US-China Foreign Language, February 2022, Vol. 20, No. 2, 55-62
doi:10.17265/1539-8080/2022.02.003

An Analysis of Reading Comprehension in Comprehensive English Course (Philosophy and Civilization) Test

LIN Xiangyu
University of Shanghai for Science and Technology, Shanghai, China

As a school-based course, Comprehensive English course has been designed and offered by a large number of universities and colleges. It is an integrative course which tests students' basic English skills and English proficiency. In this study, 55 sophomores participated in Comprehensive English final examination. This paper chooses reading comprehension part and briefly discusses its specifications. First, this study is to understand reading construct and evaluate it through CLA model. Second, the main purpose of this study is to evaluate reading comprehension part through statistical analysis of experimental data and item analysis. Besides, the validity and reliability of it are testified.

Keywords: reading comprehension, test specifications, statistical and item analysis, validity, reliability

Introduction

Comprehensive English course has received more and more attention since it was developed in the universities and colleges several decades ago, and the evaluation and assessment of this course is of great importance as well. Reading comprehension test serves as a way to examine and evaluate language learners' reading ability. In the past decades, many investigations concerning reading comprehension part test have been carried out. Zou, Zhang, and Zhou (2002) explored relationships among test takers' reading strategies, question types, and their reading test scores in TEM-4. Hou (2012) evaluated content validity of TEM-4 reading. Jin and Yang (2018) talked about implications from the development of CET in the past three decades in light of the validity theory of language testing. However, there are fewer studies investigating the reading comprehension part of school-based tests. Therefore, based on former research in reading comprehension test, this study mainly focuses on the discussion of test specifications, analysis of scores and items as well as the study and analysis of the validity and reliability.

The Analysis of Reading Comprehension of Comprehensive English Course Test

Test Specifications

In Zou Shen's *Language Testing* (2012), it is known that theoretical construct of reading is mainly reading ability, and test specifications of reading tests include many elements. The major purpose of reading comprehension of Comprehensive English course test is to examine test takers' reading ability and reading skills. Reading comprehension includes Section A and Section B. Section A is Multiple Choices Questions

LIN Xiangyu, Master, College of Foreign Language, University of Shanghai for Science and Technology, Shanghai, China.

(MCQs) and Section B is Short Answer Question (SAQs). In Section A, there are altogether three passages with 12 questions, and one point for each question. In Section B, there are four short answer questions on the basis of passages of Section A, and every question is assessed by two points. The time duration is about 30 minutes in total. The content of reading comprehension of Comprehensive English course test is displayed in Table 1.

Table 1

The Content of Section A

Test tasks	Genre	Topic	Text length
Passage 1	Narration	Color bar	418
Passage 2	Biography	Socrates and his view of philosophy	559
Passage 3	Argumentation	Standard of English	464

Theoretical Framework in Language Testing

Bachman (1999) put forward Communicative Language Ability, which is a model for analyzing the tests. Under the guidance of Bachman's Communicative Language Ability Model (shown in Figure 1), Bachman (1999) pointed out that language competence consists of a set of knowledge components used in communication. It includes organizational competence and pragmatic competence. For organizational competence, it can be further categorized into grammatical competence and textual competence. This requires language learners to be familiar with the usage of the language. To be specific, language learners are required to create grammatically correct sentences or utterances, to know how sentences and texts are organized and to recognize how words, clauses, and sentences are connected. In addition, language learners are able to identify grammatical errors, while pragmatic competence means the use of language and it includes both illocutionary competence and sociolinguistic competence. This requires language learners to know how to use language appropriately in certain contexts to achieve their goal of communication. The functions of communication should be recognized by language users by context of situation and psychological mechanisms of speakers. Strategic competence is an ability to use language knowledge and different strategies for communication. Anderson (1991) argued that readers who use more strategies will tend to score higher on reading comprehension tasks. Reading comprehension examines the reading ability to obtain the required information in an effective and efficient way. In reading comprehension section, test takers need to get to know vocabularies, paragraphs, and texts, and to grasp the main idea of the passages. Furthermore, they need to be equipped with some background knowledge and context knowledge for better understanding the passages. In Short Answer Question section, test takers ought to have the ability to generalize the information and to correctly read and write.

Qualitative Study of Construct Validity Based on CLA Model

Validity refers to appropriateness of a test as a measure of assessment. Weir (2005) defined validity as to whether a test has actually examined what it is examining and whether a test has achieved its designed purpose. In language testing, construct validity refers to language ability or competence under question. It can include ability, proficiency, and skills, etc., and it is related to score interpretation. If the tester's intention to use paper to examine or evaluate test taker's language ability is generally consistent with their language proficiency, the construct validity of the test paper can be said to be relatively high.

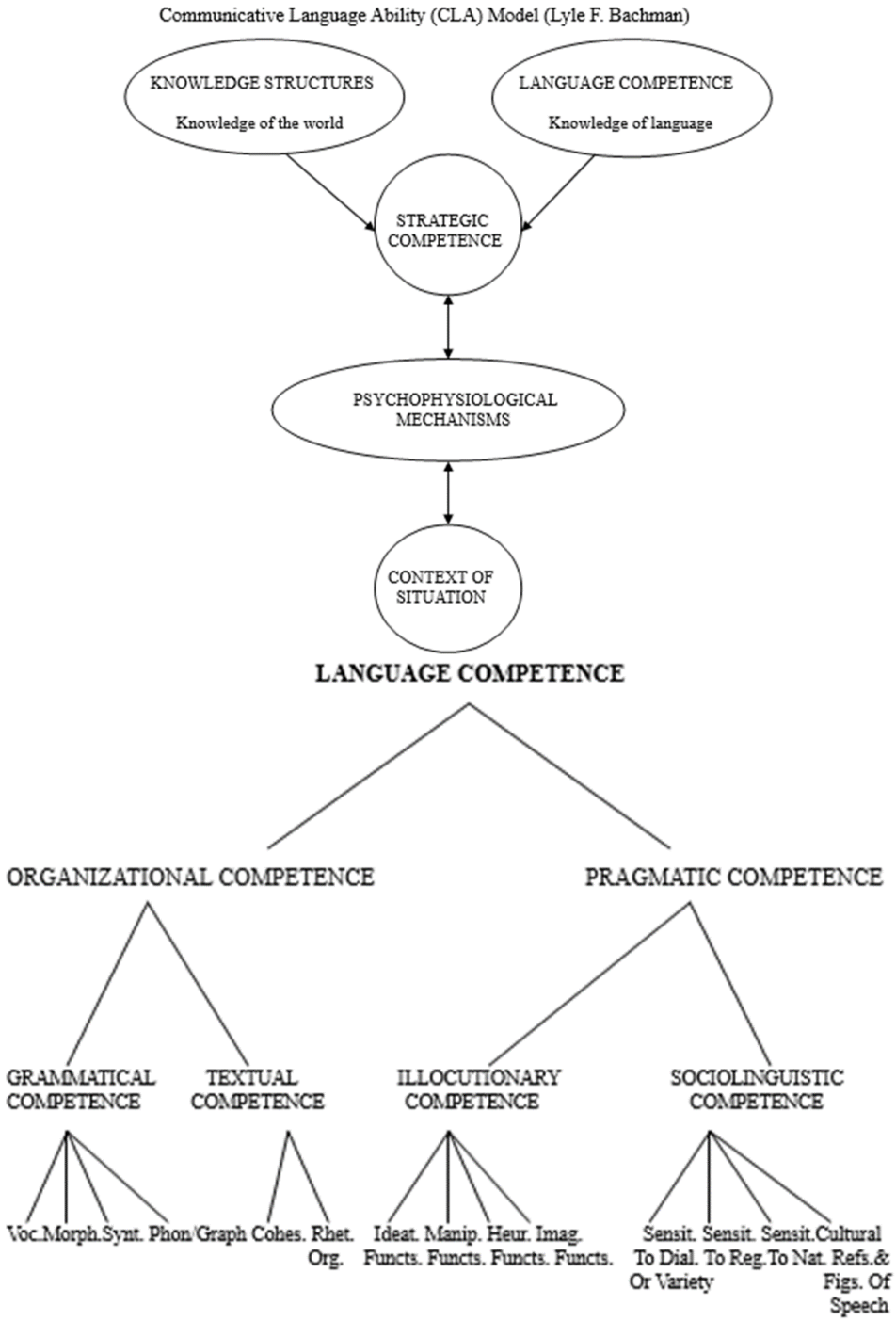

Figure 1. CLA model (Lyle F. Bachman).

To conclude, reading construct can be described in the following aspects:

a: Recognizing word and phrase meanings;

b: Obtaining word, phrase, and sentence meanings from the context;

c: Drawing some inferences from the situation of context in text;

d: Grasping the topic and the main idea of the passage;

e: Identifying an author's attitudes towards certain topic and an author's intentions of writing;

f: Getting the answers to questions by recognizing and analyzing specific details;

g: Generalizing information and drawing conclusions.

All the mentioned abilities can be considered as a set of reading constructs. In reading comprehension section, different kinds of formats and questions are used to measure test taker's reading ability. During this test, if these reading constructs are generally verified, it turns out that Comprehensive English course test has relatively high construct validity and achieves its testing purpose.

Statistical Analysis and Item Analysis of Reading Comprehension Test

Participants

Having been randomly selected from the University of Shanghai for Science and Technology, 55 sophomores majoring in English (Sino-US Joint Program) participated in this research to finish Comprehensive English course test. This test has the similar layout of the testing with TEM-4 test. These 55 sophomores are going to attend the TEM-4 test on 19, June, 2021, so to some degree they were deeply motivated to take part in this examination. In addition, they were well formed about the test content, test types as well as time duration.

Descriptive Analysis of the Scores

The score data are analyzed in SPSS 24.0. Figure 2 indicates the distribution of participants' scores. It is shown that the highest score is 19 while the lowest is 6.5. The average score of this test is 13.69. The std. deviation is 2.67. The number of participants who get the score over (including) 15.5 is 17, accounting for about 31% of the total population. By contrast, there are 17 participants whose score is lower than (including) 12.

Statistics

OverallScore

N	Valid	55
	Missing	0
Mean		13.6909
Std. Deviation		2.67099
Range		12.50

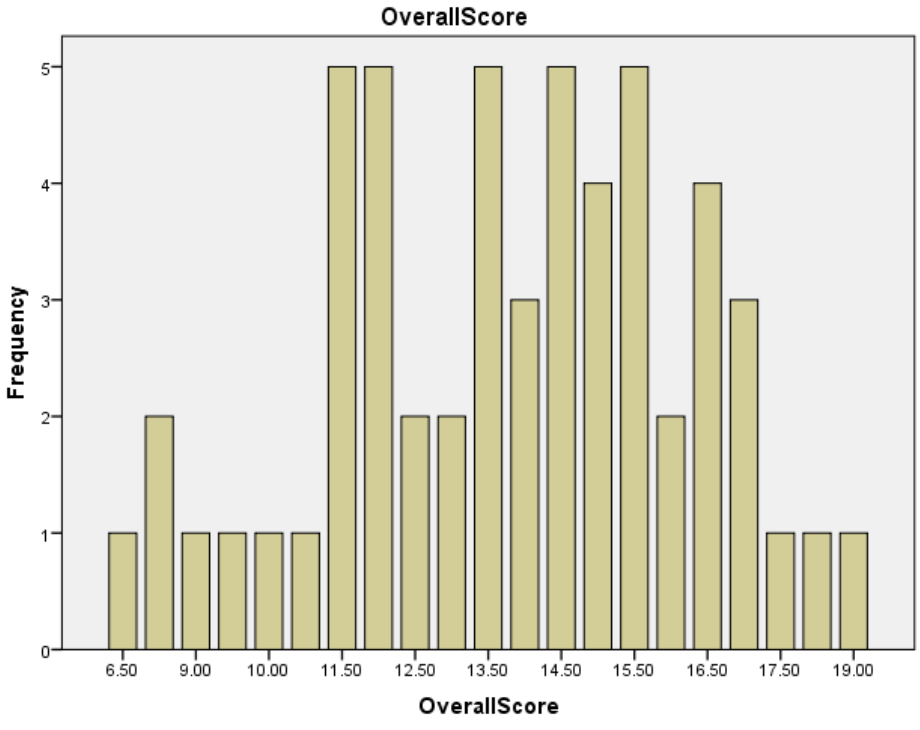

Figure 2. Score distribution of the participants.

Facility Value (F.V.) or Difficulty Index (D.I.)

Facility Value (F.V.) is a term used to indicate the difficulty of the test items, which is only applied for objective items. The lower F.V. is, the more difficult the item is. The ideal range of F.V. is between 0.3 and 0.7. The F.V.s of all the test items are calculated. The results show that there is a total of five items whose F.V.s are over 0.7 (shown in Table 2). It is further found that four of them spread over passage one. And the other seven items whose F.V.s are within the ideal range. It seems that the items of passage one are so easy for test takers to identify. To figure out the reason, it is found that both vocabularies and sentences of passage one are easy for test takers to grasp the main idea of the passage and to get the detailed information from the passage. Take item one and Item nine as examples:

Item one: “C olor bar” in the first paragaph comes closest in meaning to
A. a bar which is painted in different colors.
B. the fact that white and black customers ar eserved separately.
C. a store selling goods of different colors.
D. a counter where people of different colors are served with beer.

Item one (F.V. = 0.96) is too easy for test takers to identify and should not be designed for a proficiency test. Due to few new words and simple sentences, test takers can easily understand what the “color” and “bar” mean after reading the whole passage, thus understanding the meaning of word group “color bar”.

Item nine: Which of the following is TRUE according to the passage?
A. Mutiple choice questions are objective because people cannot be trained any techniques.
B. Sweden emphasizes the teaching of English without paying attention to other counties.
C. We have to depend on our own impression to judge the English standards.
D. Compositions are useful to test people's English ability because people have to write out their viewpoints.

In a similar way, Item nine is too easy for test takers as well. In contrast to other items such as type of predication or inference, there is no uncertainty or ambiguity at all. Test takers can easily identify which option is correct.

Table 2

Calculating F.V. of 12 Items

Item 1	Item 2	Item 3	Item 4	Item 5	Iten 6
F.V. = 0.96	F.V. = 0.85	F.V. = 0.84	F.V. = 0.91	F.V. = 0.70	F.V. = 0.34
Item 7	Iterm 8	Iten 9	Iterm 10	Item 11	Iterm 12
F.V. = 0.58	F.V. = 0.53	F.V. = 0.8	F.V. = 0.58	F.V. = 0.58	F.V. = 0.65

Discrimination Index (D.I.)

D.I. is of great importance in maintaining test reliability. The range of D.I. is between -1 and +1, while the ideal range is above 0.35. Calculating D.I. of all items of three passages, the results are presented in the following graphs (Table 3).

Table 3

Calculating D.I. of 12 Items

Item 1	Iterm 2	Item 3	Item4	Iterm 5	Iterm 6
D.I. = 0.06	D.I. = 0.29	D.I. = 0.24	D.I. = 0.06	D.I. = 0.35	D.I. = 0.18
Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
D.I. = 0.59	D.I. = 0.59	D.I. = 0.18	D.I. = 0.53	D.I. = 0.35	D.I. = 0.18

From the above table, it is shown that D.I. of all items has fallen within the range between -1 and +1. Among them, the D.I. of five items is over (including) 0.35, which is in the ideal range.

The Study and Analysis of Validity and Reliability of Reading Comprehension

Correlations

		MCQs	SAQs	OverallScore
MCQs	Pearson Correlation	1	.234	.827**
	Sig. (2-tailed)		.085	.000
	N	55	55	55
SAQs	Pearson Correlation	.234	1	.741**
	Sig. (2-tailed)	.085		.000
	N	55	55	55
OverallScore	Pearson Correlation	.827**	.741**	1
	Sig. (2-tailed)	.000	.000	
	N	55	55	55

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 3. Correlation between MCQs, SAQs, and overall score.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.468
Bartlett's Test of Sphericity	Approx. Chi-Square	66.612
	df	66
	Sig.	.456

Figure 4. KMO and Bartlett's test.

It can be seen from the Figure 3 that there is a weak or not significant correlation between MCQs and SAQs ($r = 0.234$, $P > 0.05$). According to KMO and Bartlett's Test, Figure 4 shows that KMO = 0.468 < 0.7. It may be concluded that this reading comprehension test has relatively low validity. Generally speaking, MCQs and SAQs are two different formats and they deal with different aspects or skills in terms of reading ability; in this way, two formats cannot be related with each other. To explain why reading comprehension test is in possession of lower validity further, the possible reason lies in three aspects. First, in the process of reading, test takers are not likely to completely follow the beginning-end structure of reading a passage. For example, when answering SAQs, test takers will turn to the specific points or a certain paragraph to locate related information instead of reading the whole passage. Second, it is likely for test takers to answer MCQs just by randomly guessing. Moreover, in response to SAQs, they may simply copy the original text or answer the questions without understanding them at all. Third, the relatively small and limited sample of participants maybe one factor influencing the validity of reading comprehension test.

Figure 3 reveals that there is a strong correlation between MCQs and overall score ($r = 0.827$, $P < 0.01$). There is also a strong correlation between SAQs and overall score ($r = 0.741$, $P < 0.01$). Different test formats of reading comprehension will to some extent influence reading scores. Compared to the correlation between

MCQs and overall score, there is a stronger correlation between MCQs and overall score. This indicates that MCQs is a better way to examine or reflect reading proficiency.

Reliability is a measure to assess the same trait. Bachman and Palmer (1996) defined reliability as "consistency of measurement", that is, the consistency of test scores. Cronbach (1951) thinks that Cronbach's alpha can be regarded as the most widely used measure of internal consistency reliability. To examine the reliability of the scores, Cronbach's alpha can be used to testify it.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.379	.386	12

Figure 5. Reliability statistics of MCQs.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.295	.246	4

Figure 6. Reliability statistics of SAQs.

Reliability Statistics

Cronbach's Alpha	Part 1	Value	.114
		N of Items	6[a]
	Part 2	Value	.201
		N of Items	6[b]
	Total N of Items		12
Correlation Between Forms			.330
Spearman-Brown Coefficient	Equal Length		.496
	Unequal Length		.496
Guttman Split-Half Coefficient			.480

a. The items are: Q1, Q2, Q3, Q4, Q5, Q6.

b. The items are: Q7, Q8, Q9, Q10, Q11, Q12.

Figure 7. Split-half method.

Figure 5 and Figure 6 describe reliability statistics, and it shows that Cronbach alpha (α) = 0.379. It indicates that the consistency reliability of MCQs format seems not to be acceptable. Figure 6 shows Cronbach alpha (α) = 0.295, which reveals that the reliability of SAQs format is not acceptable, either. But the reliability of MCQs format is relatively higher than SAQs. Figure 7 shows the value of split-half coefficient is relatively lower. In a word, reading comprehension enjoys low reliability.

Conclusion

Reading comprehension test including both MCQs and SAQs formats in Comprehensive English course test is a favorable way to evaluate language learner's English reading proficiency. This study is to understand reading construct in reading comprehension test that test takers are required to equipped with based on CLA model. Statistical analysis results indicate reading comprehension test as a whole enjoys low reliability and validity. This research is also trying to shed some light on the revision of Comprehensive English course or other school-based courses and give some insights for college English teaching. One thing needs to be mentioned is that the results of the statistics, especially the analysis for validity and reliability should be taken cautiously because of the limited and small sample.

References

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*(4), 460-472.

Bachman, L. F. (1999). *Fundamental considerations in language testing*. Shanghai: Shanghai Foreign Language Education Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Hou, Y. P. (2012). Content validity assessment of TEM 4 reading comprehension in the past seven years. *Journal of Hebei University (Philosophy and Social Sciences Edition)*, (4), 142-147.

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

Weir, C. J. (2005). *Language testing and validation*. London: Palgrave Macmillan.

Zou, S. (2012). *Language testing* (2nd ed.). Shanghai: Shanghai Foreign Languages Education Press.

Zou, S., Zhang, Y. L., & Zhou, Y. M. (2002). The relationship between item types, strategies and scores in reading tests—A study on the validity of TEM 4 reading items. *Foreign Language and Their Teaching*, (5), 19-22.