AN EVALUATION OF THE CET READING COMPREHENSION TEST BAND-4 RESPONSE FORMATS OF CHINESE EFL STUDENTS

XIE SHUANGMEI

SULTAN IDRIS EDUCATION UNIVERSITY

2025

AN EVALUATION OF THE CET READING COMPREHENSION TEST BAND-4 RESPONSE FORMATS OF CHINESE EFL STUDENTS

XIE SHUANGMEI

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

THESIS PRESENTED TO QUALIFY FOR A DOCTOR OF PHILOSOPHY

FACULTY OF LANGUAGES AND COMMUNICATION
SULTAN IDRIS EDUCATION UNIVERSITY

2025

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

UPSI/IPS-3/BO 32
Pind : 00 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS

Please tick (√)
Project Paper
Masters by Research
Master by Mixed Mode
PhD

✓

INSTITUTE OF GRADUATE STUDIES

DECLARATION OF ORIGINAL WORK

This declaration is made on the ...6th.....day of.....May.....2025.....

i. Student's Declaration:

I, Xie Shuangmei, P20211000833, Faculty of Languages and Communication (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled An Evaluation of The CET Reading Comprehension Test Band-4 Response Formats of Chinese EFL Students is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written forme by another person.

Xie Shuangmei

Signature of the student

ii. Supervisor's Declaration:

I Charanjit Kaur a/p Swaran Singh (SUPERVISOR'S NAME) hereby certifies that the work entitled An Evaluation of The CET Reading Comprehension Test Band-4 Response Formats of Chinese EFL Students (TITLE) was prepared by the above named student, and was submitted to the Institute of Graduate Studies as a * partial/full fulfillment for the conferment of PHD (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work.

14th., May, 2025

Date

Charanji

Signature of the Supervisor

UPSI/IPS-3/BO 31
Pind.: 01 m/s:1/1

INSTITUT PENGAJIAN SISWAZAH /
INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: An Evaluation of The CET Reading Comprehension Test Band-4 Response Formats of Chinese EFL Students

No. Matrik /*Matric's No.*: P20211000833

Saya / *I* : Xie Shuangmei

(Nama pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
   The thesis is the property of Universiti Pendidikan Sultan Idris
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
   Tuanku Bainun Library has the right to make copies for the purpose of reference and research.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
   The Library has the right to make copies of the thesis for academic exchange.
4. Silatandakan (√) bagi pilihan kategoridi bawah */Please tick (√) for category below:-*

☐ **SULIT/*CONFIDENTIAL***
Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / *Contains confidential information under the Official Secret Act 1972*

☐ **TERHAD/*RESTRICTED***
Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / *Contains restircted information as specified by the organization where research was done.*

☑ **TIDAK TERHAD / *OPEN ACCESS***

Xie Shuangmei
(Tandatangan Pelajar/ Signature)

Tarikh: 13th., May, 2025

Charanjit
(Tandatangan Penyelia / *Signature of Supervisor*)
& (Nama & Cop Rasmi / *Name & Official Stamp*)
PARAMJIT CHARANJIT KAUR A/P SWARN SINGH
Timb. Dekan (Akademik & Antarabangsa)
Fakulti Bahasa & Komunikasi
Universiti Pendidikan Sultan Idris

Catatan: Jika Tesis/Disertasi ini **SULIT** @ **TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who have supported me throughout my doctoral journey. First and foremost, I extend my sincere thanks to my supervisor, Professor Dr. Charanjit Kaur a/p Swaran Singh, for her invaluable guidance, insightful feedback, and unwavering encouragement. Your expertise and dedication have been instrumental in shaping this research and in helping me grow as a scholar. A special thank you goes to my colleagues, and friends, for their camaraderie, collaboration, and for creating an intellectually stimulating environment. The discussions, debates, and shared experiences have been an essential part of my academic and personal growth. On a personal note, I am deeply indebted to my family for their endless love, patience, and understanding. Your unwavering belief in me has been a constant source of strength and motivation. This achievement would not have been possible without your support. Finally, I dedicate this dissertation to all those who have inspired, encouraged, and challenged me along the way. Thank you for being part of this journey.

ABSTRACT

The aim of the study is twofold. First it investigates the impact of different test response formats on the reading comprehension performance of Chinese EFL learners. Second, it explores how their reading comprehension performance can be improved by employing some test taking strategies. The study employed the mixed-methods research approach. The sample consisted of 170 sophomores from Panzhihua University, majoring in non-English fields and whose first language is Chinese. Both students and teachers were selected using a purposive sampling technique. Eight teachers participated in the interviews. The research instruments comprised three types of test papers with varying formats, questionnaire, and semi-structured interview protocol. Quantitative data were analyzed by SPSS 26.0, involving descriptive statistics, T-tests and two-way ANOVA. Qualitative data from interviews were transcribed and analyzed thematically using NVivo software. Findings showed that there is a significant difference in the reading comprehension performance of Chinese EFL learners ($F=244.890$; $p=0.000$; $P<0.05$) when different types of response formats are utilized in the CET-4. An interaction effect between test types and academic background was significant, while no significant interaction was observed between test types and proficiency levels. MCQs led to the highest performance, while SAQs resulted in the lowest. Accuracy rates varied across item types and response formats, with significant differences detected. Survey results indicated that students relied on contextual clues for MCQs and found them less challenging. T-tests revealed no significant differences in students' attitudes based on gender, academic backgrounds or study duration. The qualitative findings highlight the necessity to incorporate various response types for a more balanced test design. The teacher interviews reveal a variety of structured test-taking strategies adopted to enhance EFL learners' reading comprehension performance. This study offers valuable insights into the assessment of reading comprehension, and provides some implications for EFL language testing, learning and teaching.

EVALUASI FORMAT JAWAPAN UJIAN PEMAHAMAN BACAAN CET BAND-4 BAGI PELAJAR EFL CHINA

ABSTRAK

Tujuan kajian ini adalah dua hala. Pertama, ia menyiasat kesan pelbagai format respons ujian terhadap prestasi pemahaman membaca pelajar EFL Cina. Kedua, ia meneroka bagaimana prestasi pemahaman membaca mereka boleh dipertingkatkan dengan menggunakan beberapa strategi mengambil ujian. Kajian ini menggunakan pendekatan penyelidikan kaedah campuran. Sampel terdiri daripada 170 pelajar tahun kedua dari Universiti Panzhihua, yang mengambil jurusan bukan Bahasa Inggeris dan bahasa pertama mereka adalah bahasa Cina. Pelajar dan guru dipilih menggunakan teknik persampelan bertujuan. Lapan orang guru mengambil bahagian dalam wawancara. Instrumen penyelidikan terdiri daripada tiga jenis kertas ujian dengan format yang berbeza, soal selidik, dan protokol temu bual separa berstruktur. Data kuantitatif dianalisis menggunakan SPSS 26.0. Data kualitatif daripada temu bual ditranskripsi dan dianalisis secara tematik menggunakan perisian NVivo. Penemuan menunjukkan bahawa terdapat perbezaan yang signifikan dalam prestasi pemahaman bacaan pelajar EFL China ($F=244.890$; $p=0.000$; $P<0.05$) apabila pelbagai jenis format respons digunakan dalam CET-4. Kesan interaksi antara jenis ujian dan latar belakang akademik adalah signifikan, manakala tiada interaksi signifikan diperhatikan antara jenis ujian dan tahap kemahiran. Soalan pilihan berganda (MCQ) membawa kepada prestasi tertinggi, manakala soalan jawapan pendek (SAQ) menghasilkan prestasi terendah. Kadar ketepatan berbeza mengikut jenis item dan format respons, dengan perbezaan signifikan yang dikesan. Keputusan soal selidik menunjukkan bahawa pelajar bergantung kepada petunjuk kontekstual untuk MCQ dan mendapati ia kurang mencabar. Ujian t menunjukkan tiada perbezaan yang signifikan dalam sikap pelajar berdasarkan jantina, latar belakang akademik, atau tempoh pengajian. Penemuan kualitatif menekankan keperluan untuk memasukkan pelbagai jenis respons bagi reka bentuk ujian yang lebih seimbang. Temu bual dengan guru mendedahkan pelbagai strategi pengujian yang tersusun untuk meningkatkan prestasi pemahaman bacaan pelajar EFL. Kajian ini menawarkan pandangan yang berharga dalam penilaian pemahaman bacaan dan memberikan beberapa implikasi untuk ujian bahasa EFL, pembelajaran, dan pengajaran.

CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EFL	English as a Foreign Language
IELTS	International English Language Testing System
TOEFL	Test of English as a Foreign Language
GRE	Graduate Record Examination
CET	College English Test
TEM	Test for English Majors
SECC	State Education Commission of China
SAQ	Short-answer Questions
MCQ	Multiple-choice Questions
T/F/NG	True/False/Not Given

APPENDIX LIST

CHAPTER 1

INTRODUCTION

1.1 Introduction

Language testing has been a longstanding priority of study for its significance in education and society. The test may assess individuals in their academic and professional endeavours, it may provide a ranking of test-takers for decision-making purposes, and it may also offer empirical evidence to improve teaching methods. Due to its powerful educational and social consequences, researchers have been diligently studying this topic in order to continuously improve language testing theories and operations.

In China, there are two national-wide examinations that assess the English proficiency of college students: CET-4 (College English Test Band 4) and CET-6 (College English Test Band 6). The latter test is more challenging than the former.

Both of these examinations consist of identical elements for assessment, including hearing, reading, writing, and translation, however they vary in terms of difficulty levels. Therefore, the speaking proficiency of test takers is evaluated through distinct oral examinations, namely CETS-4 (College English Spoken Test Band 4) and CETS-6 (College English Spoken Test Band 6). Only the students who pass the written tests have the quality to participate in the corresponding oral examinations and obtain certified qualifications.

The CET is a standardized English test in China that was established in 1987. It is a high-stake, large-scale exam that assesses the English proficiency of undergraduate students. The test is carried out by the Bureau of Higher Education of the Ministry of Education. Its main purpose is to ensure that students meet the basic or intermediate level of English proficiency specified in the College English Curriculum Requirements. The implementation of the CET has had a significant impact on both teachers and students, providing them with strong motivation and encouragement.

From the perspective of high stakes and large scale, the College English Test Band 4 (CET-4) in China annually involves the participation of millions of students. In the present era, the significance of the CET-4 has greatly increased. The results of this test are directly linked to a student's eligibility for obtaining a bachelor's degree in many colleges and universities. Furthermore, in the society, particularly in the job market, possessing a CET-4 certificate is considered crucial in demonstrating an individual's level of English ability. Each year, countless students take the CET-4and CET-6 exams. This is because it is widely recognized that when students begin searching for jobs, their scores on these exams are considered a highly significant

measure of their language proficiency by many employers.

From the aspect of standardized examination, CET refers to an exam which provides a recognized objective criterion by which to measure students' applied language ability through spot exam in language applying. Usually, standardized examination involves hundreds of thousand students, even millions of students. Hence, the criteria for this examination are crucial, particularly in terms of its reliability and validity. In language testing, the CET-4 is a standardized language examination that use educational measurement methodologies to assess students' English language proficiency. As a result, the examination must adhere to rigorous criteria (Fan, Frost, & Jin, 2022). As a criterion-related norm-reference test, the score of an examinee in CET-4 is determined by the comparison with the scores of the norm-reference group. The standard norm group of CET-4 is made up of approximately ten thousand undergraduates from China's six prominent universities: Beijing University, Tsinghua University, Shanghai Jiaotong University, Fudan University, Chinese Science and Technology University, and Xi'an Jiaotong University (Jin, 2011). It is doubtful to what extent such a norm-reference group can be representative in such a big country as China where the higher education is so unevenly developed.

In China, the CET is conducted twice a year in China, specifically on one weekend in June and December. In addition, it offers both negative and positive feedback regarding the teaching and learning of college-level English reading. On one hand, the CET has facilitated the advancement of English education and learning in China. It can provide constructive comments to different higher education institutions, prompting them to implement necessary actions to enhance their teaching quality. On

the other hand, it has negative aspects, such as its low validity and inappropriate item types, and has brought negative backwash on English teaching and learning, leading to intense disagreement on the matter (Cheng, & Curtis, 2004; Li, Zhong, & Suen, 2012). It is ordinary that teachers often spend a significant amount of time analysing original papers and instructing students on testing strategies to help them pass exams. Simultaneously, it compels the students to devote greater focus towards English in that period.

So, from this point, the CET has a certain degree of positive impact on English education and learning. And then, the main testing method is still multiple-choice, which possesses a lot of demerits in language testing. Testing experts have also noted that overuse of this format in a language test may result in test takers excessively focusing on test abilities and neglecting the fundamental content of language. If test takers focus too much on the test-taking skills during the process of acquiring a second language has a detrimental effect on the teaching of the English language (Cheng, 2004).

The CET has consistently improved in both quality and scale since its implementation in 1987. It underwent revisions in 2006 and 2015, respectively. According to the College English Syllabus, CET should be communicative in nature. As the Syllabus (2015) says that language is a tool for communication, and the ultimate aim of language teaching is to foster students' ability to communicate both orally and through written channels. The goal of language teaching should encompass not only the enhancement of students' linguistic skills but also the cultivation of their ability to effectively communicate. The overall success of the test is heavily

influenced by the performance of test-takers in the reading section. This is because the revised reading comprehension section now accounts for 35% of the total score in the latest version of the CET. As a result, it continues to receive significant attention from both English teachers and students.

Upon examining the response types of the College English Test (CET), it can be observed that, with the exception of the writing and translation sections, all other items are in the form of multiple-choice questions. People in the daily life is merely to choose one answer from four items, and the objective questions are easily affected by various factors, making it difficult to accurately assess a person's language proficiency. Consequently, test takers may resort to guessing when selecting the correct answer. The CET, which has been present in China for over three decades, it receives great attention from every layer of the society and exerts.

1.2 Background of the study

Reading has been one of the central concerns of applied linguists for decades. Reading abilities are essential for individuals to keep up with the rapid flow of information and constantly evolving technologies in a modern civilized society. In China, reading comprehension, as a fundamental component to examine test-takers' English proficiency, has been used in almost every large-scale English test. It also attempts to accurately assess college students' ability of using English comprehensively. However, the most prevalent testing format in reading comprehension test is MCQ. This practice encourages both teachers and students to

work over test skills and countermeasures in test preparation, which disrupts regular classroom instruction, promotes a teaching approach centred around test performance, and consequently impairs students' comprehensive acquisition of essential English knowledge and integrated skills, as well as inhibits the development of their communicative competence. However, in life study or communication environment, there are no designed options for reference or making a choice.

In order to provide an appropriate and fair basis for decisions and inferences as to test-takers' language competence, language testing researchers have embarked on an effort in examining factors affecting test performance. It was found that both the test-takers' language ability and the response format used to measure it had a significant influence on test performance (Bachman, 1990). In this light, response format has emerged as a significant focus of research. It is regarded as a vehicle that represents test designers' conception of language abilities to be tested, as well as a platform through which test-takers can demonstrate their language competence. Under such circumstances, the intense debates on the use of multiple-choice questions (MCQs) as the primary exam method in the CET-4 are raised.

1.2.1 The status quo of CET-4reading comprehension sub-test in China

The current reading sub-test of CET-4 consists of three response types: Banked Cloze, Matching Information, and skimming and scanning. However, all of them are exclusively in selected-response formats. Below is the specific introduction to them.

1.2.1.1 Banked Cloze

In this task type, there is a short passage with 10 omitted words and a list of 15 words that are available below in a word bank. Test takers must choose a single word from the provided list to fill each blank, and they cannot use any term more than once.

1.2.1.2 Matching Information

Test takers must correctly associate the given information with the corresponding paragraph, and they have the option to select a paragraph multiple times if needed. Matching is a testing method under which candidates have to match the related statements with the items or paragraphs in the other fist through given clues in passages. Matching is a highly complex question type because it requires candidate to carefully repeat reading in order to match the items one by one. This task type involves a lengthy paragraph accompanied by 10 associated sentences. Each statement corresponds to one of the paragraphs labelled with an uppercase letter. Test takers need to match the specific paragraph with that information and may choose a paragraph more than once if necessary.

1.2.1.3 Skimming and Scanning.

In this part, there are two short passages, each consisting of approximately 200-250 words, ranging from literature to science, technology, and so on. The main technique

in this section is MCQ items, which consists of a question or statement posed as a stem followed by four alternative answers. Out of these options, only one is correct. There has been much work done rebutting its defects in a long time. The main reason for the overflow of multiple-choice questions in CET is their convenience for machine grading. If this convenience is to be kept at the expense of the test validity, which in turn leads to misleading college English teaching, it violates the original aim of the test. The major criticism of the multiple-choice item, however, is that it often fails to facilitate the assessment of language as a means of communication. Responses to different stimuli in everyday settings are generated within a specific context, rather than being selected from multiple options. It is crucial to understand that context plays a vital role in face-to-face communication. Note that context is of the utmost importance in real life interaction. To have a visual picture of the current response formats and communicative approaches, the researcher lists a table to make the information more accessible in Table 1.1.

Table 1.1

Response Formats Comparison

Response Format	**Description**	**Limitations**	**Potential Communicative Approaches**
Banked Cloze	A test where participants fill in missing words from a list.	- Limited to vocabulary testing. - May not fully reflect real-life communication. - Can be too artificial.	- Use context-rich texts that mimic real-life situations. - Include tasks that involve reasoning and deduction.
Matching Information	A format where participants match items (e.g., questions to answers or statements).	- May not measure deeper comprehension. - Could be too mechanical or rigid. - May not reflect natural interaction.	-Add situational contexts to make matching more relevant. - Pair matching with follow-up questions to deepen understanding.

Response Format	Description	Limitations	Potential Communicative Approaches
Multiple Choice Questions (MCQs)	A format that presents a question with several possible answers.	- Can promote guessing. - Might not assess complex cognitive skills. - Less effective for measuring production skills.	- Use more realistic scenarios where choices reflect real-world decision-making. - Combine with tasks requiring production of language to enhance engagement.

1.2.2 Problems with CET-4's (College English Test) current reading comprehension test

The current reading comprehension section of the CET-4 has several issues that limit its effectiveness in accurately assessing students' real English proficiency, particularly their communicative competence Although various response formats, such as MCQs, T/F/NGs, and SAQs, are used in the test, they all belong to the selected-response type. These formats, while useful for testing basic comprehension, fail to replicate real-life situations and do not fully reflect a student's ability to use English as a practical tool in everyday life or professional contexts.

A significant issue with the CET-4 reading test is its heavy reliance on objective items, which limits students' ability to demonstrate deeper language skills. The focus on MCQs and other objective response types encourages students and teachers to adopt an "exercise-stuffed" approach, aiming primarily to pass exams rather than develop real communicative competence. This results in many students, even those with CET-4 or CET-6 certificates, struggling to understand practical English, such as instructions for medicine or domestic appliances.

Furthermore, the absence of titles in the reading passages of the CET-4 is another major flaw. In real-life reading materials, such as newspapers or magazines, titles serve to introduce the main theme and engage the reader's curiosity. The lack of titles in the CET-4 test reduces its authenticity and makes it harder for students to establish connections between the text and its content, which in turn hinders reading comprehension. This lack of context may also lead to students relying on test-taking strategies, such as guessing or eliminating options, without fully understanding the text to skew the test results.

Another critical issue is that the reading comprehension subtest plays a significant role in determining the overall score of the CET-4. As reading is essential for both personal and professional growth, a test that fails to accurately measure reading comprehension may have a far-reaching impact on students' opportunities in higher education and the job market. Given the centrality of reading skills in language acquisition, it is crucial for the CET-4 to evolve in order to better assess these skills in a way that reflects real-world usage.

To address these issues, it is proposed that the CET-4 decrease the number of objective items and increase the proportion of subjective response types, such as open-ended questions. This shift would allow students to demonstrate a wider range of skills, including critical thinking and the ability to communicate in English more effectively. Additionally, the introduction of a domestic version of the IELTS could provide a more comprehensive assessment of students' reading abilities, ultimately improving national English language teaching standards.

While the multiple-choice questions (MCQs) in the CET-4 reading comprehension section have notable limitations, it is important to acknowledge the benefits that MCQs offer in language testing. One of the main advantages of MCQs is objectivity. They have clear, standardized grading, and reduce the potential for bias in scoring. This ensures that all test-takers are evaluated according to the same criteria, making the test results more reliable and consistent. The ease of scoring is another significant benefit, as MCQs can be quickly and efficiently graded, saving time and resources for both test administrators and students.

Additionally, MCQs are highly efficient for assessing a wide range of content within a limited time. They allow for the testing of a broad array of language skills, such as vocabulary, main idea identification, and detail recognition, within a single test. This makes MCQs particularly useful for large-scale assessments like the CET-4, where logistical constraints require a quick, efficient method of evaluating a large number of students.

However, while MCQs provide these advantages, it is crucial to balance their use with other response formats that allow for a deeper and more holistic assessment of students' language proficiency. While MCQs can test basic comprehension and factual recall, they do not fully assess higher-order skills such as critical thinking, problem-solving, or the ability to apply language in real-world contexts. Therefore, incorporating a greater variety of response types, such as open-ended or subjective questions, could complement the benefits of MCQs and more accurately reflect students' communicative competence in English. This approach would strike a balance between the practicality and efficiency of MCQs and the need for a more

comprehensive, authentic evaluation of language proficiency.

1.2.3 The Necessity of Applying the Communicative Approach in CET-4 Reading Test

Language testing has undergone significant evolution since the 1980s, transitioning through three main stages: pre-scientific testing, structuralist-psychometric testing, and communicative language testing. Early language tests focused primarily on objectivity and reliability, while the communicative approach emphasizes broader dimensions such as reliability, construct validity, authenticity, interactivity, feedback, and practicality. Since the mid-1980s, the communicative approach has gained prominence because it better aligns with the real-world use of language, providing a more holistic measure of a learner's language proficiency.

In China, while the communicative approach has been discussed in detail in the field of language testing, its application in the CET-4 remains limited. The CET-4 reading test, with its rigid structure and reliance on objective MCQs, fails to adequately assess students' ability to use English in real-life communication. This gap underscores the need for reform to better align the test with the skills required for effective communication in both academic and professional contexts.

There are two primary reasons for adopting a communicative approach to the CET-4 reading test. First, reading is communicative. It involves an interactive process between the reader and the text, requiring not only comprehension but also engagement. A communicative approach would ensure that the reading test reflects

this interactive nature, assessing students' ability to process, analyze, and interpret information as they would in real-world situations. Second, the CET-4 is an achievement-proficiency test, intended to measure English proficiency. Given the weight of the reading section in CET-4, it should focus on assessing communicative competence, which includes the ability to understand and engage with complex texts, as well as the capacity to use language effectively for communication.

The current dominance of MCQs in the CET-4 reading test is problematic. While MCQs are efficient for marking, they encourage test-takers to rely on guessing when faced with time constraints, resulting in inaccurate assessments of reading comprehension. This reliance on objective, selected-response formats fail to capture the full range of reading skills necessary for real-world communication, such as critical thinking and the ability to engage with complex ideas. A shift towards more diverse response types that require practical application of reading comprehension would allow the test to assess a broader range of skills, making it more aligned with the skills students need to succeed in real-life communication.

In recent years, there has been a growing shift in China towards more comprehensive and valid methods of assessing language ability. The over-reliance on MCQs in the CET-4 reading section undermines the test's quality and relevance. To address this, it is essential to explore alternative response formats that reflect students' actual reading abilities and communicative competence.

This paper proposes to study the effect of different response formats on students' performance in the CET-4 reading section, using Bachman's test methods

facets. The aim is to explore how different response types influence test performance, identify which formats yield the best results, and assess students' attitudes towards various types of questions. By examining these factors, the study seeks to recommend the most effective response formats that can accurately assess students' language abilities, thereby guiding improvements in language teaching and learning practices.

To ensure a more accurate and comprehensive assessment of language ability, it is crucial to reduce the influence of extraneous factors, such as response formats. A more balanced approach, incorporating both objective and subjective methods, including MCQs and other open-ended responses, would provide a more holistic measure of reading comprehension. As reading comprehension plays a central role in both teaching and learning, such an approach would offer valuable insights into improving the CET-4 reading test and aligning it with the principles of communicative language testing.

Adopting a communicative approach would prioritize the development of real-world language skills, moving beyond mere comprehension to enable students to engage with language in authentic contexts. This shift is crucial for building the communicative competence necessary for success in language learning. Furthermore, incorporating a communicative approach into the CET-4 would bring it closer to global language proficiency standards, such as those set by IELTS and TOEFL. These internationally recognized tests assess not only reading comprehension but also the ability to critically analyze, synthesize, and engage with texts in real-world contexts. Shifting away from an over-reliance on MCQs and incorporating a variety of response types to require practical application would enhance the test's depth and relevance,

ensuring it better reflects students' ability to use English effectively in everyday situations.

Moreover, this change would better prepare students for future academic pursuits abroad, where language proficiency is evaluated not just on reading comprehension but also on the ability to interact with complex ideas, engage critically with texts, and communicate effectively across various contexts.

In conclusion, integrating a communicative approach into the CET-4 reading test would lead to tangible improvements in students' language proficiency. It would enhance their ability to use English meaningfully in real-life situations and align the CET-4 with global standards, preparing students for academic and professional success on an international scale.

1.2.4 Practical Challenges of Implementing Communicative Testing Approaches in CET-4 Reading Comprehension Subtest

Implementing a communicative approach in the CET-4 reading comprehension subtest presents several challenges that need to be carefully considered in order to achieve a balanced and effective solution. These challenges primarily involve issues related to cost, the training of test designers, and grading scalability. Each of these factors must be addressed to ensure that the shift towards a communicative testing model is both feasible and sustainable.

1.2.4.1 Cost

One of the most significant challenges in implementing communicative testing approaches is the associated cost. Unlike traditional multiple-choice question (MCQ) formats, which are cost-effective and efficient to administer, communicative approaches require the development of more complex assessment tools. These may include tasks that assess a student's ability to analyse and engage with real-world content, such as open-ended questions, critical thinking tasks, or interactive scenarios. To implement these tasks on a large scale, substantial resources are needed for test development, administration, and scoring. These financial demands could place a strain on institutions, especially for large-scale assessments, such as CET-4.

1.2.4.2 Training of Test Designers

Shifting towards a communicative testing model also requires a significant investment in the training and development of test designers. Unlike traditional tests that primarily assess linguistic knowledge, communicative tests require designers to create tasks that reflect the dynamic and interactive nature of language use in real-world contexts. This shift necessitates a deep understanding of communicative competence, which includes not only linguistic proficiency but also the ability to process and respond to information in meaningful ways. Test designers would need to be trained in both communicative language testing theory and practical test construction. Such training would likely require workshops, courses, and collaboration with language teaching professionals, thus adding additional time and resource costs to the process

for test designers.

1.2.4.3 Grading Scalability

Another key challenge is the scalability of grading in a communicative approach. While traditional MCQs offer easy and efficient grading, communicative tasks, particularly those that involve open-ended responses, require more subjective evaluation. For instance, assessing reading comprehension beyond superficial understanding, such as evaluating a student's ability to critically analyze and synthesize information requires human graders to apply nuanced criteria. This subjectivity can lead to issues with consistency and reliability in scoring, especially when thousands of students are taking the test. Moreover, grading such tasks often requires specialized skills in evaluating higher-order cognitive processes, further complicating the scaling of the grading process. To address this challenge, automated grading systems or the use of trained raters might be necessary, but both solutions involve additional costs and infrastructure.

While the adoption of a communicative approach in the CET-4 reading comprehension subtest offers several advantages in terms of assessing real-world language use, the implementation comes with notable challenges. These include the financial cost of developing and administering more complex tasks, the need for extensive training of test designers, and the difficulties associated with scaling the grading process. To overcome these challenges, it would be essential to carefully plan the integration of communicative assessments, ensuring that sufficient resources are

allocated to training, technology, and infrastructure. A balanced approach, incorporating both traditional and communicative formats, could provide a more holistic and practical measure of students' language proficiency, while managing these practical implications effectively.

1.3 Definition of Terms

1.3.1 Definition of Response Formats

Response formats refer to the specific modes through which test-takers provide their answers in an assessment, and they play a critical role in shaping the nature of the expected responses. According to Bachman and Palmer (1996), response formats are a key component of test methods that directly influence test-takers' performance in language tests. They are typically categorized into selective response formats (e.g., multiple-choice questions) and productive response formats (e.g., short-answer or essay questions). These formats are not just mechanical choices; they fundamentally affect how comprehension and other cognitive traits are assessed, as they guide the test-takers' interaction with the test material and their cognitive engagement.

In the context of reading comprehension assessments, response formats act as indirect indicators of the mental processes involved in comprehension, which cannot be directly observed. Since comprehension occurs dynamically during reading, test developers rely on response formats to infer cognitive abilities, such as problem-solving, inference-making, and information synthesis. However, a challenge arises in

selecting the most appropriate response format, as the format chosen can impact the test-takers' performance and, consequently, the validity of the inferences made about their comprehension abilities.

From a trait theory perspective, response formats should ensure that performance across different formats remains comparable under equivalent conditions. This suggests that different formats, despite varying in structure, should assess the same underlying cognitive traits if other factors are controlled. In contrast, the interactionalist view considers response formats as an integral part of the construct being measured, meaning that they are not merely tools for assessment but are essential components that shape how the construct is defined. According to this perspective, response formats must be designed to reflect the context in which the reading task takes place, and they should not introduce systematic errors in measurement.

Ultimately, the goal of selecting and designing response formats is to ensure that they provide a valid and reliable measure of the test-taker's abilities. This includes ensuring that the chosen response format aligns with the test's objectives, supports accurate inferences about the test-takers' reading processes, and reflects the real-world use of reading skills. By carefully considering the impact of response formats on test performance, test developers can create assessments that provide meaningful, reliable insights into reading comprehension abilities.

1.3.2 Reading Comprehension Competence

Reading comprehension competence is selected as the target trait in this research mainly because it is the fundamental competence for a person's survival in his or her daily life and it has witnessed most productive research findings in the academic circle. Reading comprehension competence is an outstanding achievement of human brain, a foundational and essential skill for academic achievement in school and success in life and career. Vast majority of the information that people obtain every day is from reading activities. Naturally, an accurate testing of such an essential competence has a great influence on all stakeholders concerned. Indeed, how to assess reading comprehension competence has always been the concern of language testing and assessment researchers and practitioners. However, the testing of reading comprehension competence is by no means an easy task. Actually, it has been one of the most controversial issues in language testing and assessment for its lack of an overt process to be observed directly, unlike writing or speaking competence which will produce a concrete performance of test-takers' for researchers to examine all kinds of indicators of it carefully. The traditional way to investigate validity of a reading comprehension test is to observe the response of a test-taker elicited through an indirect test method and make an inference about the nature of his/her reading comprehension processes and whether he or she possesses the target reading comprehension competence or not. As discussed before, the logic of indirect inferring from product (result) to process (reason) is fallible and a conclusion drawn from a statistical analysis is not always reliable. Therefore, it is high time to visit validity issue of reading comprehension tests from a process-focused perspective to explore the possibility of discovering new and valuable findings.

Research on reading comprehension competence started much earlier than that on writing, speaking or listening competence and people believe that they are comparatively better equipped with knowledge of the nature of reading comprehension competence. A much-studied field rather than a less explored one is preferred in the current research since the researcher believes that it is more meaningful to compare potential new findings obtained through a new instrument or method with the old and time-tested findings obtained through old ones. It is also more likely for the researcher to resort to one among a great number of existing theories or models in a much-studied field to account for unusual phenomena in case mixed results may occur unexpectedly.

1.3.3 The Definition of Test-taking Strategies

Readers do exert a significant level of active control over their reading comprehension processes through strategies (Williams, & Moran, 1989). Test-taking strategies refer to the deliberate, conscious techniques and processes employed by test-takers to effectively manage and regulate their performance during assessments. These strategies are extensions of general learning strategies, specifically adapted to the context of testing, and are integral to optimizing test outcomes. Unlike reading comprehension skills, strategies are deliberate, conscious and effortful actions employed by readers to regulate their comprehension and to monitor their performances throughout each reading activity. A better understanding of reading comprehension strategies can help researchers understand the way that readers manage their interaction with written texts and how their choice of strategies

influences their comprehension of the text (Cohen, & Upton, 2006).

The concept of strategies implies selection. Using a limited number of strategies in a response to an item may, at times, display a true control over the item on the assumption that these strategies are well-chosen and used effectively. At other times, a true control over an item requires using a great number of strategies. Any test-taking strategy can be a good or a poor choice for a given task. It depends on how a test-taker employs the strategy at a given moment on a given task with his/her particular cognitive style or degree of cognitive flexibility, language knowledge and repertoire of test-taking strategies. Some test-takers may get by with the use of a very limited number of strategies that they use well for the most part while others may be aware of an extensive number of strategies but use very few effectively. Successful test-takers differ from less successful ones in both the quantity and the quality of strategy use. Usually, successful test-takers are more able to adjust their strategies to the type of the prompt text that they are reading to their purpose, more capable of monitoring their comprehension, more aware of the strategies that they need and are using, and use the strategies more flexibly than less successful test-takers (Block, 1986).

Gagne, et al. (1993) argued that test-takers purposefully employed strategies to enhance their performances in tests. In testing situations, strategy use can be summoned to retrieve necessary declarative (knowing what), procedural (knowing how) and conditional (knowing when) knowledge in test-takers' long-term memory so as to solve task difficulty. In some cases, test-takers may use test-wiseness to circumvent the need to tap their actual language knowledge or the lack of it.

Cohen (1998) provided a foundational definition of test-taking strategies, describing them as extensions of general learner strategies applied specifically to assessment contexts. He emphasized that these strategies were integral to the test-taking process and reflected broader learning behaviours. This perspective was supported by subsequent research, which highlighted the dynamic and context-dependent nature of test-taking strategies.

Cohen and Upton (2006) offered a more detailed exploration of test-taking strategies, defining them as the conscious processes that respondents selected and utilized during tests. This definition underscored the element of awareness, suggesting that test-takers are, to some extent, be awareness of the strategies they deploy. This awareness allowed them to adapt their approaches based on the demands of the test and their own strengths and weaknesses, thereby optimizing their performance.

Nguyen and Kim (2022) described test-taking strategies as the processes by which test-takers use to effectively address and answer test items. This definition underscored the strategic aspect of these behaviours, suggesting that effective test-takers were those who can leverage their understanding of response formats and question types to maximize their scores. They emphasized the role of adaptability and skill in manoeuvring through different sections and types of questions. This definition highlighted the dynamic nature of test-taking strategies, which required both preparation and real-time decision-making during the test.

Test-taking strategies are dynamic, context-dependent, and may involve multiple stages, such as planning, monitoring, and evaluating. They encompass

processes like planning the approach to the test, selecting and applying appropriate strategies, adjusting tactics in real-time, and reflecting on the effectiveness of those strategies. These strategies also enable test-takers to make decisions about how to allocate cognitive resources, such as when to prioritize certain types of tasks or when to use shortcuts for efficiency. Ultimately, successful test-takers demonstrate both a broader range of strategies and a higher level of adaptability, flexibility, and awareness compared to less successful test-takers, leading to better management of the test-taking process and, often, improved performance.

1.3.4 The Definition of Construct validity

Construct validity can be defined in various ways. Bachman (1990) described it as a unitary concept and the core of score interpretation. Bachman emphasized that the primary issue in construct validity was the extent to which we could infer hypothesized abilities based on test performance.

Chapelle and Chapelle (2001) viewed construct validity from the perspective of second language acquisition, where testing serves as an elicitation device for abilities tied to specific constructs, such as language proficiency or communicative competence. This emphasizes that the test must reflect both the theoretical constructs and the actual language learning process.

Wood (2001) noted that construct validity involved gathering data and testing hypotheses, without a specific validity coefficient associated with it. The principle

was that measures of the same trait by different methods should correlate more highly than measures of different traits by the same method. As some testers regarded construct validity as the super-ordinate concept encompassing all other forms of validity, various approaches had been used to assess it. For example, one might hypothesize that reading ability includes several sub-abilities, such as the ability to infer the meaning of unknown words from context. Consequently, a reading test should include multiple-choice or paraphrase questions concerning the definition of some unfamiliar words.

Heaton (2004) stated that it measured specific characteristics according to a theory of language behaviour and learning. This type of validity presumed the existence of certain learning theories or constructs that underlie the acquisition of abilities and skills. For example, a speed-reading test based on a short comprehension passage would be considered an inadequate measure of reading ability (and thus have low construct validity) unless it was believed that speed reading short passages was closely related to the ability to quickly and efficiently read a book, and was a proven factor in reading ability.

Cohen et al. (2017) defined construct validity as the appropriateness of inferences made from test scores regarding an individual's standing on a specific variable. The key idea is that the test should accurately represent an individual's ability in relation to the theoretical construct.

In the Standards (2014), which stated that construct validity was used to indicate that test scores reflected the test takers' standing on the psychological

construct being measured. A construct was a theoretical variable inferred from various types of evidence, including the interrelations of test scores with other variables, observations of response processes, internal test structures, and the content of the test.

Cohen et al. (2017) described construct validity as a judgment regarding the appropriateness of inferences made from test scores about an individual's standing on a variable. Anthony (2020) suggested that construct validity was based on the extent to which the items in a test reflected the essential aspects of the theory upon which the test was based (i.e., the construct). For example, the greater the demonstrated relationship between a test of communicative competence in a language and the underlying theory of communicative competence, the higher the construct validity of the test.

1.4 Problem statement

Reading comprehension, an essential skill for language proficiency, plays a central role in both academic success and real-world communication. The CET-4, as a high-stakes assessment, measures Chinese students' English proficiency, with a particular focus on their reading comprehension abilities. However, the assessment format predominantly employs MCQs, which have raised concerns regarding their adequacy in fully capturing the diverse dimensions of reading comprehension skills. While MCQs are convenient for large-scale testing and are commonly used in educational assessments worldwide, they tend to focus on superficial comprehension, often failing to assess deeper cognitive processes involved in reading, such as critical thinking,

inference-making, and text interpretation.

The reliance on MCQs in the CET-4 reading comprehension section has prompted a need for further investigation into whether this format truly measures students' comprehensive reading abilities. Some research suggests that alternative response formats, short-answer questions (SAQs), true/false/not given (T/F/NG), and essay-type responses, may provide a more holistic assessment by engaging students in more complex cognitive tasks. Such response formats encourage students to engage more deeply with the text, as they are required to demonstrate their understanding and interpretation in more expressive ways. However, while there is a growing body of research on the effects of response formats in reading comprehension tests (e.g., Kobayashi, 2002; Brantmeier, 2005), much of the empirical work has been conducted outside of China, with limited attention paid to Chinese EFL students' experiences in the CET-4 context.

This gap in research is significant because understanding how response formats impact reading comprehension performance is crucial for improving the fairness and accuracy of the CET-4. The current reliance on MCQs may not fully represent the range of skills required for effective reading comprehension, and thus it is necessary to investigate how other response formats might provide a more holistic measure of students' reading abilities. Furthermore, it is essential to explore how different response formats influence students' engagement with the reading material, their test-taking strategies, and their overall performance. Response formats may affect not only how students process and understand the text, but also their motivation, strategy use, and confidence during the test.

This research aims to fill the gap in literature by providing empirical evidence on the impact of different response formats on Chinese EFL learners' performance in the CET-4 reading comprehension test. By analysing how various response formats, such as MCQ, SAQ, T/F/NG can affect test-takers' comprehension, this study will provide insights into the potential benefits and limitations of each format in measuring reading proficiency. Specifically, the study will investigate whether MCQs are sufficient for capturing the full range of students' reading abilities or if alternative formats would provide a more comprehensive measure of comprehension, including inferencing, interpretation, and the integration of information across the text.

Additionally, attitudes towards different response formats and their associated test-taking strategies play a crucial role in improving reading comprehension performance. Research suggests that students' attitudes toward test types significantly impact their motivation and approach to reading tasks (Biggs, 1996). Understanding how students perceive these formats can help educators design more engaging assessments. Furthermore, teaching strategies that prepare students to tackle different response formats effectively could enhance their performance, particularly in tests like the CET-4, which are crucial for academic success. English teachers can play a pivotal role in equipping students with the necessary skills and strategies to navigate various test formats and improve their reading comprehension.

In conclusion, this research aims to investigate the impact of different response formats on Chinese EFL learners' performance in the CET-4 reading comprehension section, while considering how these formats affect students' attitudes and test-taking strategies. Previous research has shown that students from different

backgrounds may approach reading comprehension tasks in unique ways, and that proficiency levels can influence how well students perform under different response formats. By examining these factors, the study will contribute valuable insights into how the CET-4 can be improved to more accurately assess reading proficiency and support the development of effective teaching practices. The findings will also offer practical recommendations for teachers to enhance their students' performance on reading comprehension tasks by utilizing appropriate strategies for different response types.

Finally, this research will contribute to improving the validity and reliability of the CET-4 reading comprehension test by offering recommendations for the inclusion of diverse response formats. By ensuring that the test measures a broader range of reading abilities and cognitive skills, the findings will help educators and test developers design more effective assessments. The study will also provide insights for English teachers on how to support students in preparing for different response formats, ultimately helping to enhance students' reading comprehension skills and test performance.

In conclusion, the problem of response formats in reading comprehension assessments, especially in the CET-4, is an important issue that affects not only the validity of the test but also the educational outcomes of students. As the CET-4 remains a crucial measure of English proficiency for Chinese students, it is essential to critically examine the response formats used in the test and consider whether alternative formats could lead to a more accurate and holistic assessment of reading comprehension. This study will help bridge the gap in research, offer practical

insights for improving assessment design, and contribute to the development of more effective and equitable language testing practices.

1.5 Research objectives

As an empirical study, this thesis bases on Bachman's framework of test method facets (Bachman, 1990), and focuses on the fourth category, expected response, to further investigate on the following six research objectives according to the corresponding research questions:

1. To explore the performance differences among Chinese EFL learners from different academic backgrounds on reading comprehension tasks in the CET-4 when using different types of response formats.

2. To analyse the extent to which Chinese EFL learners of different proficiency levels are affected by different response formats in CET-4 reading comprehension tests.

3. To identify the response format in which Chinese EFL learners perform the strongest and the weakest in CET-4 reading comprehension assessments.

4. To investigate whether there is a significant difference in the test scores of Chinese EFL learners on CET-4 reading comprehension tasks when measured by different item types (main idea, inference, detail, and interpretation).

5. To explore the attitudes of Chinese EFL learners towards the three different response formats in the CET-4 reading comprehension test.

6. To explore how English teachers can enhance their students' performance on CET-4 reading comprehension tests by employing test-taking strategies for different types of response questions.

1.6 Research questions

Many scholars in linguistics (Smith, 2015; Johnson & Lee, 2018' Brown & Williams, 2019) have begun to doubt whether MCQs can truly measure test-takers' reading ability. They are eager to construct appropriate reading comprehension tests, obtain an objective and accurate evaluation of the test-takers' reading ability, and then predict their mastery of this foreign language. Therefore, it is significant to conduct a study to analyse the differences of various response formats on test-takers' test performance, so the present study will address the following research questions:

1. Do Chinese EFL learners from different academic backgrounds perform differently on reading comprehension tasks in the CET-4 when using different types of response formats?

2. Are Chinese EFL learners of different proficiency levels affected differently by different response formats in CET-4 reading comprehension tests?

3. In which type of response format do Chinese EFL learners perform strongest and weakest in CET-4 reading comprehension assessments?

4. Is there a significant difference in the test scores of Chinese EFL learners on CET-4 reading comprehension tasks when measured by different item types (main idea, inference, detail, and interpretation)?

5. What is the level of Chinese EFL learners' attitudes towards the three different responses types of the CET-4 reading comprehension test?

6. How can English teachers enhance their students' performance on CET-4 reading comprehension tests by employing test-taking strategies for different types of response questions?

1.7 Research hypothesis

Based on the above research questions, the following null hypotheses are therefore formulated:

For research question 1, there are three hypotheses listed as follows:

Hypothesis 1a: There is no significant difference in the reading comprehension performance of Chinese EFL learners from different academic backgrounds.

Hypothesis 1b: There is no significant difference in the reading comprehension performance of Chinese EFL learners when different types of response formats are utilized in the CET-4.

Hypothesis 1c: There is no significant interaction between test and group in the reading comprehension performance of Chinese EFL.

For research question 2, there are three hypotheses listed as follows:

Hypothesis 2a: There is no significant difference in the reading comprehension performance of Chinese EFL learners from different proficiency levels.

Hypothesis 2b: There is no significant difference in the reading comprehension performance of Chinese EFL learners when different types of response formats are utilized in the CET-4.

Hypothesis 2c: There is no significant interaction between test and proficiency levels in the reading comprehension performance of Chinese EFL.

For research question 3, there is one hypothesis listed as follows:

Hypothesis 3: The Chinese EFL learners equally perform in CET-4 reading comprehension subtest when measured by different response formats.

For research question 4, there is one hypothesis listed as follows:

Hypothesis 4: There is no significant difference among testing scores measured by different item types (main idea, inference, detail, and interpretation).

1.8 Theoretical Framework

In the 1970s, the communicative approach to language testing emerged. The reading process itself is communicative in nature, a reading test, a method to examine readers' comprehension, should take the communicative testing approach as the main focus, so that the testing of reading can be in accordance with the nature of reading. Based on the research objectives, this thesis is guided by the theories of Bachman's communicative language testing theory (1990) when conduct this research.

As to CLT (communicative language testing), its application developed together with its theoretical framework in the 1980s, which was different from the traditional testing. A good language test should meet the criteria of communicative language testing, which refers to have a good balance between reliability, validity, practicality, wash-back, authenticity, and interactive-ness. The present communicative language testing has many outstanding features, which is concerned primarily with how language is used in communication, and the evaluation of real communication in the second language, rather than the knowledge of this system. It suggests that language testing should be carried out in an authentic language context with authentic language input so that a test taker's language proficiency could be evaluated through

observation of their performance in the test. Communicative language testing mainly includes two parts: communicative competence and test method facets, which are the main inspiration of this study. This part will give a clear picture of these two sub-theories.

1.8.1 Communicative competence

When refers to the communicative language testing, it is emphasized to demonstrate the history of Communicative competence. The communicative language testing mainly originated from the "communicative competence" proposed by Hymes (1972), after which Canal and Swain (1980) developed Hymes's view of communicative language, Bachman (1990) advocated a new framework of language communicative competence, which consisted of linguistic competence, strategic competence and psychological mechanisms. This research sticks to Bachman's framework of communicative language competence. After 20 years' development in language testing and assessment, Bachman (1990) advanced CLA (Communicative Language Ability) in language testing field on the basis of communicative language testing. Bachman (1990) defined communicative language competence as the ability to create and interpret meaning by combining linguistic knowledge with the features of a language-use situation. Language use is a dynamic process in which multiple knowledge, skills and mental processes are intertwined and interact with each other. Bachman and Palmer (1996) then further improved their model into an interactive model, which has achieved a further step for the notion of communicative competence.

Bachman (1990) argued that CLA (Communicative Language Ability) comprised three key components: language competence, strategic competence, and psycho-physiological mechanisms, which was regarded as the most thorough and ideal model and was applied to many kinds of teaching and testing. For these three parts, language competence was comprised of specific language knowledge, while strategic competence served as a means to apply this knowledge in real-world communication. Psycho-physiological mechanisms, on the other hand, pertained to the neurological and psychological processes involved in the actual execution of language as a physical phenomenon.

(1) Language competence: Bachman (1990) thought language competence could be composed by language organizing ability and language using ability. The formal one referred to the ability to master the structure of language when producing or deciphering the grammar correct sentences, and grasping the main idea of passages. This ability was also considered to be generalized by two kinds---grammatical competence and discourse competence. While the latter one referred to the ability to make sure how the discourse, purpose and the language context interact each other, which was also considered to be generalize by two kinds---functional competence and socio-linguistic competence.

The process of measuring linguistic competency involved three essential steps: theoretical definition of constructs, operational definition of constructs, and qualification of observations. When considering construct validation in reading, two aspects should be considered: the definition of reading ability and the methods used to elicit it. In proficiency reading test, the theoretical construct was built on experts'

definition of reading skills, the skill lists proposed by scholars (Weir, 1993). In achievement tests, the reading construct was determined by both teaching syllabus and testing syllabus. The reading construct in achievement tests was influenced by both the teaching syllabus and the testing syllabus. The realization of construct validity required the protection of statistical validation and the use of scientific methods for item writing and scoring (Zou, 2005).

(2) Strategic competence: the concept of this competence proposed by Bachman was not the same one by Canale and Swain. According to Bachman, the real essence of the strategic competence was the psychological cognitive process. In any cases of language communication, the language competence was closely related with the psychological cognitive process. It described the cognitive ability to apply the various components of language competence in real-life communicative situations. This model, therefore, facilitated the connection between language competencies and the specific contextual features in which language was used, as well as the user's knowledge structures, such as sociological and real-world knowledge (Bachman, 1990).

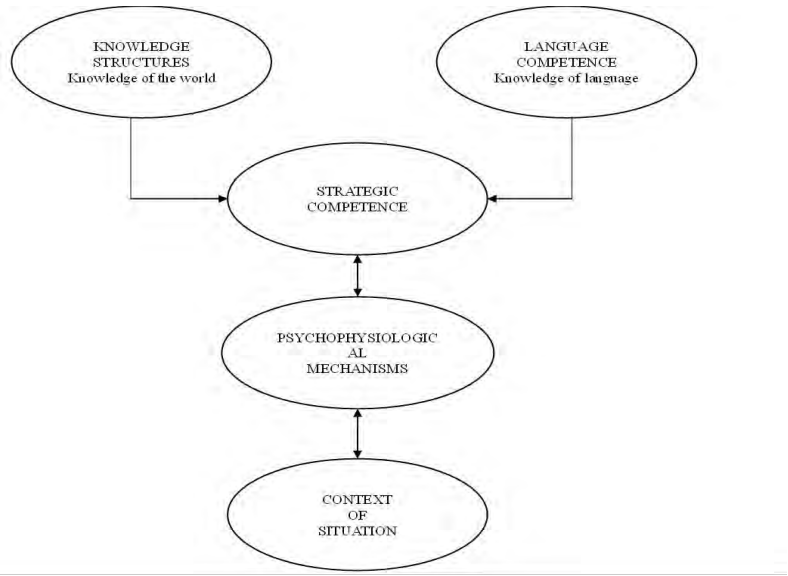
Strategic competence referred to a collection of strategies that involve goal-setting, assessment, and planning. These strategies are considered to be higher-level executive processes (Bachman, 1990). Setting goals involved determining what tasks or objectives one intended to accomplish. Assessment entails identifying the necessary resources, evaluating what was available, and measuring the effectiveness and appropriateness of the actions taken. Planning involved deciding how to effectively utilize the resources what one has.

(3) Psycho-physiological mechanisms encompassed the neurological and psychological processes involved in the execution of language. Bachman involved this competence into the communicative language competence, highlighting the importance of language proficiency. Meanwhile, it elucidated the independence of the four methods (listening, speaking, reading and writing) in actual language processing (Bachman, 1990).

The components of communicative language ability in communicative language use were vividly described in the following figure. (Figure 1.1). From Figure 1.1, we knew that the measurement of language ability, especially communicative language ability, dealt with many complicated components. Bachman's model of communicative competence was quite prevalent in the field of language testing. It served as a theoretical guideline for designing communicative language tests.

Figure 1.1

Components of Communicative Language Ability in Communicative Language Use



Source: Bachman, 1990

As choosing Bachman's Communicative Language Testing Theory (1990) as the theoretical framework for this thesis, there are many reasons as follows. First, Bachman's theory covers multiple dimensions of communicative competence, including language competence, strategic competence, and psycho-physiological mechanisms, which aligns perfectly with this study on the CET reading comprehension test. Bachman's emphasis on linguistic competence, including grammatical and discourse competence, corresponds directly to the language knowledge and contextual understanding required for reading comprehension. Additionally, his concept of strategic competence, which involves goal-setting, planning, and assessment, helps explain the strategies students use when taking the test, such as skimming, scanning, and guessing. Furthermore, Bachman's focus on psycho-physiological mechanisms addresses the psychological and neurological processes involved in language execution, which is critical for understanding factors like test anxiety and cognitive load that influence student performance. Finally, Bachman's emphasis on construct validity is essential for evaluating whether the CET reading comprehension test's response formats accurately measure the intended language abilities. Overall, Bachman's theory offers a comprehensive framework for analysing the components of the CET reading comprehension test and students' language performance.

1.8.2 Bachman's model of test method facets

It was known that test performance was affected by the characteristics of the methods used to elicit test performance (Bachman, 1990). There were numerous elements that

influence the performance of test-takers. Hence, to ensure the reliability of a test, test constructors must carefully select the suitable response styles to accurately assess students' genuine proficiency. This thesis aims at investigating the difference of different response formats on reading comprehension test performance in CET-4, so Bachman's framework of test method facets is the main inspiration that guides this thesis.

1.8.2.1 The nature of Bachman's test method facets

As stated by Bachman (1990), although test takers' performance on a language test was largely determined by their communicative language ability, test performance was also affected by the characteristics of the methods used to elicit test performance. These facets of test methods were the essential components that determine the approach used in language testing. They played a crucial role in the design, development, and utilization of language tests, since they were the aspects that we could potentially influence. For example, some test takers who felt intimidated by a cloze test may at the same time excel on a test consisting of multiple choices. In the realm of language testing, an effective language test should offer a precise assessment of test-takers' language abilities, so test designers or developers must identify and minimize the influence of potentially intervening factors on test-takers' performance in reading comprehension test (Anthony, 2020).

It was widely acknowledged that numerous factors could either positively or negatively impact a test taker's performance on test tasks of any types. Therefore,

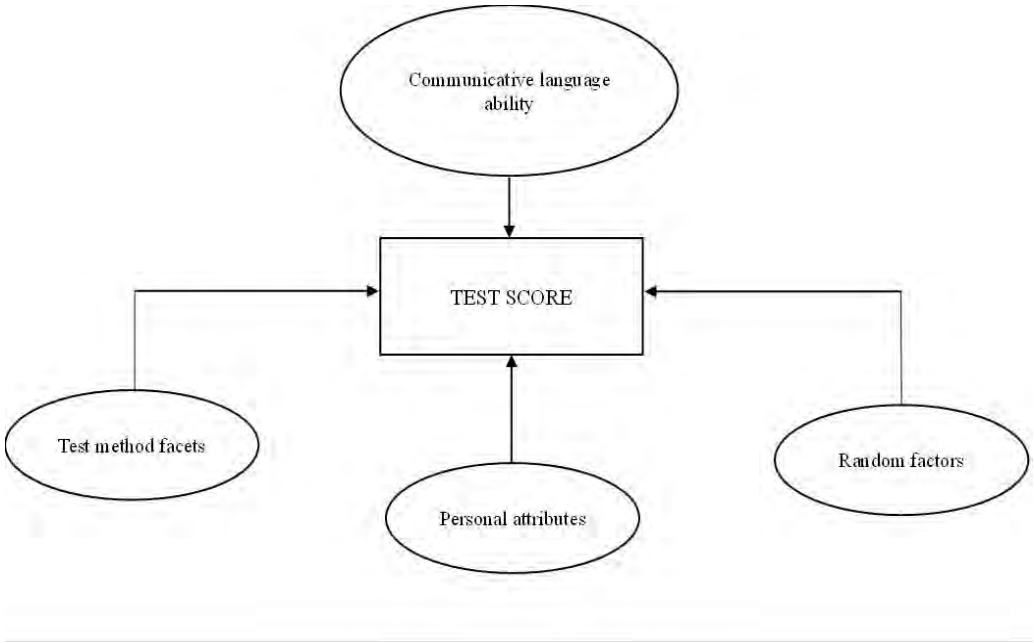
Bachman (1990) set out several factors besides communicative language ability which influenced language test scores, (1) test method factors, (2) attributes of the test-takers that were not directly related to the language abilities being measured, and (3) random factors that were largely unpredictable and temporary.

In this framework, test method facets were systematic to the extent that they were uniform from one test administration to the next. For instance, if the test used a multiple-choice format, this would not change regardless of whether the test was administered in the morning or afternoon, or if the test aimed to measure grammatical or illocutionary competence. Attributes of individuals that were not related to language ability include individual characteristics such as cognitive style and knowledge of specific content areas, and group characteristics like sex, race, and ethnic background. In the sense, these attributes were also systematic that they were likely to affect a given individual's test performance.

In addition to these systematic sources of error, an individual's test score could also be impacted by unsystematic or random factors. These included unpredictable and temporary conditions, such as the test-taker's mental alertness or emotional state. Test designers and developers must strive to control these intervening factors. The influence of these various factors on a test score could be depicted in figure 1.2.

Figure 1.2

Factors that affect language test scores (Bachman, 1990)



Source: Bachman, 1990

The figure illustrated that the performance of test takers on language tests was influenced by various sources of variance, as categorized by Bachman (1990) into four main groups: CLA (Communicative language ability), TMF (test method facets), PA (personal attributes), and RF (random factors). The primary objective of language testing was to assess an individual's communicative language capacity (CLA). However, this goal was sometimes undermined by the influence of three other factors: test method sensitivity (TMS), proficiency assessment (PA), and response format (RF). Test developers had limited influence over random factors, but they could exert considerable control over the test method facets and test-taker attributes. Researchers and educators employed diverse testing methods and evaluations to gain insight into the reading abilities of language learners, based on the component of testing methods (Dolosic, 2018).

The results of the effects of all these factors mean that individuals are unlikely to perform equally well on a language test, leading to variations in their scores. When we consider the total variation in scores among different individuals on a given test, we can attribute different proportions of this variation to the different factors discussed above. This is because the different factors will affect different individuals differently, just as individuals vary in their level of language ability, so the result will vary in the extent to which they are affected by different methods of testing. Some individuals, for example, may excel at multiple-choice tests but struggle with composition tests. test.

Therefore, test-taker's performance is influenced by numerous factors, among which the test method used to elicit test performance is of great importance. Among them, test method is the most important factor. In order to more comprehensively examine the differences in language test performance, especially the effects of test methods on test performance, Bachman proposed his own framework of test method facets in 1990 by constructing a framework that can be used to design, develop and compare language tests based on the relevant aspects. He emphasized that this framework should be seen as a guide for empirical research, rather than a definitive or exhaustive list. Bachman hoped that this framework would contribute to a better understanding of how these facets impact performance on language tests, and also encourage the identification of additional facets that were not initially included.

The framework of test method facets proposed by Bachman in 1990 included the below five major categories of test method facets: (1) the testing environment; (2) the test rubric; (3) the nature of the input the test taker receives; (4) the nature of the

expected response to that input, and (5) the relationship between input and responses.

According to Bachman's (1990) framework, the initial category of test method facets was the test environment. This category encompassed factors such as the familiarity of the location and equipment used during the test, the individuals involved in people who administer the test (both administrators and other test takers), and the timing and conditions of the test administration.

The second category, test rubric, referred to the guidelines that specify the process of taking a test. This included the organization of the test, the allocation of time, and the instructions provided. The rubric consisted of the test organization, time allocation, and instructions. The test organization specifies the number of sections in the test, their sequence, and the characteristics of each section. Time allocation determined how much time is allotted for each section. By understanding the time allocation, students could effectively manage their time and complete all the tasks. The instructions provided details about the testing procedures and the nature of the tasks. It was crucial for test-takers to be familiar with the rubric as it aligned with the test specifications.

Input and expected response constituted two other distinct sets of test method facets. Nevertheless, they were interconnected and possessed certain shared characteristics. In broad terms, input could be characterized in terms of the channel, mode, form, vehicle, language, problem identification, and speed. Expected response facets, on the other hand, included channel, mode, type of response (selected or constructed), form of response, the nature of the language used and the types and

degree of restrictions on response.

The last category of test method facets involved the relationship between input and response in language tests. This relationship could be reciprocal, meaning there was an interactive exchange of meaning. It could also be nonreciprocal, meaning there was no interaction between input and response. Lastly, it could be adaptive, meaning the input was influenced by the test taker's response, but without providing feedback to the test taker.

The present study focused on the fourth category of test method facets, expected response, by manipulating different response formats in the reading comprehension experiment in order to study the effect exerted by different response types on test-takers' performance. Specifically, the influence of the following three response formats, MCQ, T/F/NG, and SAQ would be compared and analysed in the present research. According to the above discussed, the testing method facets of CLT (communicative language testing), especially the test environment, test rubric, and input format contributed a lot to the paper designing. Consideration on testing environment meant a lot in designing the testing environment of reading sub-test in CET-4, and it could avoid unnecessary measurement errors. The rubric corresponded to test specification, facilitating analyses of the language competence in the reading sub-test. Features of input suggested language abilities completely and directly. All in all, testing method facets were an indispensable part of CLT (communicative language testing) and it was worth considering when testifying language test's construct validity.

1.8.2.2 Characteristics of the expected response

According to Bachman and Palmer (1996), test activities or tasks tended to fall under the three types of response: selected, limited production and extended production response.

Selected response was a type where the test taker was required to choose one answer from a set of options, commonly seen in multiple-choice activities. The most obvious advantage of multiple-choice was that scoring can be perfectly reliable, rapid and economical. In addition, more items were likely to be included than would otherwise be possible in a given period of time. However, if there was a lack of fit between at least some candidates' productive and receptive skills, then performance on a multiple-choice test may give a quite inaccurate picture of those candidates' ability. In addition, the probability of correctly guessing the answer may have a significant yet unknowable effect on test scores. While it was not necessary to completely eliminate multiple-choice items from examinations, it was important to refrain from using this technique excessively, without discrimination, and in a way that could potentially cause harm.

The limited production response was a concise and brief reply, typically consisting of only one word or phrase, and could be as short as a single sentence or utterance. This type of response was characteristic of what are commonly referred to as short completion items. While these items shared similarities with open-ended questions in reading exams, they were generally considered to fall under the objective category of test items. Typically, completion items necessitated test takers to provide a

single word or a brief statement. If there was not a clear and singular proper response, the process of marking the test would be challenging when the test taker was presented with a range of answers that vary in acceptability.

An extended production response referred to a response that was longer than a single sentence or utterance. It could vary in length, ranging from two sentences or utterances to a more extensive composition, whether it was spoken or written. This type of answer was characteristic of composition tests. The challenge in a composition test lied in formulating the rubrics. If the explanation of the situation on which the composition was based is excessively lengthy, the text transformed into more of a reading comprehension assessment. However, the rubric must contain enough information to effectively guide the composition and offer a practical foundation. Therefore, it was crucial to convey the precise amount of context using straightforward and succinct language that was clear and easily understood.

1.8.3 Construct Validity

Construct validity is a central concept in language testing and assessment, referring to the extent to which a test measures the theoretical construct it is intended to measure. It is foundational to score interpretation, ensuring that the inferences drawn from test scores accurately reflect the underlying abilities or traits being assessed. As defined by Bachman (1990), construct validity is a unitary concept, and its core lies in the degree to which hypothesized abilities can be inferred from test performance. This means that, for a test to have strong construct validity, it must effectively tap into the

specific traits, abilities, or skills that it purports to measure. Response formats in language testing directly impact the construct validity of an assessment. Different response formats — such as MCQs, T/F/NGs, and SAQs can offer varied ways to assess reading comprehension or other skills, but they must be carefully designed to align with the underlying construct.

MCQs are commonly used in large-scale language tests due to their ease of administration and scoring. However, their alignment with construct validity depends on how well the options reflect the construct. If the test is designed to measure reading comprehension, for example, MCQs must ensure that the options assess meaningful understanding rather than guessing. Additionally, if the theoretical construct includes higher-order cognitive skills like inference-making or analysis, MCQs need to be crafted to test these skills adequately.

T/F/NGs test a student's ability to comprehend specific details within a text. They are aligned with constructs related to reading comprehension but may not fully capture higher-order thinking skills like inference or synthesis. Therefore, the construct validity of T/F/NGs can be enhanced if the questions are designed to target subtle nuances in the text, which requires deeper understanding rather than simple recall.

SAQs provide an opportunity for students to generate their own responses, which can more directly reflect their ability to process and articulate information. From a construct validity standpoint, SAQs are valuable because they allow test-takers to demonstrate a broader range of cognitive processes, such as summarizing,

paraphrasing, and synthesizing information. They are aligned with constructs that measure more complex language skills but may pose challenges in terms of scoring reliability and test-taker variation.

If the goal is to assess reading comprehension, the test must measure various sub-skills such as vocabulary understanding, main idea identification, and inferential reasoning. Response formats like MCQs, TFNQs, and SAQs should be selected or designed to capture these aspects of comprehension. For example, MCQs may test vocabulary understanding, while SAQs might assess the ability to infer meaning or make judgments about the text.

Different response formats also impose varying levels of cognitive load on test-takers. For example, SAQs might require more cognitive effort as students must generate their own responses, whereas MCQs might only require recognition. A valid test should consider how cognitive load interacts with the specific construct. If a test is intended to measure reading speed, for instance, the cognitive load imposed by lengthy or complex response formats might undermine the validity of the assessment.

Additionally, each response format must be grounded in the theory of language proficiency it aims to measure. For instance, a test designed to assess communicative competence should include response formats that reflect the real-world use of language, such as tasks requiring interaction or problem-solving, rather than relying solely on isolated comprehension tasks.

In summary, construct validity is a crucial aspect of language testing that ensures a test accurately measures the theoretical constructs it is designed to assess. Response formats play a vital role in achieving this alignment by ensuring that the test items accurately capture the target abilities or skills. The selection and design of response formats should be guided by the underlying construct, with careful attention to cognitive load, relevance, and alignment with the theoretical foundations of the assessment. By doing so, the test can provide valid and meaningful inferences about the test-takers' abilities.

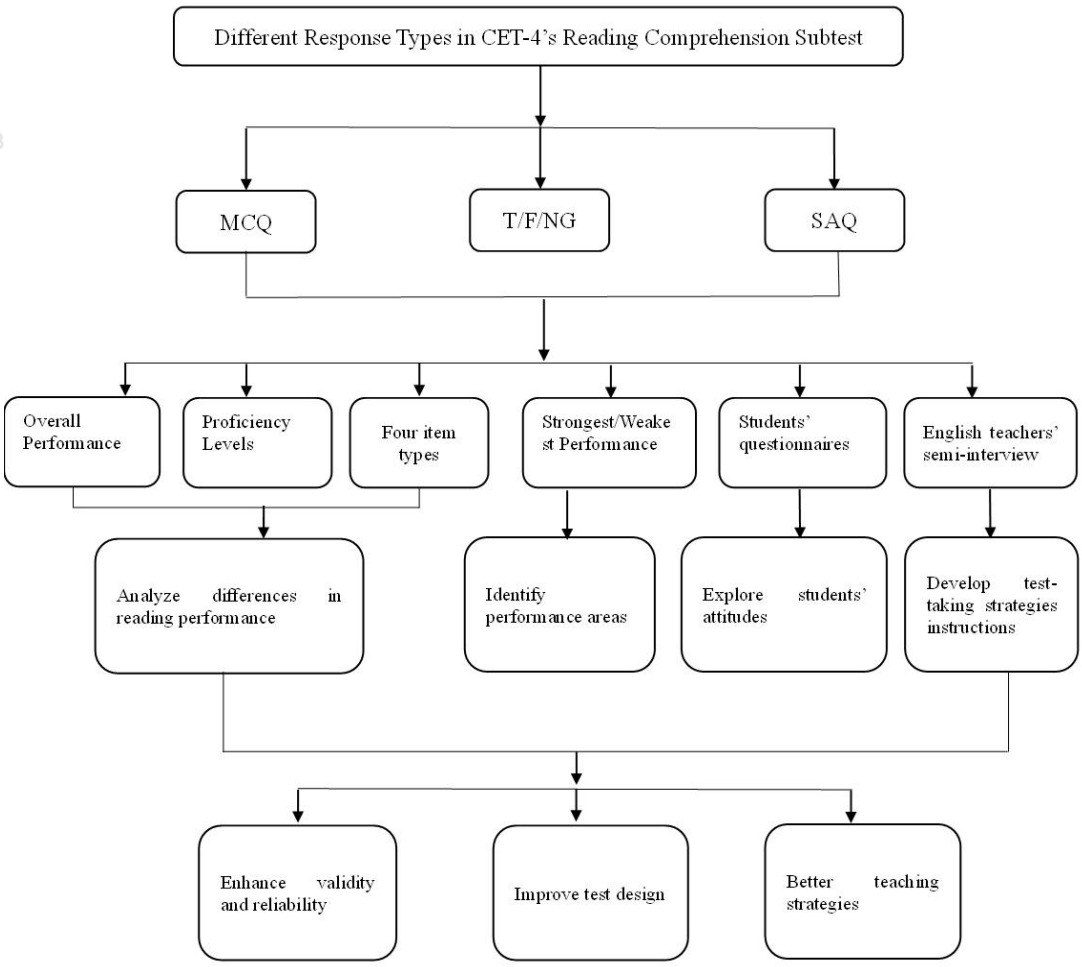
1.9 Conceptual framework

To conduct any study, the conceptual framework is essential for it can provide a guideline to the researcher. The conceptual framework for the present study, as seen in Figure 1.3. represents the variables studied and their relationships. The purpose of this study is to determine the relationship between testing response types and test scores, as mediates by test-takers' different English proficiency levels. The researcher believes that only one response type of MC (Multiple-choice questions) in testing reading comprehension cannot truly make an inference that whether he/she possesses the target reading comprehension competence or not. Therefore, the premise of the present study is that if different response types of reading comprehension sub-test test are included, more qualified the test papers are, the reference from the test performance (test scores) will be more reliable.

Since there is a relationship between different English proficiency levels and different test types, it is also the assumption of the present study that better English proficiency levels will have lower influence on different test types. Additionally, this study also aims to explore the test-takers attitudes towards the three different responses types (MC, T/F/Not Given and SAQS) adopt in this study to help truly measuring students' reading comprehension competence. The conceptual framework is listed in Figure. 1.3.

Figure 1.3

Conceptual Framework



It visually represents the key components and their relationships, focusing on the test design objective, response formats, EFL students' performance metrics, research objectives, and expected outcomes. This streamlined version is easier to follow and highlights the core aspects of your study.

1.10 Essential Characteristics and Principles of Communicative Language Testing

According to the theories of communicative language testing, there are some outstanding features should be mentioned. Bachman and Palmer (1999) supported six major features the text should have if the text needs to examine the text takers' ability of communicative language. Heaton (2004) stated that authenticity and interactive features were two distinctive characteristics of communicative tests, which distinguished communicative tests from other conventional language tests. Besides the two, there were also some other features about reliability, validity, practicality and washback. Reliability and validity were discussed heatedly in common books, journals and magazines. The other four qualities were also regarded as important in the communicative language test.

Reliability referred to the accuracy, consistency, dependability and fairness of test results. Reliability was the prior condition of validity. It was said that a test remains reliable under the condition that scores of a test under the same occasions and settings were correspondent for the same test-takers in two different tests. When the reliability was guaranteed, validity then could be assured. As for validity, a significant component of validity, it measured whether the test results could be representative of certain language ability. Validity was considered to be as important as an

indispensable quality of any language test. Of course, authenticity reflected language communication in the real world and it was crucial in CLT (communicative language testing). Interactive-ness was related to the relationship between test-takers' language ability and the test tasks. Practicality was concerned with the extent to which the ways and resources of test meet the standards. Last but not the least, washback referred to the fact that the conducted test had an effect on individual candidates, educational system and society. Next, this part gave a brief introduction of the definition of the above six major features in CLT (communicative language testing).

1.10.1 Validity

1.10.1.1 Definition of Validity

Validity is often seen as the essential quality of good assessment. Validity, which is the most crucial term in language testing, has been defined in numerous ways. According to Bachman (1990), validity was a singular idea that pertained to the sufficiency and suitability of how we interpreted and utilized test scores. He proposed that the most important consideration in designing and developing a language test was its validity, and he explained that validity pertained to the meaningfulness and appropriateness of the interpretations made on the basis of test scores.

Henning (2001) defined validity as the extent to which a test or its components accurately measured what they were intended to measure. A test was considered valid if it effectively assessed the specific construct it aimed to evaluate. From Henning's

definition of validity, validity could vary in degrees, indicating that understanding the test's purpose was crucial for ensuring its validity. This implied that users had to use their own, or somebody else's judgments when deciding on the degree of a test's validity. Henning argued that invalid tests were those containing undesirable content alongside the desired material. Consequently, a test may be valid for certain purposes but not for others.

Cohen, Manion, and Morrison (2017) defined validity as an assessment of how accurately a test or measurement tool measures what it claimed to measure. This judgement was crucial because it affected the appropriateness of the conclusions drawn and the actions taken based on the measurements.

According to Heaton (2004), the validity of a test was determined by how accurately it measured its intended purpose and nothing else. The test must strive to accurately assessed the specific abilities it was designed to evaluate.

According to Anthony (2020), validity referred to the degree to which the inferences about assessed made by the users of assessment results were justified by the evidence the assessment provided.

According to Brennan (2023), validity was the extent to which a test accurately measured what it was designed to measure or could be effectively used for its intended objectives. Various statistical procedures could be used to assess the validity of a test. Such procedures generally sought to determine what the test measured and how well it did so.

No single validity can be considered superior to another. Despite the many interpretations, this thesis will adhere to the definition provided by Standards, as this definition explicitly connects validity to the inferences that are made on the basis of test scores. Validity refers to the extent to which accumulated evidence and theory support a specific interpretation of test scores for a specific usage of a test (Standards, 2014).

1.10.1.2 Types of validity

This part presents the types of validity in detail from different perspectives, for various classifications of validity have confused many people.

(1) Face validity

Many scholars defined face validity (Bachman, 1990; Heaton, 2004; Alderson, 2011; Cohen et al., 2017; Anthony,2020; etc.). This paper adheres to the definition in the Standards (2014), which describes the extent to which accumulated evidence and theory support a specific interpretation of test scores for a given use of the test. When a test has high face validity, test takers remain motivated and put forth their best effort in completing the test tasks. Conversely, if the test items lack face validity and do not appear appropriate to the test-takers, they are less likely to fully engage, impacting the reliability of the test scores. Since the emergence of communicative language testing, there has been a growing emphasis on face validity. In fact, some advocates of communicative language testing today consider face validity to be the most crucial

aspect of all types of test validity, arguing that a well-designed test should look like something one might do “in the real world” with the language.

(2) Content validity

Many scholars defined face validity (Henning, 2001; Heaton, 2004; Cohen et al., 2017; Alderson, 2011; Anthony, 2020). This paper adheres to the definition provided in the Standards (2014), which described content validity as the aspect of validity necessary when a test user aimed to evaluate how an individual performs across the range of situations the test was designed to represent. Content validity was critically important; the higher a test’s content validity, the more accurately it measured what it was intended to assess.

(3) Concurrent validity

Many scholars defined face validity (Bachman, 1990; Wood, 2001; Cohen et al., 2017; Anthony, 2020). This research adapts Alderson (2011)’s definition, stating that concurrent validity involved comparing test scores with another measure for the same candidates taken at roughly the same time. This measure could be a well-established standardized test, scores from a parallel version of another test, or other quantifiable variables. It was pointless to compare test takers’ scores with their performance on an unreliable and invalid measure, although reliable data could be challenging to obtain in practice. If the two measures yielded similar or equivalent results, they were considered to have concurrent validity.

(4) Predictive validity

Many scholars defined face validity (Henning, 2001; Cohen et al., 2017; Anthony, 2020). According to the Standards (2014), predictive validity was a type of criterion-related validity used to infer an individual's probable standing on another variable, referred to as a criterion, based on their test score. The term criterion-related validity was updated to criterion-related evidence to emphasize that it represented one type of evidence within a unified conception of validity. Predictive evidence showed how accurately test data can forecast criterion scores obtained at a later time.

(5) Construct validity

Construct validity can be defined in various ways. Many scholars defined face validity (Bachman, 1990; Chapelle & Chapelle, 2001; Wood, 2001; Heaton, 2004; Cohen et al., 2017; Anthony, 2020). This paper adheres to the definition provided in the Standards (2014), which stated that construct validity was used to indicate that test scores reflected the test takers' standing on the psychological construct being measured. A construct was a theoretical variable inferred from various types of evidence, including the interrelations of test scores with other variables, observations of response processes, internal test structures, and the content of the test.

1.10.2 Reliability

Reliability is an essential prerequisite for any high-quality test. It is strongly correlated with the accuracy of measurement. This level of precision is demonstrated by the obtaining of similar results whether measurements are replicated across different occasions, utilizing diverse instruments, or conducted by different individuals.

Bachman (1990) asserted that reliability measures how consistently a test produced the same results. A test was deemed reliable if it yielded identical outcomes when administered at different times or by different individuals. Wood (2001) demonstrated that reliability analysis is concerned with measuring the consistency of examinee performance and quantifying that consistency or inconsistency.

Henning (2001) suggested that reliability in testing referred to the accuracy, consistency, dependability, or fairness of the scores obtained from a specific examination. In a word, reliability cared about whether a test got the same scores one day and the next under the condition of no intervening instruction. In other words, does the test produce stable and dependable scores in the sense that they will not fluctuate very much so that we may know that the score obtained by a student is pretty close to the score he would obtain if we gave the test again? If it does, the test is reliable.

Cohen et al. (2017) defined reliability as the degree to which measurements exhibit consistency or repeatability, as well as the degree to which measurements

varied across different occasions due to measurement error. Similarly, Anthony (2020) stated that reliability pertains to the extent to which an assessment generates consistent and stable results.

As for Standards (2014), reliability referred to the extent to which test scores remain consistent when the same measurement procedure was repeated. This indicated that the scores could be considered reliable and consistent for an individual test taker. Reliability also referred to the extent to which scores were free from random measurement errors within a specific group. Reliability was deemed high when the scores for each individual remain consistent throughout repeated applications of the testing procedure, whereas it was considered low if the scores do not exhibit consistency among replications. Hence, while assessing reliability, it was crucial to have a precise understanding of what constituted a replication of the testing procedure.

1.10.3 Authenticity

Authenticity is the distinctive feature of communicative tests and a primary focus in test design. Weir (1990) maintained that authenticity should be a fundamental premise for creating and evaluating a communicative test. In recent decades, there has been an increasing emphasis on aligning language assessment with real-world language usage and understanding language as a system.

Bachman (1990) defined authenticity as the extent to which the characteristics of a language test task aligned with the elements of a target-language-use (TLU) task. A test's authenticity directly correlated with its reliability and validity. Bachman categorized authenticity into two distinct types: situational authenticity and interactive authenticity. Situational authenticity referred to how well the characteristics of a test aligned with the features of specific real-life language use situations. Interactive authenticity, on the other hand, pertained to the relationship between the test and the test taker. By analysing the scores, we could evaluate the language abilities of the test takers. Authenticity offered a way to generalize test takers' performances beyond testing scenarios. Authentic test materials and tasks could positively impact test takers, helping them perform at their best. To ensure a test possesses authenticity, it should include tasks that reflect actual communicative activities. A test task that mirrored the characteristics of a Target Language Use (TLU) task was considered relatively authentic.

Anthony (2020) defined authenticity as the extent to which language teaching materials reflected the qualities of natural speech or writing. In the context of language teaching, there was a distinction between materials specifically created to demonstrate or practice particular teaching points (such as reading passages, listening texts, or model conversations) and those derived from real-world sources. As above, if a test has the feature of authenticity, it is more direct, more real, more validity.

1.10.4 Wash-back

Testing and teaching are interconnected, with each having an impact on the other. The influence of testing on teaching is recognized as backwash or wash-back. In language education and learning, a test should strive to achieve advantageous feedback, in addition to ensuring validity and reliability. The beneficial backwash results from the proper testing procedures that align with the recognized language construct, which serves as the instructional goal of a certain curriculum. If the test content and testing techniques used in the test deviate from the objectives of the course, it is probable that there may be negative backwash. In the framework of communicative language teaching, language testing has to have a corresponding communicative orientation. A well-designed test should possess validity, reliability, practicality, and positive wash-back effects, regardless of the response forms employed by test makers or designers to assess learners' language proficiency (Bachman, 1990).

Messick (1996) suggested that the term "washback" was commonly used to describe the impact of testing on the process of teaching and learning at the classroom level. According to Anthony (2020), the term refers to the impact that an assessment has on the teaching and learning process that takes place in preparation of that assessment.

This research adheres to Heaton' (2004) definition, which described the positive or negative influence of a test on classroom teaching or learning. To effect changes in teaching practices, it may be necessary to modify the tests themselves. For example, if a country's education department aimed to increase the focus on teaching

listening skills, they could incorporate a listening comprehension component into state examinations. The washback would be that teachers would allocate more class time to teaching listening skills. When teaching significantly influenced testing, this phenomenon is referred to as reverse washback.

1.10.5 Practicality

In language testing and assessment, practicality is often referred to as one important consideration in test design. A valid and reliable test turns out to be meaningless if it lacks practicality to implement.

Bachman (1990) defined practicality as the relationship between the resources that would be required in a test and the resources available for the test. Practicality was concerned with the issue of economy and ease. As for economy, the test designer should make the test as short as possible so long as it met the requirements of the criteria of validity and reliability. As for ease, two aspects should be considered. One was the ease of administration and scoring; the other was the interpretation of scores.

Anthony (2020) defined practicality as “the gap between the resources needed for the development and use of an assessment and the resources available for these activities”. It was clear that an assessment lacking sufficient resources would not be sustainable. Without adequate time, equipment, funding, or expertise, an assessment may never be implemented and will certainly not remain in use for long.

1.10.6 Interactive-ness

Another crucial aspect of communicative language testing is interactive-ness. According to Brumfit and Johnson (1979) interactive-ness referred to face-to-face oral interaction as a form of interaction that combined receptive and productive abilities, involving the change of both expression and content. Bachman (1990), described the extent and type of the test taker's personal traits involved in completing a test task. To comprehensively assess the language proficiency of test takers, the test tasks should be created to necessitate adaptable language skills.

1.11 Limitation of study

Due to the unavoidable factors, constraints, and the researcher's limited research ability, several limitations exist that may affect the reliability, generalizability, and interpretability of the research findings.

Firstly, the sample limitation is a critical issue. Although a relatively large sample size (170 non-English major university students) was selected, all participants were from the same university. This limits the representativeness of the study's results, especially in a country like China, where there are significant geographical, economic, and cultural differences. A sample from a single institution may not adequately reflect the broader reality of university students across the country. Students from different regions may experience substantial differences in English learning resources, teaching quality, and cultural background, which could influence their performance in the CET-

4 (College English Test Band-4).

Secondly, the study primarily relies on quantitative analysis, supplemented by some qualitative data. While quantitative analysis can reveal differences and trends among groups, it overlooks the complexity and diversity behind individual differences. For example, factors such as students' emotional responses, stress levels, and the use of test-taking strategies when facing different question types could significantly impact their performance, but these factors will not be deeply explored in this study due to the space and time. This oversight may limit the study's ability to explain the diversity and complexity of student performance.

Thirdly, the researcher's knowledge of the related theories and practical experience in language testing is limited. This constraint, coupled with the inaccessibility of more advanced automated statistical software, may have impacted the depth and breadth of the analysis. The study's reliance on a limited set of indices to represent key facets of reading tests means that the findings may not fully capture all the crucial dimensions of reading comprehension.

Moreover, the limitation of time is also an important constraint. The data collection and analysis in this study focused on a specific point in time, without conducting a longitudinal follow-up study. Language learning and ability development are long-term processes, and data from a single time point may not fully capture students' true learning outcomes and progress. In language ability assessments, studies spanning a longer period can better capture learners' long-term changes and developmental trends, which were not reflected in this study.

Finally, the design and implementation of this study were constrained by available resources. For example, although the study aimed to comprehensively assess students' reading comprehension abilities through multiple response formats, the limitation of research resources and time prevented the inclusion of more kinds of response formats or more complex research design. This will restrict the breadth and depth of the study to a certain extent, potentially leading to an incomplete understanding of complex language phenomena.

In summary, these limitations highlight the need for more extensive and diversified samples, more complex and multi-layered research designs, and deeper mixed methods analysis in future research to overcome the shortcomings of the current study, thereby providing more reliable and broadly applicable conclusions.

1.12 Significance of study

As stated by Bachman (1990), test performance was also affected by the characteristics of the methods used to elicit test performance. It was found that not only test-takers' language ability, but also the response format had its due impact on test. In this light, response format has become this research's focus to check whether the current CET-4's reading comprehension test can truly indicate the students' reading ability. It is imperative and meaningful to conduct an empirical study on the differences of various response types on Chinese EFL students' reading comprehension performance in CET-4, which aims to check whether test-takers perform differently if different task types are utilized to measure reading

comprehension differ greatly.

Firstly, this study contributes to the enrichment of research in reading testing field. Although numerous studies have explored the impact of different response formats on reading comprehension performance in other countries, empirical research on large-scale national exams like CET-4/6 in China remains insufficient. By deeply investigating the effects of various response formats on the reading comprehension performance of Chinese EFL students in the CET-4, this study aims to fill a critical gap in the existing literature. Numerous studies have shown that the format of a test not only affects the test-takers' language ability but also directly impacts the validity and reliability of the test results. Given the widespread use of MCQs in the CET-4 and their potential limitations, this study will empirically analyse the suitability of the current response format to determine whether it accurately reflects students' reading comprehension abilities.

Secondly, by focusing on the CET-4, a large-scale, high-stakes exam with broad educational and social implications, this study highlights its crucial role in enhancing students' academic and professional development. The findings of this research can provide empirical support for language testing policy decisions and may have far-reaching implications for improving language assessment practices in China and internationally. Additionally, this study offers valuable insights for Chinese EFL teaching, helping educators develop more effective teaching strategies, encouraging them to reflect on the validity of current reading comprehension tests, and promoting more accurate assessments of students' reading abilities. Moreover, by examining the impact of different response formats on student performance, this study helps students

better understand their learning strengths and weaknesses, enabling more targeted learning and test preparation, and ultimately improving their English reading comprehension levels.

Finally, from a methodological perspective, this study employs a mixed-method approach, combining quantitative and qualitative data to comprehensively analyse the impact of different response formats on students' test performance and attitudes. This approach not only validates the influence of response formats on test outcomes but also provides critical insights for test developers and educators in designing reading assessments that better align with the communicative nature of language use.

By thoroughly analysing the data and deeply exploring questionnaire and interview results, this research demonstrates how effectively combining quantitative and qualitative methods can advance language testing practices, enhancing the validity and reliability of assessments. Additionally, the methods and findings of this study offer valuable references for the design and implementation of future research, particularly in the comparative analysis of different English reading sub-tests.

1.13 Summary

The background section emphasizes the centrality of reading comprehension in language testing, discussing the dominance of MCQs in CET-4 and the potential drawbacks of this format. The need for more communicative approaches in reading

tests is stressed, arguing that tests should better reflect real-life language use to enhance students' communicative competence. Next, it also details various reading comprehension response formats, including MCQs, SAQs, cloze tests, and others, highlighting their respective advantages and disadvantages. The problem statement identifies gaps in the current CET-4 reading test, particularly the reliance on MCQs, especially the necessity of investigating and designing by combing different testing formats in reading comprehension part. It outlines the theoretical frameworks: Bachman's communicative language testing theory, and test methods facets. And then, the research objectives and questions are outlined, aiming to explore the differences in test performance across various response formats and to assess the impact of these formats on Chinese EFL learners. Lastly, the limitation and significance of the study are been discussed above, which can lay a sound foundation for the latter chapters of this study.