Validation Study on a College English Achievement Test in China

Lili Zhang

School of Languages and Media, Anhui University of Finance & Economics, Bengbu, 233000, China

Abstract

This study is intended to find out if the college English achievement test in a key financial university in Anhui province is appropriate in assessing what it meant to assess and then probe into some approaches to realizing more effective teaching and learning with a more appropriate assessment tool. By using the data of 2800 1st year university students' test results and focusing on the 50 objective test items of reading comprehension and vocabulary, this study conducts an exploratory factor analysis and a following confirmatory factor analysis to examine the internal structure of the assessment tool and then discusses the teaching and learning activities of college English conducted in the university. Suggestions are put forward not only on how to design more appropriate college English achievement tests but also on how to improve the teaching and learning practices accordingly.

Keywords

Validation; College English achievement test; Reading comprehension; Vocabulary; Chinese university students.

1. Introduction

1.1. College English Achievement Test

Language tests can be divided into achievement tests, aptitude tests, competency tests, diagnostic tests, and placement tests according to the usage of tests (Henning, 1987; Hughes, 1989; Liu Runqing & Han Baocheng, 2000). Among them, achievement test is a kind of syllabus-related test, which assesses "how successful an individual student, groups of students, or the courses themselves have been in achieving objectives" (Brown, 2004) To be more specific, achievement test can measure and evaluate teaching quality. Besides, Hughes (2000) points out that testing can bring backwash effects to teaching and learning. For this reason, achievement test is indispensable to and very important in college English teaching and learning.

Yuan Ping (2002) mentioned that achievement tests should be designed to provide students with useful information, and tests should also help improve students' overall language skills. In conclusion, the achievement test should be designed to measure the purpose of teaching and the language skills of students based on what have been learned in class, which is a matter of the validity of a test, that is, if a test is valid in testing what it is intended to test. However, for a long time, only those high-stake tests like CET (College English Test Band Four or Six), TEM (Test for English Majors Band Four or Eight), and TOEFL (Test for English as a Foreign Language), etc. have received wide attention from researchers and educators. Meanwhile, along with the popularization of CET in China, CET has exerted much influence on CEAT (College English Achievement Test) such as teaching plan, teaching management, and other aspects. Moreover, even achievement tests are constructed with reference to the criteria of these high-stake tests while neglecting the ultimate aim of the achievement test, which hinders realization of effective CEAT. As a result, there exists a wide gap between college English achievement tests and College English teaching, students' lack of motivation in College English learning, and teachers' failure to enhance teaching quality through the backwash effect of college English achievement test.

1.2. Validity and Validation

Validity is the quality that most affects the value of a test. A test is said to be valid if it measures exactly what it is trying to measure (Hughes, 2000). "The validity of a test can only be determined by the specific purpose for which it is used and by the specific occasion on which it is used" and "it is concerned with the relationship between test performance and the underlying ability to be measured" (Zou Shen, 2005).

The process to collect evidence to support a hypothesis or prediction is called validation. Validation is classified into three types of content-related validation, criterion-related validation, and construct-related validation (Anastasi,1988). Content-related validation is concerned with reviewing the test content systematically to find out whether it contains a representative sample of the behavior /ability domain to be measured. Criterion-related validation examines the effectiveness of a measure with relation to an external measure called a criterion and construct-related validation involves the examination of the extent to which a test measures the underlying or hypothesized trait or concept.

Messick (1989) holds that content-related and criterion-related evidence all contribute to score interpretation so that they can be viewed as aspects of construct-related evidence. In this sense, there is only one type of validity evidence left, that is, construct-related evidence. Moreover, validation centers on looking for evidence supporting construct, with other types of evidence being contributory. Thus, test validation, to some extent, is about construct validation. A construct is the latent ability of a person that cannot be measured or observed directly, and construct validity is about "the relationship between what is hypothesized (attribute/trait) and what is observed (test performance)" (Zou Shen, 2005).

In addition, although the examination of a construct is what validation aims at, this does not mean that other factors outside the test itself can be left out of our attention. Validation should go beyond the test in question to incorporate social consequences, as is stated by Messick (1989). Another point worth mentioning here is that validation is not a post-test process as is generally assumed by many people. It should start from the beginning of test development. That is to say, validation is an ongoing process. According to Cronbach and Meehl (1955), "a construct is some postulated attribute of people, assumed to be reflected in test performance".

1.3. Language Competence

As for college English achievement tests, the test construct is language competence. So defining the measurement objective is giving an operational definition of language competence. Throughout history, many scholars have discussed language competence, like Chomsky (1965), Hymes (1972), Campbell and Wales (1970), Hallidy (1973, 1978), and Michael Canale & Merrill Swain (1980).

Chomsky (1965) made a distinction between competence and performance, and according to Chomsky (1965), language competence is mainly about knowledge of grammar and other aspects of language. Hymes (1972), Campbell and Wales (1970), and Hallidy (1973, 1978) defined a broader concept of communicative competence that encompasses contextual or sociolinguistic competence as well as grammatical competence. Michael Canale & Merrill Swain (1980) further enriched language competence with communicative competence, which consists of a minimal composition of grammatical competence, sociolinguistic competence, and strategic competence.

On the basis of the results of the former researches, Bachman (1990) refined the communicative language competence framework and clarified the interrelationships between various subcategories. Bachman (1990) noted that "language competence is composed of organizational competence and pragmatic competence with interactive subcategories of grammatical competence, textual competence, illocutionary competence, and sociolinguistic competence". Bachman's (1990) communicative language competence model is commonly accepted in the current language testing field. It was also adopted in this research.

2. Methodology

This study takes a college English achievement test in a key financial university in China. It focuses on the validation study of the reading comprehension and vocabulary test items of it. There are two research questions stated as follows.

RQ 1: How do the reading comprehension and vocabulary test items in the college English achievement test fit?

RQ 2: Is the college English achievement test appropriate in testing what it meant to test?

2.1. Participants

2800 freshmen of a key financial university in Anhui province participated in this achievement test as their final test in their first semester in December 2016. Their majors cover 62 majors of economics, finance, accounting, insurance, law, advertising, and public administration, etc.

2.2. Instruments

The research tool used in this study was part of the test paper designed for "college English (I) intensive reading course". With 20 reading comprehension items from Part I and 30 vocabulary items from Part II, altogether 50 items were chosen for this research.

Since the college English achievement test in this key financial university in Anhui province is a kind of syllabus-related test, it is necessary to talk about the syllabus of this course. College English (I) syllabus in this university identified and measured acquisition of language competence based on the following six aspects of listening, speaking, reading, writing, translation, and vocabulary. The objectives measured by these aspects respectively included those interactive competencies of grammatical competence, textual competence, illocutionary competence, and sociolinguistic competence involved in language usage and communication.

Table 1. Performance objectives of College English (I) Intensive Reading Course

Category	Description
Reading	Students can read English texts in diverse styles, grasp the main ideas and understand major facts and relevant details at a speed of 100 wpm, and accuracy rate of 75%.
Vocabulary	Students can grasp 4500-5000 words and 900-1100 phrases (including those that have been grasped during their senior high school English studies), among which 2,500 words and phrases can be used actively in their speaking and writing.

Since there is an independent course named "college English listening and speaking", in which students' performance objectives of listening and speaking were measured, there were no items of listening and speaking in this college English achievement test.

2.3. Data Collection and Analysis

Convenience sampling is used in this study. Test papers were handed out to students by entrusted teachers, who are familiar with research ethics and data collection. The allocated time for the 50 items of reading comprehension and vocabulary in the achievement test was 50 minutes. After the missing data were deleted, 2800 students' data were used in this study. jMetrik program, SPSS, and AMOS were used in this study for data analysis.

3. Results

3.1. Result of Research Question 1

In order to find out how the test items of reading comprehension and vocabulary fit, the results of the equation of examinees' ability and item difficulty are made by logistic scale so that they can be compared with each other. Figure 1 shows that the majority of the items in the college English achievement test can be interpreted as items within examinees' ability range.

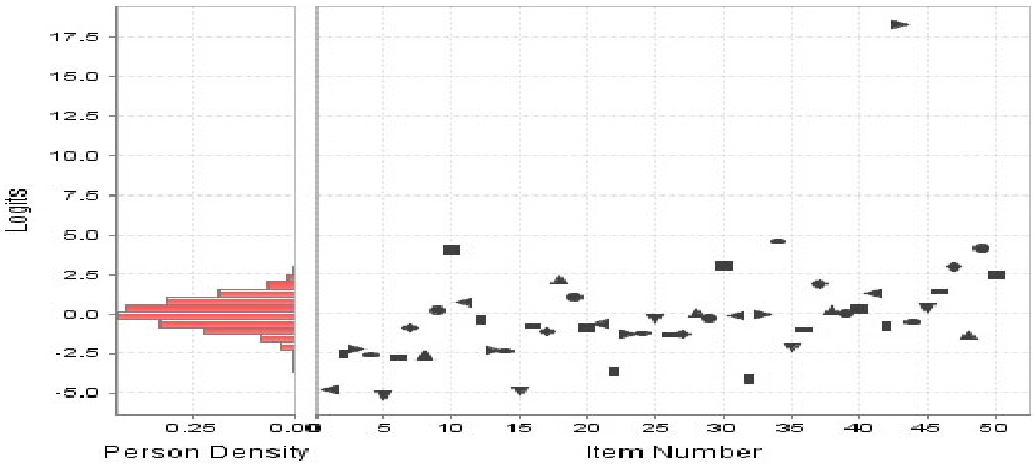

Figure 1. Equation of Examinees' Ability and Item Difficulty

According to figure 2, in this test, the ability level of the majority of students is from -1 to 1. Accordingly, it indicates that this college English achievement test has high reliability.

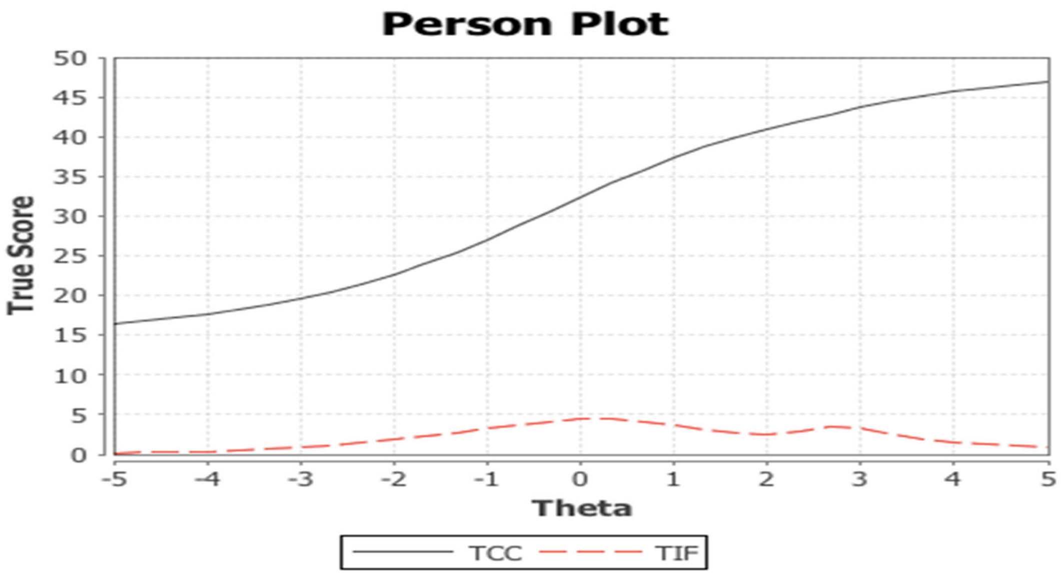

Figure 2. Test Characteristic Curve

Table 2 and Table 3 display the discrimination and difficulty distribution of individual items. From the perspective of reliability, since 66% of the items with discrimination or high discrimination and 24% of the items with medium difficulty, it indicates that this test tool is appropriate in reliability.

Table 2. Discrimination Distribution of Individual Items

Category	With none discrimination	With low discrimination	With moderate discrimination	With high discrimination	With very high discrimination	Sum
	.00-.35	.35-.65	.65-1.35	1.35-1.7	>1.7	
Items	1,3,10,11,15,22,30,48,49	6,7,8,12,14,17,19,23	2,4,5,9,13,16,18,20,21,25,26,27,28,31,32,33,34,35,36,37,38,41,42,43,44,46	24,29,39,40,45,50	47	
Number of items	9	8	26	6	1	50
Ratio	18	16%	52%	12%	2%	100%

Table 3. Difficulty Distribution of Individual Items

Category	Very easy	Easy	Medium	Difficult	Very difficult	Sum
	-2.0	-2.0--.5	-.5-.5	.5-2.0	>2.0	
	1,2,3,4,5,6,8,13,14,15,22,32,35	7,16,17,20,21,23,24,26,27,36,42,44	9,12,25,28,29,31,33,38,39,40,45,48	11,19,37,41,46	10,18,30,34,43,47,49,50	
Number of items	13	12	12	5	8	50
Ratio	26%	24%	24%	10%	16%	100%

According to the outfit criterion of 0.8-1.2, items 1, 34, 43, and 47 with outfit values of more than 1.2 are excluded from later analysis. Detailed information is presented in table 4.

Table 4. Item Characteristics Analysis and Ability Parameter Estimation

Item	D	SE	p	SE	G	SE	Infit	SE	outfit	SE	True Score	θ	Estimation Error
1	.25	.11	-4.82	9.47	.70	.42	1.04	.83	1.24	3.57	0	6.23	1.88
34	1.33	.49	4.55	.95	.05	.00	1.04	.66	1.33	4.21	33	.86	.34
43	.82	.81	18.29	18.17	.22	.01	1.12	6.15	1.26	8.93	42	2.15	.44
47	2.16	.30	2.93	.15	.06	.00	1.01	.27	1.28	4.17	46	3.17	.59

Notes: D=Discrimination; p=Difficulty; G=Guessing

Taken into account of the number of test items, test time, item difficulty distribution, there is an even distribution of items ranging from being very easy to very hard, so that ability differences among students could be distinguished, and with 50 minutes of allocated time for the 50 items, it indicates that this college English achievement test is a proper test.

3.2. Result of Research Question 2

Research question 2 is designed to find out whether the test is appropriate in testing what it meant to test. To check what kinds of language competencies the achievement test measured, first KMO and Bartlett test is used to measure whether the items are related to each other. With the value of the measure of sampling adequacy at .858, Bartlett's test of Sphericity at a value of 13354.471 (df=1225, p<.000), the results indicate it is suitable to conduct factor analysis of those items in this achievement test (Kaiser,1974).

The weighted least square mean and variance (WLSMV) method was applied to extract the basic structure of exploratory factor analysis, and square rotation was used to do factor rotation. In order to determine the number of factors through exploratory factor analysis, the steep slope point of the scree plot and the decrease scale of the non-standardized fit index were taken into consideration. From the scree plot, there is sharp slope among the first four factors, and from the fifth factor, the slope becomes smoother.

When the root mean square error of approximation (RMSEA) value is considered, the value decreases as the number of factors increases. However, it can be judged that there is almost no change in the model fit if the fit difference between models is .01 or less. From the fourth factor, the difference in the RMSEA values decreases to .01. If there is no change in the model due to the increase in the number of factors, it is desirable to explain the model with a small number of factors. So, it indicates that it is appropriate to extract four factors here.

Table 5. Factor Information of the Test Items

Factor	x^2	df	p	RMSEA
1	4082.32	1175	0.00	.023
2	2301.17	1126	0.00	.015
3	1873.37	1078	0.00	.012
4	1573.47	1031	0.00	.010
5	1402.09	985	0.00	.009
6	1271.09	940	0.00	.008

According to Bachman (1990), in factor 1, those items are about rhetorical organization and cohesion, so factor was named as “Textual Competence”. In factor 2, since the two items are about heuristic function and ideational function, factor 2 was named as “Illocutionary Competence”. Items in factor 3 are about sensitivity to naturalness, so factor 3 was named as “Sociolinguistic Competence” and items in factor are about vocabulary, so factor 4 was named as “Grammatical Competence”.

Table 6. Naming of the Four Factors

Item	Category	Factor	Label
4,35	Cohesion	1	Textual Competence
5,13,6,15,2,7	Rhetorical Organization		
19, 20	Heuristic Functions Ideational Functions	2	Illocutionary Competence
27,44,24,26,38,36,25,37,22	Sensitivity to Naturalness	3	Sociolinguistic Competence
40,45,39,33,28,31,29,41,42,46	Vocabulary	4	Grammatical Competence

Table 7 shows that the correlation coefficient between Textual Competence and Illocutionary Competence is .427; Textual competence and Sociolinguistic Competence is .496, Textual Competence and Grammatical Competence is .624, and Illocutionary Competence and Sociolinguistic Competence is .364 respectively. According to Kline (2011), when the value is much smaller than .90, discriminative validity can be confirmed, so the four factors are sufficiently discriminated.

Table 7. Four-factor Correlation Matrix

	Factor 1	Factor 2	Factor 3	Factor 4
	Textual Competence	Illocutionary Competence	Sociolinguistic Competence	Grammatical Competence
Factor 1	1.000			
Factor 2	.427	1.000		
Factor 3	.496	.364	1.000	
Factor 4	.624	.376	.601	1.000

Based on the results from exploratory factor analysis, confirmatory factor analysis is conducted to find out the relationship between measurement variables (subcategory) and the four potential variables. Besides, confirmatory factor analysis is done also for another purpose, that is, to verify if there is an integrative model that can embody the interrelationships among the four factors or a hierarchical model that can present orders of the factors. When the absolute value of the research model fit index RMSEA (Root Mean Square Error) is less than .08, TLI (Tucker-Lewis Index) is equal to or more than .90, CFI (Comparative Fit Index) is equal to or more than .90, and SRMR is less than .10, the model is fit (Browne &Cudeck,1993). According to the criterion of Browne & Cudeck (1993), both models A and B with these indices displayed in table 8 can be interpreted as models with high fitness.

Table 8. Information of the Two Models

	x^2	df	x^2/df	RMSEA	CFI	TLI
Model A	1781.99	521	3.41	.022	.927	.922
Model B	1790.98	522	3.43	.022	.927	.921

As shown in Figure 3, in model A, Textual Competence (TC), Illocutionary Competence (IC), Sociolinguistic Competence (SC), and Grammatical Competence (GC) are assumed to have interactive relationships with each other, so that model A can be constructed as an integrative model. In Model B, Textual Competence (TC) and Grammatical Competence (GC) can be grouped into Organizational Competence (OC), Illocutionary Competence (IC), and Sociolinguistic Competence (SC) can be grouped into Pragmatic Competence (PC), and model B can be hypothesized as a hierarchal model. In addition, judged from path coefficients and parameter estimates of Model A and Model B, with C. R. value >3.30 and p<0.001, the results indicate that both Model A and Model B are statistically significant.

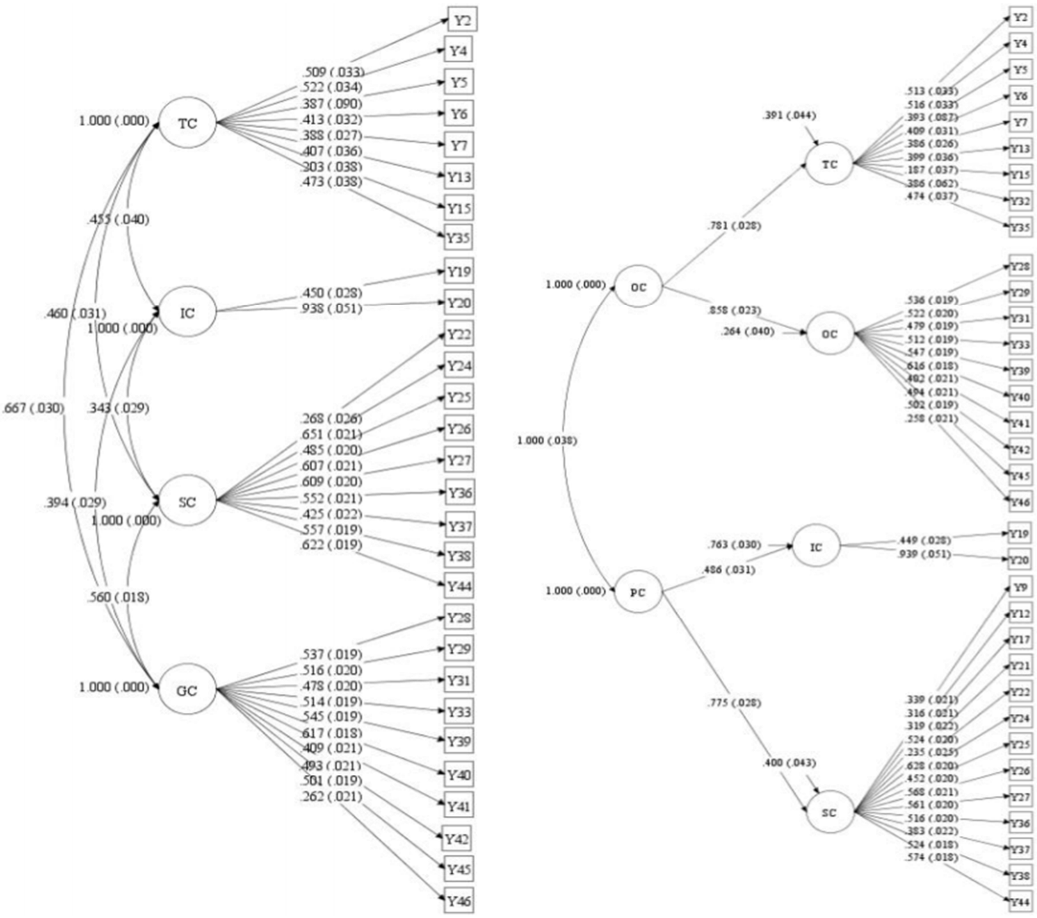

Figure 3. Model A-Integrative Approach Model & Model B-Hierarchical Approach Model

4. Discussion

As for research question 1, with four outlier items excluded from analysis, the rest of 46 items were analyzed on the basis of Item Response Theory, through comparative analysis of examinees' ability range with item difficulty, the result indicated that the test was appropriate in providing items within examinees' ability range. In addition, with 66% of the items having discrimination, 24% of the items having medium difficulty, the test was effective in distinguishing students' ability difference, and with 50 minutes for the 50 test items, this test paper is concluded as a proper test tool.

As for research question 2, 16 items with factor loadings of less than 0.3 were excluded, and then Exploratory Factor Analysis was conducted among the rest of the 34 items. 4 factors were extracted and named as Textual Competence (TC), Illocutionary Competence (IC), Sociolinguistic Competence (SC), and Grammatical Competence (GC). They can be constructed as two models as Model A-integrative model, and Model B-hierarchal model. Index values indicate that both of the two models are appropriate in distinguishing the four factors.

Judged from the correlation coefficients between different factors, grammatical competence has a high correlation both to textual competence and sociolinguistic competence, which indicates that grammatical competence is important in textual competence and sociolinguistic competence cultivation.

5. Conclusion

This study on validation of the reading comprehension and vocabulary items in the college English achievement test in this key financial university can provide enlightenment for teaching and learning. Based on the research findings, it is advisable that instruction and assessment of college English should be done in consistence with students' aptitude, teaching and learning objective, which will contribute to the realization of tailored instruction and personalized learning. Besides, more higher order thinking skills need to be integrated into college English achievement test.

In teaching and learning practice, several suggestions are put forward. First, since grammatical competence has a high correlation to textual competence, more efforts should be laid on multi-word prefabricated chunks, which consist of strong and weak collocations, lexical phrases or items, idioms, and fixed and semi-fixed expressions, through which in vocabulary teaching and learning, not only the individual words are focused, but the context of using the words will also be dealt with. Hence, textual competence will be gradually improved.

Second, since grammatical competence is closely associated with sociolinguistic competence, in language teaching and learning, incidental vocabulary learning is suggested. It can be achieved through extensive reading, watching movies, and listening to music in native English and more opportunities of using English in speaking or writing. In this way, students can have more opportunities to grasp ways of using English natively.

Third, that both the integrative approach model and hierarchal approach model are appropriate indicates that language teaching and learning is both an integrative process and also has its phases. Therefore, English language skills should not be taught or learned separately, or one by one but should be in a holistic approach. Meanwhile, strategies that match each specific language learning phase should be adopted in teaching and learning.

Fourth, those items that are not appropriate for factor analysis are about remembering or understanding factual knowledge of the meaning of vocabulary or just finding details from the context, which indicates that for improving the quality of this test, it is necessary to increase the number of the items that deal with a higher order of cognitive process like applying, analyzing, evaluating, and creating, etc.

Due to the limitations of convenience sampling in this study, if a follow-up study on results of this study is carried out through comparative study with students from other universities using the same test items, it will be of more significance. In addition, since this study focused only on reading comprehension and vocabulary items, there is a need for follow-up research on item fitness and construct validity for all parts of the English achievement test.

Acknowledgments

This study is part of the research findings of the research projects (2020jyxm0032, SK2020A0032), sponsored by Anhui Province, and the research projects (ACKYC21096, acjyzd2020038, acszjyzd2020006) sponsored by Anhui University of Finance & Economics.

References

[1] Anastasi, A. (1988). Psychological Testing (6th ed.) New York: Macmillan Publishing Company.

[2] Bachman, L. F. (1990). Fundamental considerations in Language Testing. Oxford: Oxford University Press.

[3] Becky H. Huang, Alison L. Bailey, Daniel A. Sass, & Yung-hsiang Shawn Chang. (2020). An investigation of the validity of a speaking assessment for adolescent English language learners. Language Testing, 38(3), 401-428.

[4] Brown, H. D. (2004). Language assessment: Principles and classroom practices. White Plains, NY: Pearson Education.

[5] Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing equation model fit. In Bollen, K. A., & Long, J. S. (Eds.). Testing structural equation models. Newbury Park: Sage, 136-162.

[6] Campbell, R., & Wales (1970). The study of language acquisition. In Lyons J. (Ed.). New horizons in linguistics. Harmondsworth: Penguin Books.

[7] Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. Applied Linguistics, 1, 1-47.

[8] Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, Mass.: MIT Press.

[9] Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. Psychological Bulletin, 52, 281-303.

[10] Henning, G. (1987). A guide of language Testing. Beijing: Foreign Language Teaching and Research Press.

[11] Hughes, A. (1989). Testing for language Teachers. Cambridge: Cambridge University Press.

[12] Hughes, A. (2000). Testing for Language Teachers. Beijing: Foreign Language Teaching and Research Press.

[13] Hallidy, M. A. K. (1973). Explorations in the functions of language. London: Edward Arnold.

[14] Hallidy, M. A. K. (1978). Language as Social Semiotic: The Social Interpretation of Language and Meaning. London: Edward Arnold.

[15] Hymes, D. (1972). On Communicative Competence. In Pride J. B., & Holmes A. (Eds.), Sociolinguistics: Selected Readings. Harmondsworth: Penguin.

[16] Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39, 31–36.

[17] Karami, H., Kouhpaee Nejad, M., Nourzadeh, S., & Ahmadi Shirazi, M. (2020). Validation of a bilingual version of the vocabulary size test: comparison with the monolingual version. International Journal of Bilingual Education and Bilingualism,23(4),368-380.

[18] Kline, R. B. (2011). Principles and Practice of Structural Equation Modeling. New York:Guilford Press.

[19] Liu, R. Q., & Han, B. C. (1999). Language Testing and Its Methods. Beijing: Foreign Language Teaching and Research Press.

[20] Messick, S. (1989). Validity. In Robert L. L. (Ed.). Educational Measurement (3rd ed.) New York: Macmillan Publishing Company

[21] Pelleriti, Margherita. (2020). The validation of the listening comprehension test tasks in the Certificate of English for Primary Teachers: An extension study. Language Learning in Higher Education, 10(1), 111-128.

[22] Yuan, P. (2002). Positive washback effect of achievement test on foreign language teaching. Foreign language teaching, 4, 18-21.

[23] Zou, S. (2005). An Interactive Approach to Test Validation-Reexamining the test usefulness of the TEM 4 reading component. Shanghai: Shanghai International Studies University.