ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 11, No. 5, pp. 801-808, September 2020
DOI: http://dx.doi.org/10.17507/jltr.1105.16

Analysis on the Reliability and Validity of Teachers' Self-designed English Listening Test

Zhencong Liu
School of English Language, Culture and Literature, Beijing International Studies University, China

Ting Li
School of English Language, Culture and Literature, Beijing International Studies University, China

Huiying Diao
School of English Language, Culture and Literature, Beijing International Studies University, China

***Abstract*—Language testing plays a vital role in English teaching. It can accurately reflect the teaching effect in a short period of time, and it is also an indispensable teaching method to assess the knowledge of students. The current study takes the final test of adult English listening class in School of Continual Education as an example, under the theoretical language assessment framework of Bachman and Palmer, uses the data collected by the statistical analysis software SPSS to test the reliability and validity of the listening final exam from a statistical perspective. The study found that the reliability and validity of the selected listening test were generally acceptable, the differentiation among students was obvious but it has high item difficulty. Therefore, it is necessary to improve the authenticity of the listening test and the communicative skill of the listening material. This study is conducted to find out the problems of the current listening test in the School of Continual Education, and propose specific solutions based on the basic elements of the language test. It is hoped that the research in this article will play a positive role in designing adult English listening tests.**

***Index Terms*—listening test, reliability, validity**

I. Introduction

The function of language testing is to provide important information for teaching assessment, especially in teaching process. Through the analysis and research at this point, the lecturer can deeply understand the overall knowledge proficiency of students and study whether the test questions are appropriate, so that they can conduct further teaching activities more scientifically and effectively. In order to accurately and objectively evaluate the effects of teaching and learning from the test, teachers are required to scientifically and effectively design the test questions. "An untrustworthy test score is absolutely ineffective. Reliability is a reflection of the quality of test score itself; validity is a reflection of the correct interpretation and use of the test" (Bachman, 1990). Reliability and validity are two essential elements of language testing. To improve the teaching quality of college English involves many factors, scientific analysis of students' achievements comes to the first place. Usually, studies start from the actual teaching process of college English and try to find scientifically reasonable college English test modes, so that the test can truly become an effective means of diagnostic assessment and instructional teaching.

As part of the language proficiency test, the listening test aims to measure the listening comprehension ability of participants. Traditional listening comprehension test is designed to test the language ability of participants by their actual performance in a limited sample of test tasks within a short period of time. However, not all examinations can actually and exactly reflect the real proficiency of participants. As two important elements in language testing or language assessment, it is of great significance to assess the reliability and validity of the selected test. Thus, the current study is conducted in school-based listening test and tries to analyze the reliability and validity of it under the theoretical assessment framework of Bachman and Palmer. The study aims to find out the potential problems in school-based listening tests and try to make improvements in future teaching activities and further studies.

II. Literature Review

Testing plays an important role in college English teaching. From the purpose of testing, language tests can be divided into proficiency tests, grading tests, achievement tests, and diagnostic tests. No matter which type of assessment we use, the test must be objective, purposeful and meaningful, in other words, its reliability and validity must be highly valued. As Bachman shows that reliability and validity are important qualities for the interpretation of language proficiency. "If we will explain the score of a given test as a mark of personal ability, then the score will be must be credible and valid" (Bachman, 1990). Before we stepped into the current study, it is very necessary to review previous studies on listening test assessment. This part can be divided into the studies on the validity of listening tests,

school-based assessment on listening tests and the potential problems in assessment on listening tests.

Firstly, Peng and Yuan have reviewed domestic studies on the test of English listening. The survey shows that: a. as for participants, in general, the studies mainly focuses on college students; b. most researches focus on nationwide tests rather than regional or school-based tests; c. the major content of the study involves various aspects, but there is a lack of study aligned with English curriculum standard (Peng & Yuan, 2015). Specifically, previous studies mainly focus on the validity of listening tests and very little studies pay attention to the influence of question type upon scores, limited studies focus on the assessment based on authentic English material.

To step further, previous studies on the validity of listening tests can be summarized from two aspects: theoretical studies and empirical studies. To begin with, Zhao (2000) expounds the understanding of validity and reliability from different perspectives on language acquisition and language learning. He believes that modern language testing is biased towards reliability, and language testing should focus on validity requirements and pursue as much as possible on this basis. Besides, Zhang (2003) analyzed listening strategies and test validity from two basic aspects: basic language knowledge and listening tasks. In addition, Song (2011) theoretically explores internal validity, external validity, and constructs validity to reveal the conceptual connotation of evidence of different validity.

On the other hand, empirical studies which mainly focus on how to achieve validity in the practice of listening test. These studies have focused on various exams in university. The university-level examination research mainly discusses the topics related to the validity of listening tests, such as the comparison of new and traditional CET-4 and CET-6. There are also college entrance examinations and school-based examinations.

Secondly, it is necessary to review the assessment on school-based listening tests. The current school-based test research mainly focuses on the academic achievement test of English majors and the research of listening test under the framework of criteria reference language test (CRTs). Wei (2007) conducted her study based on the analysis of Cronbach Alpha, mean, standard deviation, variance, correlation coefficient and factor analysis of a graduation test for English majors. Jiang Lan, Feng Xiaoyuan (2003) studied the validity of teachers' self-designed English examinations, and put forward 9 questions to be explored and researched in terms of propositions, examination implementation and management. Huang Ping (2001) started with diagnostic assessment in college English test and compared the scores between their college English test scores and final English test scores. According to her research, it is considered that a unified test in college English tests is necessary and feasible.

Last but not the least, previous studies have examined the problems in listening tests. Researchers have pointed out the following problems: a. unreasonable test system (Jing, 1999). It is difficult to reconcile the national unified examination standard and regional education differences; b. lack of authentic material (Niu, 2001). The authenticity of the English materials is not enough, for example, lack of titles and instructions, it will affect surface validity of listening tests; c. failed to present multiple question types (Niu 2001; Wang 2004); d. English listening test led by multiple choice questions lacks sufficient construct validity to measure students' listening proficiency effectively; e. the quality control mechanism of the raters needs to be improved (Niu, 2001); f. the school-based tests and classroom tests are not paying enough attention, and the qualities of the questions is worrying (Qian, 2004).

III. Theoretical Background and Methodology

In this part, theoretical background and methodology will be presented, including participants, research design and research questions.

A. Theoretical Background

Reliability and validity are the fundamental requirements for the quality of language tests and other educational and psychological measurements. They are also called reliability or consistency. In this part, brief introduction of reliability and validity will be introduced at the first place, then, potential factors which would influence them and necessities which can ensure their accuracy will be presented further.

Firstly, reliability refers to the degree to which the test results of a test paper are consistent, that is, the test results are not affected by external factors such as time, proctors, and classrooms. Validity refers to the extent to which a test paper can meet the purpose of the assessment. It is a matter of correctly interpreting the scores purposefully. If a test paper tests several language skills at the same time or the test content exceeds the purpose of the test, its validity will be greatly reduced. Reliability and validity are closely related and inseparable. They are related to the fundamental purpose of the test, namely how to accurately and consistently test the language ability we want to test. These two must be constrained and relatively balanced to serve the basic purpose of the test.

Secondly, language ability itself cannot be measured directly, so the concept of language ability can generally base on the observation of empirical statics or behaviors (Chen, 2011). The traditional view is that test performance is the actual language ability. In most cases, listening tests designer is most interested in assessing the listening comprehension ability of the participants through testing. The effectiveness of the listening test lies in the test itself, whether the results can truly reflect the listening comprehension ability that the participants should have in real life situation or not.

As mentioned above, reliability refers to the stability and consistency of the questionnaire results when the same method is used to study the same group of participants, in other words, whether the test results truly and objectively reflect the actual level of the participants or not. Reliability is an index that reflects whether the test is affected by

non-test factors, and reflects the objectivity and reliability of the test. There are many factors that affect the reliability of the test and they are mainly related to the two aspects of questions and scores. As far as the questions themselves are concerned, their reliability depends mainly on the scope of the test and the amount of questions. To ensure a high degree of reliability, first of all, ensure that the test paper has a certain amount questions. Generally speaking, the greater the amount of questions, the higher the reliability is. At the same time, the scores used as the test results must have a certain degree of dispersion. To meet this requirement, it means that the test paper must be highly differentiated, which can distinguish candidates at various levels, and the difficulty of the questions should be moderate. Too difficult and too easy to distinguish the level is unacceptable.

Meanwhile, validity refers to the degree to which a measure tool or means can accurately measure what needs to be measured, to see if the purpose of the test is achieved, in other words, whether the degree of accuracy and validity of the test results would be ensured. What you test should include what items, not involving irrelevant content. A set of questions must have at least surface validity, which gives first appropriate impression to people. The most important thing is content validity. The language elements and skills that should be examined must be reflected effectively. Then, construct validity which means that a set of tests should be based on a theory of language acquisition or language learning. Whether the test has achieved the purpose or whether the content of the test is what you want to test. Validity is at the core of an exam. There are many types of examinations in China, and the scale is relatively large, but most of them are based on obtaining results, and rarely consider whether these results are reliable and effective. Analysis of test results is rarely required to be explained further. Validity is as important as reliability in a test. If you omit any of them, the quality of the assessment cannot be guaranteed.

B. Participants

To begin with, the selected listening test will be introduced briefly. It is the final examination designed in the School of Continual Education for the first semester in 2019-2020 school year, with 100 points total test score, including 80% in-class knowledge and 20% extra-curricular knowledge. The whole listening test is designed with five parts and students will be required to finish all of them in one and a half hour: fill in the blanks (10 * 2'= 20'), single-choice questions (10 * 2'= 20'), true or false (10 * 2'= 20'), summarize and fill in the blanks (10 * 2'= 20'), and also fill in the blank (extra-curricular knowledge) (10 * 2'= 20'). Listening proficiency of students will be assessed from four general learning strategies and skills: vocabulary spelling, short-term memory, summary and refining, text understanding. After the examination, tests will be assessed by the English teachers of the School of Continual Education, with unified standards and special responsibility. The names and student numbers of the examinees are sealed to be objective and credible.

Therefore, the research object of the current study is the final listening test scores of 20 students in the third grade in the School of Continual Education. Their total scores and sub-question scores will be used for the reliability and validity test analysis. Their scores will be selected in this study to test the reliability and validity of the adult English listening for the final exam.

C. Research Design

This study will use statistical analysis software SPSS to collect the data required for reliability and validity analysis. The score of each subject will be input into SPSS software to test the reliability and content validity of the entire test question.

First, the reliability of the test questions is mainly analyzed by the Cronbach Alpha reliability coefficients. The dispersion of the test and the distribution of test scores are analyzed by detecting the mean, standard deviation, and variance; then the author will test the correlation coefficient, including factor analysis.

D. Research Questions

The main issue of the current study is to use the data collected by the statistical analysis software SPSS to statistically test an English listening end-of-term test. The purpose of this study is to sort and summarize the test to help teachers in designing language assessment. In the process of developing listening tests, try to reduce the impact of measurement errors as much as possible, and increase the testing intensity of the listening ability that you want to assess in order to improve the efficiency of listening tests. There are three research questions:

a. How is the reliability coefficient of the selected listening comprehension test?
b. How is the dispersion and frequency distribution of the selected listening comprehension test?
c. How is the validity of the selected listening comprehension test from content validity and factor analysis?

IV. Findings

In this part, reliability and validity of selected listening comprehension test will be analyzed in statistical software SPSS respectively. Data analysis will be presented in Chapter 4.1 and Chapter 4.2.

A. Reliability Analysis

In this part, reliability coefficient, descriptive statistics and frequency distribution will be discussed respectively.

1. Cronbach Alpha reliability coefficient

The chosen final English listening test contains 50 questions, totally 100 points, which meet the required number of tests (Li, 1997). Also, the listening test has been divided into version A and version B. These two versions share 80% the same content and the current study is conducted in version A. The topics of the test are common, suitable for students and assess comprehensive language skills for students at the same time. It would be noticed that 80% content of the test is chosen from textbook and the rest of the test is chosen from online resources. Besides, each question is accompanied by clear and concise instructions. Last but not the least, the assessment of tests is the responsibility of the English teacher of the School of Continual Education, under unified standards. Students' names and numbers will be sealed to be more objective and credible. However, as stated by Li & Shao, only by statistical analysis in scientific methods, we can obtain comprehensive, true, accurate and credible statistical Data (2003). For more detailed and accurate evaluation, the current study will make use of statistical software SPSS to measure reliability and validity of the selected test. See Table 1 and Table 2:

Table 1 Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.865	.870	48

Table 2 Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.818	.816	5

(Warnings from SPSS for Table 1: Each of the above component variables has zero variance and is removed from the scale: item7, item20)

As presented in Table 1, Cronbach's Alpha is 0.865, which shows that the selected listening test has a high reliability between 48 sub-items. Generally speaking, if the Cronbach's Alpha exceeds 0.7, the selected test is regarded as a test with high reliability. For accurately assessment, the author counts score again of each part and tests the Cronbach's Alpha again. The data shows that the Cronbach's Alpha is 0.818 which shows that the selected listening test has high reliability. In other words, the selected listening test has high consistency and the result can reflect students' actual English listening proficiency well. All in all, no matter considers 50 sub-questions or five parts of the selected English listening comprehension test, Cronbach's Alpha is high enough to prove that the listening test has good reliability among each items or each parts.

2. Descriptive statistics and frequency distribution

TABLE 3
DESCRIPTIVE STATISTICS

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Score	20	62	32	94	63.30	15.345	235.484
Valid N (listwise)	20						

When it comes to the descriptive statistics, as stated in Table 3, Mean is 63.30, which meets the standard of 100 points on average. Along with Minimum 32, Maximum 94, the range is 62. The above data shows that there is big variance among students. To step further, the selected listening test has created big variance between 20 participants. At the same time, the Std. Deviation (15.345) also shows that 20 students have very different scores of this test, that is, they have different levels in the items that the teacher wants to test. It is fair enough to infer that the chosen participants have different listening comprehension ability. To analyse further, the frequency table will also present the dispersion between students, see Figure 4:

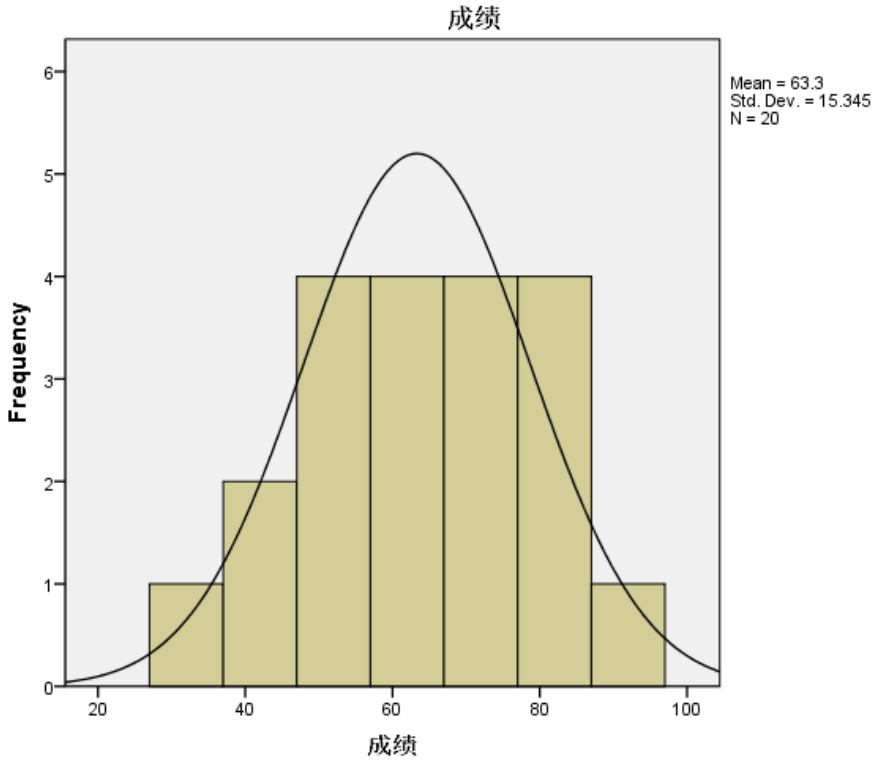

Figure 4 Frequency distribution

As shown in Figure 4, the frequency distribution of scores presents a near normal distribution at the cut-point 60, which again proves that participants have very different listening comprehension ability and the selected listening comprehension test doesn't have a balanced difficulty, which differentiates 20 participants in a bad manner, because the number of students below 60 is roughly equal to the number of students above 60. The distribution is as follows, see Table 5:

TABLE 5
THE DISTRIBUTION OF STUDENTS' TEST SCORE

	0-59 分	60-69 分	70-79 分	80-89 分	90-100 分
人数	8	5	6	0	1
比例	40%	25%	30%	0	5%

The number of students who get final score lower than 60 and higher than 80 is approximate. Also, the number of students who get 0~59 is the largest. It is not a good result for achievement test which has 40% of students who haven't met the requirement of the given test. Therefore, the selected listening comprehension test failed to keep a balanced difficulty even though it can differentiate students well.

B. Validity Analysis

In this part, content validity, correlation coefficient among five parts and factor analysis will be discussed respectively.

1. Content validity

Firstly, the instructions of five parts in the selected listening test will be introduced briefly to see what kind of listening skills would be tested under each sub question.

"Part I: Write the missing words. You will hear the recording twice. (10 items, 20%)"

"Part II: Listen to each conversation and then choose the correct answer. You will hear the recording twice. .(10 items, 20%)"

"Part III. Listen to the following conversations and decide whether the statements are true (T) or false (F). You will hear the recording twice. .(10 items, 20%)"

"Part IV. Complete the summary below according to what you hear. Use ONE word only for each blank. You will hear each conversation twice. .(10 items, 20%)"

"Part V. Listen to the following passage entitled *The Language of Air Travel*. Fill in the missing words. You will hear the recording three times. .(10 items, 20%)"

As mentioned in Chapter 3, the selected listening comprehension test is designed to assess vocabulary spelling, short-term memory, summary and refining, text understanding, etc. Generally speaking, listening comprehension skills

are basically assessed in the final term examination. More profoundly, students will not only be tested to choose and tell facts and details, they are also required to summarize certain sentences and make judgments according to their short time memory. Therefore, the selected listening test has good content validity.

To step further, we will see the differentiation among each part, see Table 6.

TABLE 6
ITEM STATISTICS

	Mean	Std. Deviation	N
part1	5.80	2.505	20
part2	7.40	1.569	20
part3	7.60	1.957	20
part4	6.35	2.110	20
part5	4.50	1.821	20

As stated in Table 6, part one and part four have higher Std. Deviation than the other three parts. By observing the instructions of each part at the beginning of this part, we can see that both part one and part four are designed to assess students' vocabulary spelling. Students need to listen to the type and then write the missing words, in part four, they are also required to do summary before they write down their answers on the answer sheet. This is a great for students in listening comprehension test than the other objective questions which need students to pick up correct answer.

2. Correlation coefficient and factor analysis

In this part, correlation coefficient among each part and factor analysis will be examined in SPSS. Correlation coefficient ranges from -1 to 1. The higher the value is the higher correlation between two parts or items.

Firstly, as stated in Table 7, part one, part three and part four have relative high correlation with other parts. However, part two and part five haven't presented high correlation coefficient with other parts. The data shows that part one (vocabulary spelling), part three (true or false) and part four (summary and word spelling) share higher correlation coefficient than that of other two parts.

For part two, it has lower correlation coefficient with part four and part five, which is 0.194 and 0.184. In this part, the ability of picking up detailed information will be emphasized, which is different from part four (summary and word spelling), part five (word spelling). The author holds the view that they have different emphasis on the abilities they want to test, which leads to lower correlation coefficient among them.

Similarly, in part five, it has lower correlation coefficient between part two (multiple choice) and part three (true or false), which is 0.184 and 0.295. The author thinks that because of the question type, part five shares higher correlation coefficient value with similar question type, like part one and part four.

TABLE 7
CORRELATION MATRIX

		part1	part2	part3	part4	part5
Correlation	part1	1.000	.450	.477	.701	.669
	part2	.450	1.000	.740	.194	.184
	part3	.477	.740	1.000	.354	.295
	part4	.701	.194	.354	1.000	.637
	part5	.669	.184	.295	.637	1.000

Next, factor analysis is designed to analyse the sub-category in the selected listening comprehension test deeply. SPSS will help us to observe the sub group of selected test. In the current study, two components are presented in factor analysis, which means that the selected listening comprehension test mainly test two kinds of abilities (see Table 8). Specifically, part one, part four and part five share high value in factor analysis, because their question types are very similar---word spelling. All of them have emphasized the vocabulary.

On the other hand, in component two, part two and part three share high values in factor analysis. From the question type, both of them are objective questions. Students are required to choose the correct answer among four descriptions and make judgments according to their listening on each conversation. Thus, the selected listening comprehension test has assessed two kinds of ability: word spelling and the ability for detailed information.

TABLE 8
COMPONENT MATRIXA

	Component	
	1	2
part1	.886	-.186
part2	.646	.689
part3	.733	.558
part4	.778	-.433
part5	.748	-.471

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

In conclusion, generally speaking, the reliability and validity of the selected listening comprehension test is good, but there are some improvements in the question types and the source of listening material. It will be discussed in Chapter 5.

V. Discussion

The results in Chapter four has shown that the selected listening comprehension test has good reliability and validity, but there are still some improvements need to be noticed. In this part, the author will discuss the result briefly and hopes to put forward some feasible suggestions, especially in the current adult English listening teaching and assessment.

At the very beginning, the given result has shown that question type should be enriched in the current listening comprehension test, not only the objective questions, see true or false and multiple choices. We have paid less attention to subjective questions. On the one hand, rich question types can be beneficial for content validity for the sake of flexible question design. Tang (2009) has summarized the advantages and disadvantages of multiple choices. She stated that the multiple choice question type is objective and economical, which can increase the amount of test questions and the coverage of the test, and the scores are more objective to give. However, researchers also doubt that multiple choices will influence student's learning style for the sake of they only pay attention to pick up correct answer. As an important part in English learning, listening comprehension ability should be assessed in real life situation and ask students to produce their answers based on real life context. Multiple choices questions would damage students' creativity. And at this point, free response question can be added in listening comprehension test (Tang, 2009). The free answer question type can greatly reduce the impact of guessing and other test skills on the test takers' true language ability.

Additionally, in the listening test, it is important to ensure that the language ability required by the test subject in real life is measured. At this point, the authenticity of English materials should be considered. To begin with, the chosen topics must be consistent with the real life situation. In addition, choosing from authentic English website or conversations can guarantee the authenticity of listening test tasks. Construct is an important guarantee for the validity of the listening test (He, 2005). As for the construct validity, communicative skills have been paid less attention in selected listening comprehension test. As stated in Yan wei and Wang Yong (2008), question type can also influence the reliability and validity in communicative listening comprehension test. Free response is regarded as an important attempt in current listening comprehension test.

As Bachman (1990) stated: "During the design and development of testing, we must consider two points: one is to reduce the impact of measurement errors, and the other is to increase the testing of the language ability we want to detect, to make the exam more complete." Looking forward to the future, the author believes that listening test research can be further expanded from the following aspects: firstly, expand and develop more participants involving high school students and college students. This study is hoped to offer suggestions to adult English listening comprehension test. Secondly, try to enrich the question type in the selected listening comprehension test and choose more authentic English materials. If authentic English conversation can be added in the listening comprehension tests along with free response questions, students' communicative skills will be assessed thoroughly.

VI. Conclusion

The current study takes the final test of adult English listening class in School of Continual Education as an example, under the theoretical language assessment framework of Bachman and Palmer, uses the data collected by the statistical analysis software SPSS to test the reliability and validity of the listening final exam from a statistical perspective. According to the test results and analysis of Cronbach's Alpha value, standard deviation, correlation coefficient, factor analysis, etc., from the requirements of the language test, the selected listening comprehension generally has good reliability and validity. However, there is still room for improvement in this test. Special attention should be paid to avoid too simple questions and add up authentic English material to it. Besides, more factors such as communicative listening skills and the ability for picking up complex information should be assessed. Due to the number of participants, there are some limitations in the study, such as the insufficient presentation of frequency distribution. It is hoped that the research in this article will play a positive role in designing adult English listening tests and it can be developed further by involving more participants.

References

[1] Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.
[2] Bachman, L. F. & Palmer, A. S. (1996). Language Testing in Practice. Oxford: Oxford University Press.
[3] Brown, J. D. (2005). Testing in language programs (New ed.). New York: McGraw-Hill.
[4] Chen Z. (2011). The Relationship between Reliability and Validity in College English Scores. *Journal of Ocean University of Guangdong. 31*(02):98-101.
[5] Guo L. (2003). The Investigation and Analysis about the Pattern of School-based English Testing. *Foreign Language World*. (02):76-80.
[6] Heaton, J.(2000). Writing English Language Tests. Beijing: Foreign Language and Research Press.
[7] Huges, A. (2003). Testing for Language Teachers. Cambridge University Press.

[8] Huang P. (2001). The Reliability and Validity of Tests for College English Majors. *Foreign Languages and Their Teaching*. (11):16-18.
[9] He, Y. B. (2005). Concept validity of listening test and its realization. *Foreign Language Teaching*. (03):72-75.
[10] Jiang L., Feng, X. Y. (2003). An investigation and Reflection on the Content Validity and Surface Validity of English test designed by Teachers Themselves. *Foreign Language Teaching*. *24*(5):85-88.
[11] Li, X. J.(1997). The Science and Techniques of Language Testing. Changsha：Education Press in Hunan Province.
[12] Meng, Z. Z. (2009). A Review of the Research on the Reliability and Validity of English Testing in China. *Journal of Baise Institute*. *22*(04):128-134.
[13] Niu Q. (2001). Problems in English Test at University. *Foreign Language Teaching and Research*. *33*(2):140-143.
[14] Peng J. (2008). The Realization of Reliability and Validity of English Listening Test Design. Out-campus Education in China (Theory). (S1):735.
[15] Peng, K. Z. & Yuan, Q. L. (2015). A Review of English Listening Test Research in Recent 15 Years in China. *Foreign Language Testing and Teaching*. (02):21-27.
[16] Qian, D. M. (2004). On the Validity and its Problems in the Integrated English Test. *Foreign language teaching abroad*. (3):8-12.
[17] Sun, R. M. (1998). On the Balance of Reliability and Validity in Listening Test. *Foreign language Teaching and Technology* (03):15-17+25.
[18] Song, Y. S. (2011). Theoretical Study on the Validity of Language Testing. *Learning Theory.* (10):230-232.
[19] Tang, P. R. (2009). Reliability and Validity of Question Design in Listening Test. *Information of Science and Technology*. (12):455.
[20] Wang, Y. Q. (2004). Problems and Solutions in Language Testing for English Majors in Teachers College. *Journal of Xianning Institute*. *24*(2):95－97.
[21] Weir, C. (2005). Language Testing and Validation: An Evidence based Approach. New York: Palgrave Macmillan.
[22] Yan W. & Wang Y. (2008). The Balance of Validity and Reliability -- a Discussion on the Question Types in Communicative English Listening Test. *Foreign language teaching theory and practice.* (01):69-74.
[23] Zhao, C. F. (2000). Discussion of the Reliability and Validity of Language Testing. *Foreign Language Teaching*. (01):11-15.
[24] Zhang, J. H. (2003). Validity and Theoretical Analysis of Foreign Language Listening Test. *Foreign Language Teaching in Basic Education* (10):123-127.
[25] Zhang, L.P., Dan, B. J. & Zeng, H. (2014). The Role of Reliability and Validity in College English Test. *Journal of Guizhou Normal University (Social Science Edition).* (01):157-160.

Zhencong Liu is currently Associate Professor in the School of English Language, Culture and Literature, Beijing International Studies University, China. He received his PH.D degree in linguistics from Beijing Foreign Studies University, China in 2007. His research interests include general linguistics, cognitive linguistics, theory and practice in English teaching. Dr. Liu has published more than 11 textbooks and 30 research papers on cognitive linguistics and applied linguistics.

Ting Li (corresponding author) was born in Shaanxi province, China in 1995. She will receive her Master's Degree in applied linguistics from Beijing International Studies University, China in 2021. Her research interests include second language acquisition, English teaching and discourse analysis.

Huiying Diao was born in Jilin province, China in 1994. She will receive her Master's Degree in applied linguistics from Beijing International Studies University, China in 2021. Her research interests include cognitive linguistics, corpus linguistics and English teaching.