DEVELOPMENT AND VALIDATION OF A READING PROFICIENCY SCALE FOR CHINESE AS A SECOND LANGUAGE

Jia Lin

A dissertation submitted to the faculty at the The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Education in Curriculum and Instruction in the School of Education

Chapel Hill
2020

Approved by:
Xue Lan Rong
Janice Anderson
Thurston Domina
Nianbo Dong
Helen H. Shen
ABSTRACT

Jia Lin: Development and Validation of a Reading Proficiency Scale for Chinese as a Second Language
(Under the direction of Xue Lan Rong)

This study develops and validates a *Chinese as a Second Language (CSL) Reading Proficiency Scale*. Reading ability is essential to the development of both basic communicative and cognitive academic language proficiency. The assessment of CSL reading needs further research. Test development starts with identifying the construct to measure. There is a long history of operationalizing theoretical constructs of language proficiency by developing proficiency scales. Existing reading proficiency scales are not directly applicable to CSL test development mainly due to the lack of theoretical underpinnings and/or empirical verification. Since Chinese uses a logographic and deep orthography, its unique linguistic features necessitate cognitive processes that do not exist in reading alphabetic languages. Very little is known about how to operationalize the construct of CSL reading for assessment purposes. Test developers need more guidance regarding: what component reading skills/knowledge should be measured and their relative levels of difficulty.

This research fills in these gaps by developing the *CSL Reading Proficiency Scale* through a data-based approach. The scale development began with modeling the construct of CSL reading ability on the basis of dimension analysis of reading ability as conceptualized in existing theoretical works. Second, an operational descriptive scheme with four descriptive parameters was established. Third, a descriptor pool was developed through comprehensively
documenting existing descriptors and creating new descriptors. Descriptors were rewritten to include these descriptive parameters and then validated qualitatively through expert review. Finally, fifty-seven surviving descriptors were compiled into a Likert-scale questionnaire. Ninety-five college-level CSL teachers evaluated 179 students’ reading ability against this questionnaire. To examine data collected by this questionnaire, confirmatory factor analysis was used to uncover the factor structure of the questionnaire and to classify descriptors into multiple subscales; Rasch analysis was then used to scale descriptors within each subscale separately and to screen descriptors that did not fit the model.

Based on the findings, the CSL Reading Proficiency Scale was represented as a multi-level progression scale with three subscales. The Scale specifies characteristics of each level on the reading proficiency continuum. Besides its significance in test development, the Scale will benefit curriculum design, classroom instruction, and textbook compiling.
To my beloved family.

Thank you for all of your support along the way
ACKNOWLEDGMENTS

My dissertation completion would not have been possible without the generous help and support of many people in every phase of the project. First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Xue Lan Rong, for her extraordinary mentorship and extensive support and patience through every stage of my academic development. I also would like to thank Dr. Helen H. Shen who has been providing me with insightful and priceless advice and innumerable opportunities for my professional and personal growth as a researcher and teacher. Sincere thanks also go to my committee members Dr. Nianbo Dong and Dr. Janice Anderson and Dr. Thurston Domina for their invaluable support, expertise, and insights that have greatly influenced my dissertation. I would like to extend my appreciation to Dr. Micheline Chalhoub-Deville, Dr. Wendan Li and Dr. Cathy Zimmer for their insightful comments and unfailing support. I must thank my colleagues, peers, and dear friends Dr. Lin Gu and Dr. Guanyan Chen. Their research on language testing has inspired me tremendously. I would like to thank the Carolina Asian Center and Educational Testing Service for funding this dissertation research and 95 teachers for participating in the survey study. My special gratitude goes to my family for their support and encouragement in my educational endeavors. I am truly grateful for their love. Finally, my most heartfelt love and never enough thanks go to my husband, Dr. Gengsong Gao.
# TABLE OF CONTENTS

LIST OF TABLES ....................................................................................................................... xii

LIST OF FIGURES ....................................................................................................................... xiv

CHAPTER 1 ..................................................................................................................................... 1

RESEARCH OBJECTIVE OVERVIEW ....................................................................................... 1

   Introduction ............................................................................................................................. 1

   Statement of the Problem ...................................................................................................... 2

   Purposes of the Study ............................................................................................................. 8

   The Research Questions ......................................................................................................... 9

   Methods ................................................................................................................................. 10

   Theoretical Framework .......................................................................................................... 11

      Communicative Language Ability Model ........................................................................ 11

      Component Skills Approach of Reading Research ......................................................... 12

      Data-based Approach of Scale Construction .................................................................. 13

   Significances and Contributions ............................................................................................ 14

   Summary ................................................................................................................................ 15

CHAPTER 2 .................................................................................................................................. 17
LITERATURE REVIEW .................................................................................................................. 17

Introduction .................................................................................................................................. 17

Method .......................................................................................................................................... 17

Relevant Literature ....................................................................................................................... 18

2.1 Evolution of Construct Definitions of Language Ability ...................................................... 18

2.2 Construct Definitions of Reading Ability ............................................................................... 28

2.3 Construct Definitions of CSL Reading: a Language-specific Perspective ......................... 45

2.4 Reading Proficiency Scales: Operationalization of Theoretical Constructs ....................... 63

2.5 Scale Development and Descriptor Writing ............................................................................ 78

Summary ........................................................................................................................................ 85

CHAPTER 3 .................................................................................................................................. 88

METHODOLOGY AND DATA ...................................................................................................... 88

Review of Theoretical Framework ................................................................................................. 88

Research Design and Procedures .................................................................................................. 90

The Instruments ............................................................................................................................... 95

Participants ...................................................................................................................................... 98

Teacher Participants of Qualitative Validation ........................................................................... 98

Teacher Participants of Quantitative Validation ......................................................................... 98

Learners Being Rated .................................................................................................................... 99
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Analysis Procedures</td>
<td>99</td>
</tr>
<tr>
<td>Cronbach’s Alpha</td>
<td>100</td>
</tr>
<tr>
<td>Confirmatory Factor Analysis</td>
<td>101</td>
</tr>
<tr>
<td>Rasch Analysis</td>
<td>107</td>
</tr>
<tr>
<td>CHAPTER 4</td>
<td>115</td>
</tr>
<tr>
<td>RESULTS-PREPARATORY STAGE</td>
<td>115</td>
</tr>
<tr>
<td>Modeling the CSL Reading Ability Construct</td>
<td>115</td>
</tr>
<tr>
<td>Establishing Descriptive Parameters</td>
<td>116</td>
</tr>
<tr>
<td>CSL Reading Proficiency Descriptor Pool</td>
<td>117</td>
</tr>
<tr>
<td>Qualitative Validation</td>
<td>122</td>
</tr>
<tr>
<td>Pilot Study Results</td>
<td>123</td>
</tr>
<tr>
<td>CHAPTER 5</td>
<td>124</td>
</tr>
<tr>
<td>RESULTS-QUANTITATIVE ANALYSIS</td>
<td>124</td>
</tr>
<tr>
<td>Confirmatory factor Analysis</td>
<td>124</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>125</td>
</tr>
<tr>
<td>Model Fit</td>
<td>126</td>
</tr>
<tr>
<td>Conclusions</td>
<td>133</td>
</tr>
<tr>
<td>Rasch Analysis</td>
<td>135</td>
</tr>
<tr>
<td>Rasch Analysis Results of the <em>Lower-level Questionnaire</em></td>
<td>136</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Rating Categories of the *CSL Reading Proficiency Questionnaire* .................................................96

Table 2. Basic Information about the *CSL Reading Proficiency Questionnaire* ...........................................97

Table 3. An Example Descriptor Item of the *CSL Reading Proficiency Questionnaire* ................................97

Table 4. Learners being Rated in Quantitative Validation .................................................................................99

Table 5. Interpretation of Cronbach's Alpha .................................................................................................101

Table 6. Three Competing Models ..............................................................................................................102

Table 7. Interpretation of Mean-square Fit Statistic Values .......................................................................110

Table 8. Parameters of Descriptors ...........................................................................................................117

Table 9. Sources of Descriptors .................................................................................................................119

Table 10. Examples of Revised Descriptors ...............................................................................................120

Table 11. Basic Information about the *CSL Reading Proficiency Descriptor Pool* .........................121

Table 12. Reliability Statistics .....................................................................................................................123

Table 13. Descriptive Statistics for the Observed Variables ......................................................................125

Table 14. Fit Indices for the Three Competing Models ..............................................................................127

Table 15. Individual Parameter Estimates for the Three Competing Models ..........................................132

Table 16. Individual Parameter Estimates of the Correlated Two-factor Model .......................................134

Table 17. *Lower-level Questionnaire* Learner Ability Measurement Report (First Round) ...............137

Table 18. *Lower-level Questionnaire* Learner Ability Measurement Report (Second Round) ...........139

Table 19. *Lower-level Questionnaire* Item Difficulty Measurement Report ........................................141

Table 20. *Lower-level Questionnaire* Item Difficulty Measurement Report (Without Misfitting Items) 142

Table 21. *Lower-level Questionnaire* Person-item Map .............................................................................144
Table 22. Lower-level Questionnaire Category Statistics ................................................................. 147
Table 23. Lower-level Questionnaire Band Levels ........................................................................ 149
Table 24. Intrim-level Questionnaire Learner Ability Measurement Report (First Round) ............... 151
Table 25. Intrim-level Questionnaire Learner Ability Measurement Report (Second Round) ......... 152
Table 26. Intrim-level Questionnaire Item Difficulty Measurement Report ..................................... 154
Table 27. Intrim-level Questionnaire Item Difficulty Measurement Report (Without Misfitting Items).... 154
Table 28. Intrim-level Questionnaire Person-item Map .................................................................... 156
Table 29. Intrim-level Questionnaire Category Statistics ................................................................. 159
Table 30. Intrim-level Questionnaire Band Levels ........................................................................... 160
Table 31. Higher-level Questionnaire Learner Ability Measurement Report (First Round) ............. 162
Table 32. Higher-level Questionnaire Learner Ability Measurement Report (Second Round) ......... 163
Table 33. Higher-level Questionnaire Item Difficulty Measurement Report .................................... 164
Table 34. Higher-level Questionnaire Person-item Map .................................................................. 166
Table 35. Higher-level Questionnaire Category Statistics ................................................................. 168
Table 36. Higher-level Questionnaire Band Levels ........................................................................... 170
Table 37. CSL Reading Proficiency Scale: Lower-level Decoding .................................................... 172
Table 38. CSL Reading Proficiency Scale: Intrim-level Textbase Construction .................................... 176
Table 39. CSL Reading Proficiency Scale: Higher-level Situation-model Building ............................ 180
Table 40. Alignment between the ACTFL Chinese Proficiency Guidelines and the CSL Reading Proficiency Scale ........................................................................................................................................ 194
LIST OF FIGURES

Figure 1. Evolution of Construct Definitions of Language Ability .............................................. 19
Figure 2. Hymes’ Communicative Competence Model .............................................................. 21
Figure 3. Canale & Swain’s Communicative Competence Model .............................................. 22
Figure 4. Areas of Communicative Language Ability ............................................................... 24
Figure 5. Non-reciprocal Language Use .................................................................................... 26
Figure 6. Component Skills of Reading ....................................................................................... 32
Figure 7. Bernhardt’s Constructivist Model of Reading .............................................................. 37
Figure 8. CSL Reading Construct ............................................................................................. 60
Figure 9. Lower-level Processing and Knowledge of CSL Reading .......................................... 62
Figure 10. Higher-level Processing and Knowledge of CSL Reading ..................................... 63
Figure 11. Flowchart of the Research Procedure ........................................................................ 91
Figure 12. Correlated Two-factor Model .................................................................................. 129
Figure 13. Correlated Three-factor Model ............................................................................... 130
Figure 14. One-factor Model ..................................................................................................... 131
Figure 15. Lower-level Questionnaire Response Category Probability Curves ..................... 147
Figure 16. Interim-level Questionnaire Response Category Probability Curves ..................... 159
Figure 17. Higher-level Questionnaire Response Category Probability Curves ..................... 169
CHAPTER 1
RESEARCH OBJECTIVE OVERVIEW

Introduction

Two principal continua of second language development are: basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALP) (Cummins, 1981, 1984). While listening and speaking abilities closely relate to BICS, reading is important to the development of CALP. Reading ability plays an essential role in helping students move beyond basic conversational skills and achieve higher proficiency. However, reading in Chinese as a second language (CSL) is a complex process, especially to learners with an alphabetic first language (L1) background. Since Chinese language uses a logographic, deep orthography with unique linguistic features, CSL reading necessitates reading processes and skills that are different from those of reading alphabetic languages. Research of CSL reading instruction and assessment is of great significance.

A test can influence, both positively and negatively, teaching and learning from many aspects (Alderson & Wall, 1993). Teaching and assessment should be aligned with each other to improve students’ reading proficiency. While there are a number of studies on CSL reading acquisition and instruction, existing research on CSL reading assessment is very limited in depth and scope. Areas that are especially worthy of more research are the construct definition of CSL reading and how to operationalize the theoretical construct of CSL reading for the purpose of assessment. Test development starts with identifying test purposes and the constructs to measure. “Construct” refers to “the concept or characteristic that a test is designed to measure” (AERA, APA, NCME, 2014, p. 11). However, while there are some theoretical studies that explore the nature and components of CSL reading competence, we
know very little about how to operationalize CSL reading construct for assessment purposes, i.e., what reading skills should be measured for each proficiency level.

There is a long history of operationalizing theoretical constructs of language proficiency by developing proficiency scales (Alderson, 2000). However, majority of existing reading proficiency scales cannot be used directly for CSL test development mainly because: (1) some existing scales lack theoretical underpinnings and/or empirical verification; (2) proficiency scales with cross-language applicability do not address individual languages. As a result, their descriptors are excessively broad for the assessment of individual languages.

The current study aims to fill in these gaps by developing a CSL-specific reading proficiency scale, which is both theoretically solid and empirically verified, through a data-driven approach. The scale development started with modeling the construct of CSL reading ability on the basis of dimension analysis of reading ability as conceptualized in existing theoretical works. After an operational descriptive scheme with four descriptive parameters were established, the author comprehensively documented reading proficiency descriptors from theoretical works and existing proficiency scales and re wrote them into can-do statements with four predetermined descriptive parameters. The bank of descriptors was then validated qualitatively through expert review. The last step was quantitative validation. Descriptors surviving from the qualitative validation were compiled into a data collection instrument, the CSL Reading Proficiency Questionnaire. On the basis of a large-scale application of this instrument, dimensions of CSL reading ability as measured by this instrument were identified by using confirmatory factor analysis. Lastly, descriptors were calibrated by Rasch rating scale analysis. Based on data analysis results, the researcher established a multiple-level progression scale, named the CSL Reading Proficiency Scale.

Statement of the Problem

Basic interpersonal communicative skills (BICS) and cognitive academic language
proficiency (CALP) are two principal continua of second language development (Cummins, 1981, 1984). The first two years of college-level CSL courses focus more on the training of BICS. BICS develop earlier and with more ease, compared to CALP. However, CALP represents a more advanced level of language proficiency.

*American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines* (ACTFL, 2012), which plays a guiding role in foreign language instruction and assessment in the U.S., describes five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. Advanced-level foreign language learners should be able to narrate and describe with rich details and to deal with unexpected complications. They should be able to not only address topics of personal interest but also talk about topics of general interest. Superior-level learners should be able to support opinions, hypothesize, and discuss topics extensively in most formal settings. In addition to a wider range of general interest topics, they are able to talk about special fields of interest and expertise (ACTFL, 2012). Therefore, CALP is an essential proficiency component of advanced and superior-level foreign language learners.

While listening and speaking abilities closely relate to BICS, reading ability is more critical to the development of CALP. Reading is the main source of abstract vocabulary, formal expressions, and content knowledge about general interest topics, expertise, and specialized topics. Reading plays an important role in helping students move beyond basic conversational skills and achieve higher proficiency. To CSL educators and researchers, the teaching and testing of reading ability are worthy of more study.

However, research on CSL reading assessment is very scant. One fundamental issue that should be prioritized is the construct definition of CSL reading and how to operationalize it for assessment purposes. While there are a number of theoretical works that explore the nature and components of first and second language (L1 and L2) reading competence, studies
focusing on CSL reading ability are very limited. In addition, very little research has been
done to investigate how to operationalize CSL reading construct for the purpose of
assessment.

A language proficiency scale/framework/guideline is organized as a hierarchy of skill
levels and descriptors specifying characteristics of each level on the proficiency continuum.
Language proficiency scales have been used to operationalize theoretical constructs, i.e., to
describe language ability at various proficiency levels; to guide test development; and to
assist raters in judging performance (Alderson, 2000). For instance, The ACTFL Proficiency
Guidelines (ACTFL, 2012) and its European counterpart, the Common European Framework
of Reference for Languages: Learning, Teaching, Assessment (CEFR) (Council of Europe,
2001), have been playing guiding roles in foreign language teaching and testing. Other
widely used foreign language proficiency scales include: Canadian Language Benchmarks,
World-class Instructional Design and Assessment (WIDA) English Language Development
Standards, The American Council for the Teaching of Foreign Languages (ACTFL)
Proficiency Guidelines, and Interagency Language Roundtable (ILR) Language Skill Level
Descriptions.

However, most existing reading proficiency scales are not directly applicable to CSL
test development because:

(1) The majority of existing reading scales are not language-specific and are
insufficiently precise for the assessment of individual languages, especially in assessing
particular linguistic features and their associated cognitive processes. Reading a particular
writing system requires a command of the cognitive mechanisms specifically developed for
dealing with its linguistic properties (Koda, 2005). Chinese writing system adopts a
logographic and deep orthography. Comprehension of unique linguistic features of the
Chinese writing system necessitates unique reading processes and skills, which do not exist in
reading alphabetic languages.

Chinese-specific proficiency scales are very scant. One is the *ACTFL Chinese Proficiency Guidelines*. Although ACTFL published this Chinese-specific scale in 1980s, this document has not been updated after being established and is not available on the ACTFL website. Therefore, it is not well known to CSL teachers and testers. More limitations of this scale are discussed in Chapter 3. Another Chinese-specific proficiency scale is *Mandarin Chinese: Four-Year Instructional Goals, Curriculum Outline, and Instructional Measures*. This document only includes a very small number of descriptors for each language skill. Both scales have some of following problems which also exist in their non-language-specific counterparts.

(2) Some scales (e.g., the *ACTFL Proficiency Guidelines, the ACTFL Chinese Proficiency Guidelines*) are produced mainly by intuitive identification of key features at diffident proficiency levels (Alderson, 2000; North & Schneider, 1998). The lack of empirical verification in scale construction has led to controversies over their validity claims (Chalhoub-Deville & Fulcher, 2003; Liskin-Gasparro, 1996; Malone, 2003).

(3) Some existing scales (e.g., the *CEFR*) were not based on large scale SLA studies and lack a clear linguistic theory (Chalhoub-Deville, 2009; Fulcher, 2004). For instance, while the *CEFR* is better anchored in empirical studies, its theoretical underpinnings are still questionable (Fulcher, 2003).

(4) Most existing scales specify language ability from a text-based (i.e., the types of texts readers at each level can comprehend) or task-based approach (i.e., tasks that readers at each level can accomplish), but not from a cognitive processing perspective. These existing scales have very few descriptors reflect the cognitive processes underlying reading comprehension (e.g., syntactic parsing, semantic-proposition encoding, and reading strategies). However, many of the theoretical works on reading are process-based which
analyze reading ability and its component skills from a psycholinguistic perspective. The lack of process-based descriptors leads to a discrepancy between existing reading proficiency scales and the latest theoretical research findings. With regard to application, proficiency scales, which lack process-based descriptors, have huge limitations in guiding the instruction and assessment of the actual cognitive processes involved in reading comprehension.

(5) Lastly, the majority of existing scales are holistic in nature. Analytic rating scales, which are more user-friendly, are in need. According to North (1993), 90% of language proficiency scales are holistic in nature. For instance, in the ACTFL Proficiency Guidelines, descriptors for a given level are presented as loosely connected sentences or phrases. It is difficult for users to figure out the relationship among descriptors or to internalize assessment criteria. To present the underlying assessment criteria explicitly, analytic rating scales are in need, in which descriptors are classified into distinct categories (e.g., grammar knowledge, character recognition, reading strategy).

Due to the above-mentioned limitations, existing reading proficiency scales are not directly applicable to the development of CSL reading assessments. As a result, test developers often determine what component reading skills should be measured based on textbook content, experience, or intuition. This is very challenging for item writers who are not familiar with the construct of CSL reading and equally challenging for teachers with little expertise in test development.

In CSL classes, especially in college-level CSL classes, students’ achievement is measured mainly by teacher-developed in-house assessments. There are no item banks developed by professional testing companies available for teachers to use. CSL teachers need to write their own assessment tools for unit tests, mid-terms, and final exams. Some are summative assessments that are used solely for grading purposes. Others are interim assessments, which serve instructional purposes and provide information based on which
teachers can improve their instruction. Traditionally, teacher-developed assessments are regarded as low-stakes and have attracted little research effort. However, grades based on teacher-developed assessments determine students’ placement, GPA, and scholarship applications, graduate school applications, and future employment. In this sense, teacher-developed assessments are not low-stakes.

Previous studies (e.g., Alderson, 2005; Stiggins, 2002) have called into question foreign language teachers’ assessment literacy (i.e., knowledge and expertise in assessment) and the quality of teacher-developed tests. Many teachers are not prepared for the difficulties associated with classroom assessment because of their lack of assessment literacy (Stiggins, 2002). A large proportion of teacher-developed tests are not of high quality thus cannot offer much information in terms of achievement, progress, or diagnostics (Alderson, 2005). Vogt & Tsagari (2014) have argued that foreign language teachers’ assessment literacy is limited in terms of grading, placing, and certifying. They have also reported a lack of knowledge in assessing skills (especially integrative assessment), reliability, validity, and use of statistics. In addition, foreign language teachers lack the knowledge and skills to evaluate self-developed assessment tools. Teachers do not receive sufficient training on assessment literacy in pre-service programs. Instead, they learn on the job from colleagues. Teachers’ insufficient assessment literacy may impact the quality of teacher-developed assessments negatively.

In sum, research efforts are needed to operationalize theoretical constructs of CSL reading for assessment purposes by developing a reading proficiency scale. Problems and limitations in current CSL reading assessment research and practice make this study very significant. On one hand, existing language proficiency scales are not directly applicable to CSL reading test development mainly due to the lack of theoretical groundings and empirical verification. On the other hand, the fact that foreign language teachers often lack assessment literacy makes the development of a CSL reading proficiency scale an urgent concern.
Purposes of the Study

This study aims to operationalize the CSL reading construct for assessment purposes by developing a language scale, named the *CSL Reading Proficiency Scale*. The scale development and validation procedures were carefully designed in order to address the limitations of existing proficiency scales. First of all, theoretical works on L1 reading, L2 reading, and CSL reading, especially the Communicative Language Ability (CLA) Model (Bachman, 1990; Bachman & Palmer, 1996, 2010), were widely consulted to model CSL reading processes and knowledge. These theoretical works served as theoretical underpinnings of the *CSL Reading Proficiency Scale*. In addition, to adequately represent linguistic properties and cognitive mechanisms unique to Chinese language, seminal works on Chinese orthography, morphology, syntax, and discourse patterns were referred to in modeling the construct of CSL reading. Second, a data-based approach was adopted to develop and validate this *CSL Reading Proficiency Scale* empirically. Both qualitative and quantitative analysis techniques were utilized for the purpose of validation. Lastly, CSL teachers were involved in both qualitative and quantitative validation. This procedure enhanced the possibility that the *CSL Reading Proficiency Scale* is relevant and useful to test development and classroom teaching.

Before discussing the nature and primary application of the *CSL Reading Proficiency Scale*, it is necessary to review briefly two taxonomies of language proficiency scales, which will be discussed in detail in Chapter 2. First, according to the purposes of scale development, language scales can be classified into four types: user-oriented, constructor-oriented, assessor-oriented, and diagnosis-oriented scales (Alderson, 1991; Pollitt & Murray, 1993). First and foremost, the *CSL Reading Proficiency Scale* is constructor-oriented in nature. By referring to the *CSL Reading Proficiency Scale*, test developers can know what reading skills should be tested for learners at certain proficiency levels. For the development of both large-
scale standardized tests and teacher-developed in-house assessments, the CSL Reading Proficiency Scale provides empirical evidence in terms of operational construct and item difficulty. The CSL Reading Proficiency Scale can also play a guiding role in textbook writing and classroom teaching. Secondarily, the CSL Reading Proficiency Scale can be used by assessors or teachers as a rubric for evaluation or diagnostic purposes. Lastly, this scale can be used by non-specialist users (e.g., CSL learners) as a self-assessment checklist of reading proficiency.

Second, according to content and structure, language proficiency scale can be classified into five categories: brief, holistic scales of reporting overall proficiency; user scales reporting proficiency in different contexts of use; detailed, holistic rating scales; detailed, analytic rating scales; and, frameworks of syllabus content and assessment criteria for stages of attainment (North, 1993). The CSL Reading Proficiency Scale is an analytic rating scale, which describes reading proficiency from multiple aspects, including syntax, vocabulary, and higher-level processing. Because this scale is developed on the basis of teachers’ collective perception of CSL learners’ average reading proficiency and also because it provides a list of tasks/vocabulary/syntactic structures for each class level, it can be regarded as a detailed content or outcome specifications for stages of attainment.

The Research Questions

The current study adopted a data-driven approach of scale development to establish a CSL Reading Proficiency Scale. Language proficiency scales are commonly presented as descriptor tables with two axes. The horizontal axis indicates dimensions of the construct (e.g., the factor structure of reading ability) and the vertical axis denotes descriptor difficulty levels. Multiple descriptors are provided to pinpoint linguistic features and cognitive abilities at a certain proficiency level and within a given dimension.

In developing the CSL Reading Proficiency Questionnaire, after establishing a bank
of descriptors, the researcher needed to classify descriptors according to the factor structure of CSL reading ability and to scale descriptors according to their difficulty estimates. Therefore, data generated from the large-scale application of the *CSL Reading Proficiency Questionnaire*, which can be regarded as a draft of the *CSL Reading Proficiency Scale*, were analyzed statistically in order to: (1) figure out the factor structure of CSL reading as measured by descriptors in the *CSL Reading Proficiency Scale*; (2) estimate the relative ‘difficulty values’ of descriptors and cut them into multiple bands. Accordingly, following three research questions were answered:

**Dimensions of descriptors:** 1. What is the factor structure of CSL reading, as measured by descriptors in the *CSL Reading Proficiency Scale*?

**Difficulty levels of descriptors:** 2. What is the estimated level of difficulty of each descriptor? 3. How are cut-off points selected to divide descriptors into multiple bands?

**Methods**

The scale development procedures, designed on the basis of the methodology used to develop the *CEFR*, consisted of six steps. The researcher first explored the construct of CSL reading ability by reviewing theoretical works and empirical studies. Second, an operational descriptive scheme for describing reading ability with four parameters was established. The four predetermined descriptive parameters were: cognitive verb, object (of the verb), text type, and modifier. A descriptor pool was then constructed by comprehensively documenting proficiency descriptors from existing scales and by authoring new descriptors based on theoretical works. Descriptors were edited into can-do statements to include predetermined descriptive parameters. The descriptor pool was validated qualitatively through consultation with experienced CSL teachers. This qualitative validation procedure enhanced the possibility that descriptors reflected teachers’ own constructs of language proficiency and
also enhanced the possibility that descriptors were relevant and useful to teaching.

Descriptors surviving from the qualitative validation procedures were compiled into a five-point Likert-scale Questionnaire, the *CSL Reading Proficiency Questionnaire*. Ninety-five CSL teachers used this questionnaire to evaluate the reading proficiency of 179 students in total. Data generated from the large-scale application of the *CSL Reading Proficiency Questionnaire* were subjected to statistical analyses (i.e., confirmatory factor analysis and Rasch analysis) to identify the factor structure of CSL reading ability and to scale descriptors. Descriptors that did not fit the construct being measured were screened out and eliminated.

Finally, 55 surviving descriptors were compiled into a language proficiency progression scale with three subscales, named the *CSL Reading Proficiency Scale: Lower-level Decoding; CSL Reading Proficiency Scale: Interim-level Textbase Construction; and CSL Reading Proficiency Scale: Higher-level Situation-model Building*. Please refer to Chapter 3 for details on procedures, participants, instruments, data collection and analysis.

**Theoretical Framework**

**Communicative Language Ability Model**

A variety of language ability models have been suggested to articulate key components of the L2 construct. The Communicative Language Ability (CLA) model (Bachman, 1990; Bachman & Palmer, 1996, 2010) represents the state of knowledge in the L2 testing field and has been widely recognized around the world (Alderson & Banerjee, 2002). The current study conceptualizes the CSL reading ability construct in light of the CLA model.

In the CLA model (Bachman, 1990; Bachman & Palmer, 1996, 2010), L2 ability is operationalized in terms of two principal components: language knowledge and strategic competence. Language knowledge represents information stored in a language user’s memory, which is called upon to produce and/or comprehend written and/or oral
communication. Basically, language knowledge consists of organizational knowledge and pragmatic knowledge. Organizational knowledge refers to the integration of features such as grammatical, vocabulary, and textual aspects of language. Pragmatic knowledge denotes how organizational features come together to realize specific communication goals in a given language use situation with particular cultural allusions, varying registers and formalities. Strategic competence represents primarily a set of metacognitive strategies in working memory. General areas of strategic competence include goal setting, appraising, and planning (Bachman & Palmer, 1996, 2010). Language users adopt metacognitive strategies to plan, monitor, modify, regulate, and execute their cognitive processes. These metacognitive processes are relied upon to activate relevant language knowledge features and connect them with pertinent topical or content knowledge, personal preferences and characteristics, affective schemata, and cognitive strategies activated to attend to specific tasks (Bachman & Palmer, 2010). These language/language use features interact with each other in complex ways when individuals process or produce speech/texts in a given context.

Component Skills Approach of Reading Research

Various reading models have been proposed to provide specific descriptions of how reading comprehension works and what reading involves. Two major types of reading models are componential models and process models. The componential models investigate reading comprehension from a component skills approach which regards reading as the product of a complex information-processing system, involving a constellation of closely related mental operations. Component skills approach attempts to explore: the set of mental operations; the organization of the operations and the pattern of information flow among them; the means by which the system of operations is controlled and coordinated, including the stimulus conditions, strategies, and capacity demands associated with effective performance; and, parameters of the system whose variation is responsible for individual and developmental
differences in the system’s overall effectiveness and efficiency (Carr et al., 1990). As pointed out by Bachman and Cohen (1998), the multi-componential view of language ability provides a very productive theoretical basis for language testing research and practice. Due to its multiple-skill orientation, the component skills approach is widely used in assessing L2 reading development (Koda, 2005).

This study adopts the component skills approach to explore the operational construct of the CSL reading ability. Specifically, the researcher models CSL reading ability as the integration of multiple processes and knowledge. Then the researcher collected or created descriptors to represent component skills of CSL reading. Descriptors included in the CSL Reading Proficiency Scale were classified into provisional groups arrived at through literature review.

**Data-based Approach of Scale Construction**

Traditionally, scales of language proficiency were and are still mostly produced by intuitive identification of key features at different proficiency levels. For example, the ACTFL Proficiency Guidelines were developed in this approach (Alderson, 2000; North & Schneider, 1998). The lack of empirical verification in scale construction has led to controversy over the validity claims of the ACTFL Proficiency Guidelines (Chalhoub-Deville & Fulcher, 2003; Liskin-Gasparro, 1996; Malone, 2003). These concerns about the intuitive approach have led to data-based solutions to the development of proficiency scales.

The CEFR was constructed through a data-driven approach and claimed to be the best-researched scale of language proficiency (North, 2014). The scale construction started with a comprehensive documentation of over 1000 descriptors used in existing language proficiency scales. Descriptors were assigned into provisional groups. The descriptor pool was then validated qualitatively through widespread consultation with experienced foreign language teachers. Finally, descriptors were scaled through Rasch rating scale analysis
This data-based approach used to develop the CEFR is widely consulted by proficiency scale construction projects all around the world (He & Chen, 2017). The current study also adopted this approach of scale construction methodology. Please refer to Chapter 3 for more details about the scale development methodology.

Significances and Contributions

This study identified theoretical construct definitions of CSL reading scattered in literature and operationalized theoretical construct for assessment purposes through developing the CSL Reading Proficiency Scale. It verified the factor structure of CSL reading ability and provided empirical evidence for the componential models of reading ability. In addition, it answered fundamental questions in CSL reading test development, i.e., what component reading skills should be measured and the relative levels of difficulty of these component skills. This study addressed the lack of a CSL-specific reading proficiency scale, which is theoretically solid, empirically verified, and directly applicable to CSL test development.

The primary application of the CSL Reading Proficiency Scale is the development of teacher-developed CSL reading assessments in the context of U.S. college-level CSL classes. By referring to this scale, teachers can be aware of what skills/knowledge should be measured for learners at each level. This proficiency scale is also applicable to the development of large-scale standardized CSL tests. In addition, the CSL Reading Proficiency Scale can guide curriculum setting and textbook development. By consulting the CSL Reading Proficiency Scale, textbook writers can be aware of the characteristics of content appropriate for each level of CSL courses. The CSL Reading Proficiency Scale also informs curriculum designers of empirical evidence, based on which they can set reasonable teaching objectives.
Instead of using secondary data, the researcher collected primary data among experienced college-level CSL teachers across the country. This set of data well present CSL teachers’ collective perception of students’ reading proficiency, reading ability development trajectory, and difficulties of component skills. This set of primary data can also benefit future research related to CSL reading.

As an interdisciplinary study grounded on the interface of language acquisition and language testing research, the current study contributes to the cross-disciplinary exchange in terms of theoretical groundings and research methodology. On one hand, the data-based scale construction methodology origins in language testing literature. The current study introduces this methodology to scholars in the field of foreign language education and second language acquisition. It will benefit teaching, testing, and research of CSL as well as other foreign languages. On the other hand, theoretical works drawn from first and second language acquisition research underlie the modeling of CSL reading construct. Existing proficiency scales, including the widely used CEFR scale and the ACTFL Proficiency Guidelines, are criticized for lacking theoretical underpinnings (Chalhoub-Deville, 2009; Fulcher, 2003). In the present study, the theoretical construct of CSL reading is grounded in seminal works published in the field of language acquisition research (e.g., Grabe, 2009; Koda, 2005). It bridges the gap between second language acquisition theories and scale development practices. More detailed discussion on the unique contributions of this dissertation is provided in Chapter 6.

**Summary**

Research of CSL reading assessment is of great significance. More studies are needed to investigate the construct definition of CSL reading and how to operationalize theoretical CSL reading construct for the purpose of assessment. Very little is known about what component CSL reading skills should be measured for learners at each level on the
proficiency continuum. Although there are many reading proficiency scales available, they are not directly applicable to CSL testing due to a variety of problems. This study was designed to fill in this gap by developing a CSL Reading Proficiency Scale using a data-driven approach. While the primary application of this scale is to guide the development of teacher-developed in-house assessment in the context of U.S. college-level CSL teaching, it is also applicable to large-scale test development, textbook writing, and classroom instruction.
CHAPTER 2
LITERATURE REVIEW

Introduction

The literature review provides an overview of: (1) the evolution of construct definitions of language ability, with a focus on the Communicative Language Ability model (Bachman, 1990; Bachman & Palmer, 1996, 2010); (2) construct definitions of L1, L2, and CSL reading; (3) limitations of existing reading proficiency scales; and (4) methodologies for developing reading proficiency scales.

Method

Due to the interdisciplinary nature of the current study, the literature review includes relevant works in both language testing field and second language acquisition arena. The literature review is organized by the following primary themes: (1) Section 2.1: evolution of construct definitions of language ability; (2) Section 2.2 and 2.3: construct definitions of reading, L2 reading, and CSL reading; (3) Section 2.4: how L2 reading construct is operationalized in existing reading proficiency scales; and, (4) Section 2.5: methodologies for developing foreign language proficiency scales.

Section 2.1 reviews the evolution of construct definitions of language ability, which can be divided into three stages: precommunicative era, communicative competence representation, and sociocognitive argument of language ability. This section focuses on the Communicative Language Ability (CLA) model (Bachman, 1990; Bachman & Palmer, 1996, 2010), which continues to represent the state of knowledge in the L2 testing field. For section 2.2 in relation to the construct of reading, the review begins with theoretical works that discuss the nature and component skills of reading. This section continues with studies that
explore the construct of L2 reading from a cross-linguistic approach. Section 2.3, which addresses CSL reading in particular, reviews empirical studies and theoretical works that reveal the uniqueness of CSL reading in terms of linguistic features and corresponding cognitive processes. Based on Section 2.1, 2.2, and 2.3, modeling of CSL reading ability construct is completed, which serves as the theoretical ground for the development of the CSL Reading Proficiency Scale.

Section 2.4 reviews how the theoretical construct of L2 reading is operationalized for assessment purposes by composing proficiency scales or test specifications. These scales or test specifications operationalize reading construct mainly from the task perspective or the reader purpose perspective. Limitations of existing proficiency scales, in terms of theoretical groundings, cross-language applicability, and construction methodology, etc., are pointed out.

Second 2.5 is a review of methodologies for language proficiency scale development. It focuses on the application of Rasch analysis in data-driven scale construction approaches. It also reviews “can-do” style language proficiency descriptors in terms of their components and qualities.

Relevant Literature

2.1 Evolution of Construct Definitions of Language Ability

“Construct” refers to “the concept or characteristic that a test is designed to measure” (AERA, APA, NCME, 2014, p. 11). Test development starts with identifying test purposes and identifying the construct to measure. Defining the construct of language ability is the basis for the language test development and score interpretation.

While most researchers support the multidimensionality of language ability, i.e., language ability consists of distinct components, there is still no agreement on what these underlying components are and how they interact (Kunnan, 1998). Therefore, a variety of models have been proposed to represent language ability. Construct definitions of language
ability evolve with the advance in linguistics, psycholinguistics, and sociocognitive science. As illustrated in the following figure, the evolution of language ability constructs consists of three major stages: precommunicative era, communicative language competence representation, and sociocognitive argument of language ability. This section reviews the key frameworks proposed on each stage, focusing on Communicative Language Ability model (Bachman, 1990; Bachman & Palmer, 1996, 2010), which represents the mainstream understanding of language ability in current test development practice.

Figure 1.

Evolution of Construct Definitions of Language Ability

2.1.1 Precommunicative Era

In precommunicative era, language ability is modeled as an innate aptitude within individual language users. Language ability was addressed from a pure cognitive perspective. As the founder of structural linguistics, Saussure has a great impact in the precommunicative representation of language ability. Saussure (1959) differentiates between la langue (i.e., a system of signs) and la parole (i.e., the realization of the system in a particular situation). In structural linguistic view, language is regarded as a rule-governed
system, consisted of discrete segments: phonemes, lexemes and morphemes.

Chomsky (1965) takes over the distinction between linguistic competence and linguistic performance, which parallels Saussure’s *la langue* and *la parole* respectively. Linguistic competence refers to an ideal speaker’s innate knowledge of grammatical structures while linguistic performance is the actual use of language in particular situations. According to Chomsky (1965)’s Universal Grammar Hypothesis, a set of grammatical rules is pre-programmed in the human brain. In this sense, human language ability is an innate aptitude stored within individual language learners. Chomsky (1965) argues that the fundamental goal of linguistic study is to explain an ideal language user’s tacit knowledge of the grammar structures (i.e., linguistic competence) of his/her native language in a completely homogenous language use community. However, linguistic performance is peripheral and secondary because it is full of error in the actual language use.

Theoretical linguistic research findings have been driving the investigation of language ability constructs in language testing field. In the first textbook of language assessment, Lado (1961) models the skills or elements underlying test-takers’ performance within a structuralist approach. Language was conceptualized as consisting of independent elements, including pronunciation, morphology, and syntax. His perception of what should be measured aligns with Chomsky’s competence side. In addition, Lado (1961) advocated discrete-point testing which measure independent language elements, instead of integrative assessment.

**2.1.2 Communicative Language Competence**

Chomsky’s linguistic theory has been criticized by communicative language competence proponents. Although communicative language competence proponents still address language ability from a cognitive, psycholinguistic perspective, they take into consideration the role of pragmatic factors. As a pioneer in proposing communicative
competence, Hymes (1972) expands on Chomsky’s representation of linguistic competence by adding the dimensions of ability for use and sociolinguistic competence. As illustrated in the following figure, Hymes (1972) conceptualized a model of communicative competence which consists of both tacit knowledge and ability for use.

![Figure 2: Hymes’ Communicative Competence Model](image)

Hymes’ theory of communicative competence lays the foundation for subsequent development of communicative competences models and benefits the fields of second language acquisition, teaching, and testing. Canale & Swain (1980) and Canale (1983) advance Hymes’ communicative competence model by adding strategic competence and discourse competence dimensions respectively. Canale’s (1983) communicative competence model has four major components: grammatical competence, sociolinguistic competence, strategic competence, and discourse competence. Figure 3 shows details of each competence components.
Canale & Swain’s Communicative Competence Model

As pioneering works on communicative competence, Hymes’ (1972) model, Canale and Swain’s (1980) model, and Canale’s (1983) model have extended Chomsky’s representation of linguistic competence and have laid the theoretical framework for the teaching and assessment of communicative language ability. However, these models do not provide a depiction of how various components of communicative competence interact with each other and with the context of language use. Bachman (1990) and Bachman and Palmer (1996, 2010) formulate the Communicative Language Ability (CLA) model on the basis of prior works (e.g., Canale, 1983) while addressing their limitations.

In the CLA model (Bachman, 1990; Bachman & Palmer, 1996, 2010), language ability is operationalized in terms of two principal components: language knowledge and strategic competence. Figure 4 presents an adaptation of the CLA areas of knowledge and competence. Language knowledge refers to information stored in a language user’s memory, which is called upon to produce and/or comprehend written and/or oral communication. Basically, language knowledge comprises organizational knowledge and pragmatic...
knowledge: (1) Organizational knowledge references the integration of features such as grammatical, vocabulary, and textual aspects of language. (2) Pragmatic knowledge denotes how organizational features come together to realize specific communication goals in given language use context with particular cultural allusions, varying registers, and formalities.
Figure 4.

Areas of Communicative Language Ability

Strategic competence is the other principal component of language ability in CLA. Learning strategies are defined as “specific action taken by the leaner to make learning easier, faster, more enjoyable, more self-directed, more effective, and more transferable to new situations” (Oxford, 1990, p. 8). Previous studies (Block, 1986; He, 2008; Lee-Thompson, 2008; Mokhtari & Shoerey, 2002; O’Malley & Chamot, 1990; Oxford, 1990; Wenden, 1991) identified strategies or learning strategies used by language learners and also generated taxonomies of strategies from different perspectives. For instance, O’Malley & Chamot (1990) classified language learning strategies into three categories: cognitive strategies (learners’ using of mental processes to construct meaning from the text), metacognitive strategies (learners’ knowledge on how to monitor, modify, regulate, and execute their cognitive reading processes), and social/affective strategies (strategies involving the interaction between readers and other persons while affective strategies refer to strategies that involve the readers’ control over affect). Oxford (1990) added another two categories: memory strategies and compensation strategies. Memory strategies help learners to store and retrieve new information while compensation strategies are used to make up for inadequate repertoire of grammar and vocabulary. In CLA model (Bachman & Palmer, 1996, 2010), strategic competence is primarily conceived as a set of metacognitive strategies, involving goal setting, appraising, and planning.

Figure 5 presents the interaction between language ability and attributes that are not part of language ability, specifically personal attributes, topical knowledge, cognitive competence, and affective schemata. Metacognitive processes (referred to as cognitive strategies) are relied upon to activate relevant language knowledge, and connect them with pertinent topical or content knowledge, personal learning styles and characteristics, affective schemata (i.e., facilitative/inhibitive feelings), and
cognitive strategies activated to attend to specific tasks (Bachman & Palmer, 2010). These components interact with each other in complex ways when individuals process or produce speech/text in given situations. While metacognitive strategies are viewed as a focal part of language ability, cognitive strategies are perceived as language users’ peripheral attributes.

![Attributes of the Individual](image)

**Figure 5**

*Non-reciprocal Language Use*


While scholars in the field of language testing tend to support the notion that foreign language (FL) and L2 constructs are multidimensional in nature, i.e., include distinct components, there is no agreement on number and nature of underlying components are and how they interact with each other (Kunnan, 1998). A variety of language ability models have been proposed to articulate key components of the FL/L2 construct. Among these models, CLA continues to represent the state of knowledge in the L2 testing field and has been widely recognized around the world.
2.1.3 Sociocognitive View

The within-language user construct is increasingly challenged by a sociocognitive construct argument that regards the interaction among multiple language users and the interaction between language users and social context as the primary determinant of language performance (Halliday & Hasan, 1985). By using Bachman and Palmer (1996)’s model as the starting point, Chalhoub-Deville (2003) advocates ‘an ability-in language user-in context’ view of language ability construct. Grounded in Vygotsky (1978)’s work, this social interactional perspective highlights two notions: First, language ability is local and situated. Traditionally, test taker and task, have been (and still are) considered to be separate. The interaction of test taker and task has often been dismissed as error. However, empirical evidence reveals that while some ability features remain stable across tasks, other ability features are not (Chalhoub-Deville, 1995a, 1995b). Second, multiple participants of a communicative event interact with each other and co-construct a discourse. One participant’s performance is greatly shaped by other participants.

Researchers with a sociocognitive view advocate that the language ability construct examination should shift its focus from response consistency to inconsistent performance across contexts (Chalhoub-Deville, 2003; McNamara, 2003). They call for a move away from a generic and cognitive representation to local theories of language ability that elaborate ability features unique to specific contexts in future scholarship.

2.1.4 Summary

As compared with other language ability construct definitions, the CLA model represents the mainstream understanding and has been widely recognized in the field
of language testing (Alderson & Banerjee, 2002). CLA presents a comprehensive
definition of the L2 construct. Language testers, however, typically focus on one or a
few areas according to the specific purposes of assessment. CLA can be said to
provide a theoretical construct representation that can accommodate a range of L2 test
development purposes (Chalhoub-Deville, 1997). However, this representation of the
L2 construct does not guide developers in terms of salient language features to target
in a given context and/or how these features interact at various proficiency levels.
With CLA, a great deal of work is still needed in order to elaborate the L2 construct in
test development. Such shortcomings of theoretical construct definitions are often
resolved by constructing hierarchical language descriptors (i.e., language proficiency
scales) such as the *ACTFL Proficiency Guidelines* and the *CEFR*, which will be
reviewed in Section 2.4.

2.2 Construct Definitions of Reading Ability

Language ability includes four major skills: listening, speaking, writing, and
reading. The review of the evolution of construct definitions of language ability
reported in Section 2.1 provides a historical and theoretical perspective of how
language ability, as a whole, has been conceptualized. Section 2.2 focuses on the
construct definitions of reading ability.

An understanding of the construct definitions of reading ability is crucial to
the development of reading assessment. However, reading, especially second
language (L2) reading and foreign language reading, is a complex construct. Alderson
(2000) defines reading as a complex process affected by both reader variables (such as
knowledge, skills, abilities, motivation, affect, and other characteristics of the reader)
and text variables (e.g., topic and content, type and genre, text organization).

Koda (2005) summarized the definitions of reading competence (aka. reading
ability) proposed from three perspectives. First, the cognitive view, which reflects the reader-text interaction, conceptualizes reading competence as composed of three operations: decoding, text-meaning construction, and assimilation with prior knowledge. Reading comprehension is the process of decoding, constructing text meaning, and assimilating new knowledge with prior knowledge. Second, the developmental perspective models reading competence as two functionally independent clusters mastered by readers subsequently: decoding and comprehension. Third, the reading gear theory, from a functional perspective, proposes that the purposes for reading should be a factor considered in determining reading competence. While conceptualizing from different perspectives, all definitions stem from the same basic assumption that “successful comprehension emerges from the integrative interaction of derived text information and preexisting reader knowledge” (Koda, 2005, p 4).

Various reading models have been proposed to represent the reading ability theory. Reading models, most of which generated on the basis of empirical studies, describe what reading involves and how reading comprehension works. Urquhart and Weir (1998) made the distinction between two major types of reading models: componential models and process models. While componential models aim to identify componential skills or subskills involved in the reading, process models intend to depict mental operations engaged in reading comprehension.

The component skills approach regards reading as the product of a complex information-processing system, involving a constellation of closely related mental operations. Due to its multiple-skill orientation, the component skills approach has provided a very productive theoretical foundation for assessing L2 reading development and examining reading skills transfer (Koda, 2005). According to
Alderson (1990), the unitary skill approach may lead to construct underrepresentation in test development. The component skills approach is widely supported in literature (Alderson, 1990; Rumelhart, 1980). However, within the camp of componential models, there is no consensus about the number and nature of essential component skills. (Alderson, 1990; Urquhart & Weir, 1998; Alderson et al., 2015). The interrelationships among component skills are far from clear (Alderson, 1990). As reviewed by Urquhart and Weir (1998), there are several componential models of reading, including Davis (1968: four skills); Munby (1978: 18 skills); Lunzer, Waite, and Dolan (1979: eight skills); and Grabe (1991: six skills).

Process models can be classified into three categories: bottom-up, top-down, and interactive approaches. Bottom-up models (Gough, 1972; LaBerge & Samuels, 1974; Samuels & LaBerge, 1983) conceptualize reading comprehension as a part-to-whole information processing procedure. Reading is driven by text. Readers analyze text chuck by chuck in order to construct meaning. Top-down approaches (Goodman, 1967, 1986, 1994, 1996; Smith, 1971, 1983) view reading as a whole-to-part process. Reading is driven by meaning/concept and it begins with readers’ assumptions about incoming text. Inference-making and readers’ background knowledge play an essential role in top-down reading comprehension. Since both bottom-up and top-down models have incurred criticisms due to their limitations, interactive models (Kintsch, 1988, 1998; Kintsch & van Dijk, 1978; Rumelhart, 1977, 2004; Stanovich, 1980, 1986, 2000; van Dijk & Kintsch, 1983) are proposed. Interactive approaches model reading as a simultaneous involvement of both textual information and readers’ mental activities. Readers use all levels of processing simultaneously nonetheless one source of meaning can be primary at a particular time.
Section 2.2 and Section 2.3 review existing theoretical works and empirical studies on L1, L2, and CSL reading. In these two sections, the construct of reading is represented primarily from a component skills approach and secondarily from an information processing perspective.

2.2.1 Construct Definitions of Reading

The review of several theoretical works on reading (Alderson et al., 2015; Grabe, 2009; Koda, 2005) leads to the general consensus about the vital components of reading ability as well as their relationship.

Reading ability can be modeled as consisting of two levels\(^1\) of cognitive/metacognitive processing. **Lower-level processing** has three components: word recognition, syntactic parsing (using grammatical information to understand the word integration), and semantic-proposition encoding (building clause-level meaning from word meanings and grammatical information). **Higher-level processing** involves: a text model of comprehension, a situation model of comprehension, and reading skills under the command of the executive control mechanism in working memory (e.g., strategies, goals, inference, background knowledge, and comprehension monitoring) (Grabe, 2009). Please refer to Figure 6 for a visualization of component skills of reading.

---

\(^1\) Some research (e.g. Koda, 2005) conceptualized reading as composing of three hierarchical operation clusters: lower-level decoding, interim level text-information construction, and higher-level situation-model building. Some parts of the interim level text-information construction are regarded as lower-level decoding and other parts of the interim level text-information construction belong to higher-level processing in the two-level model. The relationship between the two-factor model and three-factor model will be elaborated in Chapter 3.
Figure 6

Component Skills of Reading

Note. Adapted from Grabe, 2009, p.21-59

Word Recognition. Word recognition refers to the processes of extracting lexical information from graphic displays of words (Koda, 2005, p. 29). Three methods, decoding, analogy, and prediction, are used to recognize words. Word recognition involves the interaction of activated orthographic, phonological, semantic, and syntactic processes (Perfetti & Hart, 2001). In order to recognize words, readers must recognize the word forms rapidly, link graphic form with phonological information, activate semantic and syntactic resources, recognize morphological affixation in complex word forms, and access their mental lexicon (Grabe, 2009). In addition, when readers encounter word recognition difficulty or unknown words, contextual information is heavily relied in word recognition.
**Syntactic Parsing.** Grammatical information is continuously involved in reading comprehension. The process of syntactic parsing (word integration) is essential to reading (Fender, 2001; Grabe, 2005). Syntactic information from determiners, word ordering, tense, subordinate clauses, modality, and pronominal forms, etc, provide instructions for the construction of reading comprehension (Grabe, 2005). The syntactic complexity and ambiguity have a measurable impact on reading processing time (Fender, 2001; Gernsbacher, 1990, 1997). Without syntactic parsing of clauses and sentences, semantic proposition unit, the basic meaning structure in reading comprehension, cannot be formed (Perfetti & Britt, 1995).

**Semantic Propositions.** Semantic propositions are the building blocks of text comprehension (Perfetti & Britt, 1995). A semantic proposition can be regarded as a network of smaller packs of information that are linked together in a meaning unit. Semantic propositions are formed simultaneously with word recognition and syntactic parsing. On one hand, syntactic parsing of clauses and sentences underlies the formation of semantic propositions. Semantic propositions depend on isolating the main verb (predicate), subject, object, and adjunct positions. On the other hand, the semantic proposition requires the processing of transitional devices and discourse organization markers to establish the relationship within and between propositions (Perfetti & Britt, 1995). When immediate networks are activated and added to the bigger network of activated information, the propositions are connected and the textual meaning is created (Grabe, 2009).

In lower-level processing, linguistic knowledge resources and cognitive processing are integrated for comprehension. Working memory plays an important role in lower-level reading processing: it supports orthographic, phonological, and morphological processing for word recognition; it stores and combines activated
words; it supports semantic and syntactic processing at the clause level and stores relevant information for building text comprehension. Working memory also suppresses unwanted information for readers.

In addition to lower-level processing, building up a general understanding of a longer text requires higher-level processing. Three component skills of higher-level processing are: a text model of reader comprehension, a situation model of reader comprehension, and skills under the command of the executive control mechanism in working memory (e.g., strategies, goals, inference, background knowledge, and comprehension monitoring).

**Text Model of Comprehension.** While the major function of the text model is to represent the information intended by the writer, situation model of comprehension involves readers’ individual interpretation. Building up a general understanding of a longer text requires the combination of information from the currently formed proposition. This is referred to as a text model of comprehension emerging in working memory. The text model of comprehension is operated as linking information into a network, majorly by the overlapping of elements, suppressing less-active information, making simple inference, and restructuring summary.

**Situation Model of Comprehension.** Situation model of comprehension involves readers’ individual interpretation. In situation-model comprehension, readers bring their own information to process text. The notion of a situation model is generated based on cognitive psychology studies, which review comprehension as the creation of an appropriate mental model of text. Such a mental model not only incorporates the text information but also incorporates spatial and visual interpretations as well as reader knowledge (e.g., Garnham & Oakhill, 1996; Zwaan & Rapp, 2006). Factors that influence construction of a situation model include: reader
purpose and task expectation, genre activation, similar story instances, general background knowledge resources, evaluation of the importance and value of information, attitudes toward writer, story, genre, episode, inferences needed for interpretation (Grabe, 2009, p44). L2 learners will over rely on a situation model when they have difficulty in reading comprehension due to limited L2 language proficiency.

In addition to the text and situation model, higher-level reading processing involve other skills under the command of the executive control mechanism in working memory: **background knowledge, attentional processes, goal setting, strategy use, metacognitive awareness, comprehension monitoring, and metalinguistic awareness.** These factors fall within the field of reading strategy study.

Reading strategies are deliberate, goal-directed attempts to control and modify the reader’s efforts to decode text, understand words, and construct meaning of text (Afflerbach, Pearson, & Paris, 2008, p.15). As discussed in Section 2.1.2, commonly used reading strategies/language learning strategies can be classified into five categories: cognitive strategies, metacognitive strategies, social/affective strategies, memory strategies, and compensation strategies (O’Malley & Chamot, 1990; Oxford, 1990). Using background knowledge is a typical cognitive strategy. Goal setting, attentional processes, metacognitive awareness, comprehension monitoring are all metacognitive strategies.

### 2.2.2 Construct Definitions of L2 Reading: a Cross-linguistic Analysis

Second language (L2) reading differs from first language (L1) reading and uniqueness of L2 reading may recast accepted L1 reading construct (Enright, et al. 2000). L1 reading research findings and implications should be reexamined when applied to L2 reading research, teaching, and testing (Grabe, 2009).
L2 readers have already had L1 literacy experience before learning L2. L2 learners’ reading processes involve two or more languages. While L1 readers have adequate listening and speaking proficiency before learning to read, L2 readers do not have prior oral proficiency. Their limited L2 linguistic knowledge constrains reading comprehension. Due to inadequate proficiency, L2 readers engage in less reader-text interaction than in L1 reading. Lower-level processing dominates L2 reading. Two models specific to L2 reading will be reviewed in this section, i.e., Coady’s (1979) psycholinguistic model and Bernhardt’s (1986, 1991) constructivist model.

Coady (1979) identified three components of L2 reading: conceptual ability, process strategies, and background knowledge. Conceptual ability refers to cognitive competence related to reading. Process strategies consist of both knowledge system itself and the ability to use the knowledge. Lastly, in Coady’s (1979) psycholinguistic model, readers’ background knowledge is viewed as a focal component of reading ability instead of a peripheral attribute.

Based on data generated by intermediate college-level American L2 learners of French, German, and Spanish, Bernhardt (1986, 1991) proposes a constructivist model, which conceptualizes the interactive and multi-dimensional nature of L2 reading. This model conceptualizes the interaction among three lower-level factors (word recognition, phonemic/graphemic decoding, and syntactic feature recognition) and three higher-level factors (prior knowledge, intratextual perception and metacognition). Word recognition, in this context, is the match of semantic value with forms. Phonemic/graphemic decoding refers to matching spoken language with their graphic equivalents. Intratextual perception is about the reconciliation among adjacent parts of text. Prior knowledge is the knowledge readers already have before they read new information. Metacognition or metacognitive knowledge refers to learners’
knowledge on how to monitor, modify, regulate, and execute their cognitive processes. In other words, metacognitive knowledge means “thinking about thinking” or “cognition about cognition” (Zhang, 2001). As illustrated by Figure 7, the constructivist model is circular and interactive. Components of L2 reading interact and influence each other.

![Bernhardt's Constructivist Model of Reading](image)

Figure 7.

*Bernhardt's Constructivist Model of Reading*


From a cross-linguistic approach, Koda (2005) reviews impact of L1 linguistic property, L1-L2 distance, and L1-L2 involvement on some vital componential processes/knowledge of L2 reading: word recognition, vocabulary knowledge, intraword awareness, information integration in sentence processing, discourse processing, and text-structure knowledge.

**Word Recognition.** Factors affecting L2 word recognition include: L2 knowledge, L1 reading ability, and L1-L2 distance (Koda, 2005). First, inefficient word recognition is common among low-proficiency L2 readers irrespective of their L1 backgrounds. Reading strategies studies reveal that low-proficiency L2 learners are more involved in word-level than discourse-level processing (Cziko, 1980; Horiba, 1990) and they rely on a word’s graphic cues rather than its semantic information (Chamot & El-Dinary, 1999; Anderson, 1991). The amount of L2
processing experience directly contributes to the acquisition of the highly abstract knowledge of the target language’s writing system. Second, L1 literacy experience has a long-term impact on L2 processing. The fact that alphabetic L1 readers depend on phonological information to a greater degree than logographic L1 readers further support the L1 orthographic influence on L2 word recognition. L1 processing experience is a major source of performance variations among L2 learners. Third, processing skills transfer across languages. The degrees of facilitation that derive from transferred L1 competencies will vary to the extent the two languages share similar properties. Shared orthographic knowledge provides long-term facilitation in L2 reading development, first by promoting mastery of L2 visual-information sampling skills and then facilitating information integration from multiple sources. However, L1-based facilitation only occurs when L1 and L2 processing demands are identical. In sum, typological distance not only explains overall performance differences among learners with related and unrelated L1 backgrounds, but also underscores the ways L1 experience facilitates L2 lexical processing.

**Vocabulary Knowledge.** Vocabulary knowledge is multifaceted. It composes of knowledge on form (spoken, written, word parts), meaning (form and meaning, concept and references, associates), and use (syntactic and grammatical functions, collocations, constrains on use) (Nation, 2001). Vocabulary knowledge positively relates with reading comprehension, more highly than morphosyntactic knowledge (Koda, 1989; Qian, 1999) and reading strategies (Haynes & Baker, 1993). In addition, the relationship between vocabulary knowledge and reading comprehension is reciprocal. Vocabulary knowledge facilitates reading comprehension, while at the same time, vocabulary knowledge expands through reading.
For L1 learners, vocabulary knowledge is a reliable indicator of conceptual sophistication. L1 vocabulary learning is to establish connections among meaning, symbol, and sound. For L1 kids, they are expected to develop the skills for mapping familiar sounds and concepts to graphic symbols. For L2 learners, there is an apparent gap between word knowledge and conceptual knowledge. L2 learners may have already known concepts but do not know the sounds and symbol of corresponding words. L2 word learning entails linking four lexical elements: symbol, sound, meaning and L1 equivalent. L2 vocabulary learning is operationalized explicitly or implicitly as a process of connecting L2 visual forms with their meanings through definitions of L1, L2 or both.

In terms of how L1 and L2 lexical information is stored in conjunction with conceptual knowledge (i.e., dual-language lexical organization), three models are proposed: word association model (conceptual information is accessed only through L1 words), concept mediation (direct access to conceptual storage in both L1 and L2), hybrid of the concept-mediation and word-association models (while beginning learners relay on word association, more advanced learning access meanings directly during L2 lexical processing without L1 mediation). Bilingual interactive activation (BIA) model argues that lexical information in both languages is activated whenever the input shares features (visual, phonological, and semantic) with lexical alternatives in each of the involved languages. Recent experimental findings have provided substantial support for the integrative lexical representation views: the activation of L1 information is automatic, and cannot be controlled by the learner during L2 lexical processing. These findings strongly support the continuing role of L1 in L2 reading beyond initial learning phases.
**Intraword Awareness.** Metalinguistic awareness is the ability to identify, analyze, and manipulate language forms. Intraword awareness (IA) refers to “generalized metalinguistic insights related to the capabilities underlying lexical learning and processing.” (Koda, 2005, p74) IA includes both phonological and morphological awareness. IA is the facilitative agent in analyzing intraword elements. IA facilitation takes diverse forms in incidental and intentional learning. In intentional learning, IA enhances associative list learning and definition-based word learning; in incidental learning, IA enhances lexical inference and knowledge incorporation and expansion.

The IA capabilities are shaped by how lexical information is represented in the writing system and IA is language-specific. Although L2 literacy experience is a primary force governing L2 IA development, L1 writing properties also obviously and continuously shape L2 IA formation. Metalinguistically experienced adult learners acquire IA competence in a new language more quickly than beginning L1 learners of that language. Systematical comparison of L2 IA among learners with different L1 orthographic background and among learners with different degrees of L2 print-processing experience reveals the role L1 and L2 print-processing experience plays in L2 IA formation and L2 lexical competence development (Koda, 2005).

Empirical findings suggest that L1 and L2 IA are both operative during L2 lexical processing, jointly impacting on L2 lexical-competence development. While L2 print-processing experience is a primary force governing L2 IA development, L1 orthographic properties also have a continuing, clearly identifiable imprint on L2 IA formation. In sum, L1 experience establishes the scaffolding for foundation building, and L2 input instills a linguistic base necessary for fine-tuning.
Information Integration in Sentence Processing. Isolated words are integrated into larger linguistic units (i.e., phrases and sentences), by incorporating their syntactic, semantic, and pragmatic information. Factors constrains sentence processing include syntactic factors, working memory, discourse context, and local semantic property. Syntactic factors contribute to L2 sentence-processing difficulty. There are two major types of syntax-related sources of L2 comprehension difficulty: violation of structural prototypicality and lack of structural transparency (Berman, 1984).

L2 sentence processing is heavily constrained by L1 morphosyntactic properties, L1-L2 involvement, L1- L2 typological distance and universal principles. First, L1 structural properties constrain the acquisition of L2 verb meanings, thereby significantly affecting L2 syntactic processing. Second, L2 sentence processing procedures evolve from cumulative L2 processing experience through the use of automated L1 processing procedures. L2 learners’ sentence processing procedural preference may shift from L1-based to more native-like patterns as L2 proficiency increases. Third, certain components of processing mechanisms are shared across languages. However, some major components are language-specific. Fourth, L1-L2 typological distance determines the degrees of facilitation that derive from transferred L1 sentence-processing competencies.

In terms of what constitute linguistic complexity, some L1 theorists place heavy emphasis on quantitative factors (e.g., number of words) while others place greater weight on structural variables (e.g., syntactic ambiguity). To L2 comprehension difficulties, quantitative indices are more reliable, at least among beginning learners. The complexity-difficulty relationship may diminish as L2 proficiency increases. L1 reading research commonly suggests that reading
comprehension difficulty related to syntactic complexity generally arises from a deficit in the decoding skills necessary for linking spoken language with its written forms rather than from a dearth of syntactic knowledge. L2 findings also lend support for the processing deficit hypothesis. Grammatical competence is unlikely to create sentence-processing problems for L2 learners.

**Discourse Processing.** Reading can be conceptualized as consisting of three hierarchical operation clusters: low-level decoding, interim-level textbase construction, and higher-level situation-model building. Discourse processing is involved in the latter two clusters of operations.

L2 textbase construction involves two major processes: propositional computation and coherence building. Propositional computation is achieved through analysis of the text's surface form, L2 linguistic sophistication is a key factor determining its success. In addition to linguistic knowledge, coherence building necessitates a functional master of discourse devices. Although basic requirements in coherence building do not differ between L1 and L2 discourse processing, coherence building in L1 and L2 facilitated by completely different variables. For instance, explicit coherence markers (conjunctions and enumerators) may function disparately in L1 and L2 text processing. For L2 readers, coherence markers can be used to compensate for limited lexical knowledge. In L2-reading, cross-linguistic variations in referential expressions make the identification of correct antecedents of lexically attenuated forms even more difficult. L1 literacy experience, more precisely, nonlinguistic aspects of L1 comprehension ability, such as distinguish thematic and peripheral information and identifying underlying semantic relationships among propositions, should facilitate L2 coherence building. These skills are not language specific once developed in one language, they can be applied to another.
Given the centrality of background knowledge in situation-model building, a natural assumption would be that L1 conceptual knowledge is a major source of individual differences in L2 situation-model building. L2 text comprehension can be aided to the extent that L1 knowledge is relevant and/or applicable to a particular situation depicted in an L2 text. L2 situation-model building may become more difficult as the quantity of culture-specific information in a text increases. The major characteristics of adult L2 readers include a reasonably solid conceptual based formed in L1 as well as limited L2 knowledge. Top-down conceptually driven processes presumably dominate L2 text processing, at least until adequate L2 knowledge develops. Their reliance on conceptual knowledge can have a strong facilitative impact in particular text genres. However, adult L2 readers may be more susceptible than beginning L1 readers to misunderstanding and misinterpretation.

Cross-linguistic transfer occurs in virtually every aspect of L2 processing. Much of L1 lower-level processing transfer is automated and its activation is triggered by the L2 input regardless of the reader’s intention. However, discourse-level skills are poor candidates for automaticity, because many of the operations involve conceptual manipulations. Only a few higher-level L1 competencies are transferred through automatic activation, e.g., text-structure knowledge. Although transfer, presumably evolving from automatic L1 activation, plays a relatively minor role in L2 discourse processing, many useful L1 competencies can be intentionally brought into play, e.g., metacognitive awareness, comprehension strategies.

Empirical studies suggest that sufficient L2 proficiency is needed for text-coherence building; text information is represented differently in memory when reading in L1 and L2; domain knowledge is a more powerful predictor of learning from academic texts than is L2 proficiency. The way in which conceptual knowledge
(general knowledge, culture-specific knowledge, and domain knowledge) affects comprehension are complex, interfacing, among other things, with L2 proficiency, text types, and L1 comprehension ability.

**Text-structure Knowledge.** Text structure and organizational devices vary across genres within a single language and across languages. Text-structure knowledge facilitates comprehension and memory. Differences in text structure knowledge exist among L2 readers with different L2 proficiency and L1 reading ability. Text type may have differential impacts on text memory among L2 readers with diverse L1 backgrounds and proficiency levels. L2 linguistic sophistication is directly related to structural sensitivity, thus accounting for the diverse impacts of text-structure variables on comprehension among high- and low-proficiency L2 learners. For both L1 and L2 readers, text-structure instruction engenders substantial comprehension gains, as well as improved recall of both expository and narrative texts. However, studies also reveal that structural training appears to be more complex than has been assumed, entailing elements beyond text-structure knowledge.

Both L1 and L2 text-structure knowledge play a role in L2 text comprehension. Dual-language involvement, like other aspects of reading should characterize discourse-processing behaviors among L2 readers. Two factors are particularly pertinent in the in-depth examinations of cross-linguistic interactions: L1-L2 rhetorical distance and the reader’s ability to exploit L2 text-structure information. L1-L2 rhetorical distance allows reliable estimates of the extent to which L1 text-structure schema facilitates L2 discourse processing.

In sum, beyond factor accounting for comprehension variability of L1 reading, L2 reading is shaped by following factors: transfer of L1 reading skills and strategies, L1-L2 linguistic distance, L1-L2 interaction, and processing constrain caused by
undeveloped L2 linguistic knowledge (Enright, et al. 2000). Due to inadequate proficiency, L2 readers engage in less reader-text interaction than in L1 reading. Lower-level processing dominates in L2 reading. Fluent word recognition, processing efficiency, and reading rate are of great importance in successful L2 reading (Koda, 1996, 1997; Geva, Wade-Wooley, & Shany, 1997). Assessment that is designed based solely on L1 reading theories may not be a valid reflection of L2 reading ability (Enright, et al. 2000). L2 reading assessment, especially assessment for beginning learners, should focus more on lower-level processing.

2.3 Construct Definitions of CSL Reading: a Language-specific Perspective

Reading a particular writing system necessitates a command of the cognitive mechanisms specifically designed for dealing with its structural and representational properties (Koda, 2005). The language-specific perspective of reading argues that lower-level reading processing are language specific (Bassetti, 2004; Koda, 2005). Chinese is a logographic language. Its unique linguistic and psycholinguistic features determine that reading Chinese as a foreign language is different from reading other languages. Empirical and theoretical evidence are needed to depict the construct of reading in Chinese and to reveal skills and knowledge underlying decoding, comprehension and reading fluency (Zhang, 2017).

Besides factors accounting for comprehension variability of L1 reading, L2 reading is shaped by factors like transfer of L1 reading skills and strategies, L1-L2 linguistic distance, L1-L2 interaction, and processing constrain caused by undeveloped L2 linguistic knowledge (Enright, et al. 2000). This section aims to examine CSL reading from the perspectives of L1 transfer, L1-L2 linguistic distance, and L1-L2 interaction. When we talk about CSL reading in America, in most cases readers speak English as their L1. Therefore, the discussion of this section focuses on
English-Chinese distance and its effect on CSL reading. This section reviews Chinese linguistic properties at orthography, morphology, syntax, discourse, and text structure levels as compared with English. At the meanwhile, it reveals special cognitive processes dealing with unique structural and representational properties of Chinese language. Implications for CSL reading assessment are discussed at the end.

2.3.1 Lower-level Decoding

**Orthography.** Orthography refers to the representation of the sounds of a language by written or printed symbols. Writing systems differ in two dimensions: orthographic depth (the degree of regularity in symbol-sound correspondences) and orthographic representation (the linguistic unit each graphic symbol denotes).

Orthographic depth hypothesis (Katz & Frost, 1992) argues that in shallow (transparent) orthographies, phonological information is assembled through letter-to-letter, symbol-to-sound translation. However, in deep orthographies, there is no consistent sound-form correspondence and as a result phonological information is obtained after a word has been identified based on the stored knowledge of the word. Chinese is a deep orthography and English orthography is relatively more transparent as compared with Chinese.

Orthography depth hypothesis implies that phonological recoding is the main route for word recognition in reading a shallow orthography, while graphic information is the main route for word recognition in reading a deep orthography (Perfetti & Zhang, 1991; Xu, 1991). When reading in Chinese, a language with deep orthography, phonological recoding occurs after meaning retrieval has been achieved (Hoosain 1995; Wu et al, 1993). In addition, miscue analysis studies (Ke, 1986; Wang, 2006) reveal that compared with phonetic similarity, graphic similarity is a more common cause of character-recognition miscues. In sum, graphic information
about the strokes, radicals, and characters plays a more important role than sound information in reading Chinese. Chinese and English L1 readers develop different metalinguistic awareness. While English readers develop phonetic awareness, Chinese readers form awareness about the morphological structure of Chinese characters.

According to orthographic representation, writing systems can be classified into three types: logographs, syllabaries, and alphabets. In a syllabic writing system, a written sign represents one syllable. In an alphabetic writing system, the relationships between written patterns and sound patterns vary, but the written signs cue possible pronunciations (Goodman, 1996). In a logographic language, a written graph represents a morpheme, the smallest meaningful language unit.

English is an alphabetic language written in Roman letters. In English, 26 letters, spelling, punctuation marks, and the space indicating word boundaries comprise the written system (Hung, 2012). Chinese is a logographic language. Chinese written signs do not present sound units, although they sometimes cue possible pronunciations.

The Chinese written symbol is referred to as “zi” or character. The Chinese characters are composed of two layers of orthographic structure: stroke and radical. An integral character contains only one radical. A compound character, which is the majority, contains at least two or more radicals. Strokes are the building blocks for radicals. Radicals are the smallest meaningful orthographic components that play semantic or phonetic roles in compound characters (Shen & Ke, 2007). Radicals can be classified into two types according to their functions: phonetic radicals and semantic radicals. A phonetic radical cues the pronunciation while a semantic radical cues the meaning. However, due to historical evolution, even ignoring tonal differences, only 26% phonetic radicals can be regarded as reliable cues to
pronunciation (Fan, Gao, & Ao, 1984). Not all radicals are necessarily the smallest perceptual units (bujian 部件). Some radicals may contain more than two perceptual units (部件). For instance in 瞥, 敝 serves as the phonetic radical and 目 serves as the semantic radical. Perceptually, 瞥 consists of three units. 敝 is composed of two perceptual units: 亻 and 丄. Although 亻 is a semantic radical and has its own pronunciation, it has lost its semantic and phonetic functions in 瞥.

One character stands for one morpheme, although there are a very limited number of polysyllabic morphemes. One character represents one and only one syllable. In classical Chinese, one character represents one word. However, in modern Chinese, most words are composed of two or more characters. The same syllable can be represented by different characters. Around 60% syllables have homophones. This kind of ambiguity can contribute to the acquisition difficulty of reading Chinese as an L2.

**Morphology and Word Processing.** A Chinese word can be formed by one, two, or more characters but there are no inherent markers of lexical category or inflections. About 80% compound words are two-character words (i.e., bi-morphemic words) (Lin, 1971). Chinese bi-morphemic words can be classified into five categories: coordinative words, attributive compounds, verb-complements, verb-objects, and subject-predicates (Huang & Liao, 1981). When a word is formed by more than one character, character recognition during word decoding is roughly equivalent to identifying individual morphemes in an English multiple morphemic word (Shen, 2008).

Processing Chinese words include at least three sub-skills: character recognition, word decision, and lexical access (Shen, 2008). Character recognition refers to “identifying the graphic structure of a character, activating its pronunciation
either aloud or mentally, and accessing its meaning.” (Shen, 2008, p. 501) Word decision refers to the process of grouping adjacent characters into words in ongoing reading. Lexical access is the process of gaining an orthographic encoding of a word (meaning, syntactic properties, and pronunciation) from long term memory, which fits the context (Wu & Liu, 1996). Chinese words, especially one-character words, have multiple meanings and some characters have more than one pronunciation. Readers need to rely on context to decide the meaning and pronunciation of some words. Contextualized word processing plays an important role in successful reading comprehension (Shen, 2008).

The fundamental difference of word processing in reading between Chinese and English is that word decision is a necessary step for Chinese lexical access. In reading English, there are visible space boundaries between words and therefore word decision is not a necessary skill. However, in reading Chinese, word decision is a complicated task. While experienced L1 learners have no problem to segment words, for L2 Chinese learners, word decision is a hard process. One reason is that in Chinese text, space indicates character boundary not word boundary. The spaces among characters are equal. There is no spacing cue signaling where a word ends and another word starts. Second, Chinese words are composed of one, two, or more characters. A character can either stand alone as a word or can also join other characters to form multiple-character words with distinct meanings. Some combinations of multiple characters can be analyzed as either one word or as multiple words. Everson & Ke (1997) investigated the accountability of Bernhardt’s constructivist model for the reading experience of CSL learners. They found out that intermediate Chinese learners have difficulty isolating meaningful word units in the running text. Their study extended Bernhardt’s model by positing an orthographic
layer of difficulty in the word recognition component for CSL learners. In sum, in reading Chinese, word decision plays a key role in Chinese reading comprehension.

**Syntax.** Violation of structural prototypicality and lack of structural transparency are two major syntactic sources of L2 comprehension difficulty (Berman, 1984). L1 morphosyntactic properties, L1-L2 involvement, L1-L2 typological distance, and universal principles heavily constrain L2 sentence processing (Koda, 2005). Chinese sentence structures differ from English sentence structures in many ways. This section starts with a brief discussion of the positions and functions of verbs, objects, subjects, adverbials in Chinese sentences. Then, some unique Chinese syntactic patterns are highlighted.

The dominant Chinese sentence structure is the same as English: subject–verb–object (SVO). However, the relationship of subject and predicate in Chinese has been regarded as considerable semantic looseness (Norman, 1988). The subject of a sentence may stand in diverse logical relationships to the predicate. It may be the agent of a transitive verb, the patient or recipient of the action, an object about which some quality is predicated, or an instrument (Norman, 1988). Chao (1968) regards the grammatical meaning of subject and predicate as that of “topic” and “comment”: the predicate says something about the subject.

At the most general level, there are two types of Chinese sentences: major and minor types. A major sentence has both a subject and a predicate while a minor sentence omits the subject. The frequent omission of pronominal subjects indicates that minor sentences are more common in Chinese as compared with English (Norman, 1988).

In English, adverbials of manner, place, and time are normally put after the main verb. In Chinese, on the contrary, adverbials are put either before the main verb
or before the subject. In addition, Chinese is an analytic language. It does not use final inflections to express grammatical notions like aspect, plurality, passiveness, etc. Instead, these grammatical notions are represented by the use of particles and by word order.

Chinese and English show significant differences in nominal sentences. Chinese nominal sentences may contain no verb at all: *jīn tiān xīng qī liù* “TODAY SATURDAY- Today is Saturday.” Although the copular verb *shì* “is, are” can be used in nominal sentences, the copular verb *shì* has only a copular function and takes none of the aspect markers, which is different from its English counterpart. Sometimes, a copular predicate only comments on the topic of the sentence in a loose way.

Some Chinese patterns have no or little equivalents in English. For instance, the *shì…de* structure describes or inquires about the manner, time, place, or initiator of an action that has already happened. The *bǎ* construction (*S–bǎ–O–V– other elements*) emphasizes the subject’s impact upon the object. Due to the lack of English equivalents, these patterns may lead to the CSL reading difficulty.

2.3.2 Higher-level Processing

Lower-level reading processing are language-specific (Bassetti, 2004; Koda, 2005). Reading in Chinese and English differ at lower-level processing, especially in word processing. However, the higher-level psycholinguistic and cognitive processes are more the same than they are different (Hung, 2012; Koda, 2005).

**Text Model of Comprehension.** As discussed in previous sections, L2 textbase construction involves two major processes: propositional computation and coherence building. Propositional computation is achieved through analysis of the text’s surface form. L2 Chinese linguistic sophistication is a key factor determining its success. In addition to linguistic knowledge, a functional master of discourse devices
plays an important role in coherence building. Very few Chinese discourse devices have exact English equivalents, which are semantically and syntactically interchangeable (e.g., “if” in English and “li ru” in Chinese). Most Chinese discourse devices differ slightly from their English translations either in meaning or in function. To CSL learners speaking English as the L1, negative transfer from English knowledge may hinder the master of English discourse devices. Although basic requirements in coherence building do not differ between L1 and L2 discourse processing, coherence building in L1 and L2 facilitated by completely different variables. For L2 readers, coherence markers can be used to compensate for limited lexical knowledge. In L2 reading, cross-linguistic variations in referential expressions make the identification of correct antecedents of lexically attenuated forms even more difficult. In sum, the assessment of Chinese discourse devices and the ability to identify referential expressions should be included in CSL reading tests.

**Text Structure/Discourse Pattern.** The discourse pattern refers to the logical arrangement of ideas and is language-specific (Kaplan, 2001). The discourse pattern of an expository text will vary depending on the culture and the L1 of the writer. If readers do not know the discourse pattern, they cannot read for comprehension (Kaplan, 2001). Different languages organize information in different patterns. Discourse patterns of American English texts are deductive and linear. An article starts with thesis/topic sentences and main ideas, following by supportive details, and ends with conclusion. However, in Asian languages, including Chinese, Korean, and Japanese, thought is developed indirectly. It is considered very rude if a speaker/writer directly address the main idea or issue. Poetic hints are given and readers/listeners need to figure out the main idea through these hints. To CSL readers, the knowledge of Chinese discourse patterns is an essential component of CSL reading ability.
**Situation Model of Comprehension.** Situation model differs from text model of comprehension mainly in that the former involves readers’ individual interpretation. According to Grabe (2009), factors that influence construction of a situation model include: reader purpose and task expectation, genre activation, similar story instances, general background knowledge resources, evaluation of the importance and value of information, attitudes toward writer, story, genre, episode, inferences needed for interpretation. Among these factors, genre activation may be language-specific and should be highlighted in modeling CSL reading construct.

**Skills under the Command of the Executive Control Mechanism in Working Memory.** Besides text model and situation model of comprehension, higher-level reading processing involves other skills under the command of the executive control mechanism in working memory: background knowledge, attentional processes, goal setting, strategy use, metacognitive awareness, metalinguistic awareness, and comprehension monitoring. These factors fall within the field of reading strategy study. While using background knowledge is a cognitive strategy, the rest, except metalinguistic awareness, belong to the category of metacognitive strategies.

Cross-language transfer theory of reading argues that many reading skills developed in L1 can be transferred to L2 reading (Koda, 2005). Learners with different L1 backgrounds use different cognitive tactics during L2 reading. Findings of two empirical studies (Abbott, 2006; Bang & Zhao, 2007) support the cross-language transfer theory: L1 linguistic features and L1 literacy experience shape EFL/ESL readers’ aptitude in cognitive strategy use when they read English. Abbott (2006) compares seven Arabic- and eight Chinese-speaking intermediate EFL learners’ reading processes and found out that Arabic-speaking EFL learners
outperformed their Chinese counterparts in answering questions which required using three top-down reading strategies (skimming, connecting, and inferring). However, Chinese EFL learners did better in answering questions requiring the use of four bottom-up reading strategies (breaking a word into smaller parts, scanning, paraphrasing, and matching). Abbott (2006) attributed this difference to the distinction of Chinese and Arabic orthographies. The comprehension of Arabic writing is very sentence-oriented. To decode a word, readers must first understand the sentence in which this word appears (Abu-Rabia, 1997). Decreased dependence on local cues encourages Arabic-speaking EFL/ESL learners to develop better mastery of top-down reading strategies. However, Chinese orthography encodes language at the level of morphemes and the understanding of Chinese writing is very word-oriented. Readers should first decode words and, based on their understanding of words and syntax, they can move forward to construct the meaning of whole sentences. Through their L1 literacy experience, Chinese-speaking EFL/ESL learners become more sophisticated in bottom-up reading strategies.

Another study, Bang and Zhao (2007) examined the reading strategies use of advanced Korean and Chinese EFL learners. Particular attention was devoted to how word recognition and processing skills developed in learners’ first language might influence types of strategies used in determining meanings of unfamiliar English words. Findings revealed that Korean EFL learners tended to rely on phonological strategies, while Chinese EFL learners tended to rely on visual-orthographic strategies when reading English texts. With respect to strategy use in overcoming comprehension gaps caused by insufficient vocabulary knowledge, Korean EFL learners tried to repeatedly sound out words and sentences. The Chinese EFL readers, however, favored silent reading. When guessing the meaning of new words from the
context, Korean EFL learners broke down the new words into smaller components (syllables) and analyzed these components (syllables) to make sense of them (e.g. hyper/amnesias, neuro/transmitters). However, Chinese EFL readers did not break down words. Instead, they would look at the overall spelling of the target words and try to relate them to some other words which had similar spellings (e.g. potency/potential). Korean is written in a non-Roman alphabet writing system called Hangul. One or more consonants are combined with a vowel to form a syllable, and each syllable is written in a square-shaped block (Taylor, 1980). Native Korean speakers develop ‘compound phonemic awareness’ through exposure to syllable-blocks (Koda, 1998). Thus, Korean EFL learners transfer the ‘compound phonemic awareness’ developed through L1 literacy experience into L2 English reading. Korean EFL learners tend to rely on phonological strategies. Specifically, they often break down words, which they cannot comprehend, into consonant-vowel compounds. They analyze these components to make sense of the whole words. They also repeatedly sound out words which they cannot comprehend at first glance and try to retrieve meanings by reading over and over again. In sum, Korean EFL learners tend to rely on phonological strategies and are more likely to attend to phonemes.

Chinese is a pictographic language and Chinese language uses a deep orthography which does not have transparent phonetic-graphic correspondence. Therefore, one usually has no idea how to pronounce a word based on the shape of this word. Sounds are randomly assigned to shapes and one heavily relies on rote memory to retain the sounds of most characters. In addition, Chinese does not use an alphabetic orthography and there is no such thing as spelling vowels and consonants together. Therefore, Chinese native speakers do not develop many phonological strategies through L1 literacy experience. Instead, Chinese speakers develop holistic visual-
orthographic strategies as they read in L1 and they transfer these strategies into EFL/ESL reading. Chinese EFL/ESL readers favor silent reading when trying to decode unfamiliar words. When encountering unfamiliar English words, Chinese EFL/ESL readers pay attention to overall shape of the word rather than the phonetic components. They look at the overall spelling of English words and try to relate unfamiliar words to some other familiar words with similar spellings.

Previous studies (Abbott, 2006; Bang and Zhao, 2007) report only differences in cognitive strategy use among learners with different L1 backgrounds. According to a literature review study (Lin, 2018) which systematically examines research on L2 reading strategies published between 2000 and 2017, no difference in L2 metacognitive strategy use among readers with different L1 backgrounds has been reported. One important implication to the current study is that cognitive strategies are language-specific and highly shaped by readers’ L1 literacy experience. To English-L1 CSL readers, their strategy use is greatly influenced by English linguistic features and strategies developed in reading English. When modeling the CSL reading construct, cognitive strategies, especially strategies directly associated with unique CSL linguistic features, should be highlighted.

2.3.3 Summary

Chinese differs from English in terms of linguistic properties. Special cognitive mechanisms are developed to deal with structural and representational features of Chinese language. CSL reading differs from reading in English, especially at lower-level decoding and in the construction of text model of comprehension.

Lower-level reading processing are language specific (Bassetti, 2004; Koda, 2005). Following lower-level skills/knowledge associated with the uniqueness of Chinese writing system should be highlighted in CSL reading construct:
(1) Orthography knowledge. Chinese is a deep orthography that has no consistent form-sound correspondence. As compared with phonetic information, graphic information about the strokes, radicals, and characters plays a more essential role in reading Chinese. Knowledge on radicals and knowledge on character structures are important components of CSL orthography knowledge. A large proportion of syllables have homophones and this leads to difficulty in CSL reading. Ability to distinguish homophones contributes to successful reading compression.

(2) Word processing and morphology knowledge. At least three sub-skills are involved in processing Chinese words: character recognition, word decision, and lexical access (Shen, 2008). Character recognition, i.e., identifying graphic structure, pronunciation, and meaning of a character, relies on orthography knowledge. Word decision (i.e., the ability to group adjacent characters into words in ongoing reading) is not required in reading English and is a hard process for CSL learners. In terms of lexical access, since a Chinese word often has multiple meanings and may have more than one pronunciation, contextualized word processing plays an essential role in successful CSL word recognition.

(3) Syntactic parsing. At syntactic level, Chinese and English differ in many ways. For instance, they are different in subject-predicate relationship, the omission of pronominal subjects, the position of adverbial phrases, and the use of nominal sentences. While English uses final inflections to denote grammatical notions such as aspect, plurality, these notions are represented by particles and word order in Chinese. Some unique Chinese syntactic patterns, such as “ba” pattern, “shi…de” structure, do not exist in English and may cause difficulty in CSL reading.

As compared with lower-level processes, the higher-level psycholinguistic and cognitive processes are more the same than they are different (Hung, 2012; Koda,
Following CSL-specific features/mechanisms should be stressed:

(1) Textbase construction involves two major processes: propositional computation and coherence building. Linguistic sophistication determines the success of propositional computation while a functional master of discourse devices contributes to coherence building. Discourse devices are language-specific. Very few Chinese discourse devices have exact English equivalences, which are interchangeable in both meaning and in function. Knowledge on Chinese discourse devices should be highlighted in CSL reading construct. Identifying correct antecedents of lexically attenuated forms in L2 reading is difficult due to cross-linguistic variations in referential expressions. The ability to identify referential expressions should be regarded as a component of CSL reading ability construct.

(2) Knowledge on text structure/discourse pattern plays an essential role in reading comprehension. Discourse patterns are language-specific. Chinese discourse patterns are less linear as compared with American English discourse patterns. Knowledge on Chinese discourse pattern should be taken into account in modeling CSL reading.

(3) Cognitive strategies are language-specific and highly shaped by readers’ L1 literacy experience. When modeling the CSL reading construct, cognitive strategies, especially strategies directly associated with unique CSL linguistic features, should be included.

According to Koda (2005), successful reading comprehension involves both knowledge and skills to utilize the knowledge for meaning construction. For L2 reading, knowledge and its corresponding processing skills should be treated as two separate facets since they do not develop concomitantly. Therefore, in this study, based on synthesis of literature review (Section 2.1, 2.2, 2.3), the construct of CSL
reading is operationalized into a multidivisible system with two separate but corresponding components: language knowledge related to reading; cognitive and metacognitive processing skills related to reading. Each major component consists of multiple layers of subskills/sub-knowledge. To illustrate relationship between the language knowledge and reading processes, in Figure 8, a given cognitive processing skill and its corresponding knowledge were marked by a same symbol. For instance, character recognition necessitates orthographic knowledge. These two components were both marked by an “1” symbol. Knowledge of cohesion is involved in both semantic-proposition encoding (marked by a “4”) and text model of comprehension (marked by a “5”). Therefore, knowledge of cohesion was marked by both a “4” and a “5” symbol.

In sum, on the basis of dimension analysis of reading ability as conceptualized in existing theoretical works and dimension analysis of language ability as presented in the CLA model, the researcher defines CSL reading ability as the ability to utilize knowledge (including linguistic knowledge and pragmatic knowledge) and corresponding processing skills (involving cognitive processing skills and metacognitive mechanisms) to construct meaning from written materials.

Reading abilities are language-specific especially at lower-level processing (Bassetti, 2004; Koda, 2005). This model of CSL reading ability highlights lower-level cognitive mechanisms and linguistic knowledge which are unique to CSL reading. In Figure 8, CSL reading processes which do not exist in reading in alphabetic languages (e.g., character recognition, word decision) were put in parentheses. Processes and knowledge components (e.g., lexical access, orthographic knowledge), which are very different from reading in alphabetic languages, were italicized.
Figure 8

CSL Reading Construct
Due to CSL learners’ limited proficiency, lower-level processing predominates CSL reading. CSL reading assessment, especially assessment for novice to intermediate learners, should focus more on lower-level processing. Therefore, it is necessary to closely investigate lower-level CSL reading and highlight linguistic features and cognitive processes that are specific to CSL.

In lower-level reading processing, linguistic knowledge resources and cognitive processing are integrated for comprehension. Figure 9 presents lower-level CSL reading processes and their corresponding linguistic knowledge, as denoted by orange arrows. Within each component, only two types of sub-knowledge/processes/abilities are listed: (1) knowledge/processes/abilities (e.g., word decision) that play important roles in CSL reading but do not exist in reading English; and (2) knowledge/processes/abilities (e.g., knowledge on Chinese discourse devices) that are very different from those in reading English. However, knowledge/processes/skills which are roughly identical in CSL reading and English reading are not listed out.

Figure 10 shows higher-level cognitive/metacognitive processes and their corresponding knowledge involved in CSL reading, with features and tactics unique to Chinese language listed out. Within each component, knowledge/processes/skills that are roughly identical in CSL reading as compared with English reading are excluded from this figure.
**Figure 9**

*Lower-level Processing and Knowledge of CSL Reading*
2.4 Reading Proficiency Scales: Operationalization of Theoretical Constructs

While theoretical research aims to reveal components of reading competence exhaustively, in the design of reading assessments, it is unnecessary and impossible to test every component in every testing situation. Instead, “we focus on those parts of the theory that are relevant to our testing purpose and then define these, both theoretically and operationally, in ways that are appropriate for this purpose” (Alderson, 2000, p137). In addition, competing theories often have overlaps and test developers may adopt parts from different theories for test development (Alderson, 2000). The identification of a
suitable construct or constructs is central to test validation (Alderson, 2000).

However, the theoretical CSL reading construct does not guide test developers in terms of what salient language features should be targeted in a given situation and/or how these features interact at various proficiency levels. A great deal of work is still needed to elaborate the CSL construct for test development purposes. These shortcomings of theoretical construct definitions are supposed to be resolved by developing test specifications or language proficiency scales/frameworks.

Language proficiency scales (aka language proficiency benchmarks or frameworks) define the construct of language proficiency by dividing proficiency into stages of development. Detailed descriptions are provided for each point or stage on the continuum of proficiency to describe competence of language users hierarchically from lower to higher scales. Language proficiency scales become increasingly popular because they provide guiding principles for test construction (Alderson, 1991) and provide coherent links between assessment, curriculum planning, and materials development (North, 1991).

Language proficiency scales can be classified into different categories. Previous studies (Alderson, 1991; Pollitt & Murray, 1993; North, 1993) have generated taxonomies from different perspectives. First of all, language proficiency scales are written for different primary purposes. Scales of language proficiency can be categorized into three types according to their orientations: user-oriented, constructor-oriented, and assessor-oriented scales (Alderson, 1991). Pollitt and Murray (1993) expanded Alderson’s (1991) taxonomy by adding the forth category: diagnosis-oriented scale. User-oriented and constructor-oriented scales are primarily used to inform “what the learner can do”
(North, 2000). User-oriented scales are mainly used to generate report for non-specialist users. Constructor-oriented scales are more informative and complex than the former since their primary function is to guide test development. Assessor-oriented and diagnosis-oriented scales can provide qualitative feedback on “how well the learner performs” for evaluation or diagnosis purposes (North, 2000). To a common framework/scale which intends to be applicable to different contexts and perspectives, all four orientations should be considered relevant (North, 2000).

Second, North (1993) identified five types of language proficiency scales from the perspectives of content and structure:

a. *brief, holistic scales of reporting overall proficiency*: providing short, holistic, and user-friendly statement for each level.

b. *user scales reporting proficiency in different contexts of use*: describing language proficiency from a functional rather than skill orientation.

c. *detailed, holistic rating scales*: providing holistic, but more detailed statements as compared with the first type of scales. This category of scales, including the ACTFL Guidelines, and ILR Skill Level Descriptions, account for 90% of existing literature of language proficiency scales.

d. *detailed, analytic rating scales*: providing analytic descriptors from multiple aspects/factors (e.g. grammar, fluency, vocabulary, etc.) instead of a holistic statement.

e. *frameworks of syllabus content and assessment criteria for stages of attainment*: containing detailed content, outcome specifications and providing guidelines in the format of lists of tasks/functions/structures/vocabulary for stages of attainment
in education systems.

A number of proficiency scales have been developed for L2 teaching, learning, and testing. In the current study, in order to construct a CSL Reading Proficiency Scale, the researcher collects existing descriptors from following widespread L2 proficiency scales: Canadian Language Benchmarks, World-class Instructional Design and Assessment (WIDA) English Language Development Standards, The American Council for the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines, and Interagency Language Roundtable (ILR) Language Skill Level Descriptions. These scales claim to have cross-language applicability and do not address any specific language. In addition, two Chinese-specific scales are examined for the purpose of descriptor collection: ACTFL Chinese Proficiency Guidelines, and Mandarin Chinese: Four-Year Instructional Goals, Curriculum Outline, and Instructional Measures. For this current study, among existing scales, the ACTFL Chinese Proficiency Guidelines contributed the majority of descriptors for the scale development.

This section reviews how the construct of L2 reading comprehension is operationalized in TOEFL 2000 Reading Framework (Enright, 2000), a test specification, and three widely used L2 proficiency scales: the Common European Framework of Reference for Languages (CEFR), the American Council for the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines as well as the ACTFL Chinese Proficiency Guidelines, and the World-class Instructional Design and Assessment (WIDA) English Language Development Standards.

Three broad perspectives often used in defining the construct of reading comprehension for assessment purposes are: a processing perspective, a task perspective,
and a reader purpose perspective. Much of theoretical works on the construct of reading is processing-based, which depicts the process of reading and its component skills from the cognitive perspective. One assumption of the task perspective is that reading ability can be represented by tasks that readers can accomplish. The third way to conceptualize the construct of reading is to examine reader purposes, for instance, reading for basic information, and reading for learning.

### 2.4.1 TOEFL 2000 Reading Framework

Test of English as a Foreign Language (TOEFL) is a standardized English proficiency test for non-native speakers wishing to enroll in English-speaking universities. TOEFL 2000 Reading Framework (Enright, 2000) documents the conceptual framework of reading comprehension from a **reader purpose perspective** and it also specifies **task characteristics** and **text features** of TOEFL reading tests. Four purposes for reading in the academic context are measured in the TOEFL test: reading to find information, reading for basic comprehension, reading to learn, and reading to integrate information across multiple texts. These four purposes form a natural hierarchical scale describing a continuum of reading proficiency.

Following TOEFL reading tasks are used for each reader purpose: (1) reading to find information and reading for basic comprehension: identify, interpret; (2) reading to learn: summarize, define, describe, elaborate, illustrate; (3) reading to integrate information across multiple texts: compare/contrast/classify, problem/solution, explain/justify, persuade, narrate.

TOEFL 2000 Reading Framework (Enright, 2000) specifies **situation task characteristics** of TOEFL reading tests from four aspects: participants, setting, content,
and register. Situation task characteristics are defined as extralinguistic elements associated with language tasks (Crystal, 1991, 1992). Participants of TOEFL reading tasks are readers and writers of reading passages on the test. Setting is defined as the place where the language act happens (Jamieson et al., 1999). For reading assessment, variations in settings do not change nature of reading and difficulty of reading tasks. Therefore, this variable is not considered in the development of TOEFL reading tests. The TOEFL test covers general academic topics, such as arts, social science, and natural science, which readers can understand without specialized knowledge of a particular subject. The register is viewed as functional uses that texts serve.

TOEFL 2000 Reading Framework (Enright, 2000) discusses following features of text material used on the test: grammatical/discourse features, pragmatic features, and rhetorical features.

**Grammatical/discourse Features.** While syntactic information is regarded as a source of reading difficulties, very little research has identified particular syntactic variables that should be included as task development variables for the purpose of testing. Vocabulary can largely determine the text and item difficulty. Formality measures, type-token measures, and measures of the percentage of unknown words are indices of text readability. Other factors relating to vocabulary include: collocability, functional limits according to use and situation, syntactic behavior (parts of speech, sub-categorization, case roles, transitivity), basic forms and derivational possibilities and typical occurrences, associational patterns with other words in domains of knowledge and use, idiosyncratic features of specific words, learning difficulties of certain words (e.g., similar-looking words with different meanings), and degree of abstractness-concreteness (Richard, 1976;
Nation, 1990). Contributors to comprehension difficulties can also be found at discourse level: transitional discourse markers, the roles of themerheme structuring, given and new information, definiteness and indefiniteness, noun-predicate density in texts, the positions of main idea and topic sentences in texts.

Variation in linguistic parameters can manipulate the difficulty of tasks. Specifically, TOEFL 2000 Reading Framework (Enright, 2000) lists linguistic variables that can influence task difficulty: vocabulary, syntactic complexity, transition markers, antecedent reference, modality, amount of text and amount of time allowed, distances across text when cycle or integration is involved, competing linguistic distracters in the text environment, cohesion determiners (e.g., “We bought a camera. The lens was cracked.”), grammatical relations as referents, and cohesion.

Pragmatic/rhetorical Features. Reading passages for the TOEFL 2000 test can be classified according to their dominant pragmatic and rhetorical features. While the pragmatic features are about the primary intent of the author, the rhetorical features specify the higher level text structure. Types of text materials used in the TOEFL tests include exposition, argumentation/persuasion/evaluation, and historical/biographical/autobiographical narrative. In addition to their pragmatic intent, top-level rhetorical patterns of texts are also taken into account. Rhetorical patterns used in TOEFL reading test materials include: definition, illustration, classification, comparison/contrast, problem/solution, and analysis.

2.4.2 The CEFR Scale

The CEFR was officially launched by Council of Europe in 2001. The CEFR aims to provide a common basis for the elaboration of language syllabi, curriculum guidelines,
teaching and learning materials, and foreign language proficiency assessment across Europe (Council of Europe, 2001). The core of the CEFR consists of a set of levels and corresponding descriptors that represent communicative L2 ability features (North, 1995; North & Schneider, 1998). Three components of communicative competences (linguistic competence, sociolinguistic competence, and pragmatic competence) are addressed. The CEFR includes six major levels (A1 and A2, B1 and B2, C1 and C2) and three “plus” levels (A2+, B1+, B2+). The CEFR levels and associated descriptors denote progress in language attainment by skill (listening, writing, reading, spoken interaction, and spoken production) in four domains/contexts (educational, occupational, public, and personal domain). A companion volume (Council of Europe, 2018) has been released very recently to add new descriptor scales for mediation, plurilingual competence and to fully define “plus” levels. The CEFR descriptors have cross-language applicability and are applicable to all languages.

The CEFR scale of reading ability describes reading competence from following perspectives: length of text, content and context, genre. For instance, basic users of the language (A1) can understand words and simple sentences on posters, notices, and catalogues. Independent users (B1) are able to understand texts written in everyday language and can understand descriptions in personal letters. Proficient users (C1) can easily read all forms of materials, including abstract and complex texts (Council of Europe, 2001).

The CEFR reading scale describes reading competence mainly from the features of texts, such as length, genre, and content. However, it does not address reading ability from the perspective of cognitive processing. Specifically, the CEFR scale does not
address cognitive processes at the lower level (e.g., syntactic parsing, and semantic-proposition encoding) or the higher level (e.g., constructing a text model of comprehension, using background knowledge to generate a situation model of reader comprehension, and using cognitive strategies to facilitate reading comprehension). In addition, it is not grounded on theoretical works of L2 ability. As pointed out by Fulcher (2004), the construction of CEFR was not based on large scale SLA studies and lacked a clear linguistic theory. Therefore, while the CEFR grid provides a general can-do statement for readers’ self-evaluation, it is not specific enough for the purpose of test development. In addition, to have cross-language applicability, the CEFR Common Reference Levels were developed without referencing individual languages. As a result, the CEFR descriptors are insufficiently precise for the assessment of individual languages, especially in terms of their respective linguistic features and associated cognitive processes. Although the CEFR Reference Level Descriptions (RLD) for National and Regional Languages were developed without referencing individual languages, the RLD for National and Regional Languages are currently only available for eight European languages, which do not directly address Chinese language.

### 2.4.3 The ACTFL Proficiency Guidelines

ACTFL is a membership organization, which plays a leading role in professional foreign language activities in the U.S., including teacher development and proficiency testing. In 1986, ACTFL published the first complete edition of the ACTFL Proficiency Guidelines, based on an adaptation of the U.S. federal government Interagency Language Roundtable (ILR) Skill Level Descriptions as well as the Foreign Service Institute (FSI) Scale. The ACTFL Proficiency Guidelines (ACTFL, 1986, 1999, 2001, 2012) specify
“what individuals can do with language in terms of speaking, writing, listening, and reading in real-world situations in a spontaneous and non-rehearsed context” (Swender & Vicars, 2014, p. i). For each skill, proficiency is divided into five major levels: Distinguished, Superior, Advanced, Intermediate, and Novice. Advanced, Intermediate, and Novice are further divided into three sublevels: High, Mid, and Low (ACTFL, 2012). These proficiency levels are presented as ranges, indicating what individuals at certain levels can and cannot do with language.

The ACTFL Proficiency Guidelines-Reading (ACTFL, 2012) denote reading ability primarily from the features of text (i.e., genre, content, and length of text) and from a task perspective. For instance, Novice-level readers can understand words and phrases and their comprehension rely heavily on background knowledge and extralinguistic support. At the Intermediate level, readers can understand strings of sentences containing high-frequency vocabulary. They are more accurate when reading simple texts with a predictable pattern of presentation. They are able to understand texts that convey basic information but are not able to understand texts that are detailed. At the Advanced level, readers can understand both the main idea and supporting details of authentic narratives and descriptions. They are also able to recognize the main argument of argumentative texts with familiar subject matters. They can understand texts with a clear and predictable structure and subject matters pertaining to real-world topics. However, it is challenging for them to read texts in which issues are treated abstractly. Advanced-level readers can compensate for limitations in lexical and structural knowledge by using contextual clues and knowledge of the conventions of the language.

As declared by ACTFL, “by describing the tasks that readers can perform with
different types of texts and under different types of circumstances, the Reading Proficiency Guidelines describe how readers read texts and retrieve information” (ACTFL, 2012, para. 2). However, “these Guidelines do not describe how reading skills develop, how one learns to read, nor the actual cognitive processes involved in the activity of reading. Rather, they are intended to describe what readers are able to understand from what they read” (ACTFL, 2012, para. 2).

As compared with CEFR, the ACTFL Proficiency Guidelines-Reading (ACTFL, 2012) provide more detailed descriptions from the perspective of text feature. However, the ACTFL Guidelines-Reading cannot be used directly for the assessment of CSL reading. First, although the ACTFL Proficiency Guidelines included more details about reading comprehension as compared with the CEFR, the ACTFL Guidelines are still too broad for the development of CSL reading assessment. As pointed out by Bachman & Palmer (1996, 2010), the ACTFL Guidelines should include more detailed descriptions of language knowledge (i.e., grammatical knowledge, textual knowledge, functional knowledge, and sociolinguistic knowledge) associated with tasks at each level. Second, in the ACTFL Guidelines-Reading (ACTFL, 2012), descriptors are presented as loosely connected sentences or phrases. Descriptions at some levels are repetitive and loosely-structured. As a result, it is difficult for users to figure out the relationship among descriptors and to figure out the underlying grading criteria. Last but not the least, the construction of ACTFL Proficiency Guidelines is based on intuitive identification of key features at different proficiency levels and ACTFL did not provide empirical evidence to support the scaling of descriptors (Alderson, 2000; North & Schneider, 1998). The lack of empirical verification in scale construction has led to controversies over the validity

The ACTFL Guidelines, the same as its predecessor ILR Skill Level Descriptions, are originally designed to be applicable to all languages and lack language-specific descriptors. In 1980s, ACTFL published a series of language-specific guidelines, including the ACTFL Chinese Proficiency Guidelines. However, this series of language-specific scales are not updated with the generic ACTFL Guidelines and are not available on the ACTFL official website. As a result, ACTFL language-specific scales are not well known among foreign language educators and testers. The ACTFL Chinese Proficiency Guidelines exemplify in Chinese the characteristics of each level of the generic ACTFL Guidelines. For each level, details on Chinese-specific grammatical and textual knowledge, strategic competence are provided. As indicated in its preface, ACTFL Chinese Proficiency Guidelines are not based on a particular linguistic theory. This scale suggests samples of CSL ability, but is not an exhaustive list or a discrete set required at a particular level (ACTFL, 1987). As a result, some components of CSL reading ability construct (e.g., orthographic knowledge, word decision ability, knowledge of discourse patterns) are not well represented in the ACTFL Chinese Proficiency Guidelines. This leads to a problem of construct underrepresentation. While this Guidelines contribute a large number of Chinese-specific reading proficiency descriptors for scale development in the current study, the researcher still needs to create over around forty descriptors to depict reading skills/knowledge which are not covered in the ACTFL Chinese Proficiency Guidelines.

2.4.4 WIDA English Language Development Standards
WIDA English Language Development (ELD) Standards (WIDA, 2012)

operationalize language use and development of English language learners (ELL) in K-12 context for the purposes of instruction and assessment. The WIDA ELD Standards are represented in a five-level language proficiency progression scale with three dimensions of criteria: Linguistic Complexity, Language Forms and Conventions, and Vocabulary Usage. The criteria are framed within the sociocultural context and encompass additional criteria, such as Register, Genre/Text type, Topic, Task/Situation, and Participants’ identities and social roles. The WIDA ELD Standards include two sets of performance definitions: one set is for receptive language skills (i.e., reading and listening) and the other for productive language skills (i.e., speaking and writing).

Level 1 (Entering Level) readers can understand everyday general content-related words while Level 5 (Bridging Level) readers have a good master of technical and abstract content-area language as well as words with shades of meaning across content areas. In terms of language forms and conventions, Level 1 readers can master simple grammatical structures and common social and instructional forms. Level 5 readers have a good command of complex grammatical constructions and sentence patterns characteristic of particular content areas. At linguistic complexity dimension, Level 1 readers can understand single statements or questions and ideas expressed in words, phrases or chunks of language. Level 5 readers are able to read discourse with rich descriptions and complex sentences. They are able to understand cohesive, organized, related ideas across content areas.

Compared with the CEFR and the ACTFL Proficiency Guidelines, the WIDA ELD Standards describe continuum of proficiency in a multi-dimensional approach. The WIDA
**ELD Standards** have three linguistic-oriented categories: Linguistic Complexity, Language Forms and Conventions, and Vocabulary Usage. In addition, sociocultural criteria, such as Register and Genre are also highlighted. Due to the clear presentation of assessment criteria and structure, the *WIDA ELD Standards* are relatively more user-friendly to teachers and test developers. The *WIDA ELD Standards* is structured mainly from a text-feature perspective, although it includes some descriptors on cognitive processes of reading ability.

**2.4.5 Limitations of Existing Scales**

Existing reading proficiency scales are not directly applicable to the development of CSL reading assessment. First and foremost, existing scales describe reading ability form pragmatic approaches, such as text feature, tasks, and reader purpose perspectives. They do not address the actual cognitive processes involved in reading comprehension and skills-processing variables are not well represented in these scales. However, classroom teaching and assessment are more concerned with the cognitive processes of reading. Therefore, existing reading proficiency scales are not specific enough for assessing students’ reading processes and reading development. In addition, much of the theoretical works on reading comprehension is processing-based, which analyze reading ability and its component skills from a psycholinguistic perspective. While the text-based and task-based approaches make it easy to communicate the scales to non-specialist users, the lack of processing-based descriptors leads to a discrepancy between proficiency scale development and the latest theoretical research findings.

Second, as pointed out by previous researchers, some existing proficiency scales (e.g., the CEFR) are not based on large scale SLA studies and lacked a clear linguistic or
measurement theory (Fulcher 2004, 2010, 2012). Both the *ACTFL Proficiency Guidelines* and the *CEFR* “lack what most researchers and developers consider to be a content domain based on accumulated theoretical knowledge and grounded research” (Chalhoub-Devile, 2009, p. 250). Some proficiency scales (e.g., the *Canadian Language Benchmarks*) assume their theoretical foundations with communicative competence and language development theories. However, their theoretical ground is on general language ability theories which are applicable to all modes including listening, speaking, reading, and writing, but not focuses on reading theories (Zeng & Fan, 2017).

Third, the majority of current reading proficiency scales are not specific for CSL. They are either developed for EFL/ELL reading or claim to be applicable to all languages. Proficiency scales designed to have cross-language applicability are too broad for the assessment of CSL reading. A Chinese-specific proficiency scale that particularly addresses linguistic and cognitive features of CSL reading processes and directly associated with college-level CSL curriculum is needed.

Fourth, in terms of scale construction, existing proficiency guidelines mostly rely on intuitive identification of key characteristics at different proficiency levels ranked in a consensus order (e.g., the *ACTFL Guidelines*). More objective techniques, such as statistical analysis, should be adopted to examine the actual difficulty levels of descriptors and to calibrate descriptors empirically.

Lastly, over 90% existing scales are holistic rating scales (North, 1993). Analytic scales, which are more user-friendly and present descriptors in a more categorized format, are in need. For example, in the *ACTFL Guidelines*, descriptors for a given level are in a paragraph composed of loosely connected sentences. It is very difficult for users to
understand the relationship among descriptors or to figure out assessment criteria when using holistic scales. Descriptors should be classified into distinctive categories (e.g., grammar knowledge, character recognition, reading strategy) in order to make the underlying assessment criteria more explicit.

In sum, it is very necessary to develop a proficiency scale for the purpose of assessing CSL reading ability, which is both theoretically solid and empirically verified. Developing a reading proficiency scale that well presents CSL linguistic and processing variables will benefit instruction and assessment of CSL reading ability fundamentally.

2.5 Scale Development and Descriptor Writing

2.5.1 Data-based Scale Development Approaches

Proficiency scales present a hierarchy of descriptors for levels on the continuum of proficiency. Composing scales of proficiency is an approach to operationalizing theoretical constructs of language ability. Three types of scale development methodologies are: intuitive approach, qualitative method, and quantitative approach.

The majority of language proficiency scales have been developed by appeal to intuition (North & Schneider, 1998). For instance, the ACTFL Proficiency Guidelines were produced by an expert committee through intuitive identification of key features at each proficiency level. Commonly used qualitative methodologies include expert review of existing scales and literature review. In the quantitative approach, scale development and validation are conducted on the basis of statistical analysis of data generated from large-scale application of the scales. Every approach has pros and cons. Especially, the intuitive identification has been criticized since it lacks empirical verification and validity evidence. The intuitive approach may be appropriate for the development of scales for
low-stakes purposes. However, it is inappropriate to construct national proficiency frameworks through an intuitive approach (Bachman & Savignon, 1986; Malone, 2003).

Criticism on the intuitive approach has led to data-based solutions to proficiency scale development. In data-based approaches, empirical evidence is drawn from performance samples, test items that operationalize descriptors, or large-scale application of draft scales to verify the identification and scaling of proficiency descriptors. Among these data-based solutions, the methodology used in the development of the CEFR scale (North, 1995, 2005; North & Schneider, 1998; North, 2000) has gained popularity and has been adopting by scale development projects worldwide. For instance, CEFR-J project in Japan and the development of China Standards of English Language Ability all adopt this approach (He & Chen, 2017). This data-driven methodology adopts both qualitative and quantitative techniques and consists of three steps:

1. **Comprehensive documentation.** The researchers comprehensively reviewed existing language proficiency scales based on which to a pool of descriptors was created.

2. **Qualitative validation.** Experienced teachers were consulted to ensure that the descriptors and provisional groups are relevant and usable to teachers.

3. **Quantitative validation.** Descriptors were compiled into questionnaires and teachers rated their students’ proficiency against these questionnaires. These descriptors then were scaled through Rasch rating scale analysis of teacher ratings on questionnaires (North, 1995; North & Schneider, 1998; North, 2005).

As compared with other data-driven methodologies, this particular method is of great advantages for the current study. First, experienced teachers were involved in the
qualitative validation and the scaling of proficiency descriptors. This procedure captures teachers’ collective interpretation of the construct and of the difficulty of descriptors. Poorly written descriptors or descriptors that are irrelevant to teaching can be eliminated. It ensures comprehensibility and content relevance of the scale being constructed. To the current study, which is intended to develop a CSL reading proficiency scale with teachers as the primary users, this data collection procedure is of great advantage.

Second, the quantitative validation is explicitly grounded in a measurement model, i.e., Rasch rating scale analysis. The quantitative validation starts with a large scale application of the scale among foreign language teachers. The “difficulty value” denoted by a descriptor is then established through statistical analysis (i.e., Rasch rating scale analysis) of the way teachers actually use that descriptor. In this way, teachers’ collective interpretation of the difficulty of these descriptors is being scaled. This process guarantees that the scaling of descriptors reflects the collective interpretation of a large and representative sample of users. The scaling is objective since the difficulty of descriptors is not determined subjectively but is established on the basis of large-scale data. In addition, the Rasch analysis can help identify descriptors that do not “fit” very well with the main construct. This further enhances the relevance of descriptors.

In sum, the methodology used in the construction of the CEFR functions as an example for proficiency scale development projects worldwide. The current study also adopted this data-driven methodology since it well serves the research purpose. Details about how this methodology was applied in the present study are provided in Chapter 3.

2.5.2 Rasch Rating Scale Analysis

As discussed previously, in the development of the CEFR, the difficulty level
denoted by a descriptor is established through Rasch rating scale analysis of teachers’ collective interpretation. This section elaborates Rasch Rating Scale Model and its application in proficiency scale development.

Classical Test Theory (CTT) is not effective in predicting the performance of individuals given a certain proficiency level because CTT highly depends on item sample and test taker sample. Item Response Theory (IRT) addresses limitations of CTT and has been gradually an alternative to it. IRT model can arrive at a generalizable result which is sample independent (provided samples can be considered as sub-sets of the same statistical population). IRT is developed on the basis of probability theory. In IRT model, a test taker’s performance can be estimated through the calibration of the ability level of the test taker and the item characteristics (e.g., item difficulty, discrimination).

Rasch model (Rasch, 1960) is one of the most basic Item Response Theory (IRT) models. As indicated in the following formula, the probability that test taker \( j \) can answer item \( i \) correctly is estimated by the item difficulty \( (b_i) \) and his or her ability \( \theta \) (Rasch, 1960).

\[
P(Y_j = 1 \mid \theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}
\]

\( b_i \): the difficulty of item \( i \)
\( \theta \): ability of test taker \( j \)

In Rasch analysis, both person ability measures and item characteristics measures are calibrated on the same logit scale. The units of Rasch measurement are called “logits”. Logits express not only where an item is on the single variable being measured but also where a person is located on that same variable (Linacre & Wright, 1989). Therefore, in Rasch analysis, persons can be compared to other persons (e.g., Mary has a higher level
of proficiency than Mike); items can be compared to other items (e.g., item 1 was easier than item 10); items and persons can also be compared (e.g., a student from 4th-year Chinese class has a high likelihood of agreeing with item 1. A student from 1st-year Chinese class has a low likelihood of agreeing with item 1).

IRT is mainly used to determine the difficulty of individual test items in an item bank (Council of Europe, 2001). However, Rasch model can also be used to plot the “difficulty value” of individual descriptors on the continuum of proficiency. For instance, questionnaire items used to describe students’ reading ability (e.g., “Can understand only simplest language with high-frequency oral vocabulary and structure.” is an example of questionnaire items) can be scaled by Rasch model. The descriptors can be calibrated directly onto an arithmetic scale in the same way that test items are scaled in item banks.” (Council of Europe, 2001, p.211). As compared with other IRT models, the strength of the Rasch approach also includes that it can detect and exclude messiness in data (i.e., items or persons which do not fit the model).

Three key assumptions of Rasch analysis are: unidimensionality, local independence, and no guessing. Unidimensionality assumes that all items in an instrument measure the same latent trait. Henning (1992) argues that psychological unidimensionality and psychometric unidimensionality are two different concepts. For instance, foreign language proficiency, which involves multiple categories of knowledge and processes, would be a sufficiently robust construct to cope with the psychological multi-dimensionality. However, psychological constructs which are highly intercorrelated, such as those underlying most language tests will tend to build a psychometrically unidimensional construct (Henning, 1992). Although underlying
psychological constructs contributing to foreign language proficiency are multidimensional, results for the same individuals in the assessment data still tend to be intercorrelated, building a single psychometric dimension (Henning, 1992; McNamara, 1996). Psychometric unidimensionality not necessarily implies that the psychological constructs involved is unidimensional. It only means that test takers are progressing in all underlying psychological constructs of foreign language proficiency in a very even manner. In previous language testing research (e.g., Henning, Hudson, & Turner, 1985), foreign language proficiency proved to be a sufficiently robust construct to cope with the psychological multi-dimensionality which also displays enough psychometric unidimensionality to justify calibration of descriptors in various content strands. That is why some foreign language proficiency scales, such as the CEFR, addressing multiple modes of communication (e.g. listening, speaking, and writing) still display enough psychometric unidimensionality for the application of Rasch Model. Exploratory factor analysis or confirmatory factory analysis are often used to test if unidimensionality assumption is met before using Rasch analysis.

Local independence assumption requires that for a given learner ability estimate \(\theta\), the probability of correctly responding to two or more items is equal to the product of the separate success probabilities. In other words, items should be independent and it is not necessary to answer one question correctly in order to get another question correct. This issue is not quite relevant to the development of proficiency scales (North, 2000). Lastly, the guessing problem is mostly associated with true/false and multiple-choice items, in which test takers can make get an item correct by guessing. This problem is not very relevant to scale development (North, 2000).
2.5.3 Descriptor Writing

“Can-do” style of statements describes what a person at a certain proficiency level is able to do with language. For instance, “can appreciate aesthetic values in a literary work with rich connotations” is a typical “can-do” statement for reading ability. Since they first appeared in ALTE (Association of Language Testers in Europe) “Can Do” project in 1990s, can-do descriptors are widely used in language proficiency scales worldwide. The CEFR and China's Standards of English Language Ability (CSE) both adopt the format of can-do statement in descriptor writing.

In the CSE project, a can-do statement includes following elements: language performance, criteria, and condition (Zeng, 2017). Criteria indicate the intrinsic quality of language performance, such as accuracy, fluency, degree of ease. Condition (i.e., extrinsic quality) is about the circumstance under which the performance is completed. In the following example, “Can comprehend material rich in content through comparing with similar articles”, “can comprehend material” is language performance. “Rich in content” denotes the intrinsic quality of language performance and “through comparing with similar articles” is the condition or extrinsic quality.

According to North (2000), a good can-do descriptor has following qualities:

(1) Positiveness. Statement should be written with positive orientation.

(2) Definiteness. “Descriptors should describe concrete features of performance, concrete tasks and/or concrete degrees of skill in performing tasks.” and avoid vagueness (North, 2000, p.345). For instance, “Can understand the function of some conjunctive devices” is a vague statement.

(3) Clarity. Descriptors should be written in simple syntax, avoiding multiple
clauses and jargon.

(4) Brevity. Short sentence is preferred than lengthy paragraph. Each descriptor should be no longer 20 words. Long descriptors should be split into several shorter ones otherwise cannot realistically be referred to in testing.

(5) Independence. Independence criterion requires that descriptors should be unitary and unique. Specifically, each descriptor only describes a unitary ability and there should be no overlap among descriptors in terms of content.

Summary

This chapter has reviewed and discussed: (1) theoretical construct definitions of communicative language competence and L1/L2/CSL reading as conceptualized from a multicomponential view; (2) operational construct of L2 reading as operationalized in existing language proficiency scales or test specifications; (3) rationales and limitations of some widespread L2 reading proficiency scales; and, (4) issues related to scale development, including orientations, methodologies, descriptor writing and calibration.

Owing to the interdisciplinary nature of this dissertation, seminal works in both language testing field and second language acquisition arena are included in this literature review.

First, on top of reviewing evolution of construct definitions of language ability with a focus on the Communicative Language Ability model (Bachman, 1990; Bachman & Palmer, 1996, 2010), L1, L2 and CSL reading construct are examined one after another mainly from a component skills approach. The general consensus of the research community about vital components of L1 reading ability as well as their interrelationship is summarized and visualized in Figure 6. L2 reading differs from L1 reading in many ways. From a cross-linguistic perspective, Koda (2005) reexamined impact of L1 linguistic property, L1-L2 distance, and L1-L2 involvement on vital component
skills/knowledge of L2 reading. To this dissertation, Koda’s (2005) work provides a framework for the mining of construct definitions of CSL reading from seminal theoretical and empirical literature. Finally, CSL reading is modeled as a multicomponential concept with Chinese-specific linguistic properties and corresponding cognitive tactics highlighted (Figure 8, 9, 10). In the current study, this representation of CSL reading construct serves as theoretical underpinnings for the *CSL Reading Proficiency Scale*. On one hand, it guides descriptor collection, creation, and selection, enhancing the possibility that all component skills/knowledge are well presented by reading proficiency descriptors included in the *CSL Reading Proficiency Scale*. On the other hand, it provides a provisional classification system against which descriptors in the descriptor bank can be categorized. All in all, the literature review (reported in Section 2.1, 2.2, 2.3) on construct of communicative language ability, L1 reading, L2 reading, and especially CSL reading, lays the theoretical ground for the *CSL Reading Proficiency Scale*.

Second, the summary of theatrical construct of reading follows by a review of how theoretical construct of L2 reading is operationalized in widely-used test specifications and language proficiency scales (Section 2.4). These documents are studied from following aspects: rationales, methodologies, band levels, and descriptor classification. As compared with each other, strengths and limitations of each instrument are pointed out. The review of existing reading proficiency scale identifies gaps which need to be filled in by developing the *CSL Reading Proficiency Scale*. It also justifies the significance of this dissertation study.
Lastly, techniques used in the development of the *CEFR* scale are reexamined closely with the purpose to provide methodologies for the development of the *CSL Reading Proficiency Scale*. Scale development procedures adopted by the *CEFR* project have multiple strengths: (1) the qualitative validation involves participation of classroom teachers and this procedure enhances the content relevance of the scale to be developed; (2) the quantitative validation is grounded in a measurement model (Rasch rating scale analysis). Rationales and assumptions of the Rasch model are discussed in detail; (3) the *CEFR* descriptors are written in the format of “can-do” style of statement which specify the intrinsic quality and the extrinsic circumstance of language performance. In this dissertation, methodologies for developing the *CSL Reading Proficiency Scale* mainly follow the *CEFR* procedures with some limitations addressed. Chapter 3 provides details on the *CSL Reading Proficiency Scale* development and validation methodology.
CHAPTER 3

METHODOLOGY AND DATA

This study was designed to operationalize the CSL reading construct for assessment purposes through developing a *CSL Reading Proficiency Scale*. A data-based approach was used to develop and validate the *CSL Reading Proficiency Scale* empirically. The development of this scale started with documenting reading proficiency descriptors from existing scales and creating new descriptors based on reading theory. To classify and scale these descriptors, following research questions were examined:

- **Dimensions of descriptors:** 1. What is the factor structure of CSL reading, as measured by descriptors in the *CSL Reading Proficiency Scale*?
- **Difficulty levels of descriptors:** 2. What is the estimated level of difficulty of each descriptor? 3. How are cut-off points selected to divide descriptors into multiple bands?

Subsequently, this chapter first discusses the research design in light of the theoretical framework. The author then continues with a discussion of instruments, sampling, and data collection. Last, statistical analysis procedures are laid out.

**Review of Theoretical Framework**

Three theoretical premises form the basis of the study’s research design: a) the notion of Communicative Language Ability (CLA); b) component skills approach of reading research; and, c) data-based scale construction methodology.
The Communicative Language Ability (CLA) and component skills approach of reading research underlie the modeling of the CSL reading ability. CLA (Bachman, 1990; Bachman & Palmer, 1996, 2010) is an outstanding representative of communicative language competence models. In CLA, language ability is operationalized in terms of two principal components: language knowledge and strategic competence. Language knowledge consists not only of organizational knowledge on language itself but also of pragmatic knowledge, which enables language users to put together language forms to realize specific communication goals in specific language use situations. Strategic competence, which refers in particular to the ability to use metacognitive strategies in the CLA model, is what language learners use to plan, regulate, and modify their cognitive processes. In the current study, the construct of reading ability is modeled within communicative language competence theory (especially, the CLA model). Specifically, reading proficiency descriptors included in the CSL Reading Proficiency Scale not only give details of grammatical knowledge and textual knowledge involved in reading but also denote pragmatic knowledge.

The component skills approach of reading research conceptualizes reading as a constellation of multiple subskills. This approach provides a very productive theoretical basis for language assessment research and practice (Bachman & Cohen, 1998) and for curriculum design and classroom teaching (Zeng, 2017). In the current study, CSL reading is modeled as a multi-componential construct. Specifically, theoretical works consulted in this study all address reading ability from a multi-componential view. Descriptors are selected carefully to make all vital subskills of CSL reading well represented. Based on statistical analysis results, descriptors are then classified and scaled to denote the underlying relationship and hierarchy of component skills.
Lastly, the data-driven approach used in the development of the CEFR (North, 1995, 2005; North & Schneider, 1998; North, 2000) is regarded as a good example of scale construction methodologies worldwide (He & Chen, 2017). The CEFR scale construction started with a comprehensive documentation of existing language ability descriptors. The descriptor pool was then validated both qualitatively and quantitatively (North, 1995, 2005; North & Schneider, 1998). Language teachers were involved in both the qualitative and quantitative validation to assure that the scale reflected teacher perceptions of language proficiency. The current study adopts this approach while also addressing its limitations. Details on research design are provided in the following section.

**Research Design and Procedures**

A data-driven approach consisted of six steps was used to develop and validate the CSL Reading Proficiency Scale. This approach was designed based on the methodology used to develop the CEFR while also addressed the following limitations: (1) The CEFR was not developed on the basis of large scale SLA studies and lacked a clear linguistic theory (Fulche, 2004; North, 2014). In the current study, the construct of CSL reading was grounded in CLA (Bachman, 1990; Bachman and Palmer, 1996, 2010) and theoretical works that explored the nature of reading from a component skills approach (e.g., Alderson et al., 2015; Koda, 2005). (2) The CEFR descriptors were documented mainly from existing proficiency scales and frameworks. Most descriptors addressed reading ability from pragmatic approaches, such as text-based or task-based approaches. Very few descriptors depicted the actual cognitive processes involved in reading comprehension. In the current study, the researcher created new descriptors based on theoretical works and empirical studies on reading in addition to collecting descriptors from existing proficiency scales. This procedure enabled the linguistic features and cognitive
processes of CSL reading to be well represented.

<table>
<thead>
<tr>
<th>Scale Development Preparatory Stage</th>
<th>Step 1: Modeling CSL Reading Ability Construct</th>
<th>1(a): The researcher reviewed literature on reading ability in order to establish the theoretical base.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2: Establishing an Operational Descriptive Scheme with Four Parameters</td>
<td>2(a): The researcher elicited four parameters for describing reading ability.</td>
<td></td>
</tr>
<tr>
<td>Step 3: Developing a Descriptor Pool</td>
<td>3(a): The researcher selected 89 reading descriptors from existing scales or theoretical works. 3(b): The researcher edited descriptors into can-do statements with predetermined descriptive parameters. 3(c): The researcher classified 89 selected descriptors into nine provisional groups.</td>
<td></td>
</tr>
<tr>
<td>Step 4: Qualitative Validation</td>
<td>4(a): Four experienced CSL teachers reviewed descriptors in terms of clarity, brevity, accuracy, independence, and content relevance. Revisions were made based on their feedback. 4(b): Revised descriptors were sent to teacher participants a second time to see if they needed to make more revisions. Finally, 57 descriptors survived from the expert review. 4(c): The researcher compiled 57 surviving descriptors into the CSL Reading Proficiency Questionnaire.</td>
<td></td>
</tr>
<tr>
<td>Step 5: Pilot Study</td>
<td>5(a): The researcher distributed the Questionnaire to a small group of experienced CSL teachers. 5(b): The internal consistency of the Questionnaire was examined based on 46 responses from this small group of participants.</td>
<td></td>
</tr>
<tr>
<td>Quantitative Validation Stage</td>
<td>Step 6: Quantitative Validation</td>
<td>6(a): The researcher distributed the Questionnaire to 95 experienced CSL teachers. 6(b): 179 responses, in total, were collected from 95 teacher participants. 6(c): Confirmatory factor analysis was conducted to answer Research Question 1. 6(d): Rasch analysis was conducted to answer Research Question 2. 6(e): The CSL Reading Proficiency Scale was established based on data analysis results.</td>
</tr>
</tbody>
</table>

Figure 11
Flowchart of the Research Procedure

As illustrated in Figure 11, the scale development procedure consists of following six steps. Step one to Step five constitute the scale development preparatory stage while Step six forms the quantitative validation stage.

(1) Step one: modeling the CSL reading ability construct. To measure a given language ability, we first and foremost need to model the construct to be measured by precisely defining it (Bachman, 1990). This study began with modeling the construct of CSL reading ability on the basis of dimension analysis of reading ability examined in existing theoretical works on reading (e.g., Grabe, 2009; Koda, 2005) and CLA (Bachman, 1990; Bachman & Palmer, 1996, 2010). Modeling the CSL reading ability construct is the first and very essential step in scale development because when developing a bank of descriptors (Step three), by referring to the established model of CSL reading ability construct, the researcher will be able to: (a) eliminate descriptors irrelevant to the defined CSL reading construct; (b) identify component skills that are not well represented in existing reading proficiency scales and then create new descriptors for these skills; (3) classify descriptors into provisional groups.

(2) Step two: establishing an operational descriptive scheme with four parameters. After the CSL reading construct was defined, an operational descriptive scheme was in need to operationalize the theoretical construct. With a prior project, China Standards of English (CSE) Language Ability—Reading and Listening, as the reference, the researcher selected four parameters for the operational descriptive scheme in this study: cognitive verb, object (of the verb), text type, and modifier. Please refer to Chapter 4 for details.

(3) Step three: developing a descriptor pool. The development of the descriptor pool started with a comprehensive documentation of descriptors from existing reading proficiency scales, theoretical works, and empirical studies on reading. The first source of descriptors was
existing reading proficiency scales. The researcher reviewed raw descriptors, removing repetition, and deleting descriptors that were irrelevant to the CSL reading construct defined in Step one. In other words, if a descriptor could not be classified into any component presented in Figure 8, it would not be included. In addition, descriptors denoting abilities/knowledge obviously beyond college-level CSL learners’ proficiency were excluded. For instance, distinguish-level descriptors in the ACTFL Proficiency Guideline and Level 11 and Level 12 in the Canadian Language Benchmarks were not included.

The second source of descriptors was theoretical and empirical works on L1, L2, and CSL reading. After an initial review of descriptors collected from existing language proficiency scales, it became apparent that the coverage of CSL-specific knowledge/ability was particularly weak. CSL-specific descriptors were then created with theoretical/empirical works on reading as a reference.

(4) Step four: qualitative validation. The bank of descriptors was sent to four experienced college-level CSL teachers for qualitative validation. Please refer to the following section on participants for details on how participants were recruited. Participants reviewed each descriptor in terms of clarity, brevity, accuracy, independence, and applicability in the context of U.S. college-level CSL education. They were also asked to add their own descriptors which they thought were necessary to be included. Participants provided comments and suggestions for revision through Qualtrics, a web-survey tool. Please see Appendix I for the protocol used to elicit teachers’ comments and suggestions. This procedure also ensured that descriptors were drafted in a clear, brief, and accurate way and were mutually independent. Since the qualitative validation involved the participation of experienced CSL teachers, this procedure enhanced the possibility that descriptors reflecting teachers’ perceptions of reading ability and were relevant
and useful to college-level CSL test development and teaching.

Some descriptors were revised, deleted, or added according to teacher participants’ suggestions. The revised descriptor bank was sent to participants a second time to check if all participants agreed with revisions made. Further revisions were made toward some descriptors until all participants achieved agreement.

(5)Step five: pilot study. Experienced CSL teachers were invited to use the *CSL Reading Proficiency Questionnaire* to evaluate their students’ reading proficiency. Before distributing this instrument to a large number of participants, a pilot study with the first 46 responses was conducted to examine the reliability of this instrument. Cronbach’s alpha, a measure of internal consistency, was chosen as the reliability index. If Cronbach’s alpha is larger than 0.7, it indicates that this instrument is internally consistent. If Cronbach’s alpha is lower than 0.7, item revisions or item deletions should be made to improve the internal consistency. After necessary adjustments were made according to the pilot study results, the questionnaire was distributed to a larger group of teacher participants.

(6)Step six: quantitative validation. In total, 95 experienced college-level CSL teachers used the *CSL Reading Proficiency Questionnaire* to evaluate reading ability of their students. In US colleges/universities that grant Bachelor Degrees in Chinese, four years of Chinese language courses (i.e., 1st, 2nd, 3rd, and 4th-year courses) are offered. Each teacher participant picked class levels which she/he was familiar with and rated one medium-achieving non-heritage student from each picked level against the *CSL Reading Proficiency Questionnaire*. Depending on availability and teaching experience, teacher participants could evaluate one, two, three, or four levels. In other words, if a teacher was familiar with all four levels, she/he could evaluate all these levels if she/he was available, and consequently, four responses were collected from this
participant. Including the 46 responses used in the pilot study, 179 responses from 95 teacher
participants were collected.

Teacher participants’ evaluations of student reading proficiency were made on the basis
of coursework, class observation, reading comprehension tests, or reading comprehension
assignments. Since teaching experience played an essential role in making an accurate
evaluation, only teachers with more than three-year experience of teaching college-level CSL
courses were recruited.

When collecting data, the researcher generated a separate copy of the *CSL Reading
Proficiency Questionnaire* for each class level. In this way, the researcher could better monitor
the responses and got roughly equal numbers of responses from four levels. The *CSL Reading
Proficiency Questionnaire* was distributed by using Qualtrics. Since it was infeasible to organize
an onsite workshop to inform participants how to fill in the questionnaire, at the beginning of the
*CSL Reading Proficiency Questionnaire*, very detailed instruction was provided. It helped
participants get familiar with the rating categories and rating criteria. This procedure enhanced
the possibility that participants responded to the questionnaire appropriately and inputted valid
data.

Two quantitative analysis methods, confirmatory factor analysis and Rasch analysis, were
used to examine data generated. First, confirmatory factor analysis was adopted to examine
underlying structures of the *CSL Reading Proficiency Questionnaire* empirically. Second, Rasch
analysis was used to scale descriptor items to form a hierarchy of reading ability descriptors and
to identify descriptor items that did not fit the model expectation. Details on statistical analysis
are provided in the last section of this chapter.

**The Instruments**
The CSL Reading Proficiency Questionnaire was used to collect data for quantitative validation in Step six. This instrument consisted of 57 items of descriptors on reading proficiency in addition to four demographic items. It is a Likert-scale questionnaire, in which each descriptor item was drafted as a “can-do” reading proficiency descriptor. Teacher participants used this questionnaire to evaluate student reading ability on a 5-point Likert scale ranging from zero (Can not be expected to perform like this) to four (this describes a performance which is clearly below his/her level. Could perform better than this). Table 1 presents a full description of each rating category. All items, as well as instructions, were written in both English and Chinese. Descriptors were classified into nine provisional groups as shown in Table 2. Table 3 shows an example descriptor item selected from the CSL Reading Proficiency Questionnaire as well as rating categories. In this dissertation, descriptors and items are often used interchangeably when referring to CSL reading descriptors in the CSL Reading Proficiency Questionnaire. Please refer to Appendix II for the full version of the CSL Reading Proficiency Questionnaire.

Table 1

<table>
<thead>
<tr>
<th>Rating Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>This describes a level, which is definitely beyond his/her capabilities. Can not be expected to perform like this.</td>
</tr>
<tr>
<td>1</td>
<td>Could minimally perform like this in favorable circumstances. This describes a performance which is slightly beyond his/her level. (Favorable circumstance refers to the situation in which she/he gets help or hints, has time to prepare or think over; the topics are familiar; surroundings favor his/her performance.)</td>
</tr>
<tr>
<td>2</td>
<td>Could be expected to perform like this without support in normal circumstances. His/her ability is on this level.</td>
</tr>
</tbody>
</table>
| 3               | Could be expected to perform like this even in unfavorable circumstances with efforts. This describes a performance which is slightly
below his/her level. (Unfavorable circumstance refers to the situation in which she/he does not get help or hints, has no time to prepare or think over; the topics are unfamiliar; surroundings do not favor his/her performance.)

4 Could be expected to perform like this with ease in all circumstances. This describes a performance which is obviously below his/her level.

Table 2

Basic information about the CSL Reading Proficiency Questionnaire

<table>
<thead>
<tr>
<th>Provisional Group</th>
<th>Number of Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Character Recognition and Orthographic Knowledge</td>
<td>3</td>
</tr>
<tr>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>6</td>
</tr>
<tr>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>20</td>
</tr>
<tr>
<td>4) Semantic-proposition Encoding</td>
<td>1</td>
</tr>
<tr>
<td>5) Knowledge of Coherence</td>
<td>3</td>
</tr>
<tr>
<td>6) Knowledge of Rhetorical Organization</td>
<td>4</td>
</tr>
<tr>
<td>7) The Text Model of Comprehension</td>
<td>16</td>
</tr>
<tr>
<td>8) The Situation Model of Comprehension and Pragmatic Knowledge</td>
<td>3</td>
</tr>
<tr>
<td>9) Cognitive Strategy</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3

An example descriptor item of the CSL Reading Proficiency Questionnaire

Item 2: Can recognize different forms of a particular radical, which have learned, in characters (e.g., 笔 and 心).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0: No, cannot perform like this.</td>
<td>1: Yes, in favorable circumstances.</td>
</tr>
<tr>
<td>2: Yes, in normal circumstances.</td>
<td>3: Yes, even in unfavorable circumstances.</td>
</tr>
<tr>
<td>4: Yes, with ease in all circumstances.</td>
<td></td>
</tr>
</tbody>
</table>

The CSL Reading Proficiency Questionnaire has gone through a strict examination on validity and reliability before the large-scale application. First, because this questionnaire was constructed based on a comprehensive literature review and has been validated qualitatively
through expert review, it well represents CSL reading construct. Second, before the large-scale distribution of the *CSL Reading Proficiency Questionnaire*, a pilot study with the first 46 responses was conducted to examine the reliability of this instrument. Cronbach’s alpha, a measure of internal consistency, was chosen as the reliability index. If Cronbach’s alpha was lower than 0.7, item revisions or item deletions would be made before distributing this instrument to all participants.

**Participants**

**Teacher Participants of Qualitative Validation**

Two groups of teacher participants were recruited from experienced teachers/professors who taught college-level CSL courses in American universities. First, four experienced CSL teachers were invited to do the qualitative validation of the descriptor bank. They had both expertise in CSL reading theories and abundant CSL teaching experience. Each participant had at least five years of college-level CSL teaching experience and held a Doctoral Degree in Teaching Chinese as a Foreign Language, Applied Linguistics, or Foreign Language Education. They were recruited through convenience sampling from multiple higher education institutions across the U.S.

**Teacher Participants of Quantitative Validation**

Second, 95 experienced college-level CSL teachers were invited to participate in the quantitative validation (Step six). Specifically, they used the *CSL Reading Proficiency Questionnaire* to evaluate their students’ reading ability. Teacher participants were also referred to as “respondents” in Chapter 5, which reports data analysis results. The researcher created an advertisement with a link to this questionnaire and distributed it in following ways to recruit participants: (1) the researcher posted this advertisement on the social media of CSL Research
Special Interest Group, established by The Chinese Language Teachers Association (CLTA), which is the most influential professional organization in the field; (2) the researcher contacted CLTA members individually by email and encouraged them to fill in this questionnaire; (3) the researcher also recruited participants by reaching out to alumni who received M.A. or Ph.D. degrees in Teaching Chinese as a Second Language from the University of Iowa.

In the advertisement, the researcher made it clear that participants should have at least three years’ independent teaching experience in U.S. higher-education institutions and hold M.A. or Ph.D. degrees in Teaching Chinese as a Second Language, Applied Linguistics, or Foreign Language Education.

Learners Being Rated

Every teacher participant was asked to pick familiar class levels and evaluate a medium-achieving non-heritage learner from each picked level against the CSL Reading Proficiency Questionnaire. As explained in step 6, depending on teaching experience and availability, teacher participants evaluated one to four learners. In total, 95 teacher participants evaluated 179 learners. The table below shows the number and gender of learners being rated from each class level.

Table 4

<table>
<thead>
<tr>
<th>CSL Class Level</th>
<th>Number of Learners Being Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st year (100 level)</td>
<td>45 (Female: 27; Male: 18)</td>
</tr>
<tr>
<td>2nd year (200 level)</td>
<td>48 (Female: 27; Male: 21)</td>
</tr>
<tr>
<td>3rd year (300 level)</td>
<td>43 (Female: 24; Male: 19)</td>
</tr>
<tr>
<td>4th year (400 level)</td>
<td>43 (Female: 20; Male: 23)</td>
</tr>
<tr>
<td>Total</td>
<td>179 (Female: 98; Male: 81)</td>
</tr>
</tbody>
</table>

Statistical Analysis Procedures

Multiple statistical methods were adopted for the purpose of scale construction and
validation. Before the large-scale distribution of the *CSL Reading Proficiency Questionnaire*, a pilot study was conducted to examine the reliability of this instrument through checking Cronbach’s alpha, a measure of internal consistency. Two types of statistical analyses were then used to analyze data collected through the large-scale application of the *CSL Reading Proficiency Questionnaire*. First, confirmatory factor analysis was used to uncover the factor structure of descriptors in the *CSL Reading Proficiency Questionnaire*. Second, Rasch analysis was conducted mainly for descriptor calibration. Rationales and procedures of statistical analysis were provided below.

**Cronbach’s Alpha**

Before distributing the *CSL Reading Proficiency Questionnaire* to all participants, the researcher first piloted this questionnaire with a small group of participants. The researcher examined the reliability of this instrument based on 46 responses from these participants. Reliability indicates “the degree to which test scores are free from errors of measurement for a given group.” (AERA, APA, NCME, 2014, p. 223) Reliability can be estimated in three ways: test-retest reliability, parallel forms reliability, and internal consistency coefficients. Coefficient Alpha (Cronbach’s α or rxx’α) is a measure of internal consistency. Since Cronbach’s α is applicable to polytomously-scored items (e.g., Likert-scale items) and can be interpreted as the mean of all possible split-half coefficients, it was chosen as the reliability index to test the internal consistency of the *CSL Reading Proficiency Questionnaire*.

The researcher used IBM SPSS Statistics (IBM Corp, 2017), a data analysis program, to compute Cronbach’s alpha. As presented in the table below, if Cronbach’s alpha is larger than 0.7, it indicates that the internal consistency of an instrument is acceptable. All items in this instrument measure a common construct/ability/concept. If Cronbach’s alpha is lower than 0.7,
item revisions or item deletions should be made to improve the internal consistency.

\[ r_{xx} = \frac{n_i}{n_i - 1} \left[ 1 - \frac{\sum S_i^2}{S_x^2} \right] \]

Table 5

*Interpretation of Cronbach's alpha*

<table>
<thead>
<tr>
<th>Cronbach's alpha</th>
<th>Internal consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha \geq 0.9 )</td>
<td>Excellent</td>
</tr>
<tr>
<td>( 0.9 &gt; \alpha \geq 0.8 )</td>
<td>Good</td>
</tr>
<tr>
<td>( 0.8 &gt; \alpha \geq 0.7 )</td>
<td>Acceptable</td>
</tr>
<tr>
<td>( 0.7 &gt; \alpha \geq 0.6 )</td>
<td>Questionable</td>
</tr>
<tr>
<td>( 0.6 &gt; \alpha \geq 0.5 )</td>
<td>Poor</td>
</tr>
<tr>
<td>( 0.5 &gt; \alpha )</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>


**Confirmatory Factor Analysis**

Factor analysis is a statistical method used to describe the underlying structure of a relatively large set of observed variables by extracting a potentially lower number of unobserved variables (i.e., factors). There are two types of factor analysis methods: exploratory factor analysis and confirmatory factor analysis. Confirmatory factor analysis is used to test if researchers’ hypotheses of the underlying structure of variables can be verified empirically. For instance, confirmatory factor analysis can be used to test if assumptions of dimensions of language ability are supported by empirical evidence.

A confirmatory approach to factor analysis was adopted to respond to research question one. Analyses were performed by using Mplus (Muthén & Muthén, 2010). Specifically, a series of confirmatory factor analyses were conducted to examine whether the structure of CSL reading comprehension, measured by the *CSL Reading Proficiency Scale*, conformed to one of following factor models previously established in the literature: the correlated two-factor model, the correlated three-factor model, or the one-factor model. Table 6 shows the components of each
model as well as the relationship among three competing models.

Model 1 (the correlated two-factor model) consisted of two latent factors: lower-level processes and knowledge; higher-level processes and knowledge. Thirty observed variables (i.e., descriptor item 1-item 30) loaded on factor one and the rest twenty-seven variables loaded on factor two. The two factors were allowed to covary. Model 2 (the correlated three-factor model) had three latent variables: lower-level decoding (indicated by descriptor item 1-item 29), Interim-level textbase construction (indicated by descriptor item 30-item 53), and higher-level situation-model building (indicated by descriptor item 54 to item 57). The three factors were modeled to correlate with one another. Model 3 (one-factor model) depicted a unitary-skill conceptualization of the CSL reading.

The purposes of using confirmatory factor analysis were two-fold. First, there is no consensus about the divisibility of reading ability. Three hypotheses arrived at through literature review need to be empirically verified: (1) reading is a unitary skill; (2) reading is tri-divisible; (3) reading is bi-divisible. Second, the CSL Reading Proficiency Questionnaire consisted of 57 descriptors and they were sorted into nine provisional groups, which were overwhelming and excessive to users. For the purpose of dimension reduction, confirmatory factor analysis was used to uncover a multidimensional model that is more parsimonious.

Table 6

<table>
<thead>
<tr>
<th>Three competing models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1:</strong> Correlated two-factor model</td>
</tr>
<tr>
<td>Lower-level processes and knowledge of CSL reading</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Higher-level Processes and knowledge of CSL reading</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
</tr>
<tr>
<td>4) Semantic-proposition Encoding</td>
</tr>
<tr>
<td>5) Knowledge of Coherence</td>
</tr>
<tr>
<td>6) Knowledge of Rhetorical Organization</td>
</tr>
<tr>
<td>7) The Text Model of Comprehension</td>
</tr>
<tr>
<td>8) The Situation Model of Comprehension and Pragmatic Knowledge</td>
</tr>
<tr>
<td>9) Cognitive Strategy</td>
</tr>
</tbody>
</table>

**Level of Measure.** An appropriate level of measurement is chosen based on statistical considerations and theoretical needs. The more observed variables used in an analysis, the larger sample size is needed. Using item parcel scores (e.g., task scores which are the average or sum of all item scores based on the same tasks) would help to reduce the number of observed variables and hence to reduce the sample size needed. In this study, the *CSL Reading Proficiency Questionnaire* included 57 descriptors classified into nine provisional groups. According to Wu and Wen (2011), if the purpose of a study is to explore the factor structure of an instrument but not to analyze the structural model, one should use item scores instead of item parcel scores (i.e., group scores which are the average or sum of all item scores within single provisional groups in this case). According to Loehlin (1992), to test a model with only two to four factors, the sample size should be at least 100 and 200 is preferred. The sample size of the current study was 179. The sample size was adequate but not very large. This study still used item scores as the level of measurement.
measure because the research purpose was to explore the underlying structure of the CSL Reading Proficiency Questionnaire. In total, the number of observed variables was 57 and these observed variables were all categorically scored from zero to four.

**Estimation Method.** The default estimator for confirmatory factor analysis with continuous variables is maximum likelihood (ML) in Mplus. The use of maximum likelihood (ML) assumes that the observed variables follow a continuous and multivariate normal distribution, which is not appropriate for ordinal or categorical variables. Diagonally weighted least squares (WLSMV) is specifically designed for ordinal data or categorical data (Muthén, 1993) and is the default estimator for categorical variables in Mplus. The WLSMV does not make normal distributional assumption and is the best option for modelling categorical or ordered data (Brown, 2006). Since observed variables in the current study were all categorically scored, it was decided to implement WLSMV instead of ML.

**Assessing Model Fit.** The adequacy of models was evaluated on the basis of two criteria. Selected model fit indices were used to examine the overall model fit. In addition, individual parameter estimates were checked for significance and appropriateness. According to Kline (2005), the following model fit indices were used in this study:

1. The Degrees of Freedom ($df$). The degrees of freedom ($df$) reflects how parsimonious or saturated a model is. A model’s degrees of freedom are determined by subtracting the number of free parameters from the number of unique elements in the variance-covariance structure of the data. A free parameter refers to a parameter that is free to be estimated during model estimation. On the contrary, a parameter whose value is determined without model estimation is a fixed parameter. With the same data structure, the more free parameters a model is specified to estimate, the lower the degrees of
freedom are, and the less parsimonious the model is. When model complexity decreases, model fit usually deteriorates because in a more parsimonious model there are fewer free parameters to estimate.

(2) Chi-Square Test of Model Fit. The value of the chi-square ($\chi^2$) statistic indicates the distance in fit between the variance-covariance structure of the observed data and the model-implied variance-covariance structure. A chi-square test examines the statistical significance of this distance. The lower the value is, the closer the distance is, and therefore the better the model fits the observed data.

(3) Normed Chi-Square. The Chi-square test of model fit has its limitations because it is very sensitive to sample size. When the sample size is large, the Chi-square value tends to be high. As a result, a model whose deviation from the observed data structure is not significant might be rejected by mistake (Bollen, 1989). The sensitivity of chi-square to sample size can be reduced by dividing the chi-square value by the degrees of freedom (df) (Kline, 2005). The result is called normed chi-square ($\chi^2$/df). According to Kline (1998), the normed chi-square ($\chi^2$/df) lower than 3 is an indicator of good model fit. This criterion was adopted in this dissertation.

(4) Root Mean Square Error of Approximation (RMSEA). However, neither Chi-square nor normed Chi-square has a built-in mechanism to correct for model complexity. Generally speaking, if two models show equivalent fit to the same data, the simpler one is preferred because it is more parcimonious. The RMSEA is a parsimony-adjusted index that corrects for model complexity and it favors simpler models. A RMSEA value of zero indicates a perfect fit. The higher the value goes, the worse the fit is. RMSEA values smaller than 0.05 indicate good model fit while values between 0.05 and 0.08 can be
interpreted as a sign of reasonable approximation of error (Browne & Cudeck, 1993). This criterion was adopted in this dissertation.

(5) Comparative Fit Index (CFI) and Tucker–Lewis index (TLI). CFI compares the fit of the certain model with the fit of a baseline model that assumes zero covariances among observed variables. Because it is usually unrealistic to assume that observed variables are uncorrelated, the fit of a baseline model is usually very bad. The degree of improvement in fit from the baseline model indicates whether the specified model is a better one as compared with the baseline model. As suggested by Hu and Bentler (1999), the rule of thumb is that a CFI value larger than 0.9 indicates that the specified model has a good fit. This criterion was adopted in the current study. TLI (also called the non-normed fit index or NNFI) index was developed against the disadvantage of normed fit index regarding being effected by sample size. The TFI values over 0.9 or over 0.95 are considered acceptable (e.g., Hu & Bentler, 1999).

(6) Standardized Root Mean Square Residual (SRMR). SRMR is an absolute fit index based on the mean absolute correlation residual. The size of a correlation residual shows how an observed correlation matrix differs from the model-implied. A SRMR value of zero indicates a perfect model fit, meaning that there is no difference between the two correlation matrices. According to Kline (2005), a SRMR value less than 0.1 can be considered as a sign of acceptable fit. This criterion was adopted in this dissertation.

(7) Akaike Information Criterion (AIC). AIC is a predictive fit index which evaluates model fit in hypothetical replication samples of the same size and drawn randomly from the same population as the original sample. It favors simpler models and is a parsimony-adjusted index. The AIC index is generally used to select the best model among
competing models estimated with the same data. Compared with its competing models, the model with the smallest AIC is of better fit and fewer parameters (Kline, 2005).

(8) Individual Parameter Estimates. Parameter estimates can be examined for statistical significance and appropriateness. First, the value of an estimate divided by its standard error (referred to as the critical ratio) provides a test statistic that can be used to evaluate the significance of the estimate. Previous researchers (e.g., Wang, 2008) suggested a critical ratio value larger than 2 as significant. This criterion was adopted in this study. Second, the sign of an estimate should be checked to examine if the meaning of the estimate is theoretically sound. Third, standardized factor loadings of each item should be checked. A factor loading higher than 0.6 is regarded as appropriate. Fourth, multicollinearity among latent factors can be identified by examining their correlations. An extremely high correlation estimate indicates a linear dependency among factors. This means that the factors are not empirically distinguishable and are not distinct enough to be regarded as separate factors, which makes a model implausible. Previous studies (Sawaki et al., 2008; Stricker et al., 2005; Stricker & Rock 2008) used a value of 0.9 to screen out highly correlated factors. This criterion was adopted in this dissertation. Lastly, R-square indicates how much variation in an observed variable is explained by the latent variable. It is expected that the R-square estimates are higher than 0.5.

**Rasch Analysis**

Rasch analysis (Rasch, 1960) was conducted mainly for descriptor calibration and the evaluation of model fit. Rasch analysis, as a psychometric measurement model, is relatively new to researchers in the field of foreign language education. Therefore, the last few sections of Chapter 2 have been devoted to reviewing the rationale of this technique. The current section
focuses on the application of Rasch analysis in developing the *CSL Reading Proficiency Questionnaire*.

In Rasch analysis, both person ability measures and item characteristics measures are calibrated on the same logit scale. Learners can be compared with each other in terms of ability; items can be compared to other items in terms of difficulty; item difficulty and learner ability can also be compared (e.g., a student from 4th-year Chinese class has a high likelihood of agreeing with item 1. A student from 1st-year Chinese class has a low likelihood of agreeing with item 1). As indicated in the following formula, the probability that test taker $j$ can answer item $i$ correctly is estimated by the item difficulty ($b_i$) and his or her ability $\theta$ (Rasch, 1960).

$$P(Y_j = 1 | \theta) = \frac{\exp(\theta - b_j)}{1 + \exp(\theta - b_j)}$$

$b_i$: the difficulty of item $i$

$\theta$: ability of test taker $j$

The key assumption of Rasch analysis is unidimensionality, assuming that all items in an instrument measure the same construct. In this study, it was found out that the *CSL Reading Proficiency Questionnaire* had multiple latent factors; therefore the Questionnaire was divided into several subscales accordingly. For each subscale, Rasch analysis was performed individually by using FACETS (Linacre, 2012b), a computer program for Rasch analysis, in order to: (1) screen descriptors and respondents that do not fit the Rasch model expectations. “Fit” denotes the degree to which data conform to the Rasch model. If one descriptor/responder does not act as the Rasch model would predict, we can say this descriptor/response is misfitting. As a way of quality control, the evaluation of fit enables researchers to screen descriptors/responses that are diverging from model expectations; (2) estimate “difficulty value” of each descriptor. In Rasch analysis, learner ability and descriptor difficulty are calibrated on the same logit scale. More
difficult descriptors have higher logit values. Descriptors can be ranked according to their difficulty estimates; (3) determine the number of band levels and establish cut-off points. Descriptors can be grouped into multiple band levels divided by cut-off points. When deciding cut-off points, both statistical logit locations and practical needs should be taken into consideration.

Among a series of Rasch analysis models, Rasch rating scale analysis can better fit the research purposes of the current study. It was decided to use Rasch rating scale analysis instead of other types of Rasch analysis models in this dissertation. Detailed procedures and relevant index are elaborated in the following paragraphs.

**Model Fit**

First of all, model fit should be checked. Fit, as a concept, denotes how well data conform to the Rasch model. It helps identify divergence of data from the model expectations. If one item/person is identified to be misfitting, it means that this item/person does not act as the Rasch model would predict. An example of a “misfitting” item would be a difficult test item that is correctly answered by a large number of low performing students but fewer high achieving students. An example of a “misfitting” person would be a low performing learner who answers difficult test items very well. The evaluation of fit enables researchers to identify instances in which the items and learners are behaving in a manner that does not suggest perfect functioning of the measurement scale.

The estimation of how well data fit the Rasch model, or the degree of fit, is expressed quantitatively by several fit statistics: Infit mean square (MNSQ), Infit z-standardized (ZSTD), Outfit ZSTD, and Outfit MNSQ. Outfit MNSQ and Infit MNSQ are chi-square statistics, which quantify how well data fit the model. The ZSTD value denotes the probability of the MNSQ
value occurring by chance when the data fit the Rasch model.

The calculation of mean-squares produces an average around 1.0 (Wright & Linacre, 1994). Mean-square values greater than 1.0 show underfit, which means that there is too much unexplained variance (or noise) in the data. Mean-square values lower than 1.0 show overfit, indicating that the model overpredicts the data and causes inflated reliability statistics (Boone, Staver, & Yale, 2014). As a rule of thumb, Outfit MNSQ/Infit MNSQ values in the range between 0.5 and 1.5 suggest a reasonable fit of the data to the model and are considered as “productive for measurement” (Linacre, 2002, 2003, 2014). Please check Table 7 for details.

Since the ZSTD value is based on the MNSQ statistics, researchers often first check the MNSQ statistics for the evaluation of fit. According to Boone, Staver, and Yale (2014), as long as the MNSQ value lies within an acceptable range of fit, it is not necessary to investigate the ZSTD value.

Table 7

<table>
<thead>
<tr>
<th>Interpretation of parameter-level mean-square fit statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;2.0</td>
</tr>
<tr>
<td>1.5-2.0</td>
</tr>
<tr>
<td>0.5-1.5</td>
</tr>
<tr>
<td>&lt;0.5</td>
</tr>
</tbody>
</table>

*Note.* Reprinted from Wright & Linacre, 1994

However, Infit MNSQ and Outfit MNSQ statistics are different in nature. The Outfit statistic is short for “outlier-sensitive fit statistic”. It is sensitive to “outlying” unexpected responses (e.g. guessing or thoughtless errors) (Eckes, 2015). The Infit statistic, which is short for “information weighted fit statistic” is an information-weighted indicator of fit. It focuses more on “inlying” unexpected responses and is sensitive to unusual patterns near a given item difficulty (or person ability). Some researchers (Bond & Fox, 2015; Myford & Wolfe, 2003)
argue that since Infit statistic is generally associated with higher estimation precision, it is commonly considered more important than Outfit statistic in judging fit. However, other researchers (e.g., Boone, Staver, & Yale, 2013; Linacre, 2012a) advised that the Outfit statistic should be first checked. Because of its sensitivity to outliers, the Outfit statistic can identify and fix issues of fit more easily. According to Linacre (2012a), only Outfit statistics need to be reported; if the data are heavily contaminated with irrelevant outliers, then it may be appropriate to report Infit statistic. In the current study, only outfit statistics were checked.

**Parameter Estimates and Descriptor Calibration**

In this study, Rasch analysis was used to estimate the “difficulty value” of individual descriptors and plot them on the continuum of difficulty. Both learner ability measures and descriptor item difficulty measures were calibrated on the same vertical scale. Learner ability measures and item difficulty measures were estimated in the same units, i.e., “logits” (Linacre & Wright, 1989). Therefore, learners could be compared with each other; items could be compared to other items; items and learners could also be compared on the same logit scale.

When reading the output of Rasch analysis results, the researcher first examined: (1) estimates and distributions of learner ability measures; (2) estimates and distributions of item difficulty measures. The researcher then investigated a Wright Map (person-item map) that showed graphically the calibrations for the learner ability facet and the item difficulty facet. In addition, for each facet, the separation ratio, separation reliability, chi-square, and significant level were checked to see if learners were significantly differentiated in terms of reading ability and if descriptors were significantly different from each other in terms of difficulty. Usually, if the separation ratio is larger than 2, the difference can be regarded as significant. The higher the separation ration and separation reliability, the more significant is the difference.
Evaluation of Rating Categories

The CSL Reading Proficiency Questionnaire was a five-point Likert-scale questionnaire. Five rating categories ranged from 0 (Can not be expected to perform like this) to 4 (this describes a performance which is clearly below his/her level. Could perform better than this). Category statistics were checked to see if rating categories were well defined and well arranged.

The frequency count and proportion of each rating category were first examined. According to Linacre (1999), if frequency counts of a rating category are lower than 10, this category should be merged with adjacent categories. Second, average measure, which is the average ability of learners who are rated to be at each rating category, was checked. To the Likert-scale questionnaire used in the current study, values of average measure for Category 0 to Category 4 were expected to increase monotonically. Third, the thresholds measure, which is also called the step calibrations measure, is the value of learner ability estimate where two adjacent rating category response functions intersect. To the current study, it was expected that values of the thresholds measures increase monotonically, which indicated that rating categories were well defined. In addition, differences among thresholds measure values should be larger than 1.4 logits but smaller than 5 logits, meaning that differences among rating categories are significant but not too huge. Fourth, Outfit mean square is another important criterion for checking the quality of rating categories. If this value is larger than 2, it means that unexplained variance is larger than explained variance (Linacre, 1999) because some categories introduce unexplained variance (or noise) into the data. These categories should be merged with adjacent categories. Lastly, category probability curves, which were in accordance with the category statistics, were checked. It was expected that each category formed a nice independent normal curve.
The Determining of Cut-off Points

When all descriptors were calibrated vertically, the next step was to set cut-off points in order to cut the vertical scale into multiple band levels. Dividing the scale into multiple levels would make the CSL Reading Proficiency Scale more users friendly since it helped users to associate target learners/test takers with the appropriate band levels.

Determining cut-off points involves subjective judgments (North, 2000). In the development of the CEFR, three ways of setting cut-offs were taken into consideration: (1) checking logit values to create a scale of equal intervals; (2) looking for gaps on the vertical scale of descriptor items which could be regarded as natural thresholds between levels; and, (3) referring to levels of descriptors in the source scales from which these descriptors were collected and also referring to the conventional levels suggested by literature (North, 2000).

In the current study, when deciding cut-off points, both statistical logit locations and practical needs had been taken into consideration. Around half descriptor items were collected from the first four levels of the source scale (i.e., Novice, Intermediate, Advanced, Superior levels of the ACTFL Chinese Proficiency Guidelines). In addition, the primary users of this scale are college-level CSL teachers and learners. Most universities or colleges in the U.S., which grant B.A. degrees in Chinese, offer four years of Chinese language courses. Therefore, a four-level scale with three cut-off points was preferred since it corresponded to the curriculum of four-year colleges and universities.

The researcher planned to first mark out three provisional cut-offs by dividing four equal distances on the logit scale. Second, the provisional cut-offs were adjusted to better coincide with locations of obvious gaps between descriptors. To make this scale more applicable and relevant to college-level CSL teaching, the provisional cut-offs were further adjusted to correspond to the
distribution of learner ability estimates.

Results of the preparatory stage of scale development, including modeling the CSL reading ability construct, establishing descriptive parameters, developing the descriptor pool, expert review, and pilot study, are reported in Chapter 4. Factor analysis and Rasch analysis results are presented in Chapter 5.
CHAPTER 4

RESULTS-PREPARATORY STAGE

This study aims to develop and validate a *CSL Reading Proficiency Scale*. Chapter 4 reports results of scale development preparatory stage which involves five steps: modeling the CSL reading ability construct, establishing descriptive parameters, developing the descriptor pool, expert review, and pilot study. Chapter 5 reports the results of quantitative validation.

**Modeling the CSL Reading Ability Construct**

The researcher defined CSL reading ability as a constellation of linguistic knowledge, pragmatic knowledge, cognitive processes, and metacognitive mechanisms required for constructing meaning from written materials. Figure 8 in Chapter 2 illustrates this model of CSL reading ability as well as its dimensions. As demonstrated in Figure 8, four dimensions of CSL reading ability are linguistic knowledge, pragmatic knowledge, higher-level reading processing, and lower-level reading processing. Within each dimension, component skills are also listed out.

Lower-level processing comprises a broad range of cognitive mechanisms required for constructing meaning from Chinese text (i.e., character recognition, word decision, lexical access, syntactic parsing, and semantic-proposition encoding). Higher-level processing involves both cognitive processes (i.e., constructing the text model and situation model of comprehension) and metacognitive processes (i.e., monitoring comprehension, setting goals). Pragmatic knowledge consists of functional knowledge and sociolinguistic knowledge. Linguistic knowledge includes two subgroups: grammar knowledge and textual knowledge. Figure 9 presents lower-level reading processes as well as their corresponding linguistic knowledge,
highlighting unique linguistic knowledge and cognitive mechanisms of reading Chinese language. Figure 10 focuses only on higher-level processes and their associated knowledge components, with Chinese-specific features and tactics underscored.

**Establishing Descriptive Parameters**

Four parameters were selected to operationalize the theoretical construct of CSL reading: cognitive verb, object (of the verb), text type, and modifier. The modifier category includes four types of modifiers: topic, linguistic features, domain, and purpose. Please see Table 8 for details. These four parameters lay out two dimensions for describing reading proficiency: quantity and quality. The first two (i.e., cognitive verb and object) quantify tasks that readers can complete. The last two (i.e., text type and other modifiers) are qualitative in nature. They specify the intrinsic quality (i.e., accuracy, fluency, degree of ease) and extrinsic quality of a language performance (i.e., the condition in which the performance is completed). In this way, reading proficiency descriptors used in the current study not only give details of linguistic features, but also denote features of specific language use situations and tasks. While the first two parameters, cognitive verb, and object (of the verb) were mandatory, the last two were optional. For beginning learners, compulsory parameters would be sufficient to describe their reading abilities. For more advanced levels, however, optional parameters were added to pinpoint descriptors to specific levels and distinguish them with descriptors from adjacent levels.

In the current study, descriptors of the *CSL Reading Proficiency Scale* were written in the format of can-do statements. A typical can-do descriptor with four descriptive parameters is like “Can summarize the major features of a scene described in an essay written in relatively complex language.” In this example, “summarize” is the cognitive verb and “the major features of a scene” is the object of “summarize.” “An essay” specifies the text type and “written in relatively
complex language” denotes the linguistic features of this essay.

Table 8

*Parameters of descriptors*

<table>
<thead>
<tr>
<th>Descriptive Parameter</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Verb</td>
<td>Understand, recognize, decode, identify…….</td>
</tr>
<tr>
<td>Text Type</td>
<td>Narration, description, exposition, argumentation…….</td>
</tr>
<tr>
<td>Modifier</td>
<td>Topic: daily life, sports, recreation…….</td>
</tr>
<tr>
<td></td>
<td>Linguistic feature: dense of new vocabulary, the complexity of language…….</td>
</tr>
<tr>
<td></td>
<td>Domain: personal, public, educational, professional.</td>
</tr>
<tr>
<td></td>
<td>Purpose: for informative social purposes…….</td>
</tr>
</tbody>
</table>

**CSL Reading Proficiency Descriptor Pool**

The development of the descriptor pool started with a comprehensive documentation of descriptors from existing reading proficiency scales, theoretical works, and empirical studies on reading. The first source of descriptors was existing reading proficiency scales. The researcher collected 43 descriptors from *The Canadian Language Benchmarks*; 10 descriptors from *The World-class Instructional Design and Assessment (WIDA) English Language Development Standards*; 25 descriptors from *The ACTFL Proficiency Guidelines*; 8 descriptors from *Interagency Language Roundtable (ILR) Language Skill Level Descriptions*, and 9 descriptors from *Mandarin Chinese: Four-Year Instructional Goals, Curriculum Outline, and Instructional Measures*. In addition to the generic guidelines which were claimed to be applicable to all languages, the ACTFL also published language-specific guidelines for non-native reading ability.
in Chinese. Forty-five descriptors were collected from *The ACTFL Chinese Proficiency Guidelines*.

One hundred and forty descriptors, in total, were collected from existing reading proficiency scales. However, such a large number of descriptors were beyond the scope of a doctoral dissertation. Since this study aimed to address CSL reading ability, the researcher decided to just focus on two CSL-specific scales: *Mandarin Chinese: Four-Year Instructional Goals, Curriculum Outline, and Instructional Measures* and *The ACTFL Chinese Proficiency Guidelines* at this stage. A closer examination revealed that all descriptors from the former were covered in the latter. Therefore, only the 44 descriptors from *The ACTFL Chinese Proficiency Guidelines* were included in the descriptor pool. The remaining eight-six descriptors collected from other scales would be saved for future research.

The second source of descriptors was theoretical and empirical works on L1, L2, and CSL reading. Based on theoretical works and empirical studies, forty-five new descriptors were written to cover various aspects of knowledge/ability unique to CSL reading: character recognition and orthographic knowledge, word decision, vocabulary knowledge and lexical access, syntactic parsing and syntax knowledge, semantic-proposition encoding, knowledge of cohesion, and knowledge of the rhetorical organization. For instance, based on Koda (2005), descriptors that depicted L2 reading ability from a cross-linguistic approach were created. Studies on Chinese language and CSL reading, such as Hung (2012), Norman (1988), and Shen (2008), were sources of descriptors that described unique CSL linguistic features and corresponding cognitive processes. In total, 89 descriptors (44 from existing scales and 45 created based on theory) were included in the descriptor pool. The table below summarizes descriptors collected and finally selected from each source.
Table 9

Sources of Descriptors

<table>
<thead>
<tr>
<th>Documents</th>
<th>Descriptors Collected</th>
<th>Descriptors Finally Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing Scales of Language Proficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The American Council for the Teaching of Foreign Languages (ACTFL)</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Proficiency Guidelines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The World-class Instruction Design and Assessment (WIDA) English Language Development Standards</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Interagency Language Roundtable (ILR) Language Skill Level Descriptions – Reading</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Canadian Language Benchmarks</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>Mandarin Chinese: Four-Year Instructional Goals, Curriculum Outline, and Instructional Measures.</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>The American Council for the Teaching of Foreign Languages (ACTFL) Chinese Proficiency Guidelines</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>Theoretical and empirical works on L1, L2, and CSL reading</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koda (2005)</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Grabe (2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alderson et al. (2015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hung (2012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shen (2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norman (1988)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>184</td>
<td>89</td>
</tr>
</tbody>
</table>

The researcher revised descriptors to make them all written in the format of the “can-do” statement with positive orientation and with four parameters: cognitive verb, the object of the cognitive verb, text type, and defining words. The first two parameters were compulsory while
the last two were optional. Descriptors were reexamined to make sure that (a) the ability/knowledge described by every descriptor is unique and unitary; (b) descriptors were worded to be clear, concise, and precise. Jargon and less commonly used linguistic terms were avoided. In addition, descriptors were revised to meet the following criteria suggested by North (2000): positiveness, definiteness, clarity, brevity, and independence. Please see the following table for some examples of revised descriptors as well as parameters.

Table 10

*Examples of Revised Descriptors*

<table>
<thead>
<tr>
<th>Examples of Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Can understand, {fully and with relative ease}, keywords and cognates, as well as formulaic phrases [across a range of highly contextualized texts].</td>
</tr>
<tr>
<td>2: [Where vocabulary has been learned], can understand predictable language and messages [such as those found on train schedules, roadmaps, and street signs].</td>
</tr>
<tr>
<td>3: Can typically derive meaning [from short, non-complex texts that convey basic information for which there is contextual or extralinguistic support].</td>
</tr>
<tr>
<td>4: Can recognize {a limited number of} characters in a character-based language.</td>
</tr>
<tr>
<td>5: Can identify {a number of highly contextualized} words and phrases {including cognates and borrowed words}.</td>
</tr>
<tr>
<td>6: Can {rarely} understand material {that exceeds a single phrase}.</td>
</tr>
<tr>
<td>7: Can recognize {a limited number of} letters, symbols or characters.</td>
</tr>
<tr>
<td>8: Can identify {high-frequency} words and/or phrases {when strongly supported by context}.</td>
</tr>
</tbody>
</table>

*Note.* Cognitive verbs are italicized and the objects of verbs are underlined. [ ] indicates text type and { } marks other modifiers.
Revised descriptors were provisionally labeled as one of the components of CSL reading demonstrated in Figure 8. Components closely related to each other were combined as one overarching group. For instance, descriptors belonging to the “Character Recognition” and “Orthographic Knowledge” groups were combined and put under one large category named “Character Recognition & Orthographic Knowledge”. In total, eighty-nine descriptors are classified into nine provisional groups. The following table shows names of provisional groups and the number of descriptors placed into each group.

Strategic competence is one major component of reading ability. As elaborated in Chapter 2, reading strategies can be classified into three types: metacognitive, cognitive, and social/affective strategies. Among them, cognitive strategies relate very closely to other descriptors in this descriptor pool, which mainly address the cognitive processes of reading. To make this doctoral dissertation focused and feasible, only cognitive strategies were represented in this descriptor pool. Please refer to Appendix I for the full version of the CSL Reading Proficiency Descriptor Pool sent out for qualitative validation, which is discussed in the next section.

Table 11

Basic Information about the CSL Reading Proficiency Descriptor Pool

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Character Recognition and Orthographic Knowledge</td>
<td>10</td>
</tr>
<tr>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>15</td>
</tr>
<tr>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>22</td>
</tr>
<tr>
<td>4) Semantic-proposition Encoding</td>
<td>3</td>
</tr>
<tr>
<td>5) Knowledge of Coherence</td>
<td>7</td>
</tr>
<tr>
<td>6) Knowledge of Rhetorical Organization</td>
<td>8</td>
</tr>
<tr>
<td>7) The Text Model of Comprehension</td>
<td>18</td>
</tr>
<tr>
<td>8) The Situation Model of Comprehension and Pragmatic Knowledge</td>
<td>3</td>
</tr>
<tr>
<td>9) Cognitive Strategy</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
</tr>
</tbody>
</table>
Qualitative Validation

Four experienced CSL education experts reviewed the CSL Reading Proficiency Descriptor Pool and provided comments in terms of applicability, clarity, brevity, independence, and accuracy. Based on teacher participants’ suggestions and feedback, the researcher deleted, revised, or combined some descriptors. A few examples are provided below to explain what changes have been made to the descriptor pool.

First, teacher participants picked out wordings which were not clear enough. For instance, they thought that descriptor Item 4 “Can recognize different forms of a particular radical in characters.” was not very clear. Examples were then added to explain the meaning of “different forms of a particular radical”. Second, some items which were not applicable to the teaching of CSL reading were removed. For example, descriptor Item 1 “Can decode numbers 1-1000 written in Chinese characters.” was questioned in terms of applicability by teacher participants and was removed. Third, the descriptor bank included some items which all addressed the same skills but differentiated accuracy rates or degrees of mastery. Item 8-10 in the descriptor pool, for instance, aimed to quantify learners’ ability to distinguish graphically-similar characters, ranging from 60%-90% accuracy. However, all teacher participants thought that it was unfeasible to quantify character recognition ability in this way. Therefore, Item 8-10 were combined as one item, removing accuracy rates but adding an attributive clause. The finalized item was phrased as “Can distinguish graphically-similar characters, which have learned”. Last, teacher participants were asked to add their own descriptors which were necessary but were not covered in the descriptor pool. Based on their feedback, items about discourse structures were created and added to the descriptor pool.

The revised descriptor bank was sent to participants a second time to see if more
revisions were needed. All participants agreed with revisions made. Finally, 57 descriptors were kept, renumbered, and compiled into a Likert-scale questionnaire, named the *CSL Reading Proficiency Questionnaire* (Appendix II).

**Pilot Study Results**

The *CSL Reading Proficiency Questionnaire* was sent to CSL teachers across the U.S. for the purpose of data collection. However, before being distributed to all participants, the *CSL Reading Proficiency Questionnaire* was piloted with the first 46 responses to check the reliability. Coefficient Alpha (Cronbach’s α or rxx’α) was picked as the measure of internal consistency. It turned out that the Cronbach’s alpha was 0.990 (Table 12), larger than 0.9, indicating an excellent degree of internal consistency. The *CSL Reading Proficiency Questionnaire* was ready to be distributed, without further revisions, to a larger group of participants for data collection.

**Table 12**

<table>
<thead>
<tr>
<th>Reliability Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach's Alpha</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

In total, 95 experienced CSL teachers used the *CSL Reading Proficiency Questionnaire* to evaluate 175 learners’ reading proficiency. Data generated from this large-scale application were subjected to quantitative analyses, the results of which are reported in Chapter 5.
CHAPTER 5
RESULTS-QUANTITATIVE ANALYSIS

The previous chapter reports the results of the preparatory stage of scale development, including the qualitative validation and pilot study. After the pilot study, the data collection instrument, *CSL Reading Proficiency Questionnaire*, were distributed to a larger number of experienced CSL teachers. In total, 95 teacher participants evaluated 179 CSL learners against this questionnaire. Data collected were analyzed to: (1) explore the factor structure of CSL reading ability; (2) calibrate descriptors on a vertical scale and divide them into several band levels. Accordingly, this study answered following research questions:

Dimensions of descriptors: 1. What is the factor structure of CSL reading, as measured by descriptors in the *CSL Reading Proficiency Scale*?

Difficulty levels of descriptors: 2. What is the estimated level of difficulty of each descriptor? 3. How are cut-off points selected to divide descriptors into multiple bands?

Chapter 5 reports the results of quantitative analysis to answer research questions. The author first reports factor structure of reading ability revealed by the confirmatory factor analysis. The number of difficulty levels justified by Rasch Analysis as well as the difficulty estimate of each descriptor is then presented. Finally, based on data analysis results, the *CSL Reading Proficiency Scale*, a multiple-level progression scale with three subscales is established.

**Confirmatory factor Analysis**
Descriptive Statistics

Table 13 reports the descriptive statistics for the observed variables (i.e., descriptor item 1 to item 57), including possible score range, mean, median, and standard deviation. Item 44 and item 51 ranged from 0 to 3, while other items ranged from 0 to 4. The medians of most items were 2. The averages of item scores varied from 0.82 to 3.06.

Table 13

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 2</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.04</td>
<td>2.00</td>
<td>.959</td>
</tr>
<tr>
<td>Item 3</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.04</td>
<td>2.00</td>
<td>.866</td>
</tr>
<tr>
<td>Item 4</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.42</td>
<td>2.00</td>
<td>.941</td>
</tr>
<tr>
<td>Item 5</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.17</td>
<td>2.00</td>
<td>.911</td>
</tr>
<tr>
<td>Item 6</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.30</td>
<td>2.00</td>
<td>.754</td>
</tr>
<tr>
<td>Item 7</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.16</td>
<td>2.00</td>
<td>.990</td>
</tr>
<tr>
<td>Item 8</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.63</td>
<td>2.00</td>
<td>.893</td>
</tr>
<tr>
<td>Item 9</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.74</td>
<td>2.00</td>
<td>.855</td>
</tr>
<tr>
<td>Item 10</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.68</td>
<td>3.00</td>
<td>.922</td>
</tr>
<tr>
<td>Item 11</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2.85</td>
<td>3.00</td>
<td>.937</td>
</tr>
<tr>
<td>Item 12</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.64</td>
<td>3.00</td>
<td>1.037</td>
</tr>
<tr>
<td>Item 13</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.81</td>
<td>3.00</td>
<td>1.037</td>
</tr>
<tr>
<td>Item 14</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.46</td>
<td>2.00</td>
<td>.931</td>
</tr>
<tr>
<td>Item 15</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.11</td>
<td>2.00</td>
<td>.957</td>
</tr>
<tr>
<td>Item 16</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.15</td>
<td>2.00</td>
<td>1.041</td>
</tr>
<tr>
<td>Item 17</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.28</td>
<td>2.00</td>
<td>1.086</td>
</tr>
<tr>
<td>Item 18</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.19</td>
<td>2.00</td>
<td>1.198</td>
</tr>
<tr>
<td>Item 19</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.97</td>
<td>2.00</td>
<td>1.211</td>
</tr>
<tr>
<td>Item 20</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.71</td>
<td>3.00</td>
<td>1.073</td>
</tr>
<tr>
<td>Item 21</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.20</td>
<td>2.00</td>
<td>1.008</td>
</tr>
<tr>
<td>Item 22</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.21</td>
<td>2.00</td>
<td>1.000</td>
</tr>
<tr>
<td>Item 23</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.26</td>
<td>2.00</td>
<td>1.024</td>
</tr>
<tr>
<td>Item 24</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.08</td>
<td>2.00</td>
<td>.902</td>
</tr>
<tr>
<td>Item 25</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.18</td>
<td>2.00</td>
<td>1.028</td>
</tr>
<tr>
<td>Item 26</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.41</td>
<td>2.00</td>
<td>1.069</td>
</tr>
<tr>
<td>Item 27</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.41</td>
<td>2.00</td>
<td>1.069</td>
</tr>
<tr>
<td>Item</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.54</td>
<td>1.00</td>
<td>1.087</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.50</td>
<td>1.00</td>
<td>1.088</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.96</td>
<td>2.00</td>
<td>.985</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2.61</td>
<td>2.00</td>
<td>.996</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.13</td>
<td>2.00</td>
<td>1.035</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.02</td>
<td>2.00</td>
<td>1.025</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.54</td>
<td>2.00</td>
<td>1.138</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.21</td>
<td>1.00</td>
<td>1.069</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.23</td>
<td>1.00</td>
<td>.995</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>.87</td>
<td>1.00</td>
<td>.926</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3.06</td>
<td>3.00</td>
<td>.955</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.50</td>
<td>2.00</td>
<td>1.093</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.12</td>
<td>2.00</td>
<td>1.067</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.08</td>
<td>2.00</td>
<td>1.089</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.69</td>
<td>2.00</td>
<td>1.092</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.16</td>
<td>2.00</td>
<td>1.102</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.58</td>
<td>2.00</td>
<td>.965</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.12</td>
<td>1.00</td>
<td>1.074</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.33</td>
<td>1.00</td>
<td>1.059</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.85</td>
<td>2.00</td>
<td>.862</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.22</td>
<td>1.00</td>
<td>1.056</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>.88</td>
<td>1.00</td>
<td>.969</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.04</td>
<td>1.00</td>
<td>1.019</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>.82</td>
<td>1.00</td>
<td>.915</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>.88</td>
<td>1.00</td>
<td>.928</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>.83</td>
<td>1.00</td>
<td>.939</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>2.30</td>
<td>2.00</td>
<td>1.015</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.79</td>
<td>2.00</td>
<td>1.100</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>.98</td>
<td>1.00</td>
<td>1.025</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>1.65</td>
<td>1.00</td>
<td>1.040</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* N = 179

**Model Fit**

According to literature, three competing factor models could be adequate at accounting for the underlying structure of CSL reading ability: a correlated two-factor model (Figure 12), a correlated three-factor model (Figure 13), and a one-factor model (Figure 14). All three models were tested for fit with the *CSL Reading Proficiency Questionnaire* data. As summarized in Table
14, following criteria (shown in the first column in the table above) pre-determined on the basis of the relevant literature were used to evaluate overall model fit.

Table 14

*Fit Indices for the Three Competing Models*

<table>
<thead>
<tr>
<th>Fit indices</th>
<th>Criteria</th>
<th>One-factor model</th>
<th>Correlated two-factor model</th>
<th>Correlated three-factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model df</td>
<td></td>
<td>1539</td>
<td>1538</td>
<td>1536</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>The smaller the better</td>
<td>3632.936</td>
<td>3147.688</td>
<td>3132.177</td>
</tr>
<tr>
<td>P-value of $\chi^2$</td>
<td>Not significant</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>$\chi^2/df$</td>
<td>$&lt;3$ (the closer to 1 the better)</td>
<td>2.36</td>
<td>2.05</td>
<td>2.04</td>
</tr>
<tr>
<td>CFI</td>
<td>$&gt;0.9$</td>
<td>0.963</td>
<td>0.971</td>
<td>0.971</td>
</tr>
<tr>
<td>TLI (NNFI)</td>
<td>$&gt;0.9$</td>
<td>0.961</td>
<td>0.970</td>
<td>0.970</td>
</tr>
<tr>
<td>RMSEA</td>
<td>$&lt;0.05$ (good model fit)</td>
<td>0.087</td>
<td>0.076</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>$&lt;0.08$ (reasonable approximation of error)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRMR</td>
<td>$&lt;0.1$</td>
<td>0.078</td>
<td>0.067</td>
<td>0.067</td>
</tr>
<tr>
<td>AIC</td>
<td>The smaller the better</td>
<td>554.936</td>
<td>71.688</td>
<td>60.177</td>
</tr>
</tbody>
</table>

First, the second line in Table 14 reports models’ degrees of freedom. The correlated three-factor model with 1536 degrees of freedom was the most saturated model among the three models. The one-factor model was the most parsimony one with 1539 degrees of freedom. As reported in the following paragraph, some fit indices deteriorated when model complexity decreased from the correlated three-factor model to the one-factor model.

Second, chi-square values ($\chi^2$) of all three models were significant ($p=0.000$), which put model fit in doubt. However, as discussed previously, the value of the model chi-square is highly sensitive to sample size and therefore should be interpreted with caution. The chi-square values were divided by the degrees of freedom to reduce its sensitivity to sample size. The normed chi-
square ($\chi^2$/df) values were all below 3, indicating that all three models fit the data well. Third, the values of CFI and TLI were all larger than 0.9. This meant that the fit of all three models improved significantly as compared with their respective baseline models assuming no covariances among the variables. Fourth, the values of root mean square error of approximation (RMSEA) for the two more saturated models (correlated two-factor model and correlated three-factor model) were below 0.08. However, the value for the most parsimonious one (the one-factor model) was higher than 0.08, indicating that the model fit was not satisfactory. The values of standardized root mean square residual (SRMR) for three models were all below 0.1. The correlation matrices were not significant different from the model-implied ones. The values of AIC of three models, which should be the smaller the better, were 554.936, 71.688, and 60.177, indicating that Model 3 was the best model.

In sum, as compared with the most parsimonious model (the one-factor model), the correlated two-factor model and the correlated three-factor model were considered to have good fit. The selected fit indices for the latter two models were satisfactory except for the model chi-square values. On the global level, both the correlated two-factor and the correlated three-factor model demonstrated reasonable fit to the data. However, between these two models, Model 3 fit the data slightly better.
Figure 12

Correlated Two-factor Model
Figure 13

*Correlated Three-factor Model*
Figure 14

One-factor Model
Next, individual parameter estimates were examined for significance and appropriateness. As shown in the table below, signs of estimates in three models were all positive. The critical ratios of three models were all significant. Standardized factor loadings of all observed variables were higher than 0.6. Regarding the correlated two-factor model, it was detected that the correlation between the two factors was estimated as 0.889, strictly lower than the 0.9 acceptance level. In the corrected three-factor model, the correlation estimates among three factors were 0.891, 0.807, 0.892, all lower than the 0.9 acceptance level, however they were very close to the cut-off. A correlation lower than the 0.9 acceptance level indicated that factors were distinct enough to be considered as separate factors.

In terms of the measurement model R-square, in the correlated three-factor model, only descriptor Item 2 had R-square value lower than the 0.5 acceptance value. In other two models, both descriptor Item 2 and Item 7 had R-square values lower than 0.5. However, R-square values for descriptor Item 2 and Item 7 in all three models ranged between 0.4 and 0.5, which were very close to the 0.5 acceptance level.

Table 15

*Individual Parameter Estimates for the Three Competing Models*

<table>
<thead>
<tr>
<th>Individual parameter estimates</th>
<th>Criteria</th>
<th>One-factor model</th>
<th>Correlated two-factor model</th>
<th>Correlated three-factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signs of Estimates</td>
<td>Theoretically round</td>
<td>All positive</td>
<td>All positive</td>
<td>All positive</td>
</tr>
<tr>
<td>Critical Ratio</td>
<td>&gt;2</td>
<td>significant</td>
<td>significant</td>
<td>significant</td>
</tr>
<tr>
<td>Standardized factor loading</td>
<td>&gt;0.6</td>
<td>appropriate</td>
<td>appropriate</td>
<td>appropriate</td>
</tr>
<tr>
<td>Estimated</td>
<td>&lt;0.9</td>
<td>N.A.</td>
<td>0.889</td>
<td>0.891; 0.807;</td>
</tr>
</tbody>
</table>
Correlation between factors: 0.892

<table>
<thead>
<tr>
<th>Measurement model R-square</th>
<th>All except item 2 and item 7 are appropriate</th>
<th>All except item 2 are appropriate</th>
<th>All except item 7 are appropriate</th>
</tr>
</thead>
</table>

Conclusions

Taking both model fit criteria and parameter estimates criteria into consideration, component-skill models fit the data better than the unitary-skill model. Two component-skill models (i.e., correlated two-factor model and correlated three-factor model) provided good explanations of the data. However, correlated three-factor model was slightly better. This indicated that a correlated three-factor model could better delineate the underlying structure of CSL reading comprehension. CSL reading comprehension could be conceptualized as consisting of three inter-correlated components: lower-level decoding; interim-level textbase construction; higher-level situation-model building.

Accordingly, fifty-seven descriptor items in the CSL Reading Proficiency Questionnaire were divided into three subscales: the Lower-level Questionnaire which measures lower-level processes and knowledge (indicated by descriptor Item 1-Item 29); the Interim-level Questionnaire which measures skills and knowledge involved in interim-level textbase construction (indicated by descriptor Item 30-Item 53); the Higher-level Questionnaire that measures skills and knowledge involved in higher-level situation-model building (indicated by Item 54-Item 57). Individual parameter estimates, including factor loading estimates and R-square estimates, of the correlated three-factor model, are shown in the following table.
### Table 16

**Individual Parameter Estimates of the Correlated Three-factor Model**

<table>
<thead>
<tr>
<th>LOWERLEV BY</th>
<th>Factor Loadings Estimate</th>
<th>S.E.</th>
<th>Est./S.E.</th>
<th>Two-Tailed P-Value</th>
<th>R-square Estimate</th>
<th>S.E.</th>
<th>Est./S.E.</th>
<th>Two-Tailed P-Value</th>
<th>Residual Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITEM1</td>
<td>0.739</td>
<td>0.038</td>
<td>19.511</td>
<td>0</td>
<td>0.546</td>
<td>0.056</td>
<td>9.755</td>
<td>0</td>
<td>0.454</td>
</tr>
<tr>
<td>ITEM2</td>
<td>0.655</td>
<td>0.044</td>
<td>14.758</td>
<td>0</td>
<td>0.429</td>
<td>0.058</td>
<td>7.379</td>
<td>0</td>
<td>0.571</td>
</tr>
<tr>
<td>ITEM3</td>
<td>0.752</td>
<td>0.035</td>
<td>21.565</td>
<td>0</td>
<td>0.566</td>
<td>0.052</td>
<td>10.782</td>
<td>0</td>
<td>0.434</td>
</tr>
<tr>
<td>ITEM4</td>
<td>0.829</td>
<td>0.028</td>
<td>30.082</td>
<td>0</td>
<td>0.687</td>
<td>0.046</td>
<td>15.041</td>
<td>0</td>
<td>0.313</td>
</tr>
<tr>
<td>ITEM5</td>
<td>0.833</td>
<td>0.024</td>
<td>34.51</td>
<td>0</td>
<td>0.693</td>
<td>0.04</td>
<td>17.255</td>
<td>0</td>
<td>0.307</td>
</tr>
<tr>
<td>ITEM6</td>
<td>0.776</td>
<td>0.032</td>
<td>24.213</td>
<td>0</td>
<td>0.602</td>
<td>0.05</td>
<td>12.106</td>
<td>0</td>
<td>0.398</td>
</tr>
<tr>
<td>ITEM7</td>
<td>0.699</td>
<td>0.038</td>
<td>18.398</td>
<td>0</td>
<td>0.488</td>
<td>0.053</td>
<td>9.199</td>
<td>0</td>
<td>0.512</td>
</tr>
<tr>
<td>ITEM8</td>
<td>0.762</td>
<td>0.031</td>
<td>24.37</td>
<td>0</td>
<td>0.581</td>
<td>0.048</td>
<td>12.185</td>
<td>0</td>
<td>0.419</td>
</tr>
<tr>
<td>ITEM9</td>
<td>0.868</td>
<td>0.023</td>
<td>37.735</td>
<td>0</td>
<td>0.754</td>
<td>0.04</td>
<td>18.868</td>
<td>0</td>
<td>0.246</td>
</tr>
<tr>
<td>ITEM10</td>
<td>0.81</td>
<td>0.029</td>
<td>28.121</td>
<td>0</td>
<td>0.657</td>
<td>0.047</td>
<td>14.06</td>
<td>0</td>
<td>0.343</td>
</tr>
<tr>
<td>ITEM11</td>
<td>0.888</td>
<td>0.021</td>
<td>41.741</td>
<td>0</td>
<td>0.788</td>
<td>0.038</td>
<td>20.87</td>
<td>0</td>
<td>0.212</td>
</tr>
<tr>
<td>ITEM12</td>
<td>0.842</td>
<td>0.023</td>
<td>36.319</td>
<td>0</td>
<td>0.709</td>
<td>0.039</td>
<td>18.159</td>
<td>0</td>
<td>0.291</td>
</tr>
<tr>
<td>ITEM13</td>
<td>0.897</td>
<td>0.017</td>
<td>51.953</td>
<td>0</td>
<td>0.805</td>
<td>0.031</td>
<td>25.976</td>
<td>0</td>
<td>0.195</td>
</tr>
<tr>
<td>ITEM14</td>
<td>0.852</td>
<td>0.021</td>
<td>40.572</td>
<td>0</td>
<td>0.726</td>
<td>0.036</td>
<td>20.286</td>
<td>0</td>
<td>0.274</td>
</tr>
<tr>
<td>ITEM15</td>
<td>0.849</td>
<td>0.023</td>
<td>37.243</td>
<td>0</td>
<td>0.72</td>
<td>0.039</td>
<td>18.622</td>
<td>0</td>
<td>0.28</td>
</tr>
<tr>
<td>ITEM16</td>
<td>0.903</td>
<td>0.015</td>
<td>60.895</td>
<td>0</td>
<td>0.815</td>
<td>0.027</td>
<td>30.447</td>
<td>0</td>
<td>0.185</td>
</tr>
<tr>
<td>ITEM17</td>
<td>0.898</td>
<td>0.015</td>
<td>57.934</td>
<td>0</td>
<td>0.806</td>
<td>0.028</td>
<td>28.967</td>
<td>0</td>
<td>0.194</td>
</tr>
<tr>
<td>ITEM18</td>
<td>0.91</td>
<td>0.015</td>
<td>60.99</td>
<td>0</td>
<td>0.828</td>
<td>0.027</td>
<td>30.495</td>
<td>0</td>
<td>0.172</td>
</tr>
<tr>
<td>ITEM19</td>
<td>0.895</td>
<td>0.016</td>
<td>55.228</td>
<td>0</td>
<td>0.801</td>
<td>0.029</td>
<td>27.614</td>
<td>0</td>
<td>0.199</td>
</tr>
<tr>
<td>ITEM20</td>
<td>0.779</td>
<td>0.03</td>
<td>26.214</td>
<td>0</td>
<td>0.607</td>
<td>0.046</td>
<td>13.107</td>
<td>0</td>
<td>0.393</td>
</tr>
<tr>
<td>ITEM21</td>
<td>0.889</td>
<td>0.017</td>
<td>52.608</td>
<td>0</td>
<td>0.79</td>
<td>0.03</td>
<td>26.304</td>
<td>0</td>
<td>0.21</td>
</tr>
<tr>
<td>ITEM22</td>
<td>0.886</td>
<td>0.016</td>
<td>55.652</td>
<td>0</td>
<td>0.785</td>
<td>0.028</td>
<td>27.826</td>
<td>0</td>
<td>0.215</td>
</tr>
<tr>
<td>ITEM23</td>
<td>0.894</td>
<td>0.015</td>
<td>59.57</td>
<td>0</td>
<td>0.8</td>
<td>0.027</td>
<td>29.785</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>ITEM24</td>
<td>0.787</td>
<td>0.028</td>
<td>28.241</td>
<td>0</td>
<td>0.619</td>
<td>0.044</td>
<td>14.121</td>
<td>0</td>
<td>0.381</td>
</tr>
<tr>
<td>ITEM25</td>
<td>0.835</td>
<td>0.024</td>
<td>34.308</td>
<td>0</td>
<td>0.697</td>
<td>0.041</td>
<td>17.154</td>
<td>0</td>
<td>0.303</td>
</tr>
<tr>
<td>ITEM26</td>
<td>0.871</td>
<td>0.02</td>
<td>43.504</td>
<td>0</td>
<td>0.759</td>
<td>0.035</td>
<td>21.752</td>
<td>0</td>
<td>0.241</td>
</tr>
<tr>
<td>ITEM27</td>
<td>0.841</td>
<td>0.024</td>
<td>35.613</td>
<td>0</td>
<td>0.708</td>
<td>0.04</td>
<td>17.806</td>
<td>0</td>
<td>0.292</td>
</tr>
<tr>
<td>ITEM28</td>
<td>0.814</td>
<td>0.029</td>
<td>28.248</td>
<td>0</td>
<td>0.663</td>
<td>0.047</td>
<td>14.124</td>
<td>0</td>
<td>0.337</td>
</tr>
<tr>
<td>ITEM29</td>
<td>0.875</td>
<td>0.023</td>
<td>38.007</td>
<td>0</td>
<td>0.766</td>
<td>0.04</td>
<td>19.003</td>
<td>0</td>
<td>0.234</td>
</tr>
</tbody>
</table>

### INTERIML BY

| ITEM30      | 0.831                    | 0.024| 34.811    | 0                  | 0.69             | 0.04  | 17.405    | 0                  | 0.31              |
| ITEM31      | 0.875                    | 0.022| 40.569    | 0                  | 0.765            | 0.038| 20.284    | 0                  | 0.235             |
| ITEM32      | 0.893                    | 0.018| 50.885    | 0                  | 0.797            | 0.031| 25.442    | 0                  | 0.203             |
| ITEM33      | 0.901                    | 0.016| 57.674    | 0                  | 0.812            | 0.028| 28.837    | 0                  | 0.188             |
| ITEM34      | 0.893                    | 0.018| 50.366    | 0                  | 0.797            | 0.032| 25.183    | 0                  | 0.203             |
As mentioned previously, in total, 95 experienced CSL teachers rated the reading proficiency of 179 CSL learners against the CSL Reading Proficiency Questionnaire, consisting of the Lower-level Questionnaire, the Interim-level Questionnaire and the Higher-level Questionnaire. For the purpose of item calibration, data collected by using

| ITEM35 | 0.935 | 0.013 | 70.095 | 0 | 0.874 | 0.025 | 35.048 | 0 | 0.126 |
| ITEM36 | 0.922 | 0.014 | 67.19 | 0 | 0.849 | 0.025 | 33.595 | 0 | 0.151 |
| ITEM37 | 0.915 | 0.016 | 58.7 | 0 | 0.837 | 0.029 | 29.35 | 0 | 0.163 |
| ITEM38 | 0.912 | 0.022 | 40.799 | 0 | 0.831 | 0.041 | 20.399 | 0 | 0.169 |
| ITEM39 | 0.879 | 0.02 | 43.757 | 0 | 0.773 | 0.035 | 21.878 | 0 | 0.227 |
| ITEM40 | 0.837 | 0.022 | 37.388 | 0 | 0.701 | 0.037 | 18.694 | 0 | 0.299 |
| ITEM41 | 0.849 | 0.014 | 69.856 | 0 | 0.853 | 0.024 | 34.928 | 0 | 0.147 |
| ITEM42 | 0.928 | 0.012 | 75.691 | 0 | 0.86 | 0.023 | 37.845 | 0 | 0.14 |
| ITEM43 | 0.866 | 0.018 | 48.488 | 0 | 0.75 | 0.031 | 24.244 | 0 | 0.25 |
| ITEM44 | 0.917 | 0.015 | 60.766 | 0 | 0.841 | 0.028 | 30.383 | 0 | 0.159 |
| ITEM45 | 0.938 | 0.012 | 80.841 | 0 | 0.879 | 0.022 | 40.42 | 0 | 0.121 |
| ITEM46 | 0.908 | 0.014 | 63.993 | 0 | 0.824 | 0.026 | 31.997 | 0 | 0.176 |
| ITEM47 | 0.73 | 0.033 | 21.989 | 0 | 0.533 | 0.048 | 10.995 | 0 | 0.467 |
| ITEM48 | 0.897 | 0.016 | 55.265 | 0 | 0.805 | 0.029 | 27.632 | 0 | 0.195 |
| ITEM49 | 0.947 | 0.012 | 80.876 | 0 | 0.898 | 0.022 | 40.438 | 0 | 0.102 |
| ITEM50 | 0.928 | 0.012 | 74.616 | 0 | 0.861 | 0.023 | 37.308 | 0 | 0.139 |
| ITEM51 | 0.954 | 0.011 | 88.007 | 0 | 0.91 | 0.021 | 44.003 | 0 | 0.09 |
| ITEM52 | 0.914 | 0.014 | 64.807 | 0 | 0.835 | 0.026 | 32.403 | 0 | 0.165 |
| ITEM53 | 0.92 | 0.014 | 63.755 | 0 | 0.847 | 0.027 | 31.878 | 0 | 0.153 |

**Rasch Analysis**

As mentioned previously, in total, 95 experienced CSL teachers rated the reading proficiency of 179 CSL learners against the CSL Reading Proficiency Questionnaire, consisting of the Lower-level Questionnaire, the Interim-level Questionnaire and the Higher-level Questionnaire. For the purpose of item calibration, data collected by using
the questionnaire was analyzed by FACETS (Linacre, 2012), a computer program for Rasch analysis. Lower-level Questionnaire data, Interim-level Questionnaire data, and High-level Questionnaire data were analyzed separately and results are reported in following three sections one after another.

In each section, model fits of two facets (i.e., learner ability and descriptor item difficulty) are first reported. Second, after measurement estimates of two facets are reported separately, calibrations for two facets are presented. Third, rating category statistics and the category probability curves are then reported. Lastly, the researcher reports how cut-off points of band levels were decided based on statistical logit locations and practical needs. The interpretation of data analysis results focuses on (1) identifying descriptor items which did not fit the construct being measured; (2) estimating the difficulty parameter of each item; and, (3) determining cut-off points of item difficulty levels.

**Rasch Analysis Results of the Lower-level Questionnaire**

**Model Fit**

Two facets of analysis in the current study were learner ability and descriptor item difficulty. The first round of Rasch analysis generated a report of fit statistics of learner ability facet. Table 17 presents an excerpt from this report, in which learners with fit values (Outfit MNSQ) out of acceptable criteria levels (0.5-1.5) are marked with asterisks. As discussed previously, fit statistics give the extent to which a learner “cannot perform" items whose difficulties are estimated to be below his/her level of ability and the extent to which he/she "can perform" items supposed to be beyond his/her level of ability. In total, fifty-three out of 179 learners were identified as “misfitting” since their
Outfit MNSQ statistics were out of the acceptable range.

Data of “misfitting” learners represent a pattern of responses that diverge from model prediction. In the current study, learners’ ability was rated by their teachers. Misfitting data could be attributed to divergences in teachers’ responses (e.g. thoughtless errors). For instance, some teachers/raters did not follow rating criteria consistently when responding to the *CSL Reading Proficiency Questionnaire*. Some raters might make thoughtless errors occasionally or frequently when rating learner ability. These 53 responses were regarded as unproductive data for measurement and were eliminated from the second round of Rasch analysis.

Table 17

*Lower-level Questionnaire Learner Ability Measurement Report (First Round)*
138


The second round of Rasch analysis (Table 18) revealed that only nine out of the remaining 126 learners had fit estimates out of the expected range. The rate of misfit was 7.1%, which was a little bit higher but very close to the threshold of 5% (Boone, Staver, & Yale, 2014). However, according to an email communication by Dr. Linacre as cited in Wang (2012), if misfit rate is lower than 10%, the overall model fit will not be affected.

Table 18

Lower-level Questionnaire Learner Ability Measurement Report (Second Round)
In terms of the facet of descriptor item difficulty, as shown in Table 19, only Item 2 was identified as misfitting since its fit estimate was higher than 1.5. This item did not fit very well with the main construct and was then removed from further analysis. This procedure enhanced the possibility that all descriptors selected were measuring the same construct, i.e., skills and knowledge involved in lower-level CSL decoding. After Item 2 was removed, the researcher ran Rasch analysis again and the item difficulty measurement is reported in Table 20.

**Table 19**

*Lower-level Questionnaire Item Difficulty Measurement Report*
Parameter Estimates and Item Calibrations

The calibrations for learner ability and descriptor difficulty are shown graphically in a Wright Map (person-item map) (Table 21). “Measr” in the first column stands for the equal-interval logit “ruler”, which provides a frame of reference for calibrating two facets (i.e., learner ability and descriptor difficulty) on the same scale. The second column is the distribution of learner ability while the third column shows the distribution of descriptor item difficulty. In the second column, “A”, “B”, “C”, and “D” stand for learners’ Chinese class levels, 1st, 2nd, 3rd, and 4th-year respectively. For example, “A1” stands for the first learner in the 1st-year group. Column three shows descriptor item ID. The Wright Map is
presented in a way that higher-performing students have higher logit values and harder items have higher logit values. Therefore, higher-located items are relatively difficult and items on the bottom are relatively easy. Item 29 was estimated as the most difficult descriptor and Item 11 was the easiest one.

As shown in Table 21, the distribution of item difficulty was very close to a normal distribution. Difficulties of all items were within the range of learner ability estimates. In other words, there was no item whose difficulty was above or below the ability range of 1st-year through 4th-year college-level CSL learners. However, there were thirty-eight learners above the highest item difficulty measure (i.e., Item 29) and seventeen learners below the lowest item difficulty measure (i.e., Item 11).

On the Wright Map (Table 21), one can see three gaps in the region of descriptor item difficulty “cut” by items. The first gap was between Item 3 and 9. The second was between Item 26 and Item 12. The third gap located between Item 10 and Item 13. It is widely accepted that a gap of one logit represents a year of learning (Boone, Staver, & Yale, 2014). A gap of more than one logit may indicate that some important concepts have been missed from the definition of the trait as expressed by descriptor items (Boone, Staver, & Yale, 2014). In this Lower-level Questionnaire Wright map, gaps were all smaller than one logit, meaning that the construct of lower-level CSL reading proficiency was well represented by descriptors in this instrument.
To investigate the learner ability facet, the researcher resorts to the learner ability measurement report (Table 18). The second column shows the average observed score each learner gets. The fifth column displays the student ability measure on the logit scale: the larger the measure, the higher the ability. The fit statistics are shown in columns seven to ten, where Outfit mean-square estimates between 0.5-1.5 are expected. As reported previously, in the second round of Rasch analysis, only nine out of remaining 126 learners had fit mean-square estimates out of the expected range. Since the percentage of misfitting learners was lower than 10%, it did not affect overall model fit (Wang, 2012). The separation ratio (6.94), separation reliability (0.98), chi-square (5635.2), and
significant level (p<0.05) all indicated that 126 learners were significantly differentiated in terms of reading ability.

The item parameter estimates and statistics are summarized in the item difficulty measurement report (Table 19 and Table 20). As shown in Table 19, the fit statistics in columns seven to ten indicated a really good model fit: all except one item (Item 2) had Infit mean-square values between 0.5-1.5. All items except Item 2 measured the same construct and the data was productive for the Rasch analysis. The estimated discrimination of each item is shown in the eleventh column, where 1.0 is expected and values ranging from 0.5 to 1.5 provide a reasonable fit to the Rasch model. Only one item (i.e., Item 2 again) had estimated discrimination out of this range. Item 2 was then eliminated and the number of surviving descriptor items in the Lower-level Questionnaire reduced from 29 to 28.

After Item 2 was deleted, the researcher ran Rasch analysis again with the remaining 28 items and the item parameter estimates were reported in Table 20. The item difficulty estimates shown in the fifth column, ranged from 1.33 logits for Item 29 to -2.58 logits for Item 11. The average difficulty estimate was -0.65 logit with a standard error 0.15. Lastly, the separation ratio (6.39), separation reliability (0.98), Chi-square (1128.2), and significance level (p<0.05) all suggested a significant difference among descriptors in terms of difficulty.

**Evaluation of Rating Categories**

The instrument used for data collection, the CSL Reading Proficiency Questionnaire, had five rating categories (0-4). Table 1 in Chapter 3 presents a full description for each rating category. Category statistics are reported in Table 22. The
second column and third column show the frequency count and proportion of each rating category. If frequency counts of a category are lower than 10, this rating category should be merged with adjacent categories (Linacre, 1999). In the current study, frequency counts of all categories were higher than 158.

The first and second rating category (i.e., Category 0 and 1) accounted for 4%, 22% of all frequency counts respectively. The fourth and fifth rating category (i.e., Category 3 and 4) took up 26% and 13%. The third category (i.e., Category 2), which was the midpoint, accounting for 35% of all ratings, was the most preferred among five categories. The fifth column, “Avg meas” is the average ability of leaners who were rated to be at each rating category. For instance, the average measure of Category 0 was -3.19. It meant that the average ability of all learners who were rated to be at Category 0 (i.e., “can not be expected to perform like this.”) was -3.19 logits. As shown in Table 22, the average measures of learners being rated at Category 0, 1, 2, 3, 4 were -3.19, -1.74, 0.07, 2.35, and 4.10 logits respectively. Values of average measure increased monotonically and this aligned with Rasch model.

The thresholds measure, which is also called the step calibrations measure, was -4.14 between the first and second category (i.e., Category 0 and Category 1), -1.30 between the second and third category (i.e., Category 1 and Category 2), 1.53 between the third and fourth category (i.e., Category 2 and Category 3), and 3.90 between the fourth and last category (i.e., Category 3 and Category 4). The values of the thresholds measure increased monotonically, indicating that rating categories of this instrument were well defined. Differences among thresholds measure values were all larger than 1.4 but smaller than 5. This indicated that differences among rating categories were significant
but not too large.

The seventh column shows Outfit mean-square. Values of Outfit mean-square of all categories were between 0.8-1.2, lower than 2, indicating that unexplained variance was not larger than explained variance (Linacre, 1999). All five rating categories were well established and did not introduce unexplained variance (or noise) into the data.

Table 22

*Lower-level Questionnaire Category Statistics*

<table>
<thead>
<tr>
<th>DATA</th>
<th>Category</th>
<th>Counts</th>
<th>QUALITY CONTROL</th>
<th>STEP</th>
<th>EXPECTATION</th>
<th>MOST</th>
<th>Cum. Cat</th>
<th>Obsd-Expd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>Used</td>
<td>%</td>
<td>%</td>
<td>Measure</td>
<td>S.E.</td>
<td>Measure</td>
<td>at</td>
</tr>
<tr>
<td>0</td>
<td>158</td>
<td>4%</td>
<td>4%</td>
<td>-3.19</td>
<td>-3.42</td>
<td>1.2</td>
<td>(-5.24)</td>
<td>Low</td>
</tr>
<tr>
<td>1</td>
<td>759</td>
<td>22%</td>
<td>26%</td>
<td>-1.74</td>
<td>-1.74</td>
<td>1.1</td>
<td>(-4.14)</td>
<td>Low</td>
</tr>
<tr>
<td>2</td>
<td>1234</td>
<td>35%</td>
<td>61%</td>
<td>0.07</td>
<td>0.17</td>
<td>0.9</td>
<td>-1.30</td>
<td>Low</td>
</tr>
<tr>
<td>3</td>
<td>924</td>
<td>26%</td>
<td>97%</td>
<td>2.35</td>
<td>2.28</td>
<td>0.8</td>
<td>1.53</td>
<td>Low</td>
</tr>
<tr>
<td>4</td>
<td>453</td>
<td>13%</td>
<td>100%</td>
<td>4.10</td>
<td>4.06</td>
<td>1.0</td>
<td>3.96</td>
<td>Low</td>
</tr>
</tbody>
</table>

Figure 15

*Lower-level Questionnaire Response Category Probability Curves*

Figure 15 shows the category probability curves, which is in accordance with the category statistics reported in Table 22. The horizontal axis shows learner ability estimates. It graphically depicts the probabilities of learners with certain ability measures.
being rated to be at Category 0, 1, 2, 3, and 4. The curves for the second, third and fourth categories (i.e., Category 1, 2, and 3) were similarly thin. Each category (0, 1, 2, 3, and 4) formed a nice independent normal curve.

In sum, category statistics reported above, including the frequency count, average measure, thresholds measure, category probability curves, and category Outfit mean-square, all indicated that, rating categories of the *Lower-level Questionnaire* were well defined and appropriately designed. In terms of rating categories, this instrument was of high quality and fit in with the Rasch model.

**Determining of Cut-off Points**

Twenty-eight surviving descriptor items of the *Lower-level Questionnaire* had now been calibrated on a common logit scale across 3.91 logits, from -2.58 to 1.33 (Table 20 and Table 21). Having constructed a hierarchy of items, the next step was to divide the scale into a number of bands through setting cut-off points. As discussed in Chapter 3, a four-level system with three cut-offs is preferred due to practical needs. It is widely accepted that a gap of one logit represents a year of learning (Boone, Staver, & Yale, 2014). Therefore, 3.91 logits should be divided into three or four levels. However, by checking the person-item map (Table 21), it was noticed that difficulty estimates of descriptor items in this instrument were much lower than ability estimates of the majority of 4th-year CSL learners. In other words, the range of descriptor difficulty estimates was roughly identical to the range of ability estimates of 1st-year through 3rd-year learners. Therefore, it was decided to use three instead of four band levels.

The researcher firstly marked out three equal distances on the logit scale. Two cut-offs were -1.28, and 0.03 logits. To better coincide with the natural gap between Item
and Item 12, and to better coincide with the range of 2\textsuperscript{nd}-year learners’ ability estimates, the first cut-off was adjusted from -1.28 to -1.85 logits and the second cut-off was adjusted from 0.03 to 0.71. Accordingly, descriptor items were divided into three band levels and the difficulty increased from Level 1 to Level 3 (Table 23).

As shown in Table 21, the majority of learners with ability estimates between 1.33 to 0.71 logits (i.e., the range of Level 3 item difficulty estimates), were from 3\textsuperscript{rd}-year CSL classes. Most learners with ability between -2.58 and -1.85 logits, the range of Level 1 item difficulty estimates, were 1\textsuperscript{st}-year CSL learners. Similarly, the majority of learners with ability estimates corresponding to the range of Level 2 item difficulty estimates were 2\textsuperscript{nd}-year learners. In sum, when determining the three item difficulty band levels, the researcher tried to make them coincide with the distribution of 1\textsuperscript{st}-year, 2\textsuperscript{nd}-year, and 3\textsuperscript{rd}-year learner ability respectively. It is hoped that this procedure can enhance the applicability of the *CSL Reading Proficiency Scale* for classroom teaching and learning.

**Table 23**

*Lower-level Questionnaire Band Levels*

<table>
<thead>
<tr>
<th>Level</th>
<th>Cut-off on logit scale</th>
<th>Range on logit scale</th>
<th>Descriptor item number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 3</td>
<td>0.71</td>
<td>0.71 ~ 1.33</td>
<td>29, 28, 8, 9</td>
</tr>
<tr>
<td>Level 2</td>
<td>-1.85</td>
<td>-1.85 ~ 0.71</td>
<td>3, 19, 24, 1, 15, 7, 16, 25, 5, 22, 21, 18, 23, 6, 17, 27, 4, 14, 26</td>
</tr>
<tr>
<td>Level 1</td>
<td>-2.58 ~ -1.85</td>
<td></td>
<td>12, 20, 10, 13, 11</td>
</tr>
</tbody>
</table>

**Rasch Analysis Results of Interim-level Questionnaire**

This section reports Rasch analysis results of the *Interim-level Questionnaire* data. Model fit statistics, learner ability estimates, item difficulty estimates, calibrations for
learner and item facets, rating category statistics as well as category probability curves are reported one after another. Lastly, cut-off points of band levels are discussed.

Model Fit

Table 24 presents an excerpt from the report of fit statistics of learner ability. In this table, learners with fit values (Outfit MNSQ) out of acceptable criteria levels (0.5-1.5) were marked with asterisks. In total, forty-four out of 179 learners were identified as “misfitting”. These 44 responses were regarded as unproductive data for measurement and were eliminated. The remaining 135 responses were analyzed in the second round of Rasch analysis.
The second round of Rasch analysis (Table 25) revealed that only eight out of 135 learners had fit estimates out of the expected range. The rate of misfit was 5.9%. Since the misfit rate was lower than 10%, the overall model fit would not be affected (Wang, 2012). In terms of the item difficulty facet, as shown in Table 26, the fit statistics of all
descriptors except for Item 31 were within the 0.5-1.5 acceptable range. After Item 31 was eliminated, the item difficulty measurement was estimated again and results are shown in Table 27.

Table 25

**Interim-level Questionnaire Learner Ability Measurement Report (Second Round)**

<table>
<thead>
<tr>
<th>Obv 1</th>
<th>Obv 2</th>
<th>Obv 3</th>
<th>Fair-N</th>
<th>Model 1</th>
<th>Infit</th>
<th>Outfit</th>
<th>Estimat.</th>
<th>Obs Learner ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>20</td>
<td>19</td>
<td>-1.34</td>
<td>-5.06</td>
<td>-1.54</td>
<td>2.02</td>
<td>2.2</td>
<td>1.55</td>
</tr>
<tr>
<td>20</td>
<td>18</td>
<td>19</td>
<td>-1.34</td>
<td>-5.06</td>
<td>-1.54</td>
<td>2.02</td>
<td>2.2</td>
<td>1.55</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>22</td>
<td>-4.80</td>
<td>-1.48</td>
<td>-0.60</td>
<td>1.40</td>
<td>1.8</td>
<td>1.58</td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>32</td>
<td>-4.80</td>
<td>-1.48</td>
<td>-0.60</td>
<td>1.40</td>
<td>1.8</td>
<td>1.58</td>
</tr>
<tr>
<td>21</td>
<td>22</td>
<td>20</td>
<td>-1.24</td>
<td>-1.06</td>
<td>-1.27</td>
<td>1.06</td>
<td>2.4</td>
<td>1.03</td>
</tr>
<tr>
<td>22</td>
<td>23</td>
<td>21</td>
<td>-1.24</td>
<td>-1.06</td>
<td>-1.27</td>
<td>1.06</td>
<td>2.4</td>
<td>1.03</td>
</tr>
<tr>
<td>35</td>
<td>36</td>
<td>37</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>37</td>
<td>38</td>
<td>39</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>39</td>
<td>38</td>
<td>37</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>32</td>
<td>-1.24</td>
<td>-1.06</td>
<td>-1.27</td>
<td>1.06</td>
<td>2.4</td>
<td>1.03</td>
</tr>
<tr>
<td>32</td>
<td>31</td>
<td>30</td>
<td>-1.24</td>
<td>-1.06</td>
<td>-1.27</td>
<td>1.06</td>
<td>2.4</td>
<td>1.03</td>
</tr>
<tr>
<td>40</td>
<td>41</td>
<td>42</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>42</td>
<td>41</td>
<td>40</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>44</td>
<td>45</td>
<td>46</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>46</td>
<td>45</td>
<td>44</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>48</td>
<td>47</td>
<td>48</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>48</td>
<td>47</td>
<td>46</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>50</td>
<td>51</td>
<td>52</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>52</td>
<td>51</td>
<td>50</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>54</td>
<td>55</td>
<td>56</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
<tr>
<td>56</td>
<td>55</td>
<td>54</td>
<td>-0.16</td>
<td>-1.34</td>
<td>-0.89</td>
<td>1.57</td>
<td>1.7</td>
<td>1.58</td>
</tr>
</tbody>
</table>

152
Table 26

*Interim-level Questionnaire Item Difficulty Measurement Report*

<table>
<thead>
<tr>
<th>Obsvd</th>
<th>Obsvd</th>
<th>Obsvd</th>
<th>Fair-M</th>
<th>Model</th>
<th>Infit</th>
<th>Outfit</th>
<th>Estim.</th>
<th>% Item difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>349</td>
<td>135</td>
<td>21</td>
<td>2.62</td>
<td>-1.91</td>
<td>.16</td>
<td>1.53</td>
<td>3.6</td>
<td>1.51</td>
</tr>
<tr>
<td>245</td>
<td>135</td>
<td>41</td>
<td>1.82</td>
<td>0.53</td>
<td>.15</td>
<td>1.33</td>
<td>2.5</td>
<td>1.35</td>
</tr>
<tr>
<td>200</td>
<td>135</td>
<td>32</td>
<td>1.91</td>
<td>0.21</td>
<td>.15</td>
<td>1.25</td>
<td>1.2</td>
<td>1.26</td>
</tr>
<tr>
<td>200</td>
<td>135</td>
<td>1.5</td>
<td>3.54</td>
<td>1.42</td>
<td>.15</td>
<td>1.30</td>
<td>1.4</td>
<td>1.37</td>
</tr>
<tr>
<td>278</td>
<td>135</td>
<td>2.1</td>
<td>2.05</td>
<td>-0.21</td>
<td>.15</td>
<td>1.10</td>
<td>1.3</td>
<td>1.17</td>
</tr>
<tr>
<td>232</td>
<td>135</td>
<td>2.5</td>
<td>2.48</td>
<td>-1.69</td>
<td>.15</td>
<td>1.20</td>
<td>0.8</td>
<td>1.03</td>
</tr>
<tr>
<td>292</td>
<td>135</td>
<td>2.2</td>
<td>2.16</td>
<td>-0.54</td>
<td>.15</td>
<td>1.05</td>
<td>0.4</td>
<td>1.05</td>
</tr>
<tr>
<td>270</td>
<td>135</td>
<td>2.8</td>
<td>2.00</td>
<td>-0.03</td>
<td>.15</td>
<td>0.97</td>
<td>-1.1</td>
<td>0.98</td>
</tr>
<tr>
<td>173</td>
<td>135</td>
<td>1.2</td>
<td>1.15</td>
<td>2.17</td>
<td>.10</td>
<td>0.95</td>
<td>-0.7</td>
<td>1.01</td>
</tr>
<tr>
<td>113</td>
<td>135</td>
<td>0.8</td>
<td>0.30</td>
<td>0.91</td>
<td>.10</td>
<td>0.90</td>
<td>0.2</td>
<td>0.91</td>
</tr>
<tr>
<td>280</td>
<td>135</td>
<td>2.1</td>
<td>2.12</td>
<td>-0.40</td>
<td>.15</td>
<td>0.86</td>
<td>-0.6</td>
<td>0.91</td>
</tr>
<tr>
<td>109</td>
<td>135</td>
<td>0.5</td>
<td>0.09</td>
<td>0.40</td>
<td>.10</td>
<td>0.82</td>
<td>-0.2</td>
<td>0.90</td>
</tr>
<tr>
<td>159</td>
<td>135</td>
<td>1.2</td>
<td>1.13</td>
<td>2.63</td>
<td>.10</td>
<td>0.92</td>
<td>-0.6</td>
<td>0.87</td>
</tr>
<tr>
<td>160</td>
<td>135</td>
<td>1.2</td>
<td>1.14</td>
<td>2.60</td>
<td>.10</td>
<td>0.83</td>
<td>-1.4</td>
<td>0.84</td>
</tr>
<tr>
<td>279</td>
<td>135</td>
<td>2.1</td>
<td>2.07</td>
<td>-0.23</td>
<td>.15</td>
<td>0.81</td>
<td>-1.6</td>
<td>0.81</td>
</tr>
<tr>
<td>116</td>
<td>135</td>
<td>0.9</td>
<td>0.76</td>
<td>3.82</td>
<td>.17</td>
<td>0.84</td>
<td>-1.2</td>
<td>0.80</td>
</tr>
<tr>
<td>148</td>
<td>135</td>
<td>1.1</td>
<td>3.04</td>
<td>2.91</td>
<td>.16</td>
<td>0.77</td>
<td>-2.0</td>
<td>0.71</td>
</tr>
<tr>
<td>165</td>
<td>135</td>
<td>1.2</td>
<td>1.30</td>
<td>2.47</td>
<td>.10</td>
<td>0.73</td>
<td>-2.6</td>
<td>0.73</td>
</tr>
<tr>
<td>222</td>
<td>135</td>
<td>1.6</td>
<td>1.65</td>
<td>1.09</td>
<td>.15</td>
<td>0.69</td>
<td>-2.8</td>
<td>0.69</td>
</tr>
<tr>
<td>104</td>
<td>135</td>
<td>0.8</td>
<td>0.64</td>
<td>4.20</td>
<td>.10</td>
<td>0.78</td>
<td>-1.7</td>
<td>0.68</td>
</tr>
<tr>
<td>137</td>
<td>135</td>
<td>1.0</td>
<td>0.95</td>
<td>3.21</td>
<td>.17</td>
<td>0.79</td>
<td>-2.7</td>
<td>0.66</td>
</tr>
<tr>
<td>209</td>
<td>135</td>
<td>1.5</td>
<td>1.54</td>
<td>1.40</td>
<td>.15</td>
<td>0.65</td>
<td>-3.2</td>
<td>0.65</td>
</tr>
<tr>
<td>110</td>
<td>135</td>
<td>1.0</td>
<td>0.76</td>
<td>4.01</td>
<td>.18</td>
<td>0.72</td>
<td>-2.3</td>
<td>0.64</td>
</tr>
</tbody>
</table>

**Table 27**

*Interim-level Questionnaire Item Difficulty Measurement Report (Without Misfitting Items)*
Parameter Estimates and Item Calibrations

The calibrations for learner ability and descriptor difficulty are shown graphically in a Wright map (Table 28). In this Wright Map, higher-located descriptors are relatively difficult and descriptors on the bottom are relatively easy. The descriptor item difficulty estimates ranged from -3.71 to 4.32 logits. Item 51 was estimated as the most difficult descriptor while Item 38 the easiest one. The distribution of item difficulty seemed to be skewing toward the low end of the item measure range. Difficulties of all items were within the range of learner ability estimates. There was no item whose estimated difficulty was above or below the ability range of 1st-year through 4th-year college-level CSL learners.

As compared with the Lower-level Questionnaire (Table 21), the range of Interim-
level Questionnaire item difficulty estimates was much wider. Its range was also more identical with the range of learner ability distribution. However, there were still a couple of learners whose ability estimates were out of the range of item difficulty estimates.

On the Wright map, there are several very obvious gaps in the region of item difficulty “cut” by items. The gap between Item 38 and Item 39 was larger than one logit. The gaps between Item 39 and Item 43, Item 34 and Item 46 are very close to one logit.

Table 28

Interim-level Questionnaire Person-item Map

Learner ability measurement report (Table 25) was reexamined to investigate the learner ability facet. The fit statistics are shown in columns seven to ten, where Outfit mean-square estimates between 0.5-1.5 are expected. As reported previously, in the
second round of Rasch analysis, only eight out of 135 learners had fit mean-square estimates out of the expected range. Since the percentage of misfitting learners was lower than 10\%, it would not affect overall model fit (Wang, 2012). The separation ratio (6.87), separation reliability (0.98), chi-square (5988.4), and significant level (p<0.05) all imply that learners were significantly differentiated in terms of reading ability.

The item parameter estimates and statistics are summarized in the item difficulty measurement report (Table 26 and Table 27). As shown in Table 26, the Outfit MnSq value of Item 31 was out of the acceptable criteria levels (0.5-1.5) and this item was eliminated from the second round of Rasch analysis. The number of Interim-level Questionnaire items became 23. In Table 27, the item difficulty estimates shown in the fifth column, ranged from 4.32 logits for Item 51 to -3.71 logits for Item 38. The average difficulty estimate was 1.54 logit with a standard error 0.16. The fit statistics in columns seven to ten indicated a really good model fit: after Item 31 was eliminated, all items had Infit mean-square values between 0.5-1.5, indicating that all items measured the same construct and the data was productive for the Rasch model. The estimated discrimination of each item was shown in the eleventh column, where 1.0 was expected and values ranging from 0.5 to 1.5 provided a reasonable fit to the Rasch model. All items except for Item 38 had estimated discrimination values within the acceptable range. The estimated discrimination of Item 38 was 0.48, very close to the 0.5 cut-off. Lastly, The separation ratio (12.65), separation reliability (0.99), chi-square (3485.3), and significant level (p<0.05) all suggested a significant difference among descriptors in terms of difficulty.

**Evaluation of Rating Categories**

The Interim-level Questionnaire had five rating categories (0-4). Category
statistics are reported in Table 29. If frequency counts of a category are lower than 10, this rating category should be merged with adjacent categories (Linacre, 1999). In the current study, frequency counts of all categories were higher than 174.

The first and second rating category (i.e., Category 0 and Category 1) accounted for 21%, 30% of all frequency counts respectively. The fourth and fifth rating category (i.e., Category 3 and Category 4) took up 15% and 6%. The third category (i.e., 2), which was the midpoint, accounting for 28% of all ratings. The fifth column, “Ave meas” is the average ability of learners who are rated to be at each rating category. The average ability measures of learners being rated at Category 0, 1, 2, 3, 4 were -5.77, -2.91, -0.29, 2.37, and 4.94 logits respectively. Values of average measure increased monotonically and this aligned with Rasch model.

The thresholds measure was -4.70 between the first and second category (i.e., Category 0 and 1), -1.56 between the second and third category (i.e., Category 1 and 2), 1.59 between the third and fourth category (i.e., Category 2 and 3), and 4.67 between the fourth and last category (i.e., Category 3 and 4). The values of the thresholds measure increased monotonically, indicating that rating categories of this instrument were well defined. Differences among thresholds measure values were all larger than 1.4 but smaller than 5. It meant that differences among rating categories were significant but not too large.

The seventh column shows Outfit mean-square. Values of Outfit mean-square of all categories were between 0.8-1.2, lower than 2, indicating that unexplained variance was not larger than explained variance (Linacre, 1999). All five rating categories were well established and did not introduce unexplained variance (or noise) into the data.
Table 29

Interim-level Questionnaire Category Statistics

<table>
<thead>
<tr>
<th>Score</th>
<th>Category Counts</th>
<th>Cum. %</th>
<th>QUALITY CONTROL</th>
<th>STEP</th>
<th>EXPECTATION</th>
<th>MOST</th>
<th>.5 Cum. Cat</th>
<th>Obsd-Expd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>652</td>
<td>21%</td>
<td>.8</td>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td>100%</td>
</tr>
<tr>
<td>1</td>
<td>943</td>
<td>30%</td>
<td>1.91</td>
<td>-5.77</td>
<td>5.66</td>
<td>.9</td>
<td>4.70</td>
<td>.07</td>
</tr>
<tr>
<td>2</td>
<td>858</td>
<td>28%</td>
<td>-2.9</td>
<td>5.66</td>
<td>9.0</td>
<td>4.70</td>
<td>-7.12</td>
<td>.07</td>
</tr>
<tr>
<td>3</td>
<td>478</td>
<td>15%</td>
<td>-2.37</td>
<td>1.15</td>
<td>9.0</td>
<td>4.70</td>
<td>-7.12</td>
<td>.07</td>
</tr>
<tr>
<td>4</td>
<td>174</td>
<td>6%</td>
<td>4.94</td>
<td>5.17</td>
<td>1.2</td>
<td>4.67</td>
<td>11</td>
<td>.07</td>
</tr>
</tbody>
</table>

Figure 16

Interim-level Questionnaire Response Category Probability Curves

Figure 16 shows the category probability curves, which is in accordance with the category statistics reported in Table 29. It shows the probabilities that a learner with certain ability estimates to be rated at Category 0 to Category 4. Each category (0, 1, 2, 3, and 4) formed a nice independent normal curve. The curves for the second, third and fourth categories (i.e., Category 1, 2, and 3) were roughly similarly thin.

In sum, category statistics reported above, including frequency count, average measure, threshold measure, category probability curves, and category Outfit mean-square, all indicate that rating categories of the Interim-level Questionnaire were well defined and appropriately determined. In terms of rating categories, this instrument was
of high quality and fit in with the Rasch model.

**Determining of Cut-off Points**

Twenty-three surviving descriptor items of the *Interim-level Questionnaire* had now been calibrated on a common logit scale ranging from 4.32 to -3.71 logits. The next step was to set cut-offs to divide the scale into a number of bands. As discussed in Chapter 3, a four-level system with three cut-offs was preferred due to practical needs. The researcher firstly marked out four equal distances on the logit scale. As shown in Table 30, the cut-off points dividing four band levels on the logit scale were finally decided to be -1.71, 0.29, and 2.29 logits respectively. Table 30 also shows items on each band level. The difficulty increased from Level 1 to Level 4.

According to the Wright map (Table 28), which presents the correspondence between descriptor item difficulty and learner ability, the majority of learners with ability estimates between 2.29–4.32 logits (i.e., range of Level 4 item difficulty estimates), were from 4th-year CSL classes. Similarly, the majority of learners with ability estimates corresponding to the ranges of Level 1, Level 2, and Level 3 item difficulty estimates were from 1st-year, 2nd-year, and 3rd-year classes respectively. This correspondence indicated that items on Level 1 to Level 4 could roughly reflect the average reading ability of 1st-year through 4th-year CSL learners correspondingly.

**Table 30**

*Interim-level Questionnaire Band Levels*

<table>
<thead>
<tr>
<th>Level</th>
<th>Cut-off on logit scale</th>
<th>Range on logit scale</th>
<th>Descriptor item number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 4</td>
<td>2.29</td>
<td>2.29–4.32</td>
<td>51, 53, 49, 52, 37, 50, 45</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.29</td>
<td>0.29–2.29</td>
<td>35, 48, 36, 46, 34, 44, 42, 47</td>
</tr>
</tbody>
</table>
Rasch Analysis Results of the Higher-level Questionnaire

This section reports Rasch analysis results of the Higher-level Questionnaire data. Model fit statistics, learner ability estimates, item difficulty estimates, calibrations for learner and item facets, rating category statistics as well as category probability curves are reported successively. The last section discusses how to determine cut-off points and how to cut descriptors into band levels.

Model Fit

Table 31 presents an excerpt from the report of fit statistics of learner ability. In this table, learners with fit values (Outfit MNSQ) out of acceptable criteria levels (0.5-1.5) were marked with asterisks. In total, one hundred and fourteen out of 179 learners were identified as “misfitting”. Since the number of “misfitting” leaners was large, the criteria levels of fit were extended. According to an email communication by Dr. Linacre as cited in Wang (2012), fit statistics values in the range between 0.4 and 1.8 could suggest an acceptable fit of the data. Therefore, only 100 leaners whose fit values out of the range between 0.4 and 1.8 were eliminated from the second round of Rasch analysis.
### Table 31

**Higher-level Questionnaire Learner Ability Measurement Report (First Round)**

<table>
<thead>
<tr>
<th>Score</th>
<th>Count</th>
<th>Average</th>
<th>Measure</th>
<th>Std</th>
<th>Zstd</th>
<th>Zstd (Disc.)</th>
<th>Learner Ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>1.5</td>
<td>1.48</td>
<td>-51</td>
<td>.97</td>
<td>4.52</td>
<td>3.2</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2.5</td>
<td>2.46</td>
<td>3.14</td>
<td>.98</td>
<td>4.13</td>
<td>3.1</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.8</td>
<td>1.75</td>
<td>.44</td>
<td>.99</td>
<td>3.82</td>
<td>2.4</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.8</td>
<td>3.88</td>
<td>8.37</td>
<td>1.41</td>
<td>5.02</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>1.5</td>
<td>1.48</td>
<td>.51</td>
<td>.97</td>
<td>3.12</td>
<td>3.25</td>
<td>2.5</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2.8</td>
<td>2.72</td>
<td>3.96</td>
<td>.92</td>
<td>7.09</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.8</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.0</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.0</td>
<td>3.08</td>
<td>8.37</td>
<td>1.41</td>
<td>5.02</td>
<td>1.5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2.5</td>
<td>2.46</td>
<td>3.14</td>
<td>.98</td>
<td>3.82</td>
<td>2.4</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.3</td>
<td>3.38</td>
<td>5.79</td>
<td>.99</td>
<td>2.28</td>
<td>1.6</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2.8</td>
<td>2.72</td>
<td>3.96</td>
<td>.92</td>
<td>7.09</td>
<td>2.3</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.00</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3.0</td>
<td>3.01</td>
<td>4.85</td>
<td>.96</td>
<td>1.98</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-2.48</td>
<td>1.03</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td>1.0</td>
<td>1.00</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2.8</td>
<td>2.72</td>
<td>3.96</td>
<td>.92</td>
<td>7.09</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.8</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.0</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.0</td>
<td>3.08</td>
<td>8.37</td>
<td>1.41</td>
<td>5.02</td>
<td>1.5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2.5</td>
<td>2.46</td>
<td>3.14</td>
<td>.98</td>
<td>3.82</td>
<td>2.4</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.3</td>
<td>3.38</td>
<td>5.79</td>
<td>.99</td>
<td>2.28</td>
<td>1.6</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>3.3</td>
<td>3.35</td>
<td>5.68</td>
<td>.98</td>
<td>2.18</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.00</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3.0</td>
<td>3.01</td>
<td>4.85</td>
<td>.96</td>
<td>1.98</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-2.48</td>
<td>1.03</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td>1.0</td>
<td>1.00</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2.8</td>
<td>2.72</td>
<td>3.96</td>
<td>.92</td>
<td>7.09</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.8</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.0</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.0</td>
<td>3.08</td>
<td>8.37</td>
<td>1.41</td>
<td>5.02</td>
<td>1.5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2.5</td>
<td>2.46</td>
<td>3.14</td>
<td>.98</td>
<td>3.82</td>
<td>2.4</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.3</td>
<td>3.38</td>
<td>5.79</td>
<td>.99</td>
<td>2.28</td>
<td>1.6</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3.0</td>
<td>3.01</td>
<td>4.85</td>
<td>.96</td>
<td>1.98</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-2.48</td>
<td>1.03</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td>1.0</td>
<td>1.00</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2.8</td>
<td>2.72</td>
<td>3.96</td>
<td>.92</td>
<td>7.09</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.8</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.0</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.0</td>
<td>3.08</td>
<td>8.37</td>
<td>1.41</td>
<td>5.02</td>
<td>1.5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2.5</td>
<td>2.46</td>
<td>3.14</td>
<td>.98</td>
<td>3.82</td>
<td>2.4</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.3</td>
<td>3.38</td>
<td>5.79</td>
<td>.99</td>
<td>2.28</td>
<td>1.6</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>3.3</td>
<td>3.35</td>
<td>5.68</td>
<td>.98</td>
<td>2.18</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.00</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3.0</td>
<td>3.01</td>
<td>4.85</td>
<td>.96</td>
<td>1.98</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-2.48</td>
<td>1.03</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td>1.0</td>
<td>1.00</td>
<td>-2.49</td>
<td>1.04</td>
<td>2.18</td>
<td>1.4</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>2.8</td>
<td>2.72</td>
<td>3.96</td>
<td>.92</td>
<td>7.09</td>
<td>2.3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.8</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.0</td>
<td>0.75</td>
<td>-3.63</td>
<td>1.09</td>
<td>2.53</td>
<td>1.9</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.0</td>
<td>3.08</td>
<td>8.37</td>
<td>1.41</td>
<td>5.02</td>
<td>1.5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2.5</td>
<td>2.46</td>
<td>3.14</td>
<td>.98</td>
<td>3.82</td>
<td>2.4</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3.3</td>
<td>3.38</td>
<td>5.79</td>
<td>.99</td>
<td>2.28</td>
<td>1.6</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3.0</td>
<td>3.01</td>
<td>4.85</td>
<td>.96</td>
<td>1.98</td>
<td>1.2</td>
</tr>
</tbody>
</table>

162
The second round of Rasch analysis (Table 32) revealed that ten out of remaining 79 learners had fit estimates out of the expected range (0.4-1.8). The rate of misfit was 12.7%, slightly higher than the threshold of 10%. In terms of the item difficulty facet, as shown in Table 33, the fit statistics of all descriptors were within the 0.5-1.5 acceptable range. It meant that all items in this subscale were measuring the same construct.

Table 32
<table>
<thead>
<tr>
<th>Obsvd. Score</th>
<th>Obsvd. Score</th>
<th>Obsvd. Score</th>
<th>Model</th>
<th>Infit</th>
<th>Outfit</th>
<th>Estim.</th>
<th>Score</th>
<th>Nu Learner ability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>2.0</td>
<td>3.9</td>
<td>0.58</td>
<td>0.56</td>
<td>0.43</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>22</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>26</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>28</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>29</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>31</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>32</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
<tr>
<td>33</td>
<td>4</td>
<td>2.0</td>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>1.2</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 33
Parameter Estimates and Item Calibrations

The calibrations for learner ability and descriptor difficulty are shown graphically in a Wright map (Table 34). The descriptor item difficulty estimates ranged from -1.82 to 3.25 logits. Item 56 was estimated as the most difficult descriptor while Item 54 the easiest one. Difficulties of all items were within the range of learner ability estimates. There was no item whose estimated difficulty was above or below the ability range of 1st-year through 4th-year college-level CSL learners. There were a large number of learners whose ability estimates were out of the range of item difficulty estimates. In other words, no descriptors corresponded to learners with extremely low or high reading ability. On the Wright Map, there are two very obvious gaps, which were larger than 1 logit, in the region of item difficulty “cut” by items. The first gap was between Item 57 and Item 56. The second gap, located between Item 55 and Item 54.
Learner ability measurement report (Table 32) was reexamined to investigate the learner ability facet. The separation ratio (3.08), separation reliability (0.90), chi-square (785.5), and significant level (p<0.05) all imply that learners were significantly differentiated in terms of reading ability.

The item parameter estimates and statistics are summarized in the item difficulty measurement report (Table 33). The item difficulty estimates shown in the fifth column, ranged from 3.25 logits for Item 56 to -1.82 logits for Item 54. The average difficulty estimate was 0.62 logit with a standard error 0.22. The fit statistics in columns seven to
ten indicated a really good model fit: all items had Infit mean-square values between 0.5-1.5, indicating that all items measured the same construct and the data was productive for the Rasch model. The estimated discrimination of each item was shown in the eleventh column, where 1.0 was expected and values ranging from 0.5 to 1.5 provided a reasonable fit to the Rasch model. All items had estimated discrimination values within the acceptable range. Lastly, The separation ratio (8.11), separation reliability (0.99), chi-square (250.9), and significant level (p<0.05) all suggested a significant difference among descriptors in terms of difficulty.

**Evaluation of Rating Categories**

The *Higher-level Questionnaire* had five rating categories (0-4). Category statistics are reported in Table 35. If frequency counts of a category are lower than 10, this rating category should be merged with adjacent categories (Linacre, 1999). In the current study, frequency counts of all categories were higher than 22.

The first and second rating category (i.e., Category 0 and Category 1) accounted for 13%, 23% of all frequency counts respectively. The fourth and fifth rating category (i.e., Category 3 and Category 4) took up 20% and 7%. The third category (i.e., 2), which was the midpoint, accounting for 37% of all ratings. The fifth column, “Ave meas” is the average ability of learners who are rated to be at each rating category. The average ability measures of learners being rated at Category 0, 1, 2, 3, 4 were -5.94, -3.07, -0.38, 3.05, and 5.69 logits respectively. Values of average measure increased monotonically and this aligned with Rasch model.

The thresholds measure was -5.06 between the first and second category (i.e., Category 0 and 1), -2.32 between the second and third category (i.e., Category 1 and 2),
1.96 between the third and fourth category (i.e., Category 2 and 3), and 5.41 between the fourth and last category (i.e., Category 3 and 4). The values of the thresholds measure increased monotonically, indicating that rating categories of this instrument were well defined. Differences among thresholds measure values were all larger than 1.4 but smaller than 5, meaning that differences among rating categories were significant but not too large.

The seventh column shows Outfit mean-square. Values of Outfit mean-square of all categories were between 0.6-1.1, lower than 2, indicating that unexplained variance was not larger than explained variance (Linacre, 1999). All five rating categories were well established and did not introduce unexplained variance (or noise) into the data.

Table 35

Higher-level Questionnaire Category Statistics

<table>
<thead>
<tr>
<th>Category Counts</th>
<th>Cum.</th>
<th>QUALITY CONTROL</th>
<th>STEP</th>
<th>EXPECTATION</th>
<th>MOST</th>
<th>.5 Cumul.</th>
<th>Cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>Used</td>
<td>%</td>
<td>%</td>
<td>Meas</td>
<td>Exp</td>
<td>OUTFIT</td>
<td>CALIBRATIONS</td>
</tr>
<tr>
<td>0</td>
<td>41</td>
<td>13%</td>
<td>13%</td>
<td>-5.94</td>
<td>-5.75</td>
<td>.6</td>
<td>(-6.16)</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
<td>23%</td>
<td>86%</td>
<td>-3.87</td>
<td>-3.22</td>
<td>1.1</td>
<td>-5.06</td>
</tr>
<tr>
<td>2</td>
<td>117</td>
<td>37%</td>
<td>73%</td>
<td>-3.38</td>
<td>-3.32</td>
<td>.9</td>
<td>-2.32</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>20%</td>
<td>91%</td>
<td>3.05</td>
<td>2.96</td>
<td>.9</td>
<td>1.96</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>7%</td>
<td>100%</td>
<td>5.69</td>
<td>5.76</td>
<td>.9</td>
<td>5.41</td>
</tr>
</tbody>
</table>

---(Mean)---(Modal)---(Median)---
Figure 17

*Higher-level Questionnaire* Response Category Probability Curves

Figure 17 shows the category probability curves, which is in accordance with the category statistics reported in Table 35. Each category (0, 1, 2, 3, and 4) formed a nice independent normal curve. The curves for the second and fourth categories (i.e., Category 1 and Category 3) were roughly similarly thin, while the curve for Category 3 had a slightly sharper peak, indicating a slightly higher degree of discrimination.

In sum, category statistics reported above, including frequency count, average measure, threshold measure, category probability curves, and category Outfit mean-square, all indicate that rating categories of the *Higher-level Questionnaire* were well defined and appropriately determined. In terms of rating categories, this instrument was of high quality and fit in with the Rasch model.

**Determining of Cut-off Points**

Four descriptor items of the *Higher-level Questionnaire* had now been calibrated on a common logit scale ranging from 3.25 to -1.82 logits. The next step was to set cut-offs to divide the scale into a number of bands. As discussed previously, a four-level
system with three cut-offs was preferred due to practical needs. The researcher firstly marked out four equal distances on the logit scale. Three cut-offs were -0.55, 0.72, and 1.99 logits. To better coincide with the ranges of 2\textsuperscript{nd}-year and 3\textsuperscript{rd}-year learners’ ability estimates, the first cut-off was adjusted from 0.72 to 0.39 logits.

Table 36 shows items on each band level. The difficulty increased from Level 1 to Level 4. According to the Wright map (Table 34), which presents the correspondence between descriptor item difficulty and learner ability, the majority of learners with ability estimates between 1.99–3.25 logits (i.e., range of Level 4 item difficulty estimates), were from 4\textsuperscript{th}-year CSL classes. Similarly, the majority of learners with ability estimates corresponding to the ranges of Level 1 and Level 3 item difficulty estimates were from 1\textsuperscript{st}-year and 3\textsuperscript{rd}-year classes respectively. This correspondence indicated that items on Level 1 to Level 4 could roughly reflect the development trajectory of 1\textsuperscript{st}-year through 4\textsuperscript{th}-year CSL learners. However, as shown in Table 34, there was no item whose difficulty estimates could correspond to 2\textsuperscript{nd}-year learners’ ability estimates.

Table 36

<table>
<thead>
<tr>
<th>Level</th>
<th>Cut-off on logit scale</th>
<th>Range on logit scale</th>
<th>Descriptor item number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 4</td>
<td>1.99</td>
<td>1.99–3.25</td>
<td>56</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.39</td>
<td>0.72–1.99</td>
<td>57, 55</td>
</tr>
<tr>
<td>Level 2</td>
<td>-0.55</td>
<td>-0.55–0.72</td>
<td>N.A.</td>
</tr>
<tr>
<td>Level 1</td>
<td>-1.82– -0.55</td>
<td></td>
<td>54</td>
</tr>
</tbody>
</table>

Conclusions

Based on Rasch analysis results, the CSL Reading Proficiency Scale was finalized
and shown in the three tables below as well as in the Appendix III.

The Lower-level Subscale (Table 37) has 28 items. This scale describes processes and knowledge related to lower-level decoding, such as character recognition, lexical access, syntactic parsing, and knowledge on orthography, vocabulary, and syntax. This scale is divided into three progressive band levels, which roughly correspond to the decoding skill development of 1st-year to 3rd-year CSL learners.

The Interim-level Subscale (Table 38) depicts skills and knowledge related to interim-level textbase construction, such as semantic-proposition encoding, text model of comprehension, knowledge on coherence and rhetorical organization. This scale consists of 23 items divided into four band levels, which depict the development trajectory of higher-level reading processing from 1st-year through 4th-year CSL learners.

The Higher-level Subscale (Table 39) has four descriptors which denote skills and knowledge involved in higher-level reading processes, including the situation model of comprehension and cognitive strategies. This subscale consists of four band levels, which depict the higher-level reading processes and knowledge that can be mastered by 1st-year, 3rd-year, and 4th-year CSL learners.

As shown in Table 37, Table 38, and Table 39, items were further classified into nine provisional groups arrived at through literature review in order to further delineate the factor structure of the CSL reading ability. Specifically, the first factor, lower-level decoding, is indicated by items which depict decoding skills/knowledge from three aspects: orthography, lexical, and syntax. The second factor, interim-level textbase construction, is represented by items which describe semantic-proposition encoding, knowledge of coherence and rhetorical organization, and the text model of
comprehension. The last factor, higher-level situation-model building, is indicated by four items placed into two provisional groups: the situation model of comprehension and pragmatic knowledge, cognitive strategy.

Table 37

**CSL Reading Proficiency Scale: Lower-level Decoding**

<table>
<thead>
<tr>
<th>Level</th>
<th>Measure (Logits)</th>
<th>Item</th>
<th>Provisional group</th>
<th>Level the its original scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 3</td>
<td>Item 29: Can have emerging control of high-frequency grammatical constructions common to expository prose. 开始表现出能掌握说明文中常见的高频语法结构的能力。</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Superior) ***</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 28: Can comprehend text successfully when the number of features and structural devices from the literary/classical stratum is limited. 当来自文学作品、文言文的特色和语法结构很少时，能够成功地理解文本。</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Superior) ***</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 8: Can use knowledge on word formation to help decode unknown words. 能用构词法知识帮助识别不认识的词。</td>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 9: When reading text on his/her level, can contextually process polysemes (words with multiple meanings). 读适合他/她阅读水平的文本时，能根据语境理解多义词在特定语境中的准确意思。</td>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 3: Can distinguish graphically-similar characters, which have learned. 对于已经学过的形近字，能够区分。</td>
<td>1) Character Recognition and Orthographic Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 19: Can comprehend sentences with following basic</td>
<td>3) Syntactic Parsing and</td>
<td>ACTFL-CHIN (Advanced-</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 24: Can be aware of the appropriate position of adverbial phrases in Chinese sentences.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>--------------------------</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 1: Can use knowledge of radicals to help decode characters.</td>
<td>1) Character Recognition and Orthographic Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item 15: Can understand roughly the relationship of multiple modifiers to element being modified, but may have some confusion.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-High)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item 7: Can analyze the intra-word structure.</td>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item 16: Can comprehend sentences with following basic patterns: Relative clause modification.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus) **</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item 25: Can comprehend sentences with following unique Chinese syntactic features: Nominal sentences (i.e., sentences contain no verbs)</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item 5: Can derive meaning of...</td>
<td>2) Lexical</td>
<td>ACTFL-CHIN</td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>Item</td>
<td>Can comprehend sentences with following basic patterns:</td>
<td>Syntactic Parsing and Syntactic Knowledge</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------------------------------------------------------</td>
<td>------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>22</td>
<td>Resultative compounds.</td>
<td>ACTFL-CHIN (Advanced-plus) **</td>
<td></td>
</tr>
<tr>
<td>-0.58</td>
<td></td>
<td>能理解含有动补短语的句子的意思。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>21</td>
<td>Various aspect markers.</td>
<td>ACTFL-CHIN (Advanced-plus) **</td>
<td></td>
</tr>
<tr>
<td>-0.67</td>
<td></td>
<td>能读懂含有下列语法的句子: 各种体标记（e.g., 着，了，过）</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>18</td>
<td>bei-passive.</td>
<td>ACTFL-CHIN (Advanced-plus) **</td>
<td></td>
</tr>
<tr>
<td>-0.69</td>
<td></td>
<td>能读懂含有下列语法的句子: 被字句</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>23</td>
<td>Directional compounds.</td>
<td>ACTFL-CHIN (Advanced-plus) **</td>
<td></td>
</tr>
<tr>
<td>-0.78</td>
<td></td>
<td>能读懂含有下列语法的句子: 方向补语</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>6</td>
<td>When reading text on his/her level, can group adjacent characters into words successfully.</td>
<td>Lexical Access and Vocabulary Knowledge</td>
<td></td>
</tr>
<tr>
<td>-0.84</td>
<td></td>
<td>读合适他/她阅读水平的文本时，能成功地切词。（切词指的是将一个汉字序列切分成一个一个单独的词）</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>17</td>
<td>ba-disposal.</td>
<td>ACTFL-CHIN (Advanced-plus) **</td>
<td></td>
</tr>
<tr>
<td>-0.89</td>
<td></td>
<td>能读懂含有下列语法的句子: 把字句</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>27</td>
<td>Chinese syntactic features: Topic-comments structure</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td>-1.02</td>
<td></td>
<td>能读懂含有下列语法的句子: 主题-评论句</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>4</td>
<td>Can derive meaning of</td>
<td>Lexical Access and Vocabulary Knowledge</td>
<td></td>
</tr>
<tr>
<td>-1.20</td>
<td></td>
<td>2)</td>
<td>ACTFL-CHIN</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 14: Can derive meaning of sentences with the aspect markers (such as le), although still makes errors.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced) *</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 26: Can comprehend sentences with following unique Chinese syntactic features: “Shi…de” structure.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>Created based on theory</td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>Item 12: Can understand simple Noun+Noun modification with <em>de</em> and Stative Verb + Noun modification with <em>de</em>.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-Mid)</td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>Item 20: Can comprehend sentences with following basic patterns: exclusive usages of question words.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus) ***</td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>Item 10: Can understand limited types of sentence structures, including Subject-Predicate constructions, simple SVO constructions, all in affirmative, negative, and simple question form.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-Low)</td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>Measure (Logits)</td>
<td>Item</td>
<td>Provisional group</td>
<td>Level in the original scale</td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>------</td>
<td>-------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Level 1</td>
<td>-2.40</td>
<td>Item 13: Can comprehend preliminarily high-frequency connectors forming complex sentences (e.g., de shí hou, suī rán... kě shì). 能初步理解常见的关联词（例如：……的时候，虽然……可是）。</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-High)*</td>
</tr>
<tr>
<td>Level 1</td>
<td>-2.58</td>
<td>Item 11: Can control certain structural sets, which have a limited number of items, including interrogatives (shuí, shén me), specifiers (zhè, nà), and very common noun measures (kuài, běn) etc. 能掌握那些本身只含有少数子项目的结构，包括简单的疑问词（谁，什么）、指示代词（这，那）、常见的名量词（块，本）等。</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-Mid)</td>
</tr>
</tbody>
</table>

Table 38

*CSL Reading Proficiency Scale: Interim-level Textbase Construction*

<table>
<thead>
<tr>
<th>Level</th>
<th>Measure (Logits)</th>
<th>Item</th>
<th>Provisional group</th>
<th>Level in the original scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 4</td>
<td>4.32</td>
<td>Item 51: Can comprehend standard newspaper items addressed to the general public, routine correspondence, reports and technical material in field of interest. 能理解面向公众的标准新闻报道、日常信件、以及感兴趣的领域的报道和技术性文章。</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Superior)**</td>
</tr>
<tr>
<td>Level 4</td>
<td>4.16</td>
<td>Item 53: Can understand minimally, yet with significant difficulties, prose which is more characteristic of the literary/classical style. 能在极小程度上理解具有文学/文言色彩的文章，但面临重大的困难</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Superior)**</td>
</tr>
<tr>
<td>Level</td>
<td>Item</td>
<td>Description</td>
<td>Knowledge of Rhetorical Organization</td>
<td>ACTFL-CHIN</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-------------</td>
<td>--------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>4</td>
<td>4.13</td>
<td>Item 49: Can comprehend, not just decode, a narrow range of authentic, expository material, including areas of professional interest, without the use of a dictionary.</td>
<td>7) The Text Model of Comprehension</td>
<td>(Superior)**</td>
</tr>
<tr>
<td>4</td>
<td>4.03</td>
<td>Item 52: Can decode, with a dictionary and with substantial effort and moderate error, popular novels, essays, and most literature for the general public.</td>
<td>7) The Text Model of Comprehension</td>
<td>(Superior)**</td>
</tr>
<tr>
<td>4</td>
<td>3.94</td>
<td>Item 37: Can have some appreciation of formulaic rhetorical devices common to more stylized writing.</td>
<td>6) Knowledge of Rhetorical Organization</td>
<td>ACTFL-CHIN (Advanced-plus)*</td>
</tr>
<tr>
<td>4</td>
<td>3.31</td>
<td>Item 50: Can comprehend, with some dictionary use, authentic material over a wide range of subject matter and topics.</td>
<td>7) The Text Model of Comprehension</td>
<td>(Superior)**</td>
</tr>
<tr>
<td>4</td>
<td>3.01</td>
<td>Item 45: Can decode authentic prose (including newspapers and magazines) for general ideas.</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Advanced)</td>
</tr>
<tr>
<td>3</td>
<td>2.71</td>
<td>Item 35: Can be aware of most Chinese text structures/discourse patterns.</td>
<td>6) Knowledge of Rhetorical Organization</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level</td>
<td>Item</td>
<td>Description</td>
<td>Score</td>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>3</td>
<td>48</td>
<td>能理解大部分种类的中文文章结构和语篇模式。</td>
<td>2.68</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>用篇章结构的知识来帮助阅读理解。</td>
<td>2.55</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>能搜集出一些特定的事实,但是会出现错误。</td>
<td>2.34</td>
<td>ACTFL-CHIN (Advanced)*</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>能识别很常见的中文文章结构和语篇模式。</td>
<td>1.47</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>能把握叙事主线。</td>
<td>1.44</td>
<td>ACTFL-CHIN (Intermediate-High)</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>在使用字典的情况下,能理解比口语更难的真实语料的字面意思,找到关键信息。</td>
<td>1.13</td>
<td>ACTFL-CHIN (Intermediate-High)</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>能够迅速、熟练地摄取关键的细节。</td>
<td>0.55</td>
<td>ACTFL-CHIN (Advanced-)</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 30: When reading text on his/her level, can successfully identify referential expressions/antecedents. 当阅读适合他/她水平的文本时，能够成功地找出指代表达/先行词。</td>
<td>4) Semantic-proposition Encoding Created based on theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 33: Can use knowledge on cohesive devices to facilitate reading comprehension. 能用关联词知识来帮助理解文章。</td>
<td>5) Knowledge of Coherence Created based on theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 40: Can comprehend sufficiently specially-prepared discourse for informative purposes. 能充分理解学习者语料，来实现获取信息的目的。</td>
<td>7) The Text Model of Comprehension ACTFL-CHIN (Intermediate-Mid)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 41: With use of a dictionary, can comprehend sufficiently main ideas and some facts in authentic material paralleling oral language. 在使用字典的情况下阅读跟口语类似的真实语料，能够充分理解文本的主旨和一些具体事实。</td>
<td>7) The Text Model of Comprehension ACTFL-CHIN (Intermediate-Mid)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 32: Can be aware of commonly used cohesive devices within and across paragraphs. 能识别段落内部以及段落间常用的关联词。（关联词包括：连词和有关联作用的介词、副词和短语等）</td>
<td>5) Knowledge of Coherence ACTFL-CHIN (Advanced plus)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 43: Can decode critical elements of public announcements to determine who, when, and where regarding such subjects as public events by using a dictionary. 在使用字典的情况下，能读懂通知中的关键信息的字面意思，来确定事件的人物、时间、地点。</td>
<td>7) The Text Model of Comprehension ACTFL-CHIN (Intermediate-Mid)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Level 2 | Item 39: Can read simple connected, specially prepared material for basic survival and | 7) The Text Model of Comprehension ACTFL-CHIN (Intermediate-
social needs (Note: “specially prepared material” hereafter refers to “material specially prepared for adult second language learners of Chinese”).

能阅读简单、连贯的学习者语料，以满足基本的生存和社交需要。（本量表中“学习者语料”是指为成年中文二语学习者而特殊编辑的阅读材料，不是真实语料。）

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Item 38: Can comprehend set expressions using basic vocabulary.</th>
<th>7) The Text Model of Comprehension</th>
<th>ACTFL-CHIN (Novice-High)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.71</td>
<td>能理解由基础词汇构成的固定表达。例如“今天天气怎么样?”“请问现在几点?”等等。</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 39

*CSL Reading Proficiency Scale: Higher-level Situation-model Building*

<table>
<thead>
<tr>
<th>Level 4</th>
<th>Item 56: Can appreciate, to a limited degree, nuances or stylistics, but gaps in detail may be frequent.</th>
<th>8) The Situation Model of Comprehension and Pragmatic Knowledge</th>
<th>ACTFL-CHIN (Superior)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.25</td>
<td>能在有限程度上欣赏文章的精妙之处和文体风格，但是在细节的把握上可能常常有缺失。</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3</th>
<th>Item 57: Can use preliminarily lexical guessing strategies in dealing with authentic materials, although errors are frequent.</th>
<th>9) Cognitive Strategy</th>
<th>ACTFL-CHIN (Intermediate-High)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.65</td>
<td>能初步运用“猜测生词的意思”这一阅读策略来处理真实语料，虽然常常出现错误。</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3</th>
<th>Item 55: Can minimally detect subjective attitudes, values, and judgments.</th>
<th>8) The Situation Model of Comprehension and Pragmatic Knowledge</th>
<th>ACTFL-CHIN (Advanced-plus)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.39</td>
<td>能在极小程度上体察文章</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>Item</td>
<td>Question</td>
<td>ACTFL-CHIN (Advanced-plus)***</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>----------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>54</td>
<td>Item 54: When reading text on his/her level, can only draw simplest inferences from reading. 当阅读适合该读者水平的文章时，只能根据文章做最简单的推论。</td>
<td>8) The Situation Model of Comprehension and Pragmatic Knowledge</td>
</tr>
</tbody>
</table>
CHAPTER 6
DISCUSSION AND CONCLUSIONS

This chapter starts with a review of this study, followed by a discussion of the primary findings. Discussion focuses on two topics: (1) the factor structure of reading ability as measured by the *CSL Reading Proficiency Scale*, and (2) descriptor difficulty of the *CSL Reading Proficiency Scale*. The implications of findings for CSL test development, classroom teaching as well as textbook writing, and reading theory are summarized. After the study’s unique contributions are discussed, limitations and recommendations for future research are pointed out at the end.

**Overview of the Study**

This dissertation research developed and validated a *CSL Reading Proficiency Scale* through a data-driven approach. Following scale development and validation procedures were adopted in order to enhance the possibility that the *CSL Reading Proficiency Scale* was both theoretically solid and empirically verified. The scale development preparatory stage included: (1) on the basis of dimension analysis of reading ability as conceptualized in reading theory research, the construct of CSL reading ability was modeled as consisting of a constellation of subskills and knowledge (Figure 8); (2) an operational descriptive scheme with four descriptive parameters (i.e., cognitive verb, object of the verb, text type, and modifier) were set to operationalize the theoretical construct of CSL reading; (3) a descriptor pool was built up through comprehensively documentation of reading proficiency descriptors from theoretical works and from existing proficiency scales. Descriptors collected were edited to include above
mentioned descriptive parameters; (4) as a way of qualitative validation, four experienced college-level CSL teachers were invited to review the bank of descriptors, revising descriptors and screening descriptors unrelated to college-level CSL instruction.

Next, descriptors surviving from the preparatory stage were compiled into the *CSL Reading Proficiency Questionnaire*. Ninety-five experienced college-level CSL teachers evaluated students’ reading proficiency against this instrument. Data generated were subjected to statistical analyses to answer following research questions which guided the development of the *CSL Reading Proficiency Scale*:

Dimensions of descriptors: 1. What is the factor structure of CSL reading, as measured by descriptors in the *CSL Reading Proficiency Scale*?

Difficulty levels of descriptors: 2. What is the estimated level of difficulty of each descriptor? 3. How are cut-off points selected to divide descriptors into multiple bands?

**Summary of the Primary Findings**

The first research question inquired the dimensions of CSL reading ability as measured by descriptors in the *CSL Reading Proficiency Scale*. To answer this research question, confirmatory factor analysis was used to test three hypothesized models arrived at through literature review: one-factor model, correlated two-factor model, and correlated three-factor model. Data in the current study comforted to both the latter two models. However, the correlated three-factor model fit the data slightly better than the correlated two-factor model. This indicates that CSL reading, as measured by the *CSL Reading Proficiency Questionnaire*, should be modeled as a three-factor construct. The three inter-correlated components of CSL reading are: lower-level decoding; interim-level textbase construction; higher-level situation-
Accordingly, 57 descriptors in the CSL Reading Proficiency Questionnaire were classified into three subscales: the Lower-level Questionnaire which measured lower-level processes and knowledge of CSL reading (indicated by descriptor Item 1-Item 29); the Interim-level Questionnaire that measures knowledge and skills related to interim-level textbase construction (indicated by Item 30-Item 53); the Higher-level Questionnaire which measured higher-level processes and knowledge of CSL reading (indicated by descriptor Item 54-Item 57).

The second research question investigated difficulty estimates of descriptors in the CSL Reading Proficiency Questionnaire and how to select cut-offs to divide descriptors into multiple band levels. Rasch rating scale analysis was conducted to calibrate descriptors in the Lower-level Questionnaire, Interim-level Questionnaire, and the Higher-level Questionnaire separately. Item 2 and Item 31 were eliminated because they were identified as misfitting. Fifty five surviving descriptor items constitute the CSL Reading Proficiency Scale that has three subscales: lower-level decoding; interim-level textbase construction; higher-level situation-model building. These three subscales are also called Lower-level Subscale, Interim-level Subscale, and Higher-level Subscale for short.

The difficulty estimates of remaining 28 items (Item 1, Item 3-Item 29) in Lower-level Subscale descriptors ranged from -2.58 to 1.33 logits. Two cut-offs (i.e., -1.85, and 0.71 logits) were selected to divide calibrated descriptors into three band levels. Level 1, the easiest band level, consisted of five descriptors. Level 2 and Level 3 had nineteen and four descriptors respectively.

The difficulty estimates of 23 items (Item 30, Item 32-Item 53) in the Interim-level Subscale ranged from 4.32 logits to -3.71 logits. Three cut-offs (i.e., -1.71, 0.29, and 2.29 logits)
were selected to divide calibrated descriptors into four band levels. Level 1, the easiest band level, consisted of one descriptor. Level 2, 3, and 4 comprised seven, eight, and seven descriptors respectively.

The *Higher-level Subscale* consists of four items (Item 54-Item 57), ranging from -1.82 logits to 3.25 logits. These items were divided into four bands by three cut-offs (i.e., -0.55, 0.039, and 1.99 logits). Level 1, 3, 4 had one, two, and one item respectively. There was no item placed into Level 2. The finalized *CSL Reading Proficiency Scale* is shown in Appendix III.

**Discussion**

Discussion centers on the following two topics: (1) dimensions of CSL reading ability as measured by the *CSL Reading Proficiency Scale*, and (2) difficulty of descriptors in the *CSL Reading Proficiency Scale*.

**Dimensions of Reading Ability**

As discussed in Chapter 2, there is no consensus about whether reading competence is a unitary skill or a cluster of component skills. Data in this study comforted to two componential models, i.e., the correlated two-factor model and the corrected three-factor model. On the contrary, the one-factor model did not fit the data very well. In other words, this study does not find evidence that CSL reading is an indivisible skill.

This finding provides empirical evidence for the component skills approach of reading research, which is widely supported by previous studies (e.g., Alderson, 1990; Grabe, 2009; Koda, 2005; Rumelhart, 1980). Modeling reading ability as a multiple-component construct has provided a productive approach to studying reading skills cross-language transfer and measuring L2 reading development. It is also an effective way of teaching and training CSL reading ability.

However, the interrelationships among component reading skills are far from clear
Previous studies proposed a number of componential models of reading, such as Davis (1968: four skills); Munby (1978: 18 skills); Lunzer, Waite, and Dolan (1979: eight skills); and Grabe (1991: six skills). This study only examined two very parsimonious models: correlated two-factor model (Grabe, 2009) and correlated three-factor model (Koda, 2005). As compared with the two-factor model which conceptualizes CSL reading as composed of two levels of knowledge and skills, the three-factor model fits data slightly better. In other words, CSL reading can be perceived as a three-factor construct, consisting of three components: lower-level decoding; interim-level textbase construction; and, higher-level situation-model building.

Within both the three-factor model and the two-factor model, factors were highly related. In the two-factor model, the correlation between two factors (i.e., higher-level processes and knowledge; lower-level processes and knowledge) was 0.889. The correlations among three factors (i.e., lower-level decoding; interim-level textbase construction; and, higher-level situation-model building) were 0.807, 0.891, and 0.892. This indicates that although component skills of reading are distinct enough to be considered as separate factors, they are highly intercorrelated with one another in nature. CSL readers’ performance on one component skill is highly predictive of their mastery of other skills.

**Difficulty Levels of the CSL Reading Proficiency Scale**

The *CSL Reading Proficiency Scale* consists of three subscales: *Lower-level, interim-level, and Higher-level Subscale*. Table 37, Table 38 and Table 39 show the difficulty estimates of descriptor items of these three subscales. For each item, the three tables also show the provisional groups and the difficulty levels in the original scale (i.e., the ACTFL Chinese Guidelines). In following sections, important findings related to three subscales are discussed one after another. Lastly, to descriptors collected from the *ACTFL Chinese Proficiency*
Guidelines, ACTFL-assigned levels are compared with levels verified by the Rasch analysis in this study.

**Lower-level Subscale**

The Difficulty estimates of items in the *CSL Reading Proficiency Scale: Lower-level Decoding* range across 3.91 logits (from -2.58 to 1.33 logits). This scale is divided into three band levels, among which Level 1 is the easiest one. The descriptor item with the lowest difficulty estimate is Item 11 which describes learners’ mastery of basic structural sets. Item 29, the most difficult item, describes CSL learners’ command of grammatical constructions frequently used in expository prose.

Descriptors on lower-level decoding (Table 37) are divided into three provisional groups: character recognition and orthographic knowledge, lexical access and vocabulary knowledge, syntactic parsing and syntactic knowledge. The provisional group of character recognition and orthographic knowledge has two items (Item 1 and Item 3). Character recognition and orthographic knowledge are the foundation of decoding. They have been regarded as very basic skills/knowledge and are supposed to develop since the beginning stage of CSL learning. It is surprising that both items in this group are estimated to be upper Level 2 items, the fifth and the eighth hardest items among all items in the *Lower-level subscale*. They are even estimated to be more difficult than some lexical access processes and syntactic parsing skills.

Item 1 depicts learners’ ability to use knowledge of radicals to help decode characters. Previous studies (Jackson, Everson, & Ke, 2003; Shen & Ke, 2007; Taft & Chang, 1999) reported that CSL learners’ radical knowledge and radical awareness could affect character learning and character recognition. Many CSL educators have already recognized the importance of radical knowledge and have integrated the training of radical knowledge into their classes.
since the very beginning level of CSL classes. However, in this study, the ability to use radical knowledge to help decode characters is still rated to be a highly difficult ability, which is equivalent to the average ability of upper-level 2nd-year CSL learners. In addition, Item 3 “can distinguish graphically-similar characters, which have learned” is also estimated to be the fifth most difficult item, which is very close to the average ability of 3rd-year CSL learners. This finding suggests that although radical knowledge and character decoding skill are very basic, they are much more difficult and take a longer time to fully develop than traditionally assumed. More intensive training on radical and character decoding should be widely adopted in beginning-level CSL classes. Universities that have not imbedded the training of radicals in Chinese classes may want to reconsider its importance.

The lexical access and vocabulary knowledge group has six items. Item 8 (can use knowledge on word formation to help decode unknown words) is rated to be a Level 3 item. This item is estimated to be the most difficult one in this group and the third hardest descriptor in the Lower-level Subscale. This indicates that the ability to apply word formation knowledge in decoding unknown words is not an easy skill and is not fully developed until learners have had three years’ learning experience. Item 9 is about learners’ ability to contextually process polysemes and is also rated to be a Level 3 item. The remaining four items on lexical ability are all rated to be Level 2 items.

The group with the largest number of items is syntactic parsing and syntactic knowledge, which has 20 items. There are two, thirteen, and five items on Level 3, Level 2, and Level 1 respectively. These twenty items highlight grammar patterns which college-level CSL learners need to master and their difficulty estimates cross 3.91 logits. While some patterns/skills can be mastered by CSL learners at the very early stage of learning (i.e., the 1st-year class), others
cannot be fully mastered until the very late stage of learning. Among these 20 syntactic patterns, the command of 13 patterns happen during the 2\textsuperscript{nd}-year of CSL learning.

One problem of the \textit{Lower-level Subscale} descriptors is that the range of descriptor difficulty is obviously not identical with the range of learner ability estimates (Table 21). There are thirty-eight learners above the highest item difficulty measure (i.e., Item 29) and seventeen below the lowest item difficulty measure (i.e., Item 11). Lower-level descriptors all address basic processing skills/knowledge, such as character recognition, lexical access, and syntactic parsing. It makes sense if difficulty estimates of these descriptors are lower than the ability estimates of more advanced learners in this study (i.e., upper 3\textsuperscript{rd}-year and 4\textsuperscript{th}-year CSL learners). However, it is very necessary to add descriptors with logit values below -2.58 logits to reflect the reading ability of learners with extremely low proficiency (e.g., low-achievers in the 1\textsuperscript{st}-year classes).

\textbf{Interim-level Subscale}

Difficulty estimates of items in the \textit{CSL Reading Proficiency Scale: Interim-level} \textit{Textbase Construction} cross over 8.03 logits, ranging from –3.71 to 4.32 logits. This scale is divided into four band levels, in which Level 4 is the most difficult one. Item 51, \textit{Can comprehend standard newspaper items addressed to the general public, routine correspondence, reports and technical material in field of interest}, is estimated to be the most difficult item. The descriptor item with the lowest difficulty estimate is Item 38 which describes learners’ ability to comprehend basic set expressions. As compared with the \textit{Lower-level Subscale}, the \textit{Interim-level Subscale} items show a wider range of difficulty. It suggests that while lower-level decoding skills/knowledge are more identical in the terms of difficulty and develop more simultaneously, skills/knowledge related to interim-level textbase construction show more discrepancies in difficulty and in development trajectory. It also indicates that interim-level processes and
knowledge is a complex construct involving a variety of subskills.

There are three obvious gaps in the Wright map (Table 28), which are larger or very close to one logit. A gap of more than one logit may indicate that some important concepts have been missed from the definition of the trait as expressed by descriptor items (Boone, Staver, & Yale, 2014). In the current study, three obvious gaps were larger or very close to one logit, indicating that there might be a problem of construct underrepresentation. To further improve the Interim-level Subscale, additional items should be created to make all component skills and knowledge involved in interim-level textbase construction well represented. Similar to the Lower-level Subscale, there is no descriptor which corresponds to learners with extremely low and high ability estimates. Items with difficulty estimates higher than 4.32 logits and easier ones with difficulty estimates lower than -3.71 logits should be added in order to better evaluate CSL leaners with extremely low or high ability.

Descriptors on interim-level textbase construction, as shown in Table 38, are divided into four provisional groups: semantic-proposition encoding; knowledge of coherence; knowledge of rhetorical organization; the text model of comprehension. The group of knowledge of rhetorical organization has four items in total. This group of items are all about students’ mastery of rhetorical devices and discourse patterns/text structures. These skills are involved in meaning construction beyond sentence and paragraph levels. The four items are all rated as Level 4 or upper Level 3 items, indicating that this group of skills/knowledge are of very high difficulty and are not fully developed until the upper level of CSL learning. This finding highlights that CSL learners have huge problems in discourse-level comprehension. However, while sentence-level and paragraph-level comprehension training are systematically integrated in CSL textbook and classroom teaching and are the focus of elementary and intermediate-level classes, instruction on
discourse patterns and rhetorical devices has rarely been included even in advanced-level CSL teaching and textbook. Research on discourse-level processing is much more scant than research on sentence and paragraph-level reading comprehension. The ability to process discourse-level text is crucial for CSL learners to achieve high CSL proficiency. Without explicit training and instruction, learners may have huge challenge in developing discourse competence by incidental learning. CSL researchers, educators and textbook writers do need to recognize the underrepresentation of discourse-level processing skills/knowledge in current CSL teaching and research.

Two items indicating the group of knowledge of coherence are Item 33 and Item 32. They both are evaluated to be Level 2 items, which can be regarded as relatively easy items in this subscale. These two items quantify learners’ awareness of cohesive devices within and across paragraphs. The mastery of cohesive devices plays an essential role in understanding and producing paragraph-level text. This finding suggests that students’ ability to use and comprehend cohesive devices starts to develop since the early stage of CSL learning, that is, the second year of taking CSL classes. The text model of comprehension group has the largest number of descriptor items. Among 16 items in this group, six are rated to be Level 4 items. There are five items estimated to be Level 3 and four rated to be Level 2 items. Level 1 gets one item from this group.

**Higher-level Subscale**

Difficulty estimates of items in the *CSL Reading Proficiency Scale: Higher-level Situation-model Building* range from –1.82 to 3.25 logits. The scale with four descriptors was divided into four band levels. Item 56, *Can appreciate, to a limited degree, nuances or stylistics, but gaps in detail may be frequent*, is estimated to be a Level 4 item, the most difficult one. Item
57 and Item 55 are both estimated to be Level 3 items. Item 54, *When reading text on his/her level, can only draw simplest inferences from reading*, is estimated to be a Level 1 item, the easiest one in this subscale.

There are only four descriptors in this subscale and there are two obvious gaps in the Wright map (Table 3.4). As discussed before, a gap of more than one logit may indicate that some important concepts have been missed from the definition of the trait as expressed by descriptor items (Boone, Staver, & Yale, 2014). Since this subscale only has four descriptor items, it is very likely that the construct of higher-level textbase construction is not well represented by currently used descriptors. More descriptor items should be added in order to fully represent skills and knowledge involved in higher-level textbase construction.

In addition, as shown in the Wright map (Table 3.4), there are a couple learners whose ability estimates are out of the range of item difficulty estimates. For learners with extremely high ability estimates, more items should be added in order to better evaluate their mastery of higher-level reading processes and knowledge. As compared with the lower-level counterparts, most higher-level reading processes/knowledge are of high difficulty. Learners with extremely low ability estimates could master very few number of higher-level reading processes/knowledge. Therefore, it is not surprising if no or very few descriptors correspond to ability of low-proficiency learners.

Four descriptors in this subscale were divided into two provisional groups: situation model of comprehension; cognitive strategies. While text model of comprehension only involves understanding of authors’ intended information, situation model of comprehension requires readers’ interpretation and judgments. Following factors influence the construction of a situation model: reader purpose and task expectation, genre activation, similar story instances, general
background knowledge resources, evaluation of the importance and value of information, attitudes toward writer, story, genre, episode, inferences needed for interpretation (Grabe, 2009, p44). However, it does not mean that situation model of comprehension is more difficult and developed later than text model of comprehension. On the contrary, when learners’ language proficiency is inadequate and have difficult in construct meaning from the text, L2 learners over rely on the situation model to figure out the meaning (Grabe, 2009). In Higher-level Subscale, the group of situation model of comprehension is represented by three items. They are estimated to be Level 4, Level 3, and Level 1 and correspond to the average ability of 4th-year, 3rd-year, and 1st-year CSL learners respectively.

There is only one descriptor item representing the group of cognitive strategy: can use preliminarily lexical guessing strategies in dealing with authentic materials, although errors are frequent. This cognitive strategy item is rated to be Level 3, which 3rd-year CSL learners should be able to have fully mastery of.

**Comparisons with Original ACTFL Levels**

The last column of Table 37, Table 38, and Table 39 shows levels of descriptors in their original scale (i.e., the *ACTFL Chinese Proficiency Guidelines*), which include four major proficiency levels: novice, intermediate, advanced, and superior. Novice and intermediate major levels can be further divided into three sublevels: low, mid, and high. Advanced level is further divided into two sublevels: advanced and advanced plus. In total, the *ACTFL Chinese Proficiency Guidelines* have nine sublevels: novice-low, novice-mid, novice-high, intermediate-low, intermediate-mid, intermediate-high, advanced, advanced-plus, and superior.

The following table shows the alignment between college Chinese class grade levels and ACTFL proficiency levels. It also shows the linking between Chinese class grade levels and the
levels of the *CSL Reading Proficiency Scale*. The alignment between class grade levels and *ACTFL Chinese Proficiency Guidelines* levels was officially approved by the Chinese Program of the University of North Carolina at Chapel Hill, whose Chinese learners could be regarded as typical examples of college-level CSL learners in the U.S.

Table 40

*Alignment between the ACTFL Chinese Proficiency Guidelines and the CSL Reading Proficiency Scale*

<table>
<thead>
<tr>
<th>CSL Reading Proficiency Scale Levels</th>
<th>College Grade Levels</th>
<th>Class</th>
<th>ACTFL Chinese Proficiency Guidelines Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;-Year</td>
<td></td>
<td>Novice, Intermediate-low/mid</td>
</tr>
<tr>
<td>Level 2</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;-Year</td>
<td></td>
<td>Intermediate-mid/high</td>
</tr>
<tr>
<td>Level 3</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-Year</td>
<td></td>
<td>Intermediate-high</td>
</tr>
<tr>
<td>Level 4</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;-Year</td>
<td></td>
<td>Advanced</td>
</tr>
</tbody>
</table>

As mentioned previously, the *ACTFL Chinese Proficiency Guidelines* were developed by an expert committee through intuitive identification of key features at each proficiency level. The *CSL Reading Proficiency Scale* was developed and validate empirically through a data-based approach. The researcher compared the ACTFL-assigned level and the *CSL Reading Proficiency Scale* level of each descriptor item. In Table 37, Table 38 and Table 39, items whose ACTFL proficiency levels are different from the *CSL Reading Proficiency Scale* levels were marked with asterisks. For instance, according to the *ACTFL Chinese Proficiency Guidelines*, Item 29 is a superior level item which is beyond the ability scope of 4<sup>th</sup>-year CSL learners. However, according to the *CSL Reading Proficiency Scale*, it is a Level 3 item, which is equivalent to intermediate-high ACTFL level. To item 29, the level assigned by the ACTFL committee is three sublevels higher the level verified empirically in this study. Therefore, three asterisks were marked beside this item. Similarly, one asterisk indicates that the level assigned by the ACTFL committee is one sublevel lower or higher than the empirically verified level in the *CSL Reading Proficiency Scale*. Two asterisks indicate a two-sublevel discrepancy.
In the Lower-level Subscale, three items (Item 29, 28, and 20) are scaled three sublevels lower than ACTFL-assigned levels. For instance, Item 28 was evaluated to be a superior-level item by the ACTFL Guideline committee, indicating that this item describes a reading ability which generally cannot be mastered by students who take college-level classes less than five years. However, the Rasch analysis in this study revealed that it was a Level 3 item which was equivalent to Intermediate-high ACTFL level. Students who have taken three years of Chinese class are very likely to be able to master this ability. Seven items were empirically scaled two sublevels lower than their ACTFL levels. Item 14 and Item 13 were scaled one sublevel lower than their assigned ACTFL level.

Overall, the difficulty levels of syntactic skill items have been over estimated by the ACTFL committee. Especially, syntactic patterns (indicated by descriptor Item 16-Item 23), including bei-passive, ba-disposal, relative clause, resultative compounds, aspect markers, directional compounds, indefinite and exclusive usages of question words, were all evaluated to be advanced or advanced-plus items by the ACTFL committee. However, they have been estimated to be much easier by the Rasch analysis in this study. For example, bei-passive (Item 18) and ba-disposal (Item 19) were both rated to be Advanced-plus level items by the ACTFL committee, indicating that CSL learners usually cannot master these two patterns until they have studied Chinese for four or five years. However, the Rasch analysis reveals that actually students’ command of these two patterns fully develops after two years of learning, much earlier than the ACTFL committee assume.

As shown in Table 38 and Table 39, among 22 items collected from the ACTFL Chinese Proficiency Guidelines, one was estimated to be three sublevels lower than the ACTFL-assigned level. Ten items in total were estimated to be two sublevels lower than the levels assigned by the
ACTFL committee. Two items were verified to be one sublevel lower than their ACTFL-assigned levels and one item was scaled one sublevel higher than its ACTFL level. Specifically, nine out of 16 items, which depict the text model of comprehension, were over estimated by the ACTFL committee in term of difficulty. All three items about the situation model of comprehension and pragmatic knowledge were evaluated to be more difficult by the ACTFL committee as compared with the Rasch scaling results. Similarly, the difficulty of one item about coherence knowledge and one item about rhetorical organization were also over estimated by the ACTFL committee.

In sum, this finding suggests that there is a large discrepancy between ACTFL-assigned levels and empirically verified levels. There is a tendency for ACTFL-assigned difficulty levels to be much higher than levels verified empirically. It is not a reliable way to develop proficiency scales merely through identifying intuitively key features at each proficiency level by an expert committee. The intuitive approach of scale development has been criticized because it lacks of empirical verification and validity evidence (Chalhoub-Deville & Fulcher, 2003; Liskin-Gasparro, 1996; Malone, 2003). Data-based scale development approach should play a more important role in scale development in the future.

**Implications**

The outcomes of this study have implications for a broad range of issues associated with language testing and teaching. First, the applications of the *CSL Reading Proficiency Scale* are elaborated. After thoughts regarding implications for test development are discussed, pedagogical implications are provided for CSL educators and textbook writers. Second, implications of findings for reading research are discussed. Last, implications for scale development are addressed.
Before discussing application of the *CSL Reading Proficiency Scale*, it is necessary to review briefly two taxonomies of language proficiency scales, which have been discussed in detail in Chapter 2, in order to guide the discussion of the implications of the *Scale*. First, according to the purposes of scale development, language scales can be classified into four types: user-oriented, constructor-oriented, assessor-oriented, and diagnosis-oriented scales (Alderson, 1991; Pollitt & Murray, 1993). First and foremost, the *CSL Reading Proficiency Scale* is constructor-oriented in nature. The *Scale* provides test developers with empirically based evidence for development of CSL reading assessments. Secondarily, the *CSL Reading Proficiency Scale* can also be used by assessors or teachers as a rubric for evaluation or diagnostic purposes. Lastly, this scale can be used by non-specialist users (e.g., CSL learners) as a self-assessment checklist of reading proficiency.

Second, according to content and structure, language proficiency scale can be classified into five categories: brief, holistic scales of reporting overall proficiency; user scales reporting proficiency in different contexts of use; detailed, holistic rating scales; detailed, analytic rating scales; and, frameworks of syllabus content and assessment criteria for stages of attainment (North, 1993). The *CSL Reading Proficiency Scale* is an analytic rating scale, which describes reading proficiency from multiple aspects, including character recognition, lexical access, syntactic parsing, and higher-level processing, etc. In developing the *Scale*, CSL teachers were involved in both the qualitative and quantitative validation procedures. It reflects teacher’s collective perception of CSL learners’ average reading proficiency and can be regarded as a road map of college-level CSL reading instruction. Therefore, to college-level CSL teaching, the *Scale* can function as a detailed content or outcome specifications for stages of attainment.

**Implications for Test Development**
This dissertation developed and validated the *CSL Reading Proficiency Scale*, an operational construct of CSL reading. The *CSL Reading Proficiency Scale* is constructor-oriented in nature and the primary application of this scale is to guide test development. It is applicable to the development of both in-house teacher-developed reading assessment and large-scale standardized reading tests. The findings of the present study have several important implications with respect to test development.

First, component skills approach of reading research has provided a very productive theoretical foundation for assessing L2 reading development (Koda, 2005). The unitary skill approach may lead to construct underrepresentation in test development (Alderson, 1990). Instead of regarding the CSL reading ability as an indivisible whole, this study conceptualizes reading ability as consisting of three hierarchical operation clusters: lower-level decoding, interim-level textbase construction, and higher-level situation-model building. Vital componential skills are well represented by 55 descriptors that are further classified into nine sub-groups. By referring to the *CSL Reading Proficiency Scale*, test developers will be able to know what specific subskills/knowledge should be tested at each proficiency level. It is hoped that the *CSL Reading Proficiency Scale* can help reduce construct underrepresentation in test development.

Second, reading comprehension, especially lower-level processing, is highly language specific. Chinese-specific linguistic features and corresponding cognitive processes should be well represented in CSL reading assessment. Linguistic features and cognitive processes unique to Chinese language, which are not included in most existing proficiency scales, are highlighted in the *CSL Reading Proficiency Scale*. For instance, descriptors about Chinese syntactic patterns (e.g., *ba* pattern, *shi*...*de* structure) are included in this scale. Chinese-specific reading skills and
knowledge, including word decision, radical knowledge, are also stressed in the CSL Reading Proficiency Scale. Therefore, when developing CSL reading assessment, this scale can be used as a checklist to enhance the possibility that skills, processes, and knowledge unique to CSL reading are well represented in the measurement.

Third, in developing the CSL Reading Proficiency Scale, items were scaled empirically according to their difficulty estimates. Scaled items were cut into three or four band levels, which can roughly be mapped to grade levels of college CSL classes. By referring to this scale, test developers are able to know what skills/knowledge CSL learners at a certain level should be able to master. When developing achievement tests for a given grade level, they can better target the appropriate proficiency level, avoiding including too many test items out of test takers’ scope of ability. However, reading proficiency tests usually are not targeted to a certain level or a certain learning context. The three subscales of CSL Reading Proficiency Scale all cover descriptors at various difficulty levels. Specifically, the higher Level Scale includes descriptors with difficulty estimates across 8.03 logits. The Lower-Level Subscale descriptor difficulty estimates range from -2.58 to 1.33 logits. The difficulty estimates of Interim-level Subscale items cross 5.07 logits. Therefore, the CSL Reading Proficiency Scale enables developers of proficiency tests to write test items across a very wide range of difficulty. In addition, the Scale includes descriptors on syntactic knowledge and word recognition skills ranked according to difficulty. Therefore, the Scale can also guide the development of discrete point tests on grammar and vocabulary in terms of item difficulty.

Last, the CSL Reading Proficiency Scale provides empirical evidence for the selection and editing of reading comprehension test prompts. In reading comprehension tests, test takers read prompt articles before answering reading comprehension questions. The selection of
prompts plays an essential role in assessing reading ability accurately. The *CSL Reading Proficiency Scale* specifies the genre, length, and linguistics features of text that learners at each proficiency level can comprehend. This scale provides test developers with an evidence-based guideline regarding how to select and revise authentic materials to compose prompts.

**Implications for Teaching, Classroom Assessment, and Textbook Writing**

Language proficiency scales become increasingly popular since they not only offer guiding principles for test development but also provide coherent connections between assessment, curriculum planning, and materials development (North, 1991). When developing the *CSL Reading Proficiency Scale*, experienced CSL teachers were involved in both expert review and descriptor scaling. The *CSL Reading Proficiency Scale* represents CSL teachers’ collective perception of CSL reading ability and its development. The *Scale* is highly relevant and useful to CSL teaching, curriculum development, classroom assessment, and textbook writing.

The findings of the present study have several important instructional implications. The *CSL Reading Proficiency Scale* reveals the relative difficulty of each reading skill. This provides CSL curriculum planners with solid empirical evidence on what specific skills/knowledge should be taught in each level of classes. To CSL teachers, the *CSL Reading Proficiency Scale* can be very informative for several reasons. First, when it comes to how to train reading ability, some teachers regard reading ability as an indivisible whole. Sometime, this is not a very practical way. Some teachers have realized that the training of reading ability could be divided into the teaching of multiple separated subskills; however, they need more guidance regarding what these subskills are and their interrelationship. The *Scale* models the CSL reading ability as a three-factor construct. Each factor is indicated by multiples descriptors which are further divided into
nine groups. With the CSL Reading Proficiency Scale as a reference, teachers are able to divide reading ability training into multiple portions, including the training of vocabulary/character recognition, the training of syntactic knowledge and syntactic processing, the training of cognitive strategies, etc. Second, the Scale has mapped descriptors with grade levels empirically. It advises teachers what reading skills and knowledge should be focused in certain classes. Teachers can set appropriate teaching objectives and select appropriate reading materials accordingly. Third, since the Scale includes a large number of descriptors on grammar knowledge and character/word recognition skills, it can benefit not only CSL reading class but also the teaching of grammar and orthography as separate skills.

In addition to formal tests used for summative purposes, teachers often conduct classroom assessment for formative purposes. The classroom assessment could be in the format of pop quizzes, group projects, presentations, etc., in which self and peer assessment is used in conjunction with teacher assessment. The CSL Reading Proficiency Scale provides teachers with guidance regarding item writing and grading criteria. This scale can also be used by non-specialist users (e.g., CSL learners) as a self-assessment or peer-assessment checklist.

Last, the CSL Reading Proficiency Scale can play a guiding role in textbook writing. It is useful to the development of reading comprehension textbooks as well as textbooks for integrated language skill training. The Scale provides guidance for textbook writers in terms of: (1) the genre, linguistic difficulty, and length of text that students at each proficiency level can read; (2) grammar patterns and syntactic knowledge that students at a certain level should be able to master; (3) skills involved in lower-level decoding, interim-level textbase construction, and higher-level situation-model building that need be trained as well as their development trajectory.

Implications for Reading Research

201
First, the research community has not yet reached an agreement regarding the divisibility of reading ability. This study examines three hypotheses arrived at through literature review: (1) reading is a unitary skill; (2) reading ability is tri-divisible; (3) reading ability is bi-divisible. Results of the confirmatory factor analysis provide empirical evidence to support the divisibility of reading ability. In this dissertation, data was collected from college-level CSL readers. This set of data conformed to both the correlated three-factor model (i.e., Lower-level decoding; Interim-level textbase construction; Higher-level situation-model building) and the correlated two-factor model (i.e., Lower-level processes and knowledge of CSL reading; Higher-level processes and knowledge of CSL reading). The three-factor model fit data slightly better. This finding affirms previous studies on L1 or L2 alphabetic languages (e.g., Alderson, 2000; Grabe, 1991) that reading ability is a multi-componential construct instead of an indivisible whole.

Second, through observing the mapping between descriptor difficulty estimates and learner grade levels (Table 37, Table 38, and Table 39), one can notice that the locations of most lower-level reading descriptors are relatively lower than the locations of interim-level and higher-level reading descriptors on the vertical scale. This suggests that: (1) a great number of lower-level decoding skills develop much earlier than interim-level and higher-level processing skills; (2) the majority of skills and knowledge involved in interim-level and high-level processing are more difficult as compared with lower-level decoding skills. This finding also provides empirical support for conceptualizing reading ability as consisting of hierarchical operation clusters.

Third, Rasch analysis calibrated reading proficiency descriptors on vertical scales (Table 37, Table 38, and Table 39), in which lower-located descriptors are relatively easy and descriptors on the top of the vertical scale are more difficult for CSL readers. The CSL Reading
Proficiency Scale shows the sequence of reading skills development, indicating which skills are mastered first and which are mastered on a late stage by CSL readers. This result sheds light on the trajectory of L2 reading ability development. Researchers on grammar acquisition and vocabulary acquisition may also find this Scale informative, since it uncovers development trajectory of subskills, like syntactic processing skills and character/word recognition skills.

Last but not the least, an important theoretical implication is that, for reading Chinese as a logographic language, prior reading theory, which is developed mainly on the basis of studies on reading in alphabetic languages, could be expanded to include orthographic-level processing, including radical recognition, character recognition, and word decision. This argument echoes pervious empirical studies on CSL reading (e.g., Everson & Ke, 1997).

Implications for Scale Development

In the field of foreign language education and language testing, the mainstream scale development approach has been the intuitive approach. As pointed by North and Schneider (1998), most language proficiency scales have been constructed by appeal to intuition. For instance, the ACTFL Proficiency Guidelines and the ACTFL Chinese Proficiency Guidelines were both developed by an expert committee through intuitive identification of key features at each proficiency level. The difficulty levels of descriptors have not been verified empirically. Due to its lack of empirical verification and validity evidence, the intuitive approach of scale development has incurred numerous criticisms (e.g., Chalhoub-Deville & Fulcher, 2003; Liskin-Gasparro, 1996; Malone, 2003).

In this study, the researcher compared the ACTFL-assigned levels with the levels scaled by Rasch analysis for the same group of descriptors. It is found that ACTFL-assigned difficulty levels tend to be higher than levels verified empirically. There is a large inconsistency between
ACTFL-assigned levels and empirically verified levels. This finding confirmed previous criticisms towards the intuitive approach of scale development.

Criticism on the intuitive approach has led to data-based solutions to proficiency scale development. In data-based approaches, empirical evidence drawn from performance samples, test items that operationalize descriptors, or large-scale application of draft scales is used to verify the identification and scaling of proficiency descriptors. Among these data-based solutions, the methodology used to develop the CEFR scale (North, 1995, 2005; North & Schneider, 1998; North, 2000) becomes influential and has been widely consulted by scale development projects worldwide. This study also adopts this approach while addressing its limitations.

To the development of national or worldwide language proficiency scales with wide influence, merely relying on traditional intuitive approach may cause negative consequences. Data-based scale development approach which provides empirical evidence for descriptor identification and scaling should play a more important role in new scale development projects and in validating existing language proficiency scales.

**Unique Contributions**

This dissertation makes unique contributions in: identifying theoretical construct definition of CSL reading from existing literature; operationalizing theoretical construct of CSL reading by developing the *CSL Reading Proficiency Scale* which addresses limitations of existing scales; and, contributing to the interdisciplinary exchange between second language acquisition research and the language testing filed.

First, this dissertation identifies theoretical construct of CSL reading. Prevailing L1 and L2 reading models were investigated and well represented when the researcher identified the construct definitions of the CSL reading ability. Especially, the researcher explored and
synthesized linguistic properties and corresponding cognitive mechanisms unique to Chinese language which scatter in literature. This thorough literature review can advance audience’s understanding of construct definitions of L1, L2, and especially CSL reading.

Second, this dissertation focuses on developing and validating the CSL Reading Proficiency Scale, which operationalizes theoretical construct of CSL reading for the purposes of assessment. This scale facilitates the application of reading theories in test development. It provides guidelines for test developers in terms of what salient knowledge and skills should be targeted in the context of college-level CSL learning at certain proficiency levels. Particularly, as discussed previously in Chapter One, foreign language teachers often lack sufficient training on assessment development. The CSL Reading Proficiency Scale, which is primarily designed to guide the development of in-house assessment, meets the urgent need of foreign language teachers.

Third, the CSL Reading Proficiency Scale addresses limitations of existing proficiency scales: (1) Some existing proficiency scales have been criticized for lacking a linguistic theory base (Chalhoub-Deville, 2009; Fulcher, 2003). In the current study, theoretical works on L1 reading, L2 reading, and CSL reading, especially the Communicative Language Ability (CLA) Model (Bachman, 1990; Bachman & Palmer, 1996, 2010), serve as theoretical underpinnings of the CSL Reading Proficiency Scale. (2) To well represent linguistic properties and cognitive mechanisms unique to Chinese language, descriptors about Chinese orthography, morphology, syntax, and discourse patterns were created and included in the CSL Reading Proficiency Scale. The Scale can provide more precise and more specific description of the CSL reading ability as compared with existing scales which are not language-specific. (3) The CSL Reading Proficiency Scale has gone through rigorous empirical verification. Both qualitative and quantitative
validation procedures were conducted to screen descriptors which did not fit the construct being measured. Difficulty levels of descriptors were estimated through Rasch rating scale analysis. These verification procedures well support validity claims of the *CSL Reading Proficiency Scale*. (4) Descriptors depicting reading from a cognitive processing perspective, which are very scant in other existing scales, count for a relatively large proportion of the *CSL Reading Proficiency Scale* descriptors. Processing-based descriptors, together with text-based and task-based descriptors in the scale, provide a panorama view of CSL learners’ reading processing and development. (5) The majority of existing scales are holistic scales, which only provide general statements for each proficiency level. The *CSL Reading Proficiency Scale* is analytic in nature. In the current study, the underlying factor structure of *CSL Reading Proficiency Scale* was uncovered empirically. According, this scale is presented as an analytic scale with three distinct subscales which depicting lower-level, interim-level, and high-level reading processes respectively. In addition, descriptors in each subscale were further placed into nine sub-groups, including character recognition, lexical access, syntactic parsing, and cognitive strategies, etc.

Fourth, instead of using secondary data, this study analyzed primary data collected with 95 experienced college-level CSL teachers across the United States. Teachers’ collective perception of CSL learners’ reading proficiency and their collective perception of descriptors’ difficulty were comprehensively documented. This set of primary data can also be informative and useful to future studies related to CSL reading.

Lastly, this dissertation research is an interdisciplinary study grounded on the interface of language acquisition and language testing research. It is hoped that this dissertation can make contributions to the interdisciplinary exchange in terms of the construct definition of L2 reading ability and data-based scale development methodology. On one hand, scholars have been arguing
that some prevailing language proficiency scales, such as the CEFR scale, are not grounded on
large scale SLA studies and lack of a clear linguistic theory (Chalhoub-Deville, 2009; Fulcher,
2004). In the current study, reading theories drawn from L1 and L2 acquisition research underlie
the modeling of CSL reading construct. With dual expertise in second language acquisition and
language testing studies, the researcher applied reading theories existing in the field of language
acquisition to the practice of scale development. It is expected that this dissertation can make
contribution in bridging second language acquisition theoretical foundation with scale
development practice. On the other hand, this study adopted a data-based scale development
methodology which origins in language assessment literature and is new to foreign language
education scholars, especially CSL educators. Even the ACTFL Guidelines, the official
guidelines of foreign language education in the U.S., were developed through an intuitive
approach. The current study introduces the application of data-based scale development
methodology and Rasch analysis to the field of foreign language education. It makes contribution
to the advancement of foreign language scale development methodology.

**Limitations and Recommendations**

The present study has several limitations to be acknowledged. First, due to the constraint
of budget and resource, only a highly selected group of descriptors were used in scale
development. In this study, two sources of descriptors were: existing reading proficiency scales
and theoretical works on reading. In total, 149 descriptors were collected from following existing
reading proficiency scales: Canadian Language Benchmarks; The World-class Instructional
Design and Assessment (WIDA) English Language Development Standards; The American
Council for the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines; Interagency
Language Roundtable (ILR) Language Skill Level Descriptions, Mandarin Chinese: Four-Year
Instructional Goals, Curriculum Outline, and Instructional Measures; and, The ACTFL Chinese Proficiency Guidelines. However, it is beyond the capacity of a doctor dissertation to develop a language proficiency scale involving hundreds of descriptors. Only descriptors directly related to CSL reading were included in the descriptor bank. When having enough budgets for a larger-scale study, the researcher will investigate these remaining descriptors which are not included in this dissertation.

In scaling the descriptor items of the Interim-level Subscale and the Higher-level Subscale, obvious gaps, larger than or very close to one logit, were identified on the Wright maps (Table 28 and Table 34). Such large gaps among items suggest that there may be construct underrepresentation problem for interim-level reading construct and higher-level reading construct, which is very unlikely to exist for the lower-level reading construct in this study. The interim-level and higher-level processing are very broad constructs, involving cognitive, metacognitive processes, linguistic and sociolinguistic knowledge, etc. However, only 27 items in total were selected to represent these two constructs. As shown in Table 2, descriptor items placed into each factor were further classified into nine provisional groups arrived at through literature review in order to further delineate the factor structure of the CSL reading ability. Some provisional groups include around 20 items, however, others, such as cognitive strategy and semantic-proposition encoding, are indicated by only one item. It is likely that provisional groups with only one or few descriptor items are underrepresented in the CSL Reading Proficiency Scale. Although it is not necessary to have equal numbers of descriptors for all provisional groups, more descriptor items should be added for groups with less than five items in order to make them well represented.

Strategic competence is one major component of communicative language ability. As
discussed in Chapter 2, reading strategies can be classified into three categories: metacognitive, cognitive, and social/affective strategies. However, for the sake of space, in this study, only one descriptor about cognitive strategies was included in the descriptor pool. In subsequent projects, descriptors of metacognitive and social/affective strategies need to be investigated.

In addition, the *Higher-level Subscale, Interim-level Subscale, and the Lower-level Subscale* all lack descriptors which can match the ability of learners on two extremes. To fully present the reading ability profiles of college-level CSL learners across all levels, descriptors with extremely high or low difficulty estimates need to be included.

The second problem lies in variances in rater severity which may exist in the quantitative verification procedure. In total, 95 participants across the U.S. were recruited as raters to evaluate their students’ reading proficiency against the *CSL Reading Proficiency Questionnaire*. Teacher participants might vary in severity when evaluating students’ proficiency. One way to address differences in rater severity is to organize a rater norming workshop, in which raters get familiar with the questionnaire and receive instruction on rating criteria. The second way is to estimate rater severity by having all teacher participants rate a limited number of common samples (i.e., reading ability of a few CSL learners in this context) and then adjustment for rater severity can be made in following analysis. The second way is not very feasible since reading ability is a receptive skill. Capturing an observable sample of one’s reading ability is not as easy as getting a record of one’s speaking and writing abilities. Organizing a rater norming conference was the better solution. Due to constrain of budget, participant availability, and distance, it was hardly feasible to arrange an onsite training conference prior to the data collection. Instead, the researcher provided written instructions which were as detailed as possible at the very beginning of the online questionnaire to help teacher participants get familiar with the rating categories and
rating criteria. However, it was likely that variance in rater severity still existed and might impact findings. Future studies which aim to develop rating scales for receptive skills (i.e., listening and reading skills) need to figure out better solutions to address the impact of rater severity variance.

Third, one important assumption of Rasch analysis is unidimensionality, assuming that all items in an instrument measure the same latent trait. The instrument used in this study, the CSL Reading Proficiency Questionnaire, measures three distinct dimensions of reading ability: decoding, interim-level textbase construct, and higher-level situation-model building. Therefore, this instrument was divided into three subscales and calibrated separately by using Rasch analysis. Multidimensional IRT models can analyze multidimensional construct directly. For the current study, multidimensional IRT models can calibrate all descriptor items directly although they fall into multiple distinct dimensions. However, multidimensional IRT models require larger sample sizes (Jiang, Wang, & Weiss, 2016). Due to the limitation of the sample size, using multidimensional IRT models was not an option for this research. Future studies with larger sample sizes could consider using multidimensional IRT models to scale reading proficiency descriptors with more than one dimension.
APPENDIX 1: QUALITATIVE VALIDATION OF THE CSL READING PROFICIENCY DESCRIPTOR POOL

Dear Participants:

Attached is a bank of descriptors of Chinese as a second language (CSL) reading proficiency. You are invited to review this descriptor bank in terms of: clarity, brevity, accuracy, independence, and applicability. Please go through each descriptor and think about the following questions:

1) **Applicability**: Is this descriptor applicable to college-level CSL reading assessment? In other words, does the skill/knowledge represented by this descriptor need to be included in college-level CSL reading teaching and testing?

2) **Clarity**: Is this descriptor clear in meaning? If not, how do you want to revise it?

3) **Brevity**: Is this descriptor brief enough? If not, how do you want to condense it?

4) **Independence**: Is this descriptor overlapping with other descriptors? Please write down the item number.

5) **Accuracy**: Is this descriptor accurate in meaning? If not, how would you like to revise it?

After each descriptor, there are spaces where you can leave comments regarding above five criteria. For instance, if you think Item 1 is not applicable, you can comment after the "applicability" icon. If you think Item 1 is not clear in meaning, please comment after the "clarity" icon. If you think an item is fine, you can skip and move to the next one. You can response in English or Chinese.

**Group 1: Character recognition & Orthographic Knowledge**:

Item 1: Can decode numbers 1-1000 written in Chinese characters.

  - Applicability:
  - Clarity:
  - Brevity:
  - Independence:
  - Accuracy:

Item 2: Can use knowledge of radicals to help decode unknown characters preliminarily.

  - Applicability:
  - Clarity:
  - Brevity:
  - Independence:
  - Accuracy:

Item 3: Can use knowledge of radicals to help decode unknown characters efficiently.

  - Applicability:
  - Clarity:
Item 4: Can recognize different forms of a particular radical in characters.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 5: Can identify structures of characters with ease and high degree of accuracy. (For instance, 你 is a “left to right” character. 因 is a “full surround” character.)
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 6: Can use knowledge of character structure to help decode unknown characters preliminarily.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 7: Can use knowledge of character structure to help decode unknown characters efficiently.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 8: Can distinguish graphically-similar characters with ease and high degree of accuracy (over 90%).
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 9: Can distinguish graphically-similar characters with moderate degree of accuracy (around 60-89%).
  Applicability:
  Clarity:
  Brevity:
Independence:
Accuracy:

Item 10: Can distinguish graphically-similar characters, but the degree of accuracy is low (lower than 60%).

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

**Group 2: Lexical access & vocab Knowledge**

Item 11: Can derive meaning of sentences with coverbial phrases (Coverbs refer to a group of words which can be used as both verbs and prepositions. For instance, in “用中文写信”, “用中文” is a coverbial phrase. In “我为你做事”, "为你" is a coverbial phrase.), but his/her comprehension is based more on lexical identification than grammatical structure knowledge.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 12: Can derive meaning of sentences with verb complements (动词补语结构), but his/her comprehension is based more on lexical identification than structural knowledge.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 13: When comprehending more complex structure, can be less dependent on native language word-order strategies.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 14: Can comprehend more complex structure, but rely on lexical identification rather than on grammatical cues.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:
Item 15: When reading text on his/her level, can group adjacent characters into words with ease and high degree of accuracy (over 90%).
   
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 16: When reading text on his/her level, can group adjacent characters into words with moderate degree of accuracy (around 60-89%).
   
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 17: When reading text on his/her level, can group adjacent characters into words, but the degree of accuracy is low (lower than 60%).
   
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 18: Can have sufficient knowledge on word formation.
   
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 19: Can have some knowledge on word formation.
   
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 20: Can have limited knowledge on word formation.
   
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:
Item 21: Can use knowledge on word formation to help decode unknown words preliminarily.

Item 22: Can use knowledge on word formation to help decode unknown words efficiently.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 23: When reading text on his/her level, can contextually process polysemes (words with multiple meanings) with ease and high degree of accuracy (over 90%).

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 24: When reading text on his/her level, can contextually process polysemes (words with multiple meanings) with moderate degree of accuracy (60-89%).

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 25: When reading text on his/her level, can contextually process polysemes (words with multiple meanings) but the degree of accuracy is low (lower than 60%).

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

**Group 3: Syntactic parsing & Syntactic Knowledge**

Item 26: Can understand limited types of sentence structures, including Subject-Predicate constructions, simple SVO constructions, all in affirmative, negative, and simple question form.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:
Item 27: Can master certain small closed structural sets (closed structural sets are defined as structural sets which include only a limited number of items), including interrogatives (shuí, shén me), specifiers (zhè, nà), and very common noun measures (kuài, běn) etc.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 28: Can understand simple Noun+Noun and Stative Verb + Noun modification with de (得).

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 29: Can have preliminary ability to comprehend high-frequency connectors forming complex sentences (e.g., de shí hou, suī rán... kě shì)

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 30: Can be aware of aspect markers (such as le), but may ignore them in deriving meaning.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 31: Can understand roughly the relationship of multiple modifiers to element being modified, but may have some confusion.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 32: Can be alert to a basic repertory of high-frequency patterns which often do not parallel the reader’s native language constructions.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 33: Can comprehend, with sporadic errors, relative clause modification (关系从句作为修饰成份, e.g., “你认识的那个女孩子”).
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 34: Can have better understanding of the aspect markers (such as le), although still makes errors.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 35-40: Can master the following basic patterns: Relative clause modification (关系从句作为修饰成份, E.g., “你认识的那个女孩子”); ba-disposal (把字句); bei-passive (被字句); Indefinite and exclusive usages of question words (疑问词的定指和不定指); Various aspect markers (着,了,过); Resultative/directional compounds （结果/趋向补语）.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 41-44: Can master the following unique Chinese syntactic features: The position of adverbial phrases; Nominal sentences (i.e., sentences contain no verb.); “Shi…de” structure; Topic-comments structure.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 45: When reading patterns and usages drawn from the more strictly literary/classical stratum, can understand lexicon but not all grammar structures.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:
Item 46: Can understand text successfully when the number of features and structural devices from the literary/classical stratum is limited.
  
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 47: Can have preliminary mastery of high-frequency grammatical constructions common to expository prose.
  
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

**Group 4: Semantic-proposition Encoding**

Item 48: When reading text on his/her level, can identify referential expressions/antecedents (指称语/先行词) with ease and high degree of accuracy (over 90%).
  
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 49: When reading text on his/her level, can identify referential expressions/antecedents (指称语/先行词) with moderate degree of accuracy (60-89%).
  
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 50: When reading text on his/her level, can identify referential expressions/antecedents (指称语/先行词) but the degree of accuracy is low (lower than 60%).
  
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

**Group 5: Knowledge of Cohesion**

Item 51: Can be aware of commonly used cohesive devices (衔接手段，例如连词，有关联作
Item 52: Can have functional mastery of a limited number of Chinese cohesive devices (衔接手段，例如连词，有关联作用的副词和短语等).

Item 53: Can have functional mastery of nearly all commonly-used Chinese cohesive devices (衔接手段，例如连词，有关联作用的副词和短语等).

Item 54: Can be highly aware of differences between Chinese and English cohesive devices (衔接手段: 例如连词，有关联作用的副词和短语等) in terms of function and meaning.

Item 55: Can be moderately aware of differences between Chinese and English cohesive devices (衔接手段: 例如连词，有关联作用的副词和短语等) in terms of function and meaning.

Item 56: Can use knowledge on cohesive devices (衔接手段: 例如连词，有关联作用的副词和短语等) to facilitate reading comprehension preliminarily.
Item 57: Can use knowledge on cohesive devices (衔接手段: 例如连词，有关联作用的副词和短语等) to facilitate reading comprehension effectively.
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Group 6: Knowledge of Rhetorical Organization

Item 58: Can have rich knowledge on Chinese text structure/discourse pattern.
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 59: Can have some knowledge on Chinese text structure/discourse pattern.
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 60: Can have little knowledge on Chinese text structure/discourse pattern.
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 61: Can be highly aware of differences in text structure/discourse pattern between Chinese and English.
   Applicability:
   Clarity:
   Brevity:
   Independence:
   Accuracy:

Item 62: Can be moderately aware of differences in text structure/discourse pattern between Chinese and English.
   Applicability:
   Clarity:
Item 63: Can use text structure knowledge to facilitate reading comprehension preliminarily.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 64: Can use text structure knowledge to facilitate reading comprehension efficiently.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 65: Can have some appreciation of formulaic rhetorical devices (修辞手法) common to more stylized writing.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

**Group 7: The Text Model of Comprehension**

Item 66: Can comprehend set expressions using basic vocabulary.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 67: Can read simple connected, specially prepared material for basic survival and social needs. (Note: "specially prepared material" hereafter refers to "material specially prepared for language learners".)

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 68: Can comprehend sufficiently specially-prepared discourse for informative purposes.

Applicability:
Item 69: With use of a dictionary, can comprehend sufficiently main ideas and some facts in authentic material paralleling oral language.

Item 70: Can decode authentic materials, which are harder than oral language, for key points with use of a dictionary.

Item 71: Can identify products and prices in popular illustrated advertising such as food, clothing, work supplies, and travel with extensive reliance on a dictionary.

Item 72: Can decode critical elements of public announcements to determine who, when, and where regarding such subjects as public events by using a dictionary.

Item 73: Can follow the narrative thread in more extended, specially-prepared discourse.

Item 74: Through decoding and heavy reliance on dictionary, can sporadically identify bits and pieces of factual information in simple, authentic articles written in straightforward expository
style on topics dealing with current events and daily activities familiar to the reader.

Item 75: Can decode authentic prose (including newspapers and magazines) for general ideas.

Item 76: Can glean a range of specific facts from short authentic pieces on sports, movies, and current events, but with error.

Item 77: Can be increasingly skillful in extracting concrete, minor details.

Item 78: Can read some types of authentic expository prose without reliance on a dictionary where such prose tends to reflect the spoken language.

Item 79: Can read, not just decode, a narrow range of authentic, expository material, including areas of professional interest, without the use of a dictionary.

Item 80: Can read, with some dictionary use, authentic material over a wide range of subject
matter and topics.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 81: Can read standard newspaper items addressed to the general public, routine correspondence, reports and technical material in field of interest.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 82: Can decode, with a dictionary and with substantial effort and moderate error, popular novels, essays, and most literature for the general public.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 83: Can understand minimally, yet with significant difficulties, prose more characteristic of the literary/classical style.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

**Group 8: The Situation Model of Comprehension and Pragmatic Knowledge**

Item 84: Can only draw simplest inferences.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:

Item 85: Can minimally detect subjective attitudes, values, and judgments.
  Applicability:
  Clarity:
  Brevity:
  Independence:
  Accuracy:
Item 86: Can appreciate, to a limited degree, nuances or stylistics, but the specifics may often be lost.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

**Group 9: Cognitive Strategy**

Item 87: Can puzzle out pieces of some authentic material with considerable difficulty, as it reflects similarity to specially prepared materials.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 88: Can puzzle out pieces of some authentic material with considerable difficulty, when it uses high-frequency informal/colloquial vocabulary and structure.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Item 89: Can have preliminary ability to use the lexical guessing strategy in dealing with authentic materials, although errors are frequent.

Applicability:
Clarity:
Brevity:
Independence:
Accuracy:

Lastly, do you want to add other descriptors to this pool? In other words, is there any skill/knowledge you think that should be tested in to be included in college-level CSL reading teaching and testing?
APPENDIX 2: CSL READING PROFICIENCY QUESTIONNAIRE-100 LEVEL

The researcher is developing a Chinese as a second language (CSL) reading proficiency scale. You are invited to try out the scale for the purpose of validation. Please pick one medium-achieving non-heritage student from 1st-year Chinese classes and rate what his/her reading proficiency is after one-year college-level Chinese learning against this scale. Your evaluation can be made by recalling this specific student's performance in class activities, reading comprehension tests, or assignments. The purpose of this study is to validate the difficulty of items in the scale. It is not intended to evaluate a certain student’s proficiency. Therefore, you do not need to provide the student’s name and no identifiable information will be collected. I will send you $25 gift card to express my gratitude. To protect your identity as a research subject, the researcher will not share your information with anyone. In any publication about this research, your name or other private information will not be used. If you have any questions, please contact the investigator by calling 3194710712 or emailing jialin@email.unc.edu. If you have questions or concerns about your rights as a research subject, you may contact the UNC Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu. (IRB Study #: 18-1020). If you have no questions, you can go ahead by clicking the "I consent"!

1. How many minutes does his/her Chinese class meet per week:

________________________________________________________________

2. Student’s gender:
   Male
   Female

3. Please leave your email in order to get the gift card.

________________________________________________________________

226
4. Your affiliation:

Please rate this student for each item on this questionnaire using the following scale. 请使用本问卷，根据下列等级评价学生的阅读能力。

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>This describes a level, which is definitely beyond his/her capabilities. Can not be expected to perform like this. 0: 完全做不到。他/她的水平明显低于此条描述。</td>
</tr>
<tr>
<td>1</td>
<td>Could minimally perform like this in favorable circumstances. This describes a performance which is slightly beyond his/her level. (Favorable circumstance refers to the situation in which she/he gets help or hints, has time to prepare or think over; the topics are familiar; surroundings favor his/her performance.) 1: 在有利的情况下，勉强做到。他/她的水平稍低于此条描述。(有利的情况指的是能够得到帮助或暗示,有时间思考准备,熟悉话题,环境有利于发挥。)</td>
</tr>
<tr>
<td>2</td>
<td>Could be expected to perform like this without support in normal circumstances. His/her ability is on this level. 2: 在一般情况下，能够独立做到。他/她的水平与此条描述基本符合。</td>
</tr>
<tr>
<td>3</td>
<td>Could be expected to perform like this even in unfavorable circumstances with efforts. This describes a performance which is slightly below his/her level. (Unfavorable circumstance refers to the situation in which she/he does not get help or hints, has no time to prepare or think over; the topics are unfamiliar; surroundings do not favor his/her performance.) 3: 即使在困难的情况下，也能努力做到。他/她的水平稍高于此条描述。（困难的情况指的是得不到帮助或暗示，没有时间思考准备，不熟悉话题，环境不利于发挥。）</td>
</tr>
<tr>
<td>4</td>
<td>Could be expected to perform like this with ease in all circumstances. This describes a performance which is obviously below his/her level. 4: 在任何情况下，都能轻松完成。他/她的水平明显高于此条描述。</td>
</tr>
</tbody>
</table>

**Group 1: Character recognition & Orthographic Knowledge**

Item 1: Can use knowledge of radicals to help decode characters. 能用部首知识帮助识别汉字。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 2: Can recognize different forms of a particular radical, which have learned, in characters (e.g., 心). 对于已经学过的部首，能辨别同一个部首在汉字中呈现的不同形式（例如，
“忄”和“心”。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 3: Can distinguish graphically-similar characters, which have learned. 对于已经学过的形近字，能够区分。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Group 2: Lexical access & vocab knowledge

Item 4: Can derive meaning of sentences with coverbial phrases. 能理解含有副动词词组的句子的意思。（副动词指既可以用作动词也可以用作介词的一类词。例如，“用中文写信”中“用中文”就是副动词词组。）

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 5: Can derive meaning of sentences with verb complements. 能理解含有动补短语的句子的意思。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 6: When reading text on his/her level, can group adjacent characters into words successfully. 读适合他/她阅读水平的文本时，能成功地切词。（切词指的是将一个汉字序列切分成一个一个单独的词）

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到
Item 7: Can analyze the intra-word structure. 能够识别词的内部结构。（例如：能识别“看书”是一个动宾结构的词）

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 8: Can use knowledge on word formation to help decode unknown words. 能用构词法知识帮助识别不认识的词。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 9: When reading text on his/her level, can contextually process polysemes (words with multiple meanings). 读适合他/她阅读水平的文本时，能根据语境理解多义词在特定语境中的准确意思。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Group 3: Syntactic parsing and syntactic knowledge
Item 10: Can understand limited types of sentence structures, including Subject-Predicate constructions, simple SVO constructions, all in affirmative, negative, and simple question form.

能理解有限的几种句子结构类型：包括肯定/否定/问句形式的主谓结构、简单的主谓宾结构。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 11: Can control certain structural sets, which have a limited number of items, including interrogatives (shuí, shén me), specifiers (zhè, nà), and very common noun measures (kuài, běn) etc.

能掌握那些本身只含有少数子项目的结构，包括简单的疑问词（谁, 什么）、指示代词（这, 那）、常见的名量词（块, 本）等。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 12: Can understand simple Noun+Noun modification with de and Stative Verb + Noun modification with de.

能理解简单的“名词+的+名词”和“静态动词+的+名词”的修饰结构。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 13: Can comprehend preliminarily high-frequency connectors forming complex sentences (e.g., de shí hou, suī rán... kě shì).

能初步理解常见的关联词（例如：……的时候, 虽然……可是）。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 14: Can derive meaning of sentences with the aspect markers (such as le), although still makes errors.

能理解含有体标记的句子（例如“了”）的意思，虽然还会出现错误。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 15: Can understand roughly the relationship of multiple modifiers to element being modified, but may have some confusion. 能大致理解多重修饰语和被修饰成分之间的关系，但可能会出现混淆。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 16: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
Relative clause modification. 关系从句做修饰成分（e.g., 你认识的那个女孩子。）
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 17: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
ba-disposal. 把字句
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 18: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
bei-passive. 被字句
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 19: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
Indefinite usages of question words. 疑问词的不定指（例如在“谁都不喜欢这本书”中，“谁”
用作不定代词，意思是“任何人”。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 20: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
Exclusive usages of question words. 疑问词的特指用法（例如在“这是谁的书?”中，“谁”用作特指疑问词。）
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 21: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
Various aspect markers. 各种体标记（e.g., 着，了，过）
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 22: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
Resultative compounds. 结果补语
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 23: Can comprehend sentences with following basic patterns. 能读懂含有下列语法的句子:
Directional compounds. 方向补语
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到
Item 24: Can be aware of the appropriate position of adverbial phrases in Chinese sentences. 能意识到在中文句子中，副词短语应该放置的正确位置。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 25: Can comprehend sentences with following unique Chinese syntactic features. 能读懂含有下列语法的句子:
Nominal sentences(i.e., sentences contain no verbs) 无动词谓语句
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 26: Can comprehend sentences with following unique Chinese syntactic features. 能读懂含有下列语法的句子:
“Shi…de” structure. “是……的”结构
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 27: Can comprehend sentences with following unique Chinese syntactic features. 能读懂含有下列语法的句子:
Topic-comments structure 主题-评论句
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 28: Can comprehend text successfully when the number of features and structural devices from the literary/classical stratum is limited. 当来自文学作品，文言文的特色和语法结构很少时，能够成功地理解文本。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 29: Can have emerging control of high-frequency grammatical constructions common to expository prose. 开始表现出能掌握说明文中常见的高频语法结构的能力。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Group 4: Semantic-proposition encoding

Item 30: When reading text on his/her level, can successfully identify referential expressions/antecedents. 当阅读适合他/她水平的文本时，能够成功地找出指代表达/先行词。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Group 5: Knowledge of cohesion

Item 31: Can know meanings and functions of a limited number of cohesive devices. 只能知道有限的几个关联词的意思和用法。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 32: Can be aware of commonly used cohesive devices within and across paragraphs. 能识别段落内部以及段落间常用的关联词。（关联词包括：连词和有关联作用的介词、副词和短语等）
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 33: Can use knowledge on cohesive devices to facilitate reading comprehension. 能用关联词知识来帮助理解文章。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

**Group 6: Knowledge of rhetorical organization**

**Item 34:** Can be aware of very common Chinese text structures/discourse patterns. 能识别很常见的中文文章结构和语篇模式。

  0: No, cannot perform like this. 完全做不到
  1: Yes, in favorable circumstances. 在有利条件下可以做到
  2: Yes, in normal circumstances. 在一般情况下可以做到
  3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
  4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

**Item 35:** Can be aware of most Chinese text structures/discourse patterns. 能识别大部分种类的中文文章结构和语篇模式。

  0: No, cannot perform like this. 完全做不到
  1: Yes, in favorable circumstances. 在有利条件下可以做到
  2: Yes, in normal circumstances. 在一般情况下可以做到
  3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
  4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

**Item 36:** Can use text structure knowledge to facilitate reading comprehension. 能用篇章结构的知识来帮助阅读理解。

  0: No, cannot perform like this. 完全做不到
  1: Yes, in favorable circumstances. 在有利条件下可以做到
  2: Yes, in normal circumstances. 在一般情况下可以做到
  3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
  4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

**Item 37:** Can have some appreciation of formulaic rhetorical devices common to more stylized writing. 能够欣赏文学性写作中常见的程式化的修辞手法。

  0: No, cannot perform like this. 完全做不到
  1: Yes, in favorable circumstances. 在有利条件下可以做到
  2: Yes, in normal circumstances. 在一般情况下可以做到
  3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
  4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

**Group 7: The text model of comprehension**

**Item 38:** Can comprehend set expressions using basic vocabulary. 能理解由基础词汇构成的固定表达。

235
定表达。例如“今天天气怎么样？”“请问现在几点？”等等。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 39: Can read simple connected, specially prepared material for basic survival and social needs (Note: “specially prepared material” hereafter refers to “material specially prepared for adult second language learners of Chinese”). 能阅读简单、连贯的学习者语料，以满足基本的生存和社交需要。（本量表中“学习者语料”是指为成年中文二语学习者而特殊编辑的阅读材料，不是真实语料。）
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 40: Can comprehend sufficiently specially-prepared discourse for informative purposes. 能充分理解学习者语料，来实现获取信息的目的。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 41: With use of a dictionary, can comprehend sufficiently main ideas and some facts in authentic material paralleling oral language. 在使用字典的情况下阅读跟口语类似的真实语料，能够充分理解文本的主旨和一些具体事实。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 42: Can decode authentic materials, which are harder than oral language, for key points with use of a bilingual dictionary. 在使用字典的情况下，能理解比口语更难的真实语料的字面意思，找到关键信息。
0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到

4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 43: Can decode critical elements of public announcements to determine who, when, and where regarding such subjects as public events by using a dictionary. 在使用字典的情况下，能读懂通知中的关键信息的字面意思，来确定事件的人物、时间、地点。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 44: Can follow the narrative thread in extended, specially-prepared discourse. 在阅读篇幅较长的学习者语料时，能把握叙事主线。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 45: Can decode authentic prose (including newspapers and magazines) for general ideas. 能读懂包括报纸杂志在内的真实语料的字面意思，来了解文章大意。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 46: Can glean a range of specific facts from short authentic pieces on sports, movies, and current events, but with error. 在阅读关于体育、电影、时事方面的简短真实语料时，能搜集出一些特定的事实，但是会出现错误。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到
Item 47: When reading text on his/her level, can be skillful in extracting concrete, minor details.

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 48: Can comprehend authentic expository prose without reliance on a dictionary where such prose tends to reflect the spoken language.

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 49: Can comprehend, not just decode, a narrow range of authentic, expository material, including areas of professional interest, without the use of a dictionary.

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 50: Can comprehend, with some dictionary use, authentic material over a wide range of subject matter and topics.

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 51: Can comprehend standard newspaper items addressed to the general public, routine correspondence, reports and technical material in field of interest.

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 52: Can decode, with a dictionary and with substantial effort and moderate error, popular novels, essays, and most literature for the general public. 在使用字典的条件下，能读懂通俗小说、散文、以及大多数面向公众的文学作品的字面意思，但是得付出非常大的努力也会出现中等程度的错误。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 53: Can understand minimally, yet with significant difficulties, prose which is more characteristic of the literary/classical style. 能在极小程度上理解具有文学/文言色彩的文章，但面临重大的困难。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Group 8: The situation model of comprehension and Pragmatic Knowledge

Item 54: When reading text on his/her level, can only draw simplest inferences from reading. 当阅读适合该读者水平的文章时，只能根据文章做最简单的推论。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 55: Can minimally detect subjective attitudes, values, and judgments. 能在极小程度上体察文章的主观态度、价值标准和评判。
   0: No, cannot perform like this. 完全做不到
   1: Yes, in favorable circumstances. 在有利条件下可以做到
   2: Yes, in normal circumstances. 在一般情况下可以做到
   3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
   4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到

Item 56: Can appreciate, to a limited degree, nuances or stylistics, but gaps in detail may be frequent. 能在有限程度上欣赏文章的精妙之处和文体风格，但是在细节的把握上可能常
Group 9: Cognitive strategy

Item 57: Can use preliminarily lexical guessing strategies in dealing with authentic materials, although errors are frequent. 能初步运用“猜测生词的意思”这一阅读策略来处理真实语料，虽然常常出现错误。

0: No, cannot perform like this. 完全做不到
1: Yes, in favorable circumstances. 在有利条件下可以做到
2: Yes, in normal circumstances. 在一般情况下可以做到
3: Yes, even in unfavorable circumstances. 即使在困难的情况下也可以努力做到
4: Yes, with ease in all circumstances. 在任何情况下都可以轻松做到
## APPENDIX 3: CSL READING PROFICIENCY SCALE

### CSL Reading Proficiency Scale: Lower-level Decoding

<table>
<thead>
<tr>
<th>Level</th>
<th>Item</th>
<th>Provisional group</th>
<th>Level the its original scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 3</td>
<td>Item 29: Can have emerging control of high-frequency grammatical constructions common to expository proses. 开始表现出能掌握说明文中常见的高频语法结构的能力。</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Superior) ***</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 28: Can comprehend text successfully when the number of features and structural devices from the literary/classical stratum is limited. 当来自文学作品、文言文的特色和语法结构很少时，能够成功地理解文本。</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Superior) ***</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 8: Can use knowledge on word formation to help decode unknown words. 能用构词法知识帮助识别不认识的词。</td>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 9: When reading text on his/her level, can contextually process polysemes (words with multiple meanings). 读适合他/她阅读水平的文本时，能根据语境理解多义词在特定语境中的准确意思。</td>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 3: Can distinguish graphically-similar characters, which have learned. 对于已经学过的形近字，能够区分。</td>
<td>1) Character Recognition and Orthographic Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 19: Can comprehend sentences with following basic patterns: Indefinite usages of question words. 能读懂含有下列语法的句子：疑问词的不定指（例如在“谁都不喜欢这本书”中，“谁”用作不定代词，意思是“任何人”。）</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 24: Can be aware of the appropriate position of adverbial phrases in Chinese sentences. 能意识到在中文句子中，副词短语应该放置的正确位置。</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 1: Can use knowledge of radicals to help decode characters. 能用部首知识帮助识别汉字。</td>
<td>1) Character Recognition and Orthographic Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level</td>
<td>Item 15: Can understand roughly the</td>
<td>3) Syntactic</td>
<td>ACTFL-CHIN</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 7: Can analyze the intra-word structure.</td>
<td>Parsing and Syntactic Knowledge</td>
<td>(Intermediate-High)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 16: Can comprehend sentences with following basic patterns: Relative clause modification.</td>
<td>2)Lexical Access and Vocabulary Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 25: Can comprehend sentences with following unique Chinese syntactic features: Nominal sentences(i.e., sentences contain no verbs)</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 22: Can comprehend sentences with following basic patterns: Resultative compounds.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 21: Can comprehend sentences with following basic patterns: Various aspect markers.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 18: Can comprehend sentences with following basic patterns: bei-passive.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 23: Can comprehend sentences with following basic patterns: Directional compounds.</td>
<td>3) Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 6: When reading text on his/her level, can group adjacent characters into words successfully.</td>
<td>2)Lexical Access and Vocabulary Knowledge</td>
<td>Created based on theory</td>
</tr>
</tbody>
</table>
| Level 2 | Item 17: Can comprehend sentences with following basic patterns: ba-disposal. | 3) Syntactic Parsing and Knowledge | ACTFL-CHIN (Advanced-
<table>
<thead>
<tr>
<th>Level</th>
<th>Item</th>
<th>Can understand</th>
<th>Created based on Theory</th>
<th>Syntactic Parsing and Syntactic Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>27</td>
<td>Can comprehend sentences with following unique Chinese syntactic features: Topic-comments structure.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>Can derive meaning of sentences with coverbial phrases.</td>
<td>2) Lexical Access and Vocabulary Knowledge</td>
<td>ACTFL-CHIN (Intermediate-mid)</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>Can derive meaning of sentences with the aspect markers (such as le), although still makes errors.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced) *</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>Can comprehend sentences with following unique Chinese syntactic features: “Shi…de” structure.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>Can understand simple Noun+Noun modification with <em>de</em> and Stative Verb + Noun modification with <em>de</em>.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-Mid)</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>Can comprehend sentences with following basic patterns: exclusive usages of question words.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Advanced-plus) ***</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>Can understand limited types of sentence structures, including Subject-Predicate constructions, simple SVO constructions, all in affirmative, negative, and simple question form.</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-Low)</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>Can comprehend preliminarily high-frequency connectors forming complex sentences (e.g., de shí hou, suī rán... kě shì).</td>
<td>3)Syntactic Parsing and Syntactic Knowledge</td>
<td>ACTFL-CHIN (Intermediate-High)*</td>
</tr>
</tbody>
</table>
**CSL Reading Proficiency Scale: Interim-level Textbase Construction**

<table>
<thead>
<tr>
<th>Level</th>
<th>Item</th>
<th>Provisional group</th>
<th>Level in the original scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 4</td>
<td>Item 51: Can comprehend standard newspaper items addressed to the general public, routine correspondence, reports and technical material in field of interest.</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Superior)**</td>
</tr>
<tr>
<td>Level 4</td>
<td>Item 53: Can understand minimally, yet with significant difficulties, prose which is more characteristic of the literary/classical style.</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Superior)**</td>
</tr>
<tr>
<td>Level 4</td>
<td>Item 49: Can comprehend, not just decode, a narrow range of authentic, expository material, including areas of professional interest, without the use of a dictionary.</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Superior)**</td>
</tr>
<tr>
<td>Level 4</td>
<td>Item 52: Can decode, with a dictionary and with substantial effort and moderate error, popular novels, essays, and most literature for the general public.</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Superior)**</td>
</tr>
<tr>
<td>Level 4</td>
<td>Item 37: Can have some appreciation of formulaic rhetorical devices common to more stylized writing. 能够欣赏文学性写作中常见的程式化的修辞手法。</td>
<td>6) Knowledge of Rhetorical Organization</td>
<td>ACTFL-CHIN (Advanced-plus)*</td>
</tr>
<tr>
<td>Level 4</td>
<td>Item 50: Can comprehend, with some dictionary use, authentic material over a wide range of subject matter and topics. 在偶尔使用字典的情况下，能理解多种多样的题材和话题的真实语料。</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Superior)**</td>
</tr>
<tr>
<td>Level 4</td>
<td>Item 45: Can decode authentic prose (including newspapers and magazines) for general ideas. 能读懂包括报纸杂志在内的真实语料的字面意思，来了解文章大意。</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Advanced)</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 35: Can be aware of most Chinese text structures/discourse patterns. 能识别大部分种类的中文文章结构和语篇模式。</td>
<td>6) Knowledge of Rhetorical Organization</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 45: Can decode authentic prose (including newspapers and magazines) for general ideas. 能读懂包括报纸杂志在内的真实语料的字面意思，来了解文章大意。</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Advanced-plus)**</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 36: Can use text structure knowledge to facilitate reading comprehension. 能用篇章结构的知识来帮助阅读理解。</td>
<td>6) Knowledge of Rhetorical Organization</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 46: Can glean a range of specific facts from short authentic pieces on sports, movies, and current events, but with error. 在阅读关于体育、电影、时事方面的简短真实语料时，能搜集出一些特定的事实，但是会出现错误。</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Advanced)*</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 34: Can be aware of very common Chinese text structures/discourse patterns. 能识别很常见的中文文章结构和语篇模式。</td>
<td>6) Knowledge of Rhetorical Organization</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 44: Can follow the narrative thread in extended, specially-prepared discourse. 在阅读篇幅较长的学习者语料时，能把握叙事主线。</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Intermediate-High)</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 42: Can decode authentic materials, which are harder than oral language, for key points with use of a bilingual dictionary. 在使用字典的情况下，能理解比口语更难的真实语料的字面意思，找到关键信息。</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Intermediate-High)</td>
</tr>
<tr>
<td>Level</td>
<td>Item</td>
<td>Description</td>
<td>The Text Model</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Level 3</td>
<td>Item 47: When reading text on his/her level, can be skillful in extracting concrete, minor details.</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Advanced-CHIN)</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 30: When reading text on his/her level, can successfully identify referential expressions/antecedents.</td>
<td>4) Semantic-proposition Encoding</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 33: Can use knowledge on cohesive devices to facilitate reading comprehension.</td>
<td>5) Knowledge of Coherence</td>
<td>Created based on theory</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 32: Can be aware of commonly used cohesive devices within and across paragraphs.</td>
<td>5) Knowledge of Coherence</td>
<td>ACTFL-CHIN (Advanced-CHIN)</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 41: With use of a dictionary, can comprehend sufficiently main ideas and some facts in authentic material paralleling oral language.</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Intermediate-Mid)</td>
</tr>
<tr>
<td>Level 2</td>
<td>Item 39: Can read simple connected, specially prepared material for basic survival and social needs (Note: “specially prepared material” hereafter refers to “material specially prepared for adult second language learners of Chinese”).</td>
<td>7) The Text Model of Comprehension</td>
<td>ACTFL-CHIN (Intermediate-Low)*</td>
</tr>
</tbody>
</table>

*Level 2 Item 39: Can read simple connected, specially prepared material for basic survival and social needs (Note: “specially prepared material” hereafter refers to “material specially prepared for adult second language learners of Chinese”).
| Level 1 | Item 54: When reading text on his/her level, can only draw simplest inferences from reading. | 8) The Situation Model of Comprehension and Pragmatic Knowledge | ACTFL-CHIN (Advanced-plus)** |
| Level 1 | Item 55: Can minimally detect subjective attitudes, values, and judgments. | 8) The Situation Model of Comprehension and Pragmatic Knowledge | ACTFL-CHIN (Advanced-plus)** |
| Level 3 | Item 56: Can appreciate, to a limited degree, nuances or stylistics, but gaps in detail may be frequent. | 8) The Situation Model of Comprehension and Pragmatic Knowledge | ACTFL-CHIN (Superior)** |
| Level 3 | Item 57: Can use preliminarily lexical guessing strategies in dealing with authentic materials, although errors are frequent. | 9) Cognitive Strategy | ACTFL-CHIN (Intermediate-High) |
| Level 3 | Item 38: Can comprehend set expressions using basic vocabulary. 能理解由基础词汇构成的固定表达。例如“今天天气怎么样？”“请问现在几点？”等等。 | 7) The Text Model of Comprehension | ACTFL-CHIN (Novice-High) |

**CSL Reading Proficiency Scale: Higher-level Situation-model Building**
APPENDIX 4: DEFINITION OF TERMS

Alphabetic writing—“A writing system made up of separate letters which represent sounds. Some examples of alphabetic writing systems are Roman script, Arabic script, Cyrillic script.” (Richards, Platt, and Platt, 2002, p.18).

Assessment literacy—“Knowledge about testing that supports valid interpretation of test scores for their intended purposes, such as knowledge about test development practices, test score interpretations, threats to valid score interpretations, score reliability and precision, test administration, and use” (AERA, APA, NCME, 2014, p. 216).

Calibration—“In item response theory, the process of estimating the parameters of the item response function.” (AERA, APA, NCME, 2014, p. 216)

CEFR—the Common European Framework of Reference for Languages: Learning, Teaching, Assessment, abbreviated in English as CEFR, is a guideline used to describe foreign language abilities across Europe and, increasingly, in other countries.

Coefficient alpha—An internal-consistency reliability coefficient based on the number of parts into which a test is partitioned (e.g., items, subtests, or raters), the interrelationships of the parts, and the total test score variance. (AERA, APA, NCME, 2014, p. 217)

Component skills—“Often the (language) skills are divided into subskills, such as discriminating sounds in connected speech, or understanding relations within a sentence.” (Richards, Platt, and Platt, 2002, p.255). Subskills are also referred to as component skills.

Construct—“The concept or characteristic that a test is designed to measure” (AERA, APA, NCME, 2014, p. 11).

CSL leaners—Learners who learn Chinese as the second language.

Decoding—“The process of trying to understand the meaning of a word, phrase or sentence” (Richards, Platt, and Platt, 2002, p.122).
Descriptor—A language proficiency scale is composed of a group of descriptors define at an ascending series of levels various aspects of language proficiency.

Dimension—Aspect or feature of a situation, problem, or thing. In this study, dimensions of reading ability refer to component skills of reading ability.

Higher-level processing—Reading ability can be modeled as consisting of two levels of cognitive/metacognitive processing: Lower-level decoding and higher-level processing. Higher-level processing involves: a text model of comprehension, a situation model of comprehension, and reading skills under the command of the executive control mechanism in working memory (e.g., strategies, goals, inference, background knowledge, and comprehension monitoring) (Grabe, 2009).

Interim assessments or tests—“Assessments administered during instruction to evaluate students’ knowledge and skills relative to a specific set of academic goals to inform policy-maker or educator decisions at the classroom, school, or district level.” (AERA, APA, NCME, 2014, p. 220).

Lower-level decoding—Reading ability can be modeled as consisting of two levels of cognitive/metacognitive processing: Lower-level processing and higher-level decoding. Lower-level processing has three components: word recognition, syntactic parsing (using grammatical information to understand the word integration), and semantic-proposition encoding (building clause-level meaning from word meanings and grammatical information) (Grabe, 2009).

Language proficiency scale/framework—A hierarchical scale of descriptors that specifies proficiency characteristics at various points along the proficiency continuum.

Language skill—“The mode or manner in which language is used. Listening, speaking, reading, and writing are generally called the four language skills.” (Richards, Platt, and Platt,

Rasch Analysis—Rasch model (Rasch, 1960) is one of the most basic Item Response Theory (IRT) models. It can be used to plot the “difficulty value” of individual descriptors on the continuum of proficiency. Specifically, “descriptors can be used as questionnaire items for teacher assessment of their learners (Can he/she do X?). The descriptors can be calibrated directly onto an arithmetic scale in the same way that test items are scaled in item banks.” (Council of Europe, 2001, p.211).

Reading ability—the ability to utilize linguistic knowledge, pragmatic knowledge, cognitive processing skills, and metacognitive mechanisms to construct meaning from written materials.

Standardized test—Tests that are developed, administered, scored under uniformed conditions.

Summative assessment—“The assessment of a test taker’s knowledge and skills typically carried out at the completion of a program of learning, such as the end of an instructional unit.” (AERA, APA, NCME, 2014, p. 224)

Teacher-developed assessments—Assessments made by teachers instead of testing professionals.

Unidimensionality—Unidimensionality assumes that all items in an instrument measure the same latent trait and all point in the same direction along it. (Bond, and Fox, 2015, p 157)

Validity— “The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test.” (AERA, APA, NCME, 2014, p. 225)

Word decision—The process of grouping adjacent characters into words in ongoing reading
(Shen, 2008).
REFERENCES


Chalhoub-Deville, M. (2009). Content validity considerations in language testing


Information from lecture delivered by M. R. Montaño-Harmon, Ph. D., Professor Emeritus, California State University, Fullerton, June, 2001, based on (1) doctoral dissertation research and (2) ongoing research in four states in the United States. (Kaplan, 2001)


Koda, K. (1997). Orthographic knowledge in L2 lexical processing: A cross-
linguistic perspective. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: Rationale for pedagogy* (pp. 35-52). New York: Cambridge University Press.


Richards, J. (1976). The role of vocabulary teaching, TESOL Quarterly, 10, 77-89


Wu, J-T., & Liu, I-M. (1996). Chinese lexical access. In M. H. Bond (Eds.), *The handbook of Chinese psychology* (pp. 31-42). Hong Kong: Oxford University.


