great challenge to the integrity and coherence of the validity.

To address the lack of coherence in validity evidences, Bachman and Palmer (2010) proposed Assessment Use Argument (AUA) from the philosophical perspective based on the test usefulness model. AUA framework links the originally independent six elements (reliability, construct validity, authenticity, interactivity, impact and feasibility) in the test usefulness model into a coherent and organic whole, and lists 15 guiding questions around the four claims of consequences, decision, interpretations and assessment report (Han & Luo, 2013). The framework requires assessors to collect evidences from four aspects: test performance, interpretations of scores, decisions based on these interpretations and consequences of test use, so as to ensure the rationality of test score interpretations and test use (Bachman, 2005, 2006, 2010). It is known as a validation framework integrating flexibility, dynamic, malleability and adaptivity.

Then based on AUA framework, He (2019) established Linking Use Argument (LUA) as the validation framework for linking language tests to language standards. LUA framework brings up four propositions and describes the linear relation in the consequences of linking, linking-based decisions, linking-based interpretations, linking records and test performances. It inherits the two-way linear flow structure of AUA framework and serves two purposes of linking implementation and linking result interpretations and use. Among them, the consequences of linking emphasize the beneficence of linking result use and linking-based decisions; linking-based decisions emphasize values sensitivity and equitability; linking-based interpretations emphasize relevance, sufficiency, meaningfulness, generalizability and impartiality; linking records emphasize consistency (He, 2019). In conclusion, the consequences of linking and linking-based decisions stress the social impacts of linking results, which further illustrates the importance of validation to reduce errors. And propositions of linking-based interpretations and linking records are important evidences for judging whether the linking results are reliable.

2.2 Generalizability and Consistency

Generalizability is a prominent issue in the validity chain of linking productive language tests to language standards (Cai, 2019). Productive language skills often include subjective questions, such as writing and speaking, which can only accommodate a small number of tasks and cover limited range of goal construct (Hughes, 2003). Therefore, the construct and task types of productive language tests do not wholly cover the content of language standards. In essence, to link language tests to language standards is to connect the test-takers' scores with the descriptors of the language standard, that is, to generalize the test performances to specific language abilities described by the language standards. Therefore, to what degree the language tests cover the language standards is an important index to measure generalizability. In view of this, generalizability is an outstanding validity issue in linking tests of productive language skills to language standards.

In addition, the importance of generalizability is reflected in AUA framework and LUA framework. In the interpretation part of AUA framework, Bachman and Palmer (2010) raised a question: how can we ensure that the interpretations of test-takers' language abilities can be effectively generalized to the target language domains to be measured? That is the connotation of generalizability. In the subsequent LUA framework, He (2019) emphasized in her fourth proposition that generalizability was an important validity issue in linking language tests to language standards, and pointed out that evidences can be collected through content analysis of the examination inspection step. By comparing the similarities and differences between the test content and the language activities described in the language standards, it can be judged whether the interpretations of the linking results is generalizable. Generalizability, with beneficence, values sensitivity, equitability, relevance, sufficiency, meaningfulness, impartiality and consistency, constitute a complete chain of evidence for validity. The lack of any attribute above will pose a serious threat to validity (Cai, 2019). In this study, generalizability refers specifically to the extent to which IELTS speaking test can be generalized to the target language use domains defined by relevant descriptors in CSE.

Consistency is also a key concern when the productive language test is linked to the language standard. From AUA perspective, reliability, as one of the six elements, constitutes the rationale on which the assessment report depends. The concept of "reliability" refers to the stability of assessment report recorded upon the test performances of test-takers (Xu, 2012). As one of the four claims of AUA, assessment report raises two questions about stability: 1) how do we ensure that the assessment report is stable? 2) How will we ensure the relative stability of the assessment report to different groups of test-takers? And here, stability can be understood

as consistency. Similarly, in LUA framework, ensuring the consistency of linking records is the first step in the evidence chain of the validation of linking results; If the linking result itself is unstable, the subsequent validity evidence will lose its foundation (He, 2019; Min, 2019).

In linking research, the connotation of consistency generally includes three aspects: consistency of panelists' judgments, consistency between panelists' judgments and empirical scores, and the stability of cut scores (Cizek & Bunch, 2007; Tannenbaum & Cho, 2014). The consistency of panelists' judgments can be measured from the results of Many-Facet Rasch Model (MFRM) analysis of test scores, specifically considering the severity, fitting index and significant deviation of panelists; the consistency between panelists' judgments and empirical scores is reflected by the correlation coefficient; the stability of cut scores is reflected in the standard error of the cut scores (Cai, 2019). This study examines the consistency of the linking results from the first two aspects.

3. Research Design

3.1 Research Questions

The determination of linking results comes from panelists' judgments. However, it is the test-takers who are directly affected by the linking results. Therefore, the degree to which the test-takers recognize the linking results is absolutely an important consideration. Since students account for the vast majority in IELTS test-takers, this study takes students and teachers as participants respectively, and makes appropriate comparative analysis between the two groups.

The current study, based on the LUA framework,, which examines generalizability and consistency of the results of linking IELTS speaking test to CSE, attempts to answer the following two questions:

1) To what extent is the generalizability of the results of linking IELTS speaking test to CSE?
2) To what extent is the consistency of the results of linking IELTS speaking test to CSE?

3.2 Participants

In order to explore the generalizability of the linking results, the researcher developed a generalizability questionnaire, which covered 72 speaking-related descriptors in the CSE. 81 IELTS test-takers were invited to participate in the research, including 47 boys and 34 girls, with an average age of 22 years; and 9 College English teachers, including 2 males and 7 females, among whom there was 1 Professor, 3 associate professors, 4 lecturers and 1 teaching assistant, with an average age of 42 years.

In order to explore the consistency of the linking results, the researcher compiled a consistency questionnaire based on descriptors of the Table 82 self-assessment scale for oral expression from CSE. Since the linking results are limited to 4-8 levels in CSE, this researcher only investigated IELTS test-takers whose speaking scores were from 5.0 to 7.0. The specific number of each score is shown in Table 1. The investigation involved 113 students in total, including 58 boys and 55 girls, with an average age of 22 years; In addition, it included two college English teachers, an associate professor and a teaching assistant (the associate professor participated in the development of CSE scale).

Table 1. Specific Number of People in Each Score

IELTS test score (speaking)	Number
5.0	5
5.5	20
6.0	43
6.5	24
7.0	21

3.3 Data Collection & Analysis

3.3.1 Generalizability

Firstly, the researcher selected all speaking-related subscales from CSE, including 12 subscales in three aspects, namely oral expression, pragmatic expression and expression strategies. The follow-up steps are as follows:

1) From the 12 subscales, four subscales in CSE related to IELTS speaking test were preliminarily selected involving Table 4 Overall oral expression, Table 32 Oral description, Table 33 Oral narration and Table 36 Oral argumentation. Among the above four subscales, 72 descriptors from level 4 to level 8 were selected.

2) Based on the above 72 descriptors, a questionnaire (including five options: very relevant; moderate relevant; uncertain; less relevant; completely irrelevant) was established to investigate the participants' views on the congruence of descriptors from CSE to IELTS speaking test tasks. Finally, 98 questionnaires were collected with 90 valid questionnaires (the invalid questionnaires refer to questionnaires where participants ticked the same choice for all items), participants including 9 College English teachers and 81 IELTS test-takers.

3) The researcher collected the questionnaire data, separated the data of teachers and students, and calculated respectively the percentage of descriptors of different degrees of congruence in each subscale and each level.

Through the above steps, the congruence between IELTS speaking test and CSE descriptors from perspectives of teachers and students was obtained, so as to deduce the the degree of generalizability.

3.3.2 Consistency

Consistency research was also divided into teacher group and student group. In the teacher group, 11 record videos of IELTS speaking test officially presented by Cambridge IELTS were selected, with scores ranging from 5.0 to 7.0. On the basis of the familiarization with CSE, the two teachers rated the videos on specific CSE level. Then the consistency of the teachers' judgements and the consistency between teachers' judgments and empirical scores can be seen from the questionnaire data. In the student group, the process is more complex. Below are specific steps.

1) Based on the descriptors of level 4-8 in self-assessment scale for oral expression, a questionnaire was established (options were assigned: yes=3; uncertain=2; no=1). Participants were asked to offer their IELTS speaking test scores.

2) SPSS.26.0 was used to carry out one-sample t-test on the questionnaire data and calculate the average score of each descriptor.

Upon the completion of above steps, consistency of the linking results from perspectives of teachers and students was shown.

4. Results and Discussion

4.1 Generalizability

The degree of generalizability of the linking results can be reflected in the coverage of the IELTS speaking test tasks on the relevant descriptors in CSE. The most general indicator is the proportion of the relevant descriptors in all descriptors, and the more detailed indicator is the proportion of the relevant descriptors in each subscale or at each level. Specific statistical details are presented in Table 2 and 3.

The respondents in Table 2 are teachers and students in Table 3. The arrangement and layout of Table 2 and 3 are exactly the same. From top to bottom are the four subscales representing the target language domain; the total numbers of descriptors in each subscale; the numbers of relevant descriptors of level 4-8 in each subscale (the three relevance degree are relevant, very relevant and moderate relevant); the total numbers of relevant descriptors of three relevance degree in each subscale and their proportion in the total descriptors in corresponding subscales. The second column on the left of the two tables are the total numbers of descriptors at each level, and the two columns on the far right shows the total number of relevant descriptors of three degree of relevance at each level and their proportion in the total descriptors at their levels.

It should be noted that the relevant descriptors in the table refer to the sum of very relevant descriptors and moderate relevant descriptors. As is presented in Table 2 and 3, the proportions of relevant descriptors in total descriptors of all levels and subscales are almost equal to or close to 100%. This is because the four subscales were selected for they were, to some extent, related to IELTS speaking test, and this finding verifies that the four subscales are indeed highly related to IELTS speaking test. While in previous studies, researchers often took all relevant subscales into consideration for analysis (Cai, 2019). This research method is comprehensive and rigorous, but also time-consuming and highly-demanded on experimental conditions. Therefore, the current study focuses directly on the four relevant subscales. Except for the overall oral expression scale, the other three subscales were selected according to the oral communication activities they describe: oral narration (such as personal experience and historical events), oral description (such as natural scenes and character psychology), and oral argumentation (such as speech) (Jin & Jie, 2017). It is precisely for the above reasons that in the final statistical data, the proportion of very relevant descriptors can better explain the coverage of IELTS speaking test tasks on the descriptors in CSE, that is, the degree of generalizability.

Table 2. Numbers of Relevant Descriptors in Each Subscale and Level (Teachers)

	Total of each level	Subscales	Overall oral expression	Oral argumentation	Oral description	Oral narration	Total	Proportion
		Total of each subscale	16	25	18	13	72	
Level 4	18	Relevant	4	3	7	4	18	100%
		Very relevant	3	3	5	3	14	78%
		Moderate relevant	1	0	2	1	4	22%
Level 5	17	Relevant	3	7	4	3	17	100%
		Very relevant	2	7	4	1	14	82%
		Moderate relevant	1	0	0	2	3	18%
Level 6	16	Relevant	3	5	4	4	16	100%
		Very relevant	2	3	1	2	8	50%
		Moderate relevant	1	2	3	2	8	50%
Level 7	12	Relevant	3	5	1	1	10	83%
		Very relevant	3	1	1	1	6	50%
		Moderate relevant	0	4	0	0	4	33%
Level 8	9	Relevant	3	3	2	1	9	100%
		Very relevant	3	0	1	1	5	56%
		Moderate relevant	0	3	1	0	4	44%
Total	72	Relevant	16	23	18	13	70	97%
		Proportion	100%	92%	100%	100%	97%	
		Very relevant	13	14	12	8	47	64%
		Proportion	81%	56%	67%	62%	64%	
		Moderate relevant	3	9	6	5	23	32%
		Proportion	19%	36%	33%	38%	32%	

In teacher group (Table 2), there were 47 very relevant descriptors out of the 72 descriptors overall, accounting for 64%. It showed that from the perspective of teachers, the IELTS speaking test covered most of the descriptors and had good generalizability. However, it at the same time saw certain uneven coverage among different levels and subscales. On one hand, the proportions of very relevant descriptors at level 4-8 were all over 50%, of which proportions of level 4 (78%) and level 5 (82%) exceeded 70%; while that of level 6 (50%), level 7 (50%) and level 8 (56%) were relatively low, featuring higher extent of coverage at low level and lower extent at high level in general. The reason for this characteristic may lie in that the number of high-level descriptors was relatively small compared with that of low-level descriptors. In the four subscales, the total number of descriptors at each level is 18 at level 4, 17 at level 5, 16 at level 7, 12 at level 7 and 9 at level 8 respectively. Among them, there were only 1-2 descriptors at level 7 and level 8 in the oral description and oral narrative subscale, which covered a relatively narrow range of language use domain and the degree of generalizability could be reduced accordingly. On the other hand, the proportions of very relevant descriptors in each subscale were all above 50%,

among which the overall oral expression scale topped the list, and the other three were relatively low, with oral description (67%), oral narration (62%) and oral argumentation (56%). On the subscale facet, it had the feature of lower extent of coverage in overall scale while lower extent of coverage in subscales. Reason behind this may be that descriptors in overall scale are refined and selected on the basis of subscales, so it has better integrity and representativeness than the subscales.

Table 3. Numbers of Relevant Descriptors in Each Subscale and Level (Students)

	Total of each level	Subscales	Overall oral expression	Oral argumentation	Oral description	Oral narration	Total	Proportion
		Total of each subscale	16	25	18	13	72	
Level 4	18	Relevant	4	3	7	4	18	100%
		Very relevant	4	0	4	1	9	50%
		Moderate relevant	0	3	3	3	9	50%
Level 5	17	Relevant	3	7	4	3	17	100%
		Very relevant	2	0	2	1	5	29%
		Moderate relevant	1	7	2	2	12	71%
Level 6	16	Relevant	3	5	4	4	16	100%
		Very relevant	0	0	0	1	1	6%
		Moderate relevant	3	5	4	3	15	94%
Level 7	12	Relevant	3	6	1	1	11	92%
		Very relevant	1	0	0	0	1	8%
		Moderate relevant	2	6	1	1	10	83%
Level 8	9	Relevant	3	3	2	1	9	100%
		Very relevant	0	0	0	0	0	0%
		Moderate relevant	3	3	2	1	9	100%
Total	72	Relevant	16	24	18	13	71	99%
		Proportion	100%	96%	100%	100%	99%	
		Very relevant	7	0	6	3	16	22%
		Proportion	44%	0%	33%	23%	22%	
		Moderate relevant	9	24	12	10	55	76%
		Proportion	56%	96%	67%	77%	76%	

In student group (Table 3), the situation was quite different. In general, there were only 16 very relevant descriptors among the 72 descriptors, accounting for 22%. Most of the descriptors were classified as moderate relevant degree, with a total of 55, occupying 76%, indicating that from the perspective of students, the coverage of IELTS speaking test on CSE is relatively low compared with that of teachers.

On one hand, among all levels, only at level 4, the proportion of very relevant descriptors was more than half. Others like level 5 (29%), level 6-8 (lower than 10%) and level 8 (0%) were all poor at generalizability. Possible reason is that there were few people who had reached level 8 in this study, and level 8 asked for a quite high speaking ability with some abstract descriptors, which is not as specific, detailed and understandable as that of other levels. As a consequence, most participants cannot reach the ability at level 8, and they dare not choose

very relevant easily. The experimental results here still had the characteristics of higher extent of coverage at low level and lower extent at high level in general. On the other hand, none of the four subscale had a proportion of very relevant descriptors higher than 50%, only that of the overall oral expression taking up 44%, close to half. Other subscales like oral description (33%) and oral narration (23%) presented low generalizability, and the oral argumentation had even hit the bottom at 0%. This is possibly related to the keywords "speech" in oral argumentation subscale. Words like "speech", "argument", "elaborate", "argumentation" are very common to see in the oral argumentation subscale, which are divorced from the life-oriented context of IELTS speaking test, and dispose participants to think that they do not have such a high degree of relevance to IELTS. Besides, same to the perspective of teachers, it also had the feature of lower extent of coverage in overall scale while lower extent of coverage in subscales from the perspective of students. It's worth noting that although teachers tend to perform better than students when judging the relevance of descriptors, in the investigation there were a considerable number of examinees who had taken IELTS speaking test more than once, some twice. They had great familiarity with IELTS speaking test even beyond the teachers, involving its test contents, question types and test construct. Therefore, the students' data part should never be ignored. Then despite the relatively low proportion of very relevant descriptors compared with teachers, the degree of congruence of linking results was good on the whole in view of all the factors above and high proportion of relevant descriptors.

Generalizability is significant for the validation of the results of linking language tests to language standards, and is even the prerequisite for the linking validity. It is therefore of great value to bring generalizability into validation framework. In LUA framework, the premise to ensure the meaningfulness of linking-based interpretations is that the test constructs and task types of tests are highly correlated with the constructs and descriptors of language standards, otherwise the linking results would make no sense (Tannenbaum & Cho, 2014). Unfortunately, in previous studies, researchers tended to ignore the verification of generalizability, which was just included in the exammination inspection step (Council of Europe, 2009), rather than the validity validation, which is not conducive to the integrity of validation. Omission of congruence between language abilities defined by the descriptors in language standards and that examined in the language tests will lead to unexplained linking results and unjustified linking-based decisions made by stakeholders.

Based on the findings of the present study, the results of linking IELTS speaking test to China's Standards of English Language Ability witness good generalizability. Considering the large number of Chinese students going abroad every year, the high generalizability of the linking results will help "Chinese standard" get international recognition and lay the foundation for wider international recognition of China's English test. In addition, high congruence between IELTS speaking test and CSE give IELTS teachers and test-takers a chance to better understand specific ability requirements of IELTS test, and then refine the ability training objectives and improve their abilities with the help of sufficient relevant descriptors in CSE.

4.2 Consistency

The first facet of consistency is consistency of teachers' judgments, namely inter-rater reliability. It can be seen from table 4 that the two teachers shared the same rating results in four videos including No. 4, No. 7, No. 9 and No. 11; other 5 videos of No. 1, No. 3, No. 5, No. 8 and No. 10, had only one grade difference between the two teachers, while a big difference between Video 2 and video 6. In previous experiments, researchers mostly explored the consistency of panelists' judgments through MFRM analysis and examined whether the ratings of each panelist was consistent with model expectations (Cai, 2019). However, considering that this experiment only invited two teachers and cannot conduct large-scale quantitative research, which is basically a case study, there can inevitably be a not-small difference between the rating results. Therefore, all videos with a difference of less than one grade between the two teachers were counted as behaving well in consistency. Then there were about 9 videos with consistent rating results, accounting for about 81% in total, indicating good consistency of the teachers' rating results.

Table 4. Rating Results of Teachers

Number of videos	1	2	3	4	5	6	7	8	9	10	11
Actual IELTS score	5.0	5.0	6.0	6.0	6.0	6.5	6.5	6.5	7.0	7.0	7.0
Corresponding CSE level	4	4	6	6	6	7	7	7	8	8	8
Teacher 1 (CSE level)	4	7	6	5	6	7	5	8	4	7	5
Teacher 2 (CSE level)	5	4	5	5	5	5	5	7	4	6	5
Average (CSE level)	4.5	5.5	5.5	5	5.5	6	5	7.5	4	6.5	5

Table 5. One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
5.5	20	2.41	0.47	0.11
6.0	43	2.47	0.38	0.06
6.5	24	2.54	0.30	0.06
7.0	21	2.50	0.40	0.09

Table 6. One-Sample Test

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
5.5	-5.53	19.00	0.00	-0.59	-0.81	-0.37
6.0	-9.12	42.00	0.00	-0.53	-0.64	-0.41
6.5	-7.46	23.00	0.00	-0.46	-0.59	-0.33
7.0	-5.68	20.00	0.00	-0.50	-0.68	-0.32

The above is the analysis from the teachers' perspective. However, opinions from test-takers are also indispensable to the validation of linking results. Table 5 and table 6 give the statistical details of consistency between students' actual ability and the ability demonstrated in the self-assessment scale. Totally there were 113 valid questionnaires, of which only 5 samples of 5.0 were collected, and the amount was far from enough for statistical analysis, so the samples of 5.0 were removed, and thus only 5.5-7.0 samples were calculated in this study. From table 5, the mean score of each score segment of 5.5-7.0 was between 2.0 and 3.0, that is, the examinees' judgments of their own speaking ability level were between "uncertain" and "yes", indicating that there existed a fuzzy zone when students self-assessed their speaking ability, and their self-perception of speaking ability was relatively poor. From table 6, there was a significant difference between the mean scores of each score segment and "3" ($p<0.05$), which showed that from the perspective of students, the consistency of linking results between IELTS speaking test and CSE was not as high as that assessed by teachers. That maybe because the students' self-assessment ability is not strong enough and they often have a deviation in the understanding of the descriptors. Actually, some scholars have also analyzed students' self-assessment results before, mainly focusing on students' poor self-assessment ability and unreasonable scale design. Through correlation analysis, Wang Hua (2020) proved that teachers' evaluation ability was higher than students' in subjective evaluation, and the teachers' evaluation results were more credible. What's more, by the method of text mining, he found that the descriptors in part of the adjacent levels are not clear-cut in speaking subscales of CSE (Wang, 2020). When conducting the study on the relations of the Chinese Placement Test and CEFR (Common European Framework of Reference for Languages), Luo (2017) also found that the correlation between students' self-evaluation results and teachers' was not high, and speculated that this may be related to the design of the scale, students' motivation to fill in the self-evaluation form and insufficient self-evaluation training for students. Based on the above two points, the indistinct boundary between the descriptors of adjacent levels makes the examinee group with low self-evaluation ability easy to get perplexed and confuse and mix language abilities of adjacent levels, bringing difficulties to the accuracy of the examinees' self-evaluation. This may violate the initial purpose of setting the self-evaluation form——promoting learning, that is, to let the examinees understand their advantages and disadvantages based on corresponding descriptors of their test scores, so as to achieve the effect of feedback. And with the knowledge of specific ability gap between adjacent levels, test-takers can have specific and clear ability goals in mind. Therefore, it is advised that the self-assessment scale should be set to be closer to students' understanding ability in the statement of descriptors, and students should receive corresponding training before doing self-evaluation against the self-assessment scale.

In addition to the above two reasons, the present paper puts forward another two conjectures: 1) when dividing the level of descriptors, there is a difficulty gap between each level, leading to a certain degree of ability gap even for test-takers at the same level. 2) The construct of IELTS speaking test does not wholly match the language use domain defined by the scale descriptors, and the low degree of generalizability leads to a low degree of consistency accordingly. For the first conjecture, during the grading process of scale descriptors, the experts divided all descriptors into 9 levels according to the principle of approximate equal distance, and the spacing of each level is about 0.7 logit (Liu, 2021). That means there is a minimum difficulty level and a

maximum difficulty level at each grade of the scale, and some test-takers at the same level may have just reached the minimum ability of this level, while some may have reached the highest level of ability. When this difference is reflected in a large number of sample groups of test-takers, it will lead to different scores in self-assessment, so the average value can never fully reach the 3-point level of the full score. Moreover, some studies believed that the test-takers should be recognized to reach certain level when the self-assessment score exceeded 2.5 (Luo, 2017). Then, the average values of 6.5 (2.54) and 7.0 (2.50) both reached 2.5, while the differences between 5.5 (2.41) and 6.0 (2.47) and 2.5 points was also small. Therefore, on the whole, the degree of consistency is good. For the second conjecture, the construct of IELTS speaking test is not completely equivalent or related to the construct of descriptors in CSE, so the language use domain defined by some descriptors in the scale is not within the scope of IELTS speaking test construct. Three descriptors with the lowest scores are listed in Table 7. The common feature of these three descriptors is that they are all divorced from the daily life situation. However, the IELTS speaking test is mainly aimed at daily oral communication. On the official website of IELTS, it writes: the speaking test focuses on the ability to communicate opinions and information on everyday topics and common experiences or situations. There are not too many requirements for literary review, academic discussion or business communication. Therefore, when judging these ability descriptors, that test-takers found themselves fail to reach the relevant ability becomes understandable. Taking all above into consideration, it is reasonable to speculate that the correlation between the language test and language standards will, to some extent, affect the consistency of linking results. This finding attests Cai's (2019) research result: the consistency can be weakened when the number of relevant descriptors is small.

Table 7. Mean Scores of Descriptors

Can-do descriptors	Mean
1. I can briefly analyse literary works and explain their artistic effects.	2.21
2. I can express my viewpoints accurately and fluently on professional topics at academic seminars.	2.33
3. I can communicate and negotiate effectively on various matters in business telephone communication and dispute settlement.	2.29

In general, the consistency of linking results is high from the perspective of teachers while relatively low from the students' perspective, but in view of various reasons, the linking results perform well in consistency.

5. Conclusion

In the validation of the results of linking productive language tests to language standards, generalizability and consistency are always two key issues worthy of attention. Based on this recognition, this study takes oral skills as an example to verify the validity of the linking results of IELTS test and CSE. The innovation lies in the extra self-assessment step for students as test-takers which is adoptable in future researches. But limitations also existed. Firstly, only four subscales related to IELTS speaking test were selected, thus there may exist a deviation between the congruence degree obtained in this paper and the actual congruence degree between the IELTS speaking test and CSE descriptors; secondly, in the consistency evaluation experiment of teachers, only two teachers were invited to participate in, which was closely to a case study without large-scale teacher research. Therefore, the data obtained may be affected by randomness and had subjective color and thus the data analysis was a little bit rough. Finally, due to limited time and energy, the present study didn't verify the validity of standard error of cut scores which is a significant part in consistency. Future studies are recommended to include this point in the validation chain.

Acknowledgements

I am deeply indebted to my supervisor, Professor Wu Xuefeng. From the selection of topic to the accomplishment of whole dissertation, it was he that answered a cascade of strange and naive questions from a beginner over and over again, and it was he that carefully identified the flaws and shortcomings in the dissertation——from the depth of theory to the smallest formatting and punctuation. I often think how fortunate I was to have met such a supervisor right at the beginning of my academic journey. During interactions with Mr Wu, I never had the opportunity to feel the condescension and oppression of a so-called university professor, but only equality, sincerity and kindness.

References

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Asessment Quarterly, 1*(2), 1-34. https://doi.org/10.1207/s15434311laq0201_1

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

Cai, H. W. (2019). Validity issues in linking tests of productive language skills to language standards: Generalizability and consistency. *Modern Foreign Languages*, (5), 709-721.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412985918

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR): A manual.* Strasbourg: Council of Europe.

Dunlea, J. (2015). *Validating a set of Japanese EFL proficiency tests: Demonstrating locally designed tests meet international standards* (Unpublished doctoral dissertation). University of Bedfordshire, Luton, UK.

Han, B. C. & Luo, K. Z. (2013). The evolution of validity and validation in language assessment. *Foreign Language Teaching and Research, 45*(3), 411-425+481.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, (4), 333-362. https://doi.org/10.1080/15434303.2015.1092545

He, L. Z. (2019). A Validation Framework for Linking Language Tests to Language Standards. *Modern Foreign Languages*, (5), 660-671.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

Jie, W. (2019). Relating English speaking tests to China's Standards of English: CET-SET 4 as a case study. *Foreign Language World*, (1), 71-80.

Liu, J. D. (2021). Validating China's Standards of English Language Ability. *Modern Foreign Languages, 44*(1), 86-100.

Luo, L. (2017). A Study on the Relations of the Chinese Placement Test and CEFR. *Applied Linguistics*, (2), 110-118. https://doi.org/10.16499/j.cnki.1003-5397.2017.02.013

Martyniuk, M. (2010). *Aligning tests with the CEFR, studies in language testing 33.* Cambridge: Cambridge University Press.

Tannenbaum, R. J., & Cho, Y. (2014). Criteria for evaluating standard -setting approaches to map English language test scores to frameworks of English language proficiency. *Language Assessment Quarterly, 11*(3), 233-249. https://doi.org/10.1080/15434303.2013.869815

Wang, H. (2020). Aligning school-based English proficiency tests with China's Standards of English Language Ability: A case study. *Foreign Language World*, (5), 72-79.

Wang, H. (2020). Validating the speaking scale of China's Standards of English Language Ability:A text-mining approach. *Journal of Xi'an International Studies University, 28*(2), 69-74. https://doi.org/10.16362/j.cnki.cn61-1457/h.2020.02.014

Xu, Q. L. (2012). The AUA Framework——the Evolvement of Language Assessment Theory in the New Era. *Technology Enhanced Foreign Language Education*, (1), 37-41.

Min, S. C. (2019). Validity Issues in Linking Tests of Receptive Language Skills to Language Standards: Consistency. *Modern Foreign Languages, 42*(177), 696-708.

Bachman, L. F. (2006, April). Linking interpretation and use in educational assessments. Paper presented at *the National Council for Measurement in Education, San Francisco.*

Pan, R. J., & Wang, J. M. (2021). Aligning Examination for Preparatory Education for International Students in China and Chinese Proficiency Standard. *Examinations Research*, (5), 3-9.

Jin, Y., & Jie, W. (2017). Principles and methods of developing the Speaking Scale of the China's Standards of English. *Foreign Language World*, (2), 10-19.