VIETNAM NATIONAL UNIVERSITY, HANOI

**UNIVERSITY OF LANGUAGES AND INTERNATIONAL STUDIES**

**\*\*\*\*\*\*\*\***

**WEI YU**

**SUMMARY OF THE DOCTORAL DISSERTATION**

**A STUDY ON THE VALIDITY OF A VIETNAMESE DEVELOPED TEST OF ENGLISH READING PROFICIENCY**

**(Nghiên cứu xác trị bài thi đọc hiểu của một định dạng bài thi tiếng Anh của Việt Nam)**

**MAJOR:** ENGLISH LANGUAGE TEACHING METHODOLOGY

**CODE:** 9140231.01

**SUPERVISORS:** 1. PROF. NGUYEN HOA

2. DR. DUONG THU MAI

**HANOI, 2023**

TABLE OF CONTENTS

[LIST OF ABBREVIATIONS v](#_Toc149758022)

[CHAPTER I: INTRODUCTION 1](#_Toc149758023)

[1.1 Rationale of the study 1](#_Toc149758024)

[1.2 Objective of the study 1](#_Toc149758025)

[1.3 Scope of the study 2](#_Toc149758026)

[1.4 Clarifications of the key terms 2](#_Toc149758027)

[1.5 Context of the study 2](#_Toc149758028)

[1.5.1 Introduction of VSTEP.3-5 3](#_Toc149758029)

[1.5.1.1 Development of VSTEP.3-5 3](#_Toc149758030)

[1.5.1.2 Administration of VSTEP.3-5 3](#_Toc149758031)

[1.5.1.3 Test takers of VSTEP.3-5 3](#_Toc149758032)

[1.5.1.4 Test format of VSTEP.3-5 3](#_Toc149758033)

[1.5.1.5 Cut-off scores of VSTEP.3-5 3](#_Toc149758034)

[1.5.2 Introduction of VSTEP.3-5 reading test 3](#_Toc149758035)

[1.5.2.1 Test purpose of VSTEP.3-5 reading test 3](#_Toc149758036)

[1.5.2.2 Test format of VSTEP.3-5 reading test 3](#_Toc149758037)

[1.5.2.3 Proficiency level description of VSTEP.3-5 reading test 4](#_Toc149758038)

[1.6 Significance of the study 4](#_Toc149758039)

[1.7 Overall structure of the thesis 4](#_Toc149758040)

[CHAPTER II: LITERATURE REVIEW 4](#_Toc149758041)

[2.1 Review of L2 reading 4](#_Toc149758042)

[2.1.1 Nature of L2 reading 4](#_Toc149758043)

[2.1.2 Constructs of L2 reading 5](#_Toc149758044)

[2.1.2.1 The processing perspective 5](#_Toc149758045)

[2.1.2.2 The task perspective 6](#_Toc149758046)

[2.1.2.3 The reader purpose perspective 6](#_Toc149758047)

[2.2 Review of test validation approaches 7](#_Toc149758048)

[2.2.1 Test validation approaches from product-based perspective 7](#_Toc149758049)

[2.2.1.1 Item Response Theory 7](#_Toc149758050)

[2.2.1.2 Rasch models 8](#_Toc149758051)

[2.2.2 Test validation approaches from process-based perspective 8](#_Toc149758052)

[2.2.2.1 Think aloud protocols 8](#_Toc149758053)

[2.2.2.2 Eye tracking 8](#_Toc149758054)

[CHAPTER III: THEORETICAL FRAMEWORK 9](#_Toc149758055)

[3.1 Validity 9](#_Toc149758056)

[3.2 Studies on the validity of VSTEP.3-5 9](#_Toc149758057)

[3.3 Validation 9](#_Toc149758058)

[3.3.1 Evidence-based validation frameworks 10](#_Toc149758059)

[3.3.1.1 Bachman and Palmer’s test usefulness framework 10](#_Toc149758060)

[3.3.1.2 Weir’s Socio-cognitive framework 10](#_Toc149758061)

[3.3.2 Argument-based validation frameworks 10](#_Toc149758062)

[3.3.2.1 Kane’s interpretive argument framework 10](#_Toc149758063)

[3.3.2.2 Chapelle, Enright, and Jamieson’s interpretive argument framework 10](#_Toc149758064)

[3.3.2.3 Bachman and Palmer’s assessment use argument framework 10](#_Toc149758065)

[3.4 Empirical validation studies with interpretive argument framework 10](#_Toc149758066)

[3.5 Interpretive argument for the VSTEP.3-5 reading test 11](#_Toc149758067)

[3.5.1 Justification for the use of interpretive argument 11](#_Toc149758068)

[3.5.2 Articulation of an interpretive argument for the VSTEP.3-5 reading test 11](#_Toc149758069)

[CHAPTER IV: RESEARCH METHODOLOGY 11](#_Toc149758070)

[4.1 Ontology and epistemology for the design of the study 11](#_Toc149758071)

[4.2 The convergent mixed methods design for the study 11](#_Toc149758072)

[4.3 Participants of the study 12](#_Toc149758073)

[4.4 Instruments of the study 13](#_Toc149758074)

[4.4.1 A version of the VSTEP.3-5 reading test paper 13](#_Toc149758075)

[4.4.2 The VSTEP.3-5 reading test specifications 13](#_Toc149758076)

[4.4.3 Expert interview questions 13](#_Toc149758077)

[4.4.4 The modified framework of task characteristics for the VSTEP.3-5 reading test 13](#_Toc149758078)

[4.4.5 Expert judgment forms 13](#_Toc149758079)

[4.4.6 Think aloud protocols operation instructions 14](#_Toc149758080)

[4.4.7 Eye tracking metrics 14](#_Toc149758081)

[4.4.8 The reading processing checklist 14](#_Toc149758082)

[4.5 Data collection procedures of the study 14](#_Toc149758083)

[4.5.1 Data collection procedures for research questions 1, 2 and 3 14](#_Toc149758084)

[4.5.2 Data collection procedures for research question 4 14](#_Toc149758085)

[4.5.3 Data collection procedures for research question 5 15](#_Toc149758086)

[4.6 Tools for data analysis 15](#_Toc149758087)

[4.6.1 Compleat Lexical Tutor 15](#_Toc149758088)

[4.6.2 Readable.io 15](#_Toc149758089)

[4.6.3 WINSTEPS 5.4.1 15](#_Toc149758090)

[4.6.4 SPSS 26.0 15](#_Toc149758091)

[4.7 Data analysis procedures of the study 15](#_Toc149758092)

[4.7.1 Data analysis procedures for research question 1 15](#_Toc149758093)

[4.7.2 Data analysis procedures for research question 2 16](#_Toc149758094)

[4.7.3 Data analysis procedures for research question 3 16](#_Toc149758095)

[4.7.4 Data analysis procedures for research question 4 16](#_Toc149758096)

[4.7.5 Data analysis procedures for research question 5 16](#_Toc149758097)

[CHAPTER V: FINDINGS AND DISCUSSIONS ON VALIDITY OF VSTEP.3-5 READING TEST FROM THE PRODUCT-BASED PERSPECTIVE 17](#_Toc149758098)

[5.1 Findings and discussions for research question 1 17](#_Toc149758099)

[5.1.1 Findings from test tasks 17](#_Toc149758100)

[5.1.1.1 Comparison of test tasks in the investigated test and test specifications 17](#_Toc149758101)

[5.1.1.2 Expert interview on test tasks 17](#_Toc149758102)

[5.1.2 Findings from Rasch model analysis of test items 17](#_Toc149758103)

[5.1.2.1 Item unidimensionality 17](#_Toc149758104)

[5.1.2.2 Item estimates according to Rasch model analysis 17](#_Toc149758105)

[5.1.2.3 Item and person calibration according to Rasch model analysis 18](#_Toc149758106)

[5.1.3 Discussions 18](#_Toc149758107)

[5.2 Findings and discussions for research question 2 18](#_Toc149758108)

[5.2.1 Differential item functioning by genders 18](#_Toc149758109)

[5.2.2 Differential item functioning by majors 18](#_Toc149758110)

[5.2.3 Discussions 18](#_Toc149758111)

[5.3 Findings and discussions for research question 3 19](#_Toc149758112)

[5.3.1 Separation and reliability measures 19](#_Toc149758113)

[5.3.2 Discussions 19](#_Toc149758114)

[CHAPTER VI: FINDINGS AND DISCUSSIONS ON VALIDITY OF VSTEP.3-5 READING TEST FROM THE PROCESS-BASED PERSPECTIVE 19](#_Toc149758115)

[6.1 Findings and discussions for research question 4 19](#_Toc149758116)

[6.1.1 Expert judgement on reading subskills in the investigated test 19](#_Toc149758117)

[6.1.2 Discussions from expert judgement 20](#_Toc149758118)

[6.1.3 Think aloud protocols 20](#_Toc149758119)

[6.1.3.1 Reading strategies in think aloud protocols 20](#_Toc149758120)

[6.1.3.2 Reading subskills in think aloud protocols 20](#_Toc149758121)

[6.1.4 Discussions from think aloud protocols 21](#_Toc149758122)

[6.2 Findings and discussions for research question 5 21](#_Toc149758123)

[6.2.1 Eye tracking of successful and unsuccessful test takers across item types 21](#_Toc149758124)

[6.2.2 Discussions from eye tracking of successful and unsuccessful test takers across item types 22](#_Toc149758125)

[6.2.3 Eye tracking of test takers at different proficiency levels and results of reading processing checklists 22](#_Toc149758126)

[6.2.4 Discussions from eye tracking of test takers at different proficiency levels and results of reading processing checklists 23](#_Toc149758127)

[CHAPTER VII: CONCLUSION 23](#_Toc149758128)

[7.1 Interpretive argument for the VSTEP.3-5 reading test revisited 23](#_Toc149758129)

[7.1.1 A recap of the generalization inference 23](#_Toc149758130)

[7.1.2 A recap of the explanation inference 23](#_Toc149758131)

[7.1.3 Summary 23](#_Toc149758132)

[7.2 Implications of the study 24](#_Toc149758133)

[7.3 Limitations of the study 24](#_Toc149758134)

[7.4 Suggestions for future studies 24](#_Toc149758135)

[LIST OF PUBLICATIONS RELATED TO DISSERTATION 24](#_Toc149758136)

# LIST OF ABBREVIATIONS

AERA American Educational Research Association

ALTE Association of Language Testers in Europe

AOI Areas of Interest

APA American Psychological Association

ASEAN Association of Southeast Asian Nations

BNC British National Corpus

CAE Cambridge English: Advanced

CCC Category Characteristic Curve

CEFR Common European Framework of Reference for Languages

CEFR-VN Common European Framework of Reference for languages for Vietnamese learners

CET-4 College English Test Band 4

CTT Classical Test Theory

DIF Differential Item Functioning

DRP Degrees of Reading Power

EAP English for Academic Purposes

EPT English Placement Test

ESL English as a second language

FCE First Certificate in English

GEPT General English Proficiency Test

GRE Graduate Record Examination

ICC Item Characteristic Curve

IELTS International English Language Testing System

IRT Item Response Theory

L1 First Language

L2 Second Language

MELAB Michigan English Language Assessment Battery

MNSQ Mean Square

MSE Main Suite Examinations

NCME National Council on Measurement in Education

NFL 2020 National Foreign Language 2020 Project

OECD The Organization for Economic Co-operation and Development

PCA Principal Component Analysis

PTMA-CORR. Point-measure Correlation

PTMA-EXP. Point-measure Expected Correlation

SPSS Statistical Product and Service Solutions

TEEP Test of English for Educational Purposes

TOEFL Test of English as a Foreign Language

TOEFL iBT Foreign Language Internet-Based Test

VSTEP Vietnamese Standardized Test of English Proficiency

VSTEP.3-5 Vietnamese Standardized Test of English Proficiency 3-5

ZSTD Standardized as a Z-score

# CHAPTER I: INTRODUCTION

In this chapter, rationale of the study is firstly presented to locate this study in a broad context of language testing and assessment. The study centers on an eternal topic in language testing and assessment which is validity. Specifically, several reasons are given to justify the necessity of conducting the current study from both product-based and process-based perspectives. Objective of the study is put forward, followed by scope of the study and clarifications of key terms. And then, context of the study is illustrated in the aspects of the introduction of VSTEP.3-5 and it’s reading test. Significance of the study and overall structure of the thesis are specifically described. This chapter ends with a summary.

## 1.1 Rationale of the study

There are three important reasons that motivate the conduction of the current study. **Firstly**, it can be attributed to the important position of validity in the field of theories and practices in language testing and assessment. Theoretical exploration and empirical studies have never ceased to emerge on the validity of tests. **Secondly**, in Vietnam, one of the significant achievements in English language testing and assessment is the emergence of VSTEP.3-5. Since it is a newly developed test, like other high-stakes tests in order to evaluate validity and reliability, more research needs to be conducted. **Thirdly**, the testing of reading is by no mean an easy task. It has been one of the most controversial issues for its lack of an overt process to be observed directly, unlike writing or speaking competence which can be performed in a concrete way by test takers for researchers to examine all kinds of indicators of it carefully. The traditional way to investigate validity of a reading test is to observe the response of a test taker elicited through an indirect test method. Only from this perspective is not enough. The validation of reading test needs multidimensional evidences combining test product with test process to explore the possibility of discovering new and valuable findings.

## 1.2 Objective of the study

The objective is to make an ongoing attempt in building a systematic, transparent and reasonable section of validity argumentation for the VSTEP.3-5 reading test. To be specific, framed within an argument-based approach combining qualitative and quantitative analyses, this study aims to investigate the validity of VSTEP.3-5 reading test from both product-based and process-based perspectives comprehensively.

Thus, the main research questions structured for this study are as follows:

1. To what extent do the test tasks and test items of the VSTEP.3-5 reading test meet the requirements of the test specifications?
2. To what extent do the test takers’ genders and majors affect their reading scores?
3. To what extent are the VSTEP.3-5 reading test scores adequately reliable in measuring the test takers’ English proficiency?
4. What reading processes are applied to answer VSTEP.3-5 reading test correctly? From think aloud protocols, to what extent do the expected reading processes correspond with the processes actually engaged in by test takers while taking VSTEP.3-5 reading test?
5. How do successful and unsuccessful test takers differ in processing the different types of VSTEP.3-5 reading test items in terms of their eye movements? From eye tracking, to what extent does the VSTEP.3-5 reading test discriminate test takers at different proficiency levels in terms of cognitive reading process?

## 1.3 Scope of the study

This study was conducted in the context of the COVID-19, which increased the difficulty of collecting data. Because both Vietnam and China had severe outbreaks at that time, it was unavailable for the researcher to carry out the whole study in Vietnam. Thus, data such as the test takers’ scores of VSTEP.3-5 reading test, think aloud protocols, eye tracking etc. were collected in China for this study. The research sites were two universities in China. One was a comprehensive university; the other was a normal university. The two universities are major public tertiary education institutions in China with high reputation for quality teaching and learning. English is taught both for general and specific purposes. Both of them offered undergraduate and postgraduate programs to more than 20,000 students in different academic disciplines. Admittedly, VSTEP3-5 aims at assessing English proficiency from level 3 to level 5 according to the CEFR-VN for Vietnamese learners. But as its designers stated, it aims to be internationally recognized, thus the results of this study indicate that the test can be taken by international students.

## 1.4 Clarifications of the key terms

If the key terms in the current study are not clarified first, the investigation cannot be conducted effectively. **Validity** in the current study refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of a test, which is defined in *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Furthermore,validity defined here after is closely related not only to test products but also to the processes which give rise to the test products and the understanding of the characteristics being measured requires an in-depth understanding of the cognitive processes required by test takers to complete the test tasks. **Product-based perspective** in the current study refers to rely mainly on collected evidence from assessment products such as using observed test scores to validate a test. Another term which is opposite to product-based perspective is **process-based perspective**, which refers to rely mainly on collected evidence from the test-taking processes to validate a test. In order to draw clearer conclusions about the validity of the test, this study aims to combine evidence of test products with experimental evidence of specific cognitive processes involved in the construction of testing processes by test takers.

## 1.5 Context of the study

The context of the study includes introduction of VSTEP.3-5 and introduction of VSTEP.3-5 reading test which provide the necessary information for the context of the study.

## 1.5.1 Introduction of VSTEP.3-5

#### 1.5.1.1 Development of VSTEP.3-5

VSTEP is a 6-level program for Vietnamese students which is considered a breakthrough solution in renewing and evaluating foreign language competencies in Vietnam. The VSTEP.3-5 has been officially implemented nationwide in Vietnam since March 2015 after developing by researchers from a professional scientific research team in University of Languages and International studies, Vietnam National University-Hanoi.

#### 1.5.1.2 Administration of VSTEP.3-5

By 2021, the Ministry of Education and Training allows several universities to organize VSTEP.3-5. Basically, VSTEP.3-5 is organized in two forms: paper-based and computer-based.

#### 1.5.1.3 Test takers of VSTEP.3-5

VSTEP.3-5 is widely used as a standard to measure the comprehensive English ability of Vietnamese English learners. According to the National Foreign Language 2020 Project (NFL 2020) and related documents, the test takers of VSTEP.3-5 includes several categories.

#### 1.5.1.4 Test format of VSTEP.3-5

VSTEP.3-5 includes four subtests which are listening test, reading test, writing test and speaking test.

#### 1.5.1.5 Cut-off scores of VSTEP.3-5

The scores of each subtest in VSTEP.3-5 are applied to assess each language skill of test takers respectively. The final score is the average of the four subtests’ scores. The cut-off scores are preset on the model VSTEP.3-5, and the results are applied to all tests assuming that the tests are strictly designed according to the test specifications.

## 1.5.2 Introduction of VSTEP.3-5 reading test

#### 1.5.2.1 Test purpose of VSTEP.3-5 reading test

VSTEP.3-5 reading test is a subtest of VSTEP.3-5. The test is designed in order for assessing the English reading competence of Vietnamese English learners.

#### 1.5.2.2 Test format of VSTEP.3-5 reading test

VSTEP.3-5 reading test consist of four different passages, each of which is followed by 10 multiple-choice questions. For these 40 questions, test takers are supposed to choose the one best answer for each question. Then, the test takers should find the question number on the answer sheet and fill in the space corresponding to the selected answer letter. Test takers need to answer all questions following a passage on the basis of what is stated or implied in that passage. Test takers have 60 minutes to answer all the questions, including the time to transfer the answers to the answer sheet. The length of the four passages varies from each other, ranging from 400 words to 550 words, with the total number of words being approximately 1900 to 2050 words. The passages of VSTEP.3-5 reading test cover a wide range of texts, roughly divided into the following categories: daily life readings, natural science articles, social science articles, other disciplines articles, specialized academic or professional publications, literary writings, and background information about Asia, Association of Southeast Asian Nations (ASEAN) or Vietnam.

#### 1.5.2.3 Proficiency level description of VSTEP.3-5 reading test

VSTEP.3-5 reading test focuses on evaluating English language learners’ reading proficiency from level 3 to level 5. The detailed information about proficiency level of VSTEP.3-5 reading test are presented.

## 1.6 Significance of the study

The current study contributes theoretically and practically to the field of research on VSTEP.3-5 reading tests. Theoretically, this current study may explore the possibility of test validation from product-based and process-based perspectives which provide a referable paradigm for future studies trying to validate VSTEP.3-5 tests based on certain theoretical framework. Practically, this current study may provide the references to raise the design of VSTEP.3-5 reading test for the test designers, aiming at improving its validity. Meanwhile, this study may help the English teachers and learners well comprehend VSTEP.3-5 reading test. Therefore, it can benefit the English teachers in effective language teaching and for English learners in test preparation.

## 1.7 Overall structure of the thesis

The whole dissertation consists of seven chapters which are introduction, literature review, theoretical framework, research methodology, findings and discussions on validity of VSTEP.3-5 reading test from the product-based perspective, findings and discussions on validity of VSTEP.3-5 reading test from the process-based perspective, and conclusion.

# CHAPTER II: LITERATURE REVIEW

This chapter reviews different approaches of conceptualizing L2 reading and the ways of these approaches shape the constructs of L2 reading. Besides, some test validation approaches of the current study are also reviewed in detail.

## 2.1 Review of L2 reading

This part begins with descriptions of the nature of L2 reading. Then, by articulating three main perspectives, this part continues to explain how these perspectives provide information for the conceptualization of constructs of L2 reading.

### 2.1.1 Nature of L2 reading

Definitions of reading and differences between L1 reading and L2 reading are reviewed respectively. The diversified versions of initial introduction or description of reading were adopted in different studies. In the most general terms, it is accepted that reading is a dynamic process, involving the reader, the text and the interaction between the reader and the text (Rumelhart, 1977). To specify the process of this interaction, professional studies indicate that it is readers’ comprehension and interpretation of the symbols on the paper (Harris & Sipay, 1985; Keith & Alexander 1989; William & Fredricka, 2005) and it engages in both the cognitive operation and linguistic knowledge, with which readers construct a meaningful representation of an author’s message (Francoise, 1981).

In the absence of a comprehensive theory of L2 reading development, researchers focusing on L2 reading tended to reply on the theoretical framework established in the study of the native language and assumed that L2 reading development comprise components that are similar or identical to those identified or studied in L1 (Jolly, 1978). However, the simple inferential theories could not stand for long and abundant research with English as a second language (ESL) readers as subjects were conducted and analyzed, resulting in progressive reports of evident difference existing between L2 and L1 reading process. A definition of reading employed by *Programme for International Student Assessment* (OECD, 2016) is selected, which demonstrates the multiplicity of the nature of reading: Reading literacy is understanding, using and reflecting on written texts, in order to achieve one’s goals, to develop one’s knowledge and potential and to participate in society. In sum, the activity of reading is not simply a decoding process to extract information. Instead, it combines readers’ active contributions to constructing meaning for a variety of purposes and the activity is embedded in a sociocultural context.

### 2.1.2 Constructs of L2 reading

Enright et al. (2000) present three perspectives in thinking about the constructs of reading comprehension for the TOEFL 2000 project – the processing perspective, the task perspective and the reader purpose perspective which are particularly relevant to the purpose of the current study and are discussed below respectively.

#### 2.1.2.1 The processing perspective

From the processing perspective, general processing approaches and the cognitive processing approach are reviewed.

*General processing approaches*

From reviewing the literature in reading, the evolution of theories and studies on the processes of L2 reading have been influenced by several main approaches from L1 reading. They are the bottom-up approach, the top-down approach, the interactive approach and the sociocultural approach.

*The cognitive processing approach*

Khalifa and Weir (2009) propose a cognitive processing approach to establishing the construct measured by reading tests and examining tests’ cognitive validity.

**Figure 2.1**

*Khalifa and Weir’s cognitive model of reading (2009)*

****

The above figure illustrates the model. It is hierarchical and possible to operationalize the mechanisms or the internal structure of cognitive processes and it highlights the central role of test takers and brings to light test takers’ underlying cognitive processes that govern their reading activities. This is the reason why it will be employed to serve as the landing point for a test-taking process investigation. There are several empirical studies for investigate readers’ cognitive process in taking English reading test which are from Bax & Weir (2012), Bax (2013), Brunfaut & McCray (2015) and Bax & Chan (2019).

The merit of Khalifa and Weir’s cognitive model of reading lies in that it highlights the central role of test takers and brings to light test takers’ underlying cognitive processes that govern their reading activities. This is the reason why it would be employed to serve as the landing point for a test-taking process investigation, especially for think aloud protocols coding and eye tracking in the current study. All considered, Khalifa and Weir’s (2009) cognitive processing model of reading can be an adequate reference to analyze the mental processing involved in VSTEP.3-5 reading test.

#### 2.1.2.2 The task perspective

The second perspective uses task-based variables to define reading construct. Bachman and Palmer (1996) propose a framework of task characteristics. Figure below lists the concrete parts of the framework of Bachman and Palmer task characteristics. The test task characteristics which are taken into consideration in this study are those under the following two categories: characteristics of the input and characteristics of the expected response.

**Figure 2.2**

*Task Characteristics Framework (Bachman and Palmer, 1996)*



#### 2.1.2.3 The reader purpose perspective

The reader purpose perspective acts as another important aspect of reading comprehension. Several studies describe reading from reader purpose perspective such as Hochberg and Brooks (1970), Smith (1994), and Grabe & Stoller (2002). Enlightened from these definitions, this part elaborates the related studies on L2 reading purposes, subskills and strategies according to the literature.

*L2 reading purposes*

Different scholars hold different views on L2 reading purposes. In general, there are three main aspects which are for real-life communicative purposes, for communicative purposes in reading tasks, and for academic purposes.

*Reading subskills*

Throughout the research on reading skills, there are mainly two views. One is the unitary view; the other is multi-divisible view. The unitary view argue that reading is an indivisible, holistic skill that can be simply called reading ability (Gholamreza, 2013). Many studies support this view (Rost, 1993; Spearritt, 1972; Drahozal & Hanna, 1979; Carver, 1992; Lunzer, Waite, & Dolan, 1979; Rosenshine, 1980). The multi-divisible view argue that reading can be divided into a set of distinct subskills. Many researchers conduct studies to hold this view (Carroll, 1993; Drum, Calfee, & Cook, 1981; Pollitt et al. 1985; Davey, 1988; Anderson et al, 1991; Bachman, Vanniarjan, & Lynch, 1988; Brutten, Perkins, & Upshur, 1991; Lumley, 1993; Weakly, 1993; Weir et al., 1990).

The concept of reading subskills acts as a basis for constructing standard language proficiency tests such as TOFEL, IELTS, and so on. Is L2 reading a unitary concept or a divisible concept? As to the context of the current study, the latter is suitable. Because one of the purposes is to understand whether or not the reading subskills of each item in the investigated VSTEP.3-5 reading test is consistent with the subskills of L2 reading proficiency provided by the test specifications.

*Reading strategies*

Many researchers have given definitions and classifications of reading strategies (Pressley & Afflerbach, 1995; Carrell & Grabe, 2002; Upton & Lee-Thompson, 2001; Bachman & Palmer, 1996; Cohen & Upton, 2006; Fransson, 1984; Gagne, et al, 1993). A better understanding of reading 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.

So far, the nature and construct of L2 reading have been discussed respectively in detail. The following section of this chapter reviews the related methodological issues of the current study.

## 2.2 Review of test validation approaches

For the perspective of test products, the current study is mainly based on Item Response Theory (IRT) and adopts the suitable Rasch model to collect evidence. As to the perspective of test-taking processes, this study mainly collects evidence by using think aloud protocols and eye tracking.

### 2.2.1 Test validation approaches from product-based perspective

#### 2.2.1.1 Item Response Theory

IRT (Rasch, 1960; Lord & Novick, 1968), as an extension of Classical Test Theory (CTT) (Gulliksen, 1950), is characterized by a collection of mathematical models in which a set of parameters (discrimination and/or difficulty and/or guessing) is used to describe the probability of a given response to an item conditional on a test taker’s ability level. The biggest criticism of CTT is its inability to separate the dependency between test characteristics and test taker characteristics; however, IRT overcomes the shortcoming through its property of item parameter invariance and ability parameter invariance (Yang, 2005).

#### 2.2.1.2 Rasch models

*Rasch family of models*

After decades of development, the basic Rasch model has grown into Rasch family of models. The most commonly used include the dichotomous Rasch model (Rasch, 1960), the rating scale model (also called the polytomous Rasch model) (Andrich, 1978), equidistant model (Andrich, 1982), partial credit model (Masters, 1982), many-facet Rasch model (Linacre, 1989), and mixed Rasch model (Rost, 1990).

*Rationale of using dichotomous Rasch model*

Because VSTEP.3-5 reading test is composed of 40 multiple-choice items belonging to objective response format which scored in two strategies (1 or 0), so dichotomous Rasch model is selected for analysis in the current study.

There are many studies verify the validity of dichotomous response items (Eckes & Grotjahn, 2006; Dávid, 2007; Beglar, 2010; Pae, Greenberg, & Morris, 2012). Back to the context of the current study, some researchers use dichotomous Rasch model as a part of analysis to validate the VSTEP.3-5. For example, Nguyen Thi Phuong Thao (2018) discusses the evaluation of the relevance and the coverage of the content of a VSTEP.3-5 reading test in the article *An Investigation into The Content Validity of a Vietnamese Standardized Test of English Proficiency (VSTEP.3-5) Reading Test.* Nguyen Thi Quynh Yen (2018) discusses the cut-score validity of the VSTEP.3-5 listening test in her doctoral dissertation.

### 2.2.2 Test validation approaches from process-based perspective

A contextual and critical review of the literature focusing the two concurrent data elicitation approaches to be utilized in the current study, namely, think aloud protocols and eye tracking are enable the researcher to better evaluate the appropriateness of these two research approaches to satisfy the purpose of the current study.

#### 2.2.2.1 Think aloud protocols

There are three aspects for reviewing think aloud protocols. Rationale of using think aloud protocols (Bowles, 2010; Ericsson, & Simon, 1984; Ericsson, & Simon, 1999; Fox, Ericsson, & Simon, 2011; Green, 1998; Barkaoui, 2011), developmental stage of think aloud protocols in reading studies (Ericsson, & Simon, 1993; Someren, Barnard, & Sandberg, 1994; Bowles, 2010; Presley, & Afflerbach, 1995...) and think aloud protocols in language testing and assessment (Green, 1998; Ericsson, & Simon, 1984; Garner, 1987; Alderson, 1999b). The literature evidences that such data are helpful to strengthen the development of cognitive psychometric research for enhancing, in part, construct validity.

#### 2.2.2.2 Eye tracking

Eye tracking is the process of measuring eye movements. For the rationale of using eye tracking, many researchers state the advantages (Seagull, Wickens, & Loeb, 2001; Seagull, et al., 2001; Frenck-Mestre, 2005). For eye tracking in reading studies, reading researchers have been exploring the combination of psycholinguistic paradigms and technologies with more traditional approaches in an attempt to address central questions in reading studies (Huey, 1908; Rayner, et al, 2001; Rayner, 1998; Rayner, et al, 2012; Dussias, 2010). For the eye tracking studies in language testing and assessment, only a handful of eye movement studies have been published. From their studies, it can be concluded that through externalizing part of reading processes in the form of eye movement, researchers can explore and discover more about how reading scores are related to test takers’ actual reading behaviors with the assistance of added information from eye tracking data. In the field of large-scale assessment, eye tracking has been increasingly used in the test item design and validation, such as the studies of Paulson & Henry (2002), Tai, Loeher & Brigham (2006), Bax & Weir (2012), Bax (2013), Brunfaut & McCray (2015), Krstić, Šoškić, Ković, & Holmqvist (2018) and Bax & Chan (2019).

# CHAPTER III: THEORETICAL FRAMEWORK

This chapter starts with the brief introduction of the important concept of the current study which is validity. It then talks about five representative validation frameworks under the unitary validity view are conducive to an in-depth understanding of the advantages and disadvantages of each framework, which lays a sound foundation for the bringing up of the current research validation framework that is an interpretive argument. And then, the chapter discusses the justification for the use of the interpretive argument. Finally, it elaborates upon the articulation of an interpretive argument for VSTEP.3-5 reading test.

## 3.1 Validity

Validity is the first factor to be considered in the development, interpretation and use of language tests, because it relates to the correct and reasonable interpretation and use of test results (scores) (AERA, APA, & NCME, 1985; Bachman, 1990). For a long time, validity has been the most important indicator for judging the quality of a test and the basic starting point of language testing. The concept of validity shifts the focus from the test itself or the test score to the interpretation of the test score, which can be seen as a great advance.

## 3.2 Studies on the validity of VSTEP.3-5

There are some researchers have conducted researches on the validity of VSTEP.3-5 from different aspects, such as the validity argument, content validity, cut-score validity, comparative studies of validity, validity from scoring perspective, and validity from test-takers’ perspective. These empirical studies have made many contributions in different perspectives and have inspired the researcher on the current study. Besides, VSTEP.3-5 is a newly developed test; thus, validation is ongoing to ensure that the test is valid and reliable, and validation of a test form in this study may offer implications to the test specification and test development.

## 3.3 Validation

Validation is the process of collecting various types of evidence to ensure validity (AERA et al., 2014; Bachman, 2005, 2010; Messick, 1989). In order to solve the operational problem of test validation, different frameworks have been produced in the period of unitary validity view.

### 3.3.1 Evidence-based validation frameworks

#### 3.3.1.1 Bachman and Palmer’s test usefulness framework

Advocating the perspective of Messick’s (1989) evidence-gathering approach to validation, Bachman and Palmer (1996) develops the test usefulness framework which includes six qualities: reliability, construct validity, authenticity, interactiveness, impact and practicality.

#### 3.3.1.2 Weir’s Socio-cognitive framework

Weir (2005) proposes the “Socio-cognitive framework” which advocates that validity verification is the process of evidence collection, and the evidence to be collected is divided into five types according to the types of validity: (1) context validity; (2) theory-based validity; (3) scoring validity; (4) criterion-related validity; (5) consequential validity.

### 3.3.2 Argument-based validation frameworks

#### 3.3.2.1 Kane’s interpretive argument framework

Kane’s interpretive argument framework (2006, 2013, 2016) specifies the scoring inference that links observations as data (test takers’ performance) to observed scores (test scores), the generalization inference that links the observed scores (test scores) to universe scores (consistent scores), the extrapolation inference that links the universe scores (consistent scores) to the target scores (the predicted performance in real life) as evidence for test use and, finally, the decision-making inference that links the target score to decisions as a claim.

#### 3.3.2.2 Chapelle, Enright, and Jamieson’s interpretive argument framework

Chapelle et al. (2008) develop another interpretive argument which provides holistic, systematic, and more necessary processes for test development and validation in language testing. One strength is that having the domain description and explanation inferences, which accurately reflect language proficiency and language use task types and that evaluate the defined test constructs in a test, respectively.

#### 3.3.2.3 Bachman and Palmer’s assessment use argument framework

Bachman and Palmer (2010) develope an assessment use argument framework and believe that test development consists of two parallel processes: assessment justification and assessment production, that is, the process of test development is accompanied by the process of test validation.

Besides, this study reviews some major validity studies under Chapelle et al. (2008) interpretive argument framework, in order to provide some references and thoughts for the current study.

## 3.4 Empirical validation studies with interpretive argument framework

Several validation studies have adopted interpretive argument frameworks so far. The current study mainly reviews some major validity studies under Chapelle et al. (2008) interpretive argument framework, in order to provide some references and thoughts for the current study including Lim (2009), Chapelle, Enright, & Jamieson (2010), Gaillard (2014), Brooks & Swain (2014), Li (2015), Riazi (2016) and Sun (2016).

## 3.5 Interpretive argument for the VSTEP.3-5 reading test

### 3.5.1 Justification for the use of interpretive argument

**On the one hand**, interpretive argument gives an explicit statement of the proposed interpretation and use of test scores, which sets a concrete research agenda for test validation. **On the other hand**, interpretive argument is tolerant and can be applied to any kind of test interpretation as it does not preclude any claim or evidence. Thus, Chapelle et al.’s (2008) interpretive argument is conceived as the most appropriate framework to guide the validation process of the current study.

### 3.5.2 Articulation of an interpretive argument for the VSTEP.3-5 reading test

This study regards the generalization and explanation inferences as the main research focuses. The generalization inference involves links from observed scores to expected scores over the relevant parallel versions of tasks and test forms. It answers the first three research questions. In order to yield sufficient evidence to back these four assumptions, four different lines of research inquiry are suggested. Besides, potential rebuttals to the generalization inference of the VSTEP.3-5 reading test are also included. For the explanation inference, it is essential that the formulation of the assumptions is sufficiently informed by influential perspectives about construct validity. It answers the last two research questions. The two assumptions which are proposed to authorize the explanation inference. In order to yield sufficient evidence to back these two assumptions, two main different lines of research inquiry are suggested.

# CHAPTER IV: RESEARCH METHODOLOGY

The purpose of this chapter is to explain clearly how the research is to be operated.

## 4.1 Ontology and epistemology for the design of the study

With the deepening of the understanding of mixed methods, different classifications have emerged. Until the latest research of Creswell & Clark (2018), mixed methods is divided into three types mainly according to the different use purposes of qualitative and quantitative data: (1) Convergent design. (2) Explanatory sequential design. (3) Exploratory sequential design.

In the process of test design and development, it is necessary to adopt a variety of research methods and integrate evidence in different fields to ensure the most important quality attribute of a test-validity (AERA et al., 2014). Using the mixed research methods to study the validity is an important trend in the field of language testing. Therefore, the current study adopts convergent mixed methods design to investigate the validity of VSTEP.3-5 reading test.

### 4.2 The convergent mixed methods design for the study

The following table illustrates the overall design for the current study.

**Table 4.1**

*The convergent mixed methods design for the study*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Interpretive****argument** | Research questions | **Methods** | **Data collection**  | **Data analyses** |
| Generalizationinference | Q1 | Qualitative & quantitative  | A VSTEP.3-5 reading test paper,the VSTEP.3-5 reading test specifications, Expert interview Test takers’ scores of the VSTEP.3-5 reading test | Document-based analysisContent-based analysisThematic analysisRasch model analysis |
|  | Q2 | Quantitative  | Test takers’ scores of the VSTEP.3-5 reading test | Rasch model analysis |
| Q3 | Quantitative  | Test takers’ scores of the VSTEP.3-5 reading test | Rasch model analysis |
| Explanationinference | Q4 | Qualitative | Expert judgement Think aloud protocols  | Content-based analysisThematicanalysis |
|  | Q5 | Quantitative & Qualitative | Eye tracking Reading processing checklist  | Content-basedanalysis |

### 4.3 Participants of the study

According to the convergent mixed methods research design, two kinds of sampling techniques which were cluster sampling and purposive sampling were adapted in the current study. The brief information of participants in the study is shown in Table 4.2.

**Table 4.2**

*Sampling of participants in the study*

|  |  |  |  |
| --- | --- | --- | --- |
| **Study stage** | **Participants** | **No. of participants** | **Sampling** **techniques** |
| VSTEP.3-5 reading test | College students in different majors from two Chinese universities  | 200 | Cluster sampling |
| Expert interviewExpert judgement | College EFL teachers in Vietnam and China | 4 | Purposive sampling |
| Think aloud protocols  | College students from Chinese universities | 6 | Purposive sampling |
| Eye tracking | College students from Chinese universities  | 15 | Purposive convenient sampling |

### 4.4 Instruments of the study

#### **4.4.1 A version of the VSTEP.3-5 reading test paper**

A version of the VSTEP.3-5 reading test paper was applied as the key research instrument in the current study, which was used in all of the study stages.

#### **4.4.2 The VSTEP.3-5 reading test specifications**

The VSTEP.3-5 reading test specifications that were in confidentiality were provided by the VSTEP.3-5 development team for the research purposes.

#### **4.4.3 Expert interview questions**

Expert interview questions were developed for experts who participated in the evaluation for the test tasks of the investigated VSTEP.3-5 reading test which analyzed by the researcher. There were two sets of interview questions (A and B).

#### **4.4.4 The modified framework of task characteristics for the VSTEP.3-5 reading test**

Based on the nature of reading and characteristics of VSTEP.3-5 reading test, the current study developed the modified framework of task characteristics for analyzing the VSTEP.3-5 reading test.

**Table 4.8**

*The modified framework of task characteristics for the VSTEP.3-5 reading test*

|  |
| --- |
| 1. Characteristics of the input of VSTEP.3-5 reading test

Text lengthLanguage of input Text domainText topicText level |
| 1. Characteristics of the expected response of VSTEP.3-5 reading test

Response type  |

#### **4.4.5 Expert judgment forms**

There were two expert judgment forms (A and B) which were applied in the expert judgement process in this study. To be specific, expert judgment form A was designed to obtain the experts’ categorizations on the reading subskills needed in the reading process which were based on the VSTEP.3-5 reading test specifications and the Khalifa and Weir’s cognitive model of reading (2009). The first part gathers the experts’ basic personal information, including their name, degrees, institution, teaching age and research direction. The second part provides a table as presented in follow for experts to categorize on the reading subskills and its descriptors needed in the reading process of VSTEP.3-5 reading test.

#### **4.4.6 Think aloud protocols operation instructions**

The VSTEP.3-5 reading test think aloud protocols operation instructions were outlined. The content contained basic information of recording, basic information of participants, guidelines, training tasks, the answer sheet and supplementary questions.

#### **4.4.7 Eye tracking metrics**

Referring to the research of Brunfaut & McCray (2015), seven eye tracking metrics were comprehensively selected to analyze the eye tracking statistics and maps. The specific metrics and the corresponding technical definitions were presented.

#### **4.4.8 The reading processing checklist**

The reading processing checklist was designed according to the results of the expert judgment of reading subskills in the investigated VSTEP.3-5 reading test to complement the findings of statistical analyses of eye tracking.

### 4.5 Data collection procedures of the study

#### **4.5.1 Data collection procedures for research questions 1, 2 and 3**

For answering the first half part of research question 1, at first, a version of the VSTEP.3-5 reading test paper and the VSTEP.3-5 reading test specifications were collected from the VSTEP.3-5 development team. After that, the content of expert interview was collected from four experts, which was about the modified framework of task characteristics for VSTEP.3-5 reading test and analyses of test tasks for the VSTEP.3-5 reading test in the study. The researcher recorded all of the content of the expert interview for further analysis.

For answering the latter part of research question 1, the data of VSTEP.3-5 reading test scores were collected from the 200 test takers for VSTEP.3-5 reading test. The 200 test takers were invited to take the VSTEP.3-5 reading test at the same time in two classrooms of their own universities respectively, by traditional pen-and-paper method. Then, the researcher collected all of the 200 test papers, scored each of them, and typed the scores on the computer.

For investigating the research questions 2 and 3, the data collection was in the same procedures as the above-mentioned.

#### **4.5.2 Data collection procedures for research question 4**

For addressing the first sub-question of the research question 4, expert judgement data were collected in the following procedures. **Firstly**, the four experts respectively received the electronic editions of investigated VSTEP.3-5 reading test paper, the VSTEP.3-5 reading test specifications in English, and expert judgment form A. **Secondly**, the experts were asked to do the categorizations on the reading subskills needed in the reading process which were based on the VSTEP.3-5 reading test specifications and the Khalifa and Weir’s cognitive model of reading (2009). Then, a Zoom discussion were conducted for the consistence of the categorizations on the reading subskills needed in the reading process from the four experts. **Thirdly**, expert judgment form B was provided for each of the four experts to make identifications on the reading subskills targeted by each test item of the investigated VSTEP.3-5 reading test. Later in Zoom discussions, the identified subskills by experts were reviewed item by item and the experts were also consulted on the reasons of their choices of some of the reading subskills, which was to ensure that any disagreements were resolved before the judgment was finalized. **At last**, the final expert judgment was summarized and regarded as a standard to compare with the test takers’ actual reading processes from think aloud protocols subsequently.

#### **4.5.3 Data collection procedures for research question 5**

Eye tracking data were collected for answering the research question 5. **Firstly**, the two of the researcher’s colleagues were invited to participate in the pilot study. **Secondly**, 15 participants were invited to the eye-tracking laboratory to take part in the formal experiment. In the pre-experiment training session, the researcher first briefed the participant on the purpose of the current study, the purpose of using eye tracking method in the current study and the major procedures to follow in the formal eye tracking data collection. **Thirdly**, the participants took the VSTEP.3-5 reading test with 40 test items on a computer screen with the Eyelink Portable Duo eye tracker attached. Data Viewer 4.1.1 was employed to display the textual stimuli, collected eye tracking statistics and maps, and obtained eye tracking metrics. Immediately after completing the 10 items of each passage, the participants were required to fill out a reading processing checklist.

### 4.6 Tools for data analysis

#### **4.6.1 Compleat Lexical Tutor**

Compleat Lexical Tutor is a vocabulary profiler tool which is developed by TomCobb of Canada. The current study mainly used one of the important tools named Vocabprofile for analyzing the vocabulary of the four passages of the VSTEP.3-5 reading test. This software provided the statistical data of the inputted text based on the research from the British National Corpus (BNC) representing a vocabulary profile of K1 to K20 frequency lists.

#### **4.6.2 Readable.io**

Readable.io is an online text readability analysis software (http://www.readable.io) used to obtain the data of text difficulty, which are the Flesch Reading Ease, Flesch-Kincard Grade Level, CEFR level, readable rating and so on.

#### **4.6.3 WINSTEPS 5.4.1**

WINSTEPS 5.4.1 is a powerful, flexible and professional Rasch model software. In the current study, WINSTEPS was applied to analyze scores of 200 test takers for VSTEP.3-5 reading test.

#### **4.6.4 SPSS 26.0**

SPSS is a statistical software with in-depth analysis of data, convenient use and complete functions. In the current study, SPSS 26.0 was applied to analyze the data of eye tracking for VSTEP.3-5 reading test focusing on the Mann-Whitney U test and Kruskal Wallis H test.

### 4.7 Data analysis procedures of the study

#### **4.7.1 Data analysis procedures for research question 1**

For addressing the first half of the research question 1, an analysis was made for the test tasks of the VSTEP.3-5 reading test which comparing with the related content in the VSTEP.3-5 reading test specifications in the aspects of characteristics of the input and characteristics of the expected response and illustrating the expert interview questions for test tasks one by one.

For addressing the latter part of the research question 1, the statistics of the VSTEP.3-5 reading test items were analyzed to explore the role of each test item in the test of a group of test takers, and how appropriately they were designed in line with the difficulty level predetermined in the test specifications. Dichotomous Rasch model was applied in test items’ quality analysis of the current study, which mainly carried out according to these steps. **Firstly**, unidimensionality of the test items is checked whether the test items measure only one construct of English reading proficiency. **Secondly**, item estimates were analyzed whether the test items functioned well enough for their use according to the stipulated cutoff points for determining item fit. **Thirdly**, item and person calibration were analyzed through the Wright Map which provided a picture of test items by placing the difficulty of the test items on the same measurement scale as the ability of the test.

#### **4.7.2 Data analysis procedures for research question 2**

For exploring the answer of research question 2, DIF was detected by dichotomous Rasch model in WINSTEPS 5.4.1. The item parameters were calculated between the two groups, were put on the same scale through transformation, and then the differences between the two groups were compared. Rasch DIF contrast and Rasch-Welch t-test were used to test DIF.

#### **4.7.3 Data analysis procedures for research question 3**

In order to answer the research question 3, the current study conducts a dichotomous Rasch analysis in WINSTEPS 5.4.1 to provide helpful indexes such as separation and reliability both for items and persons, and analyzes the indexes one by one to verify whether the investigated VSTEP.3-5 reading test has a good range of reliability for a proficiency test.

#### **4.7.4 Data analysis procedures for research question 4**

For answering the research question 4, five steps were applied to analyze data. **Firstly**, the results of the expert judgement of reading subskills in the investigated STEP.3-5 reading test were summarized and analyzed item by item. **Secondly**, think aloud protocols data were transcribed. **Thirdly**, the transcripts were coded. **Fourthly**, the answers of VSTEP.3-5 reading test from the six participants were scored and the six participants were categorized into different proficiency levels accordingly. **Fifthly**, the used 13 strategies from participants’ think aloud protocols were presented in the table and analyzed one by one according to the frequency. **Lastly**, the subskills from participants' think aloud protocols were analyzed based on the main subskills that the experts believed need to be assessed for each item.

#### **4.7.5 Data analysis procedures for research question 5**

For exploring the answer of the first sub-question of research question 5, the numerical eye tracking data on different types of valid item were compared between successful test takers and unsuccessful test takers. Specifically, the Mann-Whitney U test was conducted on the quantitative eye movement data for the seven metrics of AOI. Besides, gaze plots and heat maps were examined to supply the findings that emerged from the quantitative data analyses.

For addressing the second sub-question of research question 5, several steps were employed. **At first**, the answers of VSTEP.3-5 reading test from the 15 participants of eye tracking experiment were scored and the 15 participants were categorized into different proficiency levels accordingly. **Then**, the eye movement data from each level group were conducted by using Kruskal-Wallis H test. **Finally**, frequencies of participants’ responses in the reading processing checklist were analyzed to provide evidence for investigating the extent of the VSTEP.3-5 reading test to discriminate test takers at different proficiency levels in terms of cognitive reading process.

# CHAPTER V: FINDINGS AND DISCUSSIONS ON VALIDITY OF VSTEP.3-5 READING TEST FROM THE PRODUCT-BASED PERSPECTIVE

In this chapter, the findings of data analysis for research questions 1, 2 and 3 are generated to provide backing evidence which have been proposed in the interpretive argument for VSTEP.3-5 reading test from product-based perspective. The discussions in this chapter are centered on the reflections on the study itself and similarities or differences with other relevant studies.

## 5.1 Findings and discussions for research question 1

### 5.1.1 Findings from test tasks

#### 5.1.1.1 Comparison of test tasks in the investigated test and test specifications

*Characteristics of input*

Text length, language of input, text domain and text topic of the investigated VSTEP. 3-5 reading test satisfy the requirements in the test specifications. As for text level, the results from the online tool Readable.io shows that Passage 2 (Flesch- Kincaid Grade Level: 12.1; Flesch Reading Ease: 42.3) is easier than expected with level 3, and Passage 3 (Flesch-Kincaid Grade Level: 9.3; Flesch Reading Ease: 53.1) is more difficult than expected with level 5.

*Characteristics of the expected response*

The investigated VSTEP. 3-5 reading test satisfies the requirement of the test specifications in terms of response type.

#### 5.1.1.2 Expert interview on test tasks

All of the four experts voiced their approvals for the results of the analyses in the aspects of text length, language of input, text domain and text topic. As for the aspect of text level, experts partly agreed the analysis from the online readability tool and hold different opinions which deserved further discussions.

All of the four experts support the view that the response type of the investigated test paper was in consistence with the requirement in test specifications, while some experts have more opinions on the use of the response type which merit more discussions.

### 5.1.2 Findings from Rasch model analysis of test items

#### 5.1.2.1 Item unidimensionality

The eigenvalue of the unexplained variance in 1st contrast meets the one-dimensional requirements of dichotomous Rasch model, indicating that the test items are only affected by a single factor and can be further analyzed.

#### 5.1.2.2 Item estimates according to Rasch model analysis

All 40 items meet the requirements of dichotomous Rasch model standard and function well enough for their use.

#### 5.1.2.3 Item and person calibration according to Rasch model analysis

The investigated VSTEP.3-5 reading test could measure the proficiency of the majority of this group of test takers.

Combining findings from test tasks and test items in the VSTEP.3-5 reading test above-mentioned, it can be concluded that both test tasks and test items largely meet the requirements of the test specifications, which serves as the answer of the research question 1 and supports the first two assumptions in the generalization inference.

### 5.1.3 Discussions

In the aspect of text level, combining the opinions from experts, it can be concluded that the readability tool should not be solely depended on and the results from the online tool should be just regarded as a reference for the current study.

The aspect of response type deserves more discussions, because the two experts from China mentioned that multiple-choice questions had some defects and suggested that some of multiple-choice questions can be replaced by some items which adopted limited production response and extended production response to guarantee the validity of the VSTEP.3-5 reading test. The two experts from Vietnam argued that no use other response types were because of the practicality of the scoring.

The current study supported that during the validation process of the test validity, full consideration should be given to the actual situation in the local area. Therefore, multiple-choice questions should be retained as the only existing response type of VSTEP.3-5 reading test in the context of Vietnam.

## 5.2 Findings and discussions for research question 2

## 5.2.1 Differential item functioning by genders

The difficulty of test items does not change significantly for male and female test takers. All of the 40 items do not have substantial DIF by genders.

## 5.2.2 Differential item functioning by majors

The difficulty of test items does not change significantly for humanities and social sciences and natural science test takers. All of the 40 items do not have substantial DIF by majors.

All in all, combining the findings from the above DIF analyses, it can be concluded that the test takers’ genders and majors cannot affect their VSTEP.3-5 reading scores, which provides the answer of research question 2 and supports the third assumption of generalization inference.

## 5.2.3 Discussions

For discussions for research question 2, it should be mentioned that the reasons why the study has focused on the consistency of the difficulty of the items among genders and majors. Because the 200 test takers recruited for taking the VSTEP.3-5 reading test in traditional way were college students, who shared the representative physical and experiential characteristics of genders and majors which may be the factors affect reading scores. Just as Kunnan (2010) states, only when the cross-group result of the test is stable, the decision-making based on the observed score is meaningful, reliable and fair, and the items would have higher construct validity. The findings emerged from the study that all the items in the investigated reading test do not have substantial DIF in genders and majors, which would be useful to contribute one aspect for guaranteeing the validity of the investigated VSTEP.3-5 reading test.

## 5.3 Findings and discussions for research question 3

## 5.3.1 Separation and reliability measures

From the analyses of separation and reliability, the findings show that the investigated test has a wide spread of item difficulty, and the levels of test takers are enough to confirm the item difficulty hierarchy. Thus, the VSTEP.3-5 reading test scores are adequately reliable in measuring the test takers’ English proficiency, which serves as the answer of research question 3 and provides evidence for supporting the fourth assumption in the generalization inference.

## 5.3.2 Discussions

The same with the current study, Nguyen Thi Phuong Thao (2018) has applied Rasch analysis to estimate the reliability and separation of items in one version of VSTEP.3-5 reading test. The results show that there is a very high internal consistency for the items in the reading test. Besides, Vo Ngoc Hoi (2021) draws a conclusion that high item reliability and item separation indexes indicate that the sample size is large enough to reproduce the item difficulty hierarchy, and that the items are widely spread on the measurement continuum in investigating a version of VSTEP.3-5 reading test. It can be seen that in the few studies on VSTEP.3-5 reading test validation, the reliability of the VSTEP.3-5 reading test can be ensured generally, which provides a strong support for the content validity.

# CHAPTER VI: FINDINGS AND DISCUSSIONS ON VALIDITY OF VSTEP.3-5 READING TEST FROM THE PROCESS-BASED PERSPECTIVE

In this chapter, the findings of data analysis for research questions 4 and 5 are generated to provide backing evidence which have been proposed in the interpretive argument for VSTEP.3-5 reading test from process-based perspective. The discussions in this chapter are also centered on the reflections on the study itself and similarities or differences with other relevant studies.

## 6.1 Findings and discussions for research question 4

### 6.1.1 Expert judgement on reading subskills in the investigated test

“Making inferences” is the most commonly assessed subskill, which is primarily assessed by fourteen items, followed by “understanding explicit information” (seven items) and “understanding word meanings in context” (six items). “Integrating information” is assessed by four items. “Summarizing main ideas” and “understanding explicit/implicit author’s opinion/attitude” are each assessed by three items. “Understanding cohesive devices” is assessed by two items, while “understanding text function” is targeted by one item. Thirty-five items require a combination of at least two subskills to arrive at the answer. Item 20 is judged by the experts to require test takers to engage with a maximum of five subskills. Most of the items among the 40 items require two subskills. Also, experts identify no subskills other than the proposed ones, which affirms that the set of subskills have covered the crucial processes required for the reading items of the investigated VSTEP.3-5 reading test. In short, the above analyses of results of reading subskills in the investigated VSTEP.3-5 reading test from the expert judgement can be used for arriving at the first sub-question of research question 4.

### 6.1.2 Discussions from expert judgement

**Firstly**, experts identified eight reading subskills based on the VSTEP.3-5 reading test specifications and cognitive model of reading (Khalifa and Weir, 2009). Among the eight reading subskills, only one subskill named “understanding explicit information”, was considered to be at local-level process of reading, as defined by Grabe (2009) and Khalifa and Weir (2009).

**Secondly**, based on contemporary views on the L2 reading process and empirical evidence from L2 reading assessment research (Nassaji, 2003; Rupp, Ferne & Choi, 2006), experts believed that the process of answering specific test items in the current study involved low and high levels of multiple subskills. Among the 40 test items, 35 items were judged to require at least two reading subskills. Besides, the subskill of “understanding explicit information”, which was a local-level process of reading, was judged to be involved in answering all the 40 items. This reflects the interactive perspective of L2 reading, that is, in the lower-level text-based process and the higher- level reader-based process, multiple components need to be integrated (Nassaji, 2003). The lack or ineffectiveness of local-level processes would affect the execution of global-level processes and ultimately affect the overall understanding.

### 6.1.3 Think aloud protocols

#### 6.1.3.1 Reading strategies in think aloud protocols

Thirteen strategies were identified from the think aloud protocols’ transcriptions. The detailed information about frequency of used strategies from think aloud protocols by participants at different levels including high level, middle level and low level of proficiency respectively and the frequency in total are presented.

**Table 6.3**

*Frequency of used strategies from participants’ think aloud protocols*

|  |  |  |
| --- | --- | --- |
| **Strategies** |  **Frequency** |  |
|  | High | Middle | Low  | Total |
| Reading the question (stem) | 78 |  73 |  70 |  221 |
| Reading the options | 72 |  74 |  68 |  214 |
| Emphasizing or memorizing the key word(s) in the task |  73 |  66 | 60 | 199 |
| Selecting portion(s) of the text to read |  73 |  63 |  61 |  197 |
| Rereading the relevant portion(s) |  66 |  58 |  54 |  178 |
| Rereading the options |  50 |  55 | 62 |  167 |
| Elimination |  32 |  56 |  62 |  150 |
| Selecting appropriate type of reading |  60 |  46 |  43 |  149 |
| Rereading the question (stem)  |  35 | 52 | 60 |  147 |
| Initial decision making |  20 |  28 | 29 | 77 |
| Final decision making |  18 |  28 | 30 | 76 |
| Guessing |  10 |  24 | 40 |  74 |
| Reading the instruction | 2  | 2 |  2 |  6 |

#### 6.1.3.2 Reading subskills in think aloud protocols

In addition to strategies, there were also several reading subskills which were identified from the think aloud protocols’ transcriptions. Findings in this section were presented based on the main subskills that the experts believed need to be assessed for each item, and the reading processes of the test takers at different proficiency levels when answering the item of the VSTEP.3-5 reading test. Findings shows that expert judgement and data from participants’ think aloud protocols are basically consistent in terms of primary reading subskills assessed by the test items, which answers the second sub-question of research question 4.

### 6.1.4 Discussions from think aloud protocols

**Firstly**, in contrast to the general agreement on primary readings subskills, the differences become more pronounced when it come to the potential involvement of the subskills in responding to specific items. (The degree of differences increases as participants proceed from local-level reading process to global-level reading process and with respect to different levels of proficiency. Items that primarily assess the participants’ subskills in global-level reading process such as “integrating information”, and “summarizing main ideas” induces participants to use a variety of reading subskills to infer the answers.)

**Secondly**, it is worth mentioning that participants at different levels of proficiency, especially those at two ends of the proficiency spectrum, differ widely in their patterns of subskill use. High proficiency participants appeared to use a range of subskills more in line with what experts expect, while low proficiency participants relied primarily on their understanding of individual sentences, which reflected their limited ability to simultaneously employ a range of different subskills those at the global levels of reading processing to make sense of the text.

**Thirdly**, the two test-wiseness strategies which are elimination and guessing are deserved to be discussed. Both of them were employed by majority of participants at low proficiency and some participants at middle proficiency. For elimination, most participants reported that the purpose of using this strategy was to compensate for a lack of the text understanding or to find answers quickly based on the clues in the question stem and options instead of going back to the text. For guessing, it can be categorized into two types which were guessing based on the clues in the text and blind guessing according to the participants’ think aloud protocols. However, use of these test-wiseness strategies seems to derail or interfere the reading process intended by the test designers and identified through expert judgment.

## 6.2 Findings and discussions for research question 5

### 6.2.1 Eye tracking of successful and unsuccessful test takers across item types

From the data analysis of eye tracking of successful and unsuccessful test takers across item types, it can be seen that there are differences noted between successful and unsuccessful test takers in terms of all the metrics investigated. Eye tracking data of each item type analyzed produces different measures, as for item type of understanding explicit information, there is no significant difference in terms of test takers’ eye movements, which probably due to this item type belongs to testing subskill in local-level process. Thus, it can be attributed that both successful test takers and unsuccessful test takers can employ this subskill in reading process. In some cases, like item types of understanding word meanings in context, understanding cohesive devices, making inferences, understanding explicit/implicit author’s opinion/attitude and understanding text function, there are significant differences in some of the investigated indicators. As for item types of integrating information and summarizing main ideas, there are significant differences in all of the investigated metrics. The above findings can be used to answer the first sub-question of research question 5.

### 6.2.2 Discussions from eye tracking of successful and unsuccessful test takers across item types

**Firstly**, much of the data point in the same direction, namely that successful test takers are more efficient in reading texts and tasks than unsuccessful test takers. The findings are in line with the results from many previous eye tracking studies which reveal that skilled readers show shorter fixation durations, and fewer fixations and regressions.

**Secondly**, it is more difficult for unsuccessful test takers to extract information than successful test takers; unsuccessful test takers need more time in searching for information than successful test takers; unsuccessful test takers experience a longer search process for information and have doubt or difficulty in the text reading process; unsuccessful test takers check out both test and task frequently to construct answer due to the difficulty in integrating the information contained in the text with the task in the selection of the correct answer.

By investigating cognitive processing of test takers based on eye tracking evidence, many researchers here also come to the similar conclusion that compared with unsuccessful readers, successful readers deploy the expected cognitive process on valid items, quickly locate relevant information in the text, and pay more attention to the key points like key phrases and sentences. Consistent with previous literature, the findings help to show that gaze behaviors of successful test takers differ significantly at key points from the gaze behaviors of unsuccessful students (Bax, 2013; Brunfaut & McCray 2015), which helps teachers understand how better to train test takers and help test designers know how better to assess test takers.

### 6.2.3 Eye tracking of test takers at different proficiency levels and results of reading processing checklists

There are significant statistic differences noted among high proficiency, middle proficiency and low proficiency test takers in terms of all the metrics investigated from the eye tracking data of test takers at different proficiency levels. Combined with the results of test takers’ reading processing checklists, it is clear seen from all of the reading items that they successfully lead test takers to carry out the lower-level reading processes and also reading subskills requiring the types of higher-level cognitive processing identified by experts. The findings show that item types have an impact on test takers’ reading processes, indicating the importance of selecting appropriate item types to test particular reading subskills. It is also important to note that in the reading items it is the high proficiency test takers who exhibit the target higher-level cognitive processing, meaning that the test items are effective in distinguishing them from those test takers who do not employ the relevant cognitive processes. The above findings can be used for answering the second sub-question of research question 5.

### 6.2.4 Discussions from eye tracking of test takers at different proficiency levels and results of reading processing checklists

An issue arose from the findings is test takers use of test-wiseness strategies. It is clearly shown from the information in the test takers’ reading processing checklists presented that test takers at low proficiency level tend to report the use of more test- wiseness strategies, namely guessing and elimination. All of the low proficiency test takers and some of middle proficiency test takers report using such strategies, which is in line with the findings from think aloud protocols. They are probably aware that these are often not the best strategies to use, especially on more advanced items, and may admit that the strategy they used might not be effective for the particular item. It appeared as if they were aware of the kinds of reading subskills which the more difficult items required, but that they are not always able to employ the appropriate reading cognitive processes, and instead fell back on the test-wiseness strategies. Generally speaking, combining with all of the findings above, it is inferred that the investigated reading test’s validity can be guaranteed.

# CHAPTER VII: CONCLUSION

In this chapter, interpretive argument for the VSTEP.3-5 reading test is revisited to give a brief summary of what have been investigated in the study. Implications, limitations, and suggestions for future studies are presented as follows.

## 7.1 Interpretive argument for the VSTEP.3-5 reading test revisited

Framed within an argument-based framework of language test validation (Chapelle et al., 2008), two inferences of score interpretation and use, which pertain to the most critical purposes of the current study, have been evaluated: the generalization inference and the explanation inference.

## 7.1.1 A recap of the generalization inference

From the proposed interpretive argument for the VSTEP.3-5 reading test, generalization inference is based on the warrant that observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms. Four assumptions addressing different aspects of the validity warrant have been articulated.

## 7.1.2 A recap of the explanation inference

According to the proposed interpretive argument for the VSTEP.3-5 reading test, explanation inference is based on the warrant that test takers’ scores on the VSTEP.3-5 reading test can be attributed to the construct of English reading proficiency. Three assumptions have accordingly proposed to authorize the explanation inference of the VSTEP.3-5 reading test.

## 7.1.3 Summary

Combining the findings from product-based perspective and the findings from process-based perspective under the interpretive argument proposed for the study, it is inferred that the investigated VSTEP.3-5 reading test’s validity can be guaranteed. Besides, the related discussions have provided the room for further thinking. VSTEP. 3-5 is a newly-developed test and has not yet gained wider national and international recognition. Thus, validation is really crucial for its development to gain wider recognition nationally and internationally.

## 7.2 Implications of the study

Implications of the current study can be concluded into theoretical, methodological and practical aspects.

**Theoretically**, the current study demonstrates the usefulness of interpretive argument in validating a high-stakes standardized test of English proficiency independently developed in Vietnam.

**Methodologically**, the attempt to explore test takers’ cognitive processes inVSTEP.3-5 reading test is mainly through think aloud protocols and eye tracking which illuminates some of the unobservable sides of test takers’ mental processes while they try to complete tests and demonstrates the potential wider use of this research paradigm.

**Practically**, the multiple sources of evidence from both product-based and process-based perspectives may provide some potential implications for stakeholders of the VSTEP.3-5 reading test including test takers, teachers, test designers, curriculum designers and researchers.

## 7.3 Limitations of the study

**On the one hand**, due to the confidentiality policy of test scores and the outbreak of the Covid-19, it is inaccessible for the researcher to obtain real data from the test takers in Vietnam to conducting research. **On the other hand**, Sample size of the participants in qualitative and quantitative data collection such as thinking aloud protocols and eye tracking experiment is small due to the administrative restrictions.

## 7.4 Suggestions for future studies

**Firstly**, since the test takers recruited in the current study are exclusively students in two universities, it would be meaningful and worthwhile to include more sorts of test takers such as civil servants in major and professional positions and people who work in other fields in the future studies to make more confident generalization of the findings in the current study.

**Secondly**, since the sample size of the participants in thinking aloud protocols and eye tracking experiment is not sufficiently large, it would be reasonable and convincing to increase the number of participants in future studies.

**Thirdly**, since the current study looks at just one test form that materializes the VSTEP.3-5, not the VSTEP.3-5 test as an abstraction. More validation efforts should be executed so that the test can be guaranteed validity and enjoy public trust as gate-keeping instrument.

# LIST OF PUBLICATIONS RELATED TO DISSERTATION

1. Wei, Y., & Nguyen, D. H. (2021). A validity study on the Vietnamese Standardized Test of English Proficiency (VSTEP.3-5): From test-takers’ perspectives. *Journal of China Examinations, 10,* 67-73.

2. Wei, Y. (2021). A comparative study on the validity of VSTEP.3-5 and PETS-5 reading tests. *International Graduate Research Symposium (IGRS)*,40-51.