Measurement Quality and Rating Scale Functioning of a CSL Classroom Assessment Instrument

Shuai Li
Georgia State University, sli12@gsu.edu

Yali Feng
Georgia State University, yfeng8@gsu.edu

Ting Wen

Follow this and additional works at: https://scholarworks.gsu.edu/mcl_facpub

Part of the Other Languages, Societies, and Cultures Commons

Recommended Citation
Li, Shuai; Feng, Yali; and Wen, Ting, "Measurement Quality and Rating Scale Functioning of a CSL Classroom Assessment Instrument" (2019). World Languages and Cultures Faculty Publications. 79.
https://scholarworks.gsu.edu/mcl_facpub/79

This Book Chapter is brought to you for free and open access by the Department of World Languages and Cultures at ScholarWorks @ Georgia State University. It has been accepted for inclusion in World Languages and Cultures Faculty Publications by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
Measurement Quality and Rating Scale Functioning of a CSL Classroom Assessment Instrument

Shuai Li
Georgia State University

Yali Feng
Georgia State University

Ting Wen
Beijing Language and Culture University

Abstract
It is a wide practice that Chinese language instructors develop their own instruments for classroom assessment and make important pedagogical decisions (e.g., assigning grades) accordingly. However, the quality of such instruments has rarely been discussed in the literature. This chapter focuses on the measurement quality of an instructor-developed test used as a final written exam in an undergraduate Chinese language course in the U.S. The test was designed to assess the linguistic knowledge taught in the course and contained 37 binary-scored (0/1) items and 17 constructed-response items. Two four-category rating scales were developed to evaluate the constructed responses. Examinees were 88 students enrolled in the Chinese course. Results showed acceptable overall measurement quality of the test as indicated by measures of difficulty, discrimination, reliability, and Rasch model fit. The two rating scales, however, were found to include excessive score categories, suggesting measurement redundancy. The findings of this study are intended to raise the awareness among CSL instructors of the
potential limitations of their self-developed assessment instruments.
Keywords: measurement quality, classroom assessment, Chinese, linguistic knowledge, rating scale functioning, Rasch Model
Introduction
Over the past two decades, Chinese as a second language (CSL) has gained increasing international popularity. It is estimated that over 100 million people are studying Chinese around the globe (Hanban, 2014). By the end of 2016, 67 nations have included Chinese language teaching as part of their national education systems (Liu, 2017). This growing interest in studying Chinese has led to increased demand in assessment. For example, in 2017 about 6 million examinees took the HSK (Hanyu Shuiping Kaoshi, or the Chinese Proficiency Test) (Liu, 2017). However, empirical research on L2 Chinese assessment, especially those published in English with international readership, remains rather limited (Ke, 2012; Zhang & Lin, 2017a). Notable exceptions are the edited volume by Zhang and Lin (2017b), the monograph by Meyer (2014), and a few recent journal articles (e.g., Li, 2018; Poehner, Zhang, & Lu, 2015).

Among the existing studies on CSL assessment, the focus is mainly on standardized tests such as the HSK. Other types of assessment have received scarce attention, such as classroom assessment that is closely related to teaching practices (cf. Li, 2019; Wang, 2017). Compared to standardized proficiency tests, classroom assessment (e.g., in-class exams, tasks for assessing writing) presumably serves a larger learner population because it is used in virtually every language classroom. Moreover, because classroom assessment constitutes a primary basis for various pedagogical decisions (e.g., instructional planning, program evaluation, and student grade assignment), ensuring its quality plays a critical role in language curriculum development (Brown, 1995).

In reality, however, the quality of second language classroom assessment cannot be assumed. Language classroom assessment is typically created by instructors, who may not possess the necessary knowledge, skills, and abilities to design, develop, and evaluate such assessment due to a lack of relevant training (Gardner & Rea-Dickins, 2001; Hasselgreen, Carlsen, & Helness, 2004; Jin, 2010). Research on language assessment practices among second language instructors revealed several issues, including highly diversified approaches to assessment, impressionistic judgment, and inappropriate implementation of assessment (Davison & Leung, 2009).
Unsurprisingly, the quality of instructor-developed assessment has been questioned in the field of language testing and assessment (e.g., Alderson, 2005; Leung, 2005; Vogt & Tsagari, 2014). In the field of CSL research, scholars have just begun to empirically investigate the quality of CSL classroom assessment (e.g., Li, 2019). Echoing this recent development in the field, the present study examines the measurement quality of an instructor-developed classroom assessment instrument used in an undergraduate Chinese program in the U.S.

**Research on CSL Classroom Assessment**

Language assessment aims to infer a person’s language-related characteristics (e.g., reading comprehension ability) according to test and non-test data (e.g., observation of students’ groupwork) (Purpura, 2016). Both large-scale, standardized tests and various forms of classroom assessment (e.g., portfolios, oral interviews, self- and peer-assessment, and written tests) fall under the scope of language assessment (Cheng, 2013; Cockey, 2014). Depending on the purpose, classroom assessment can be formative (i.e., assessment for learning) or summative (i.e., assessment of learning) (Rea-Dickins, 2008). Formative assessment is generally used to gauge students’ learning progress during instruction, and the results are used to inform subsequent teaching and learning. On the other hand, summative assessment typically administered after instruction to evaluate learning outcomes, with the purpose of informing decision-making (e.g., grades, placement).

Earlier research on CSL classroom assessment concentrated on describing frameworks for assessment design. For example, Ke (2006) outlined a model of formative assessment to be implemented in an undergraduate CSL program in the U.S. Conforming to the program curriculum objectives, Ke’s model proposes a task-based (i.e., simulating classroom activities) approach with criterion-referenced testing (i.e., assessment based on pre-determined learning objectives) and componential scoring (i.e., assigning separate scores according to domains of assessment, such as vocabulary and grammar). Similarly, Bachman and Palmer (2010) detailed the need and design of a speaking test to be used in an elementary-
level undergraduate CSL course in the U.S. In both studies, however, the authors did not share the actual assessment material, nor did they report results regarding the actual application of the model and the operation of the test, making it difficult for researchers and instructors to evaluate the effectiveness of the proposed model and design.

More recently, echoing the call for empirical research on effectiveness of assessment methods (Xiang, 2016), researchers have started to empirically examine the quality of CSL assessment instruments used in classrooms (e.g., Li, 2019; Wang, 2017). For example, focusing on formative assessment, Wang (2017) explored whether self- and peer-evaluation helped advanced-level adult L2 Chinese learners develop their oral presentation skills. Results showed that the learners not only helped with the development of the evaluation rubric, but also received focused training on using the rubric for self- and peer-evaluation of their own oral productions. Results showed a significant improvement in the accuracy of self- and peer-evaluation, suggesting that the reliability of formative assessment can be enhanced by involving learners in the process of assessment development. Another example is Li’s (2019) investigation into the measurement quality of the binary (0/1) scored items of an instructor-developed final written test used in a first semester undergraduate Chinese course. He reported that the 64 test items showed acceptable overall reliability (Cronbach’s α = .86), difficulty (mean = 86%), discrimination (i.e., to what extent a test or an item can differentiate higher ability examinees from their lower-ability counterparts) (averaged item discrimination statistic = 0.28) and Rasch Model fit (explained below). He also found that the test could be improved by developing more difficult items, by revising and/or removing items with below-threshold discriminatory power (i.e., items that cannot effectively differentiate high-achievers from low-achievers), and by addressing measurement redundancy (i.e., items assessing the same aspect of linguistic knowledge have similar difficulty level).

The Rasch Model, as adopted in Li’s study, is a probability-based psychometric model and has been used in L2 performance assessment (e.g., Eckes, 2015; McNamara, 1996). Its fundamental assumption is that, under an idealized assessment condition (e.g., no time pressure, no guessing), to
the same test item, a higher ability examinee should be more likely to provide a correct answer than a lower-ability examinee; on the other hand, for the same examinee, an easier test item should have a higher level of probability to be answered correctly than a more difficult item. As a means for quality check, the Rasch Model calculates fit statistics called Mean Square (MnSq) to gauge the extent that test items and examinee behaviors conform to the model’s assumption. The range of acceptable MnSq value to indicate good model fit is 0.5 to 1.5 (Boone, Staver, & Yale, 2014, p. 166; Wright & Linacre, 1994). Based on raw test scores, the Rasch Model also calculates individual item difficulty and individual examinee ability based on a shared logit scale, which has a zero point and extends to positive and negative infinity. The unit of measurement on the logit scale is called logit. A larger (and positive) logit value denotes a higher level of examinee ability or item difficulty, and vice versa. As such, direct comparisons of examinee ability and item difficulty can be conducted, and this is aided by the item-examinee map (Figure 10.1 detailed in the Results section) generated by the Rasch Model.

As Li (2019) argued, the Rasch Model can provide important information about the quality of assessment instruments and examinee behaviors, which can inform assessment decisions in a classroom context. Traditionally, instructors use overall test scores as a basis for evaluating L2 ability. This approach may be problematic in case of low-quality test items (e.g., an item that is difficult for higher ability examinees but easy for lower-ability examinees) and/or unexpected examinee responses (e.g., a higher ability examinee incorrectly answers a very easy item). The Rasch Model can help identify such problematic test items and examinee responses, thus enabling instructors to make informed pedagogical decisions based on assessment results.

Given the limited empirical evidence, the potential of the Rasch Model for informing CSL classroom assessment has not been fully explored. Li (2019), for example, used the model to examine the measurement quality of the binary-scored sections of his test, but excluded constructed-response items that were scored based on pre-determined rubrics. Scoring rubrics are widely used in CSL classrooms to evaluate language production. For
example, a rating scale with descriptions of different levels of expected performance can be used to evaluate writing. Such a common practice seemingly assumes that each scoring band represents a level of ability that is distinct from the ability levels indicated by the neighboring score bands (e.g., scores 2 and 3 indicate two distinct levels of writing ability). Because this assumption is often left unchecked, it is difficult to detect improper rating scale functioning (e.g., scores 2 and 3 cannot be empirically separated in terms of the ability being assessed). Consequently, the quality of classroom assessment based on the rating scale cannot be assumed to be sound. The present study takes a further step by examining rating scale functioning, in addition to investigating the quality of test items and examinee test responses, for an instructor-developed CSL summative assessment instrument used in an undergraduate Chinese language course in the U.S. The two research questions were:

RQ1: What is the measurement quality (i.e., reliability, difficulty, discrimination, and Rasch Model fit) of an instructor-developed test for an elementary CFL course?

RQ2: To what extent do the rating scales developed for the test function properly?

Method

Context of the Study

The test under investigation was developed for a second semester Chinese course (3 credits) offered through an undergraduate Chinese program at a public university in the U.S. There is no explicit proficiency goal of the Chinese course. The main instructional objectives of the course include linguistic knowledge (i.e., characters, vocabulary, and grammar) and language skills (i.e., listening, speaking, reading, and writing), with cultural understanding and learning strategies as the secondary goals of instruction. The course covers Chapters 6 to 10 of the textbook entitled Integrated Chinese (Level 1, Part 1, 3rd edition).
**Test Development and Structure**

The test (Appendix One) examined in this study is the written part of the final exam packet of the course. It mainly assesses the linguistic knowledge and language skills in listening, reading, and sentence-level writing. Speaking and essay writing are assessed in separate tasks of the final exam packet. The scope and content of the test is determined by the course syllabus and the textbook.

The course instructor developed the test by drawing on his teaching experience and by referring to the accompanied workbook of the textbook. The test contains nine sections (see Table 10.1 for an overview). Sections #1 and #2 (four items for each) focus on listening comprehension with true/false questions. Section #1 is based on a dialogue while Section #2 is based on a monologue. These two sections reflect the pedagogical emphasis on developing listening skills as demonstrated in interactive and non-interactive discourses.

Sections #3 (eight items) and #4 (six items) concentrate on assessing semantic radical knowledge (i.e., a component of character that indicates the meaning of the character). Section #3 asks examinees to directly provide the meaning of the targeted semantic radicals. Section #4 requires examinees to apply their knowledge of semantic radicals to inferring the meaning of compound characters. The characters included in both sections are not taught in the class, hence examinees must draw on their radical knowledge to answer the questions.

Sections #5 (eight items) and #6 (seven items) mainly assess vocabulary knowledge, with an emphasis on adverbs, adjectives, connectives, conventionalized phrases, and modal verbs. These aspects of vocabulary knowledge receive special attention in course instruction because the instructor had found them relatively difficult for his students to master. Section #5 includes multiple-choice questions (i.e., choosing one correct answer among four options). Section #6 includes fill-in-the-blank questions with a shared word bank.

Section #7 (eight items) specifically assesses grammar in terms of word order in sentences. Examinees are expected to connect individual words and phrases together to form grammatical (and meaningful) sentences.
Finally, Sections #8 (four items) and #9 (four items) focus on reading comprehension and, to a less extent, sentence-level writing. The questions in Section #8 require examinees to first comprehend individual questions and then write responses (in Chinese characters) according to their respective circumstances. Section #9 expects examinees to comprehend a 161-word text and answer questions with Chinese characters. These two question formats reflect the pedagogical emphasis on sentence-level and discourse-level reading skills as well as sentence-level writing skills.

Examinees
Examinees were students enrolled in the second semester Chinese course from Spring 2012 to Spring 2014 (i.e., three Spring semesters and two Fall semesters). Excluding those who did not show up for the test, the final number of the participants was 88. There were 37 males and 51 females. The academic backgrounds of the examinees varied, although most of them were in the humanities and social sciences.

Data Analysis
For the 37 binary-scored items (Sections #1 to #6), a correct response was coded as 1 and an incorrect response 0. The binary-scored items were analyzed with the Rasch Dichotomous Model. For the remaining 17 items (Sections #7 to #9) that were evaluated through ratings, the researcher and one course instructor jointly developed two rating scales after reviewing all examinee responses. The first rating scale was developed for evaluating responses to the word order section (Section #7). This rating scale includes four scoring categories and holistically taps sentence grammaticality and clarity in meaning expression (see below).

1. Zero points are assigned when a sentence’s grammar is completely wrong and is incomprehensible to native Chinese speakers.
2. One point is awarded when a sentence’s grammar is partially incorrect and/or incomplete due to missing words and is only partially understandable to native speakers.
3. Two points are assigned when the grammar of a sentence is slightly less well-formed (e.g., adding extra words, slightly wrong word order), yet the
grammar error does not interfere with meaning interpretation by native speakers.

4. Three points are given when a sentence is perfect in grammar and its meaning can be correctly interpreted by native speakers.

The second rating scale was developed for evaluating responses to the short answer section (Section #8) and the reading comprehension section (Section #9). The rating scale holistically taps reading comprehension (primary criterion) and grammaticality (secondary criterion). There are four scoring categories (see below):

1. Zero points are assigned to a response that is irrelevant to the question being asked, thus demonstrating a complete lack of reading comprehension.
2. One point is given when a response partially answers the question being asked, or when a response is relevant but contains serious grammar error that interferes with meaning expression.
3. Two points are awarded when a response correctly answers the question being asked, although it contains slight grammar error that does not interfere with meaning expression.
4. Three points are assigned to a response that not only correctly answers the question but also contains no grammar error.

The first author of this project rated all responses in Sections #7, #8, and #9, resulting in a total of 1,496 individual ratings (88 examinees x 17 items per examinee). The rating data was analyzed under the Rasch Rating Scale Model.

Combining all 54 test items (37 binary-scored items + 17 rated items), the score range for each examinee is 0–88. Rasch analyses were performed with the software FACETS (Version 3.71.3). Additional statistical analysis was conducted with SPSS Version 16.0.

**Criteria for Evaluating Rating Scale Quality**

The measurement quality of the two rating scales was evaluated against the widely used criteria proposed by Linacre (2002). First, each score category (i.e., a score level) on a rating scale needs to have at least 10 observations (i.e., scores) to ensure stable parameter estimation. Second, the distributional patterns of scores based on a rating scale should show regular patterns (e.g.,
even, unimodal, and bimodal distributions across score categories). Third, the average (observed) measures (indicating the average ability of examinees who receive a particular rating score) increase with rating score categories. Fourth, the value of the outfit mean square (MnSq) statistic should be below 2.0 for each score category to ensure acceptable measurement quality. Fifth, step thresholds should advance monotonically with rating score categories. Step thresholds are the intersections where two neighboring score categories are equally likely to be observed (see the crossover points where two curves intersect in Figures 10.2 and 10.3). For a properly functioning rating scale, higher scores reflect higher levels of competence; thus, each score category can be expected to take turns to be the most probable to be assigned as ability level increases. As a result, step thresholds should also advance monotonically with increasing score categories. Step disordering, that is, step thresholds do not increase with score categories, suggests that a score category is either poorly defined or too narrowly defined in that it cannot be separated from its neighboring categories in terms of the ability being assessed. Finally, the distance between adjacent step thresholds should fall into the range between 1.4 to 5.0 logits to ensure a balanced level of measurement precision.

**Results**

RQ1 focuses on the measurement quality of the test and its items. The mean test score for the 88 examinees is 65.67 (out of 88, or 74.63%) with a standard deviation of 15.12 and a score range of 28–88 (or 31.82%–100.00%). Table 10.1 shows the means and standard deviations of each section. Because there are two scoring methods (i.e., binary scoring and rating), the relative difficulty of the sections is indicated by the percentage scores in the parentheses: Section #6 (filling in the blanks, 58%) is the most difficult section, whereas Section #7 (answering individual questions, 87%) is the easiest. For individual test items, Table 10.2 shows the difficulty and discrimination statistics of the 54 items. For the 37 binary-scored items (Sections #1 to #6) with a maximum score of 1.00 for each item, individual item means range between 0.38 and 0.93 with an average of 0.75 (out of 1.00, or 75.00%). For the remaining 17 rated items with a maximum score of
3.00 for each item, individual item means range between 1.59 (out of 3, or 53.00%) to 2.82 (out of 3, or 94.00%) with an average of 2.24 (out of 3, or 74.73%).

Table 10.1 Section Means and Standardized Deviations

<table>
<thead>
<tr>
<th>Sections (item number)</th>
<th>Format</th>
<th>Scoring method</th>
<th>Mean (Percentage)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 Listening comprehension: dialogue (k = 4)</td>
<td>True/False</td>
<td>Binary 0/1</td>
<td>0.73 (73%)</td>
<td>0.11</td>
</tr>
<tr>
<td>#2 Listening comprehension: monologue (k = 4)</td>
<td>True/False</td>
<td>Binary 0/1</td>
<td>0.64 (64%)</td>
<td>0.05</td>
</tr>
<tr>
<td>#3 Radical meaning (k = 8)</td>
<td>Fill in blanks</td>
<td>Binary 0/1</td>
<td>0.81 (81%)</td>
<td>0.07</td>
</tr>
<tr>
<td>#4 Radical application (k = 6)</td>
<td>Multiple choice</td>
<td>Binary 0/1</td>
<td>0.85 (85%)</td>
<td>0.09</td>
</tr>
<tr>
<td>#5 Vocabulary: multiple choice: regular (k = 8)</td>
<td>Multiple choice</td>
<td>Binary 0/1</td>
<td>0.82 (82%)</td>
<td>0.10</td>
</tr>
<tr>
<td>#6 Vocabulary: multiple choice: word bank (k = 7)</td>
<td>Multiple choice</td>
<td>Binary 0/1</td>
<td>0.58 (58%)</td>
<td>0.15</td>
</tr>
<tr>
<td>#7 Grammar: word order (k = 8)</td>
<td>Constructed response</td>
<td>Rating 0–3</td>
<td>2.11 (70%)</td>
<td>0.41</td>
</tr>
<tr>
<td>#8 Reading &amp; writing: individual questions (k = 5)</td>
<td>Constructed response</td>
<td>Rating 0–3</td>
<td>2.60 (87%)</td>
<td>0.19</td>
</tr>
<tr>
<td>#9 Reading &amp; writing: paragraph (k = 4)</td>
<td>Constructed response</td>
<td>Rating 0–3</td>
<td>2.06 (69%)</td>
<td>0.44</td>
</tr>
</tbody>
</table>
Table 10.2 Test Item Statistics (k = 54)

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>Rasch analysis</th>
<th>Measure</th>
<th>Model S.E.</th>
<th>Infit</th>
<th>MnSq</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Point biserial correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lis.1.1</td>
<td>.83</td>
<td>.39</td>
<td>0.25</td>
<td>-0.46</td>
<td>0.29</td>
<td>1.05</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Lis.1.2</td>
<td>.68</td>
<td>.47</td>
<td>-0.01</td>
<td>0.41</td>
<td>0.25</td>
<td>1.39</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>Lis.1.3</td>
<td>.81</td>
<td>.40</td>
<td>0.07</td>
<td>-0.37</td>
<td>0.29</td>
<td>1.19</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Lis.1.4</td>
<td>.59</td>
<td>.49</td>
<td>0.38</td>
<td>0.9</td>
<td>0.24</td>
<td>1.01</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Lis.2.1</td>
<td>.66</td>
<td>.48</td>
<td>0.42</td>
<td>0.54</td>
<td>0.25</td>
<td>0.97</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>Lis.2.2</td>
<td>.60</td>
<td>.49</td>
<td>0.36</td>
<td>0.84</td>
<td>0.24</td>
<td>1.05</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Lis.2.3</td>
<td>.60</td>
<td>.49</td>
<td>0.46</td>
<td>0.84</td>
<td>0.24</td>
<td>0.93</td>
<td>-0.8</td>
<td></td>
</tr>
<tr>
<td>Lis.2.4</td>
<td>.69</td>
<td>.46</td>
<td>0.33</td>
<td>0.35</td>
<td>0.25</td>
<td>1.05</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Rad.3.1</td>
<td>.89</td>
<td>.32</td>
<td>0.28</td>
<td>-1.07</td>
<td>0.35</td>
<td>0.99</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Rad.3.2</td>
<td>.78</td>
<td>.41</td>
<td>0.17</td>
<td>-0.21</td>
<td>0.28</td>
<td>1.13</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Rad.3.3</td>
<td>.89</td>
<td>.31</td>
<td>0.26</td>
<td>-1.07</td>
<td>0.35</td>
<td>1.01</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Rad.3.4</td>
<td>.80</td>
<td>.41</td>
<td>0.17</td>
<td>-0.29</td>
<td>0.28</td>
<td>1.13</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Rad.3.5</td>
<td>.84</td>
<td>.37</td>
<td>0.32</td>
<td>-0.64</td>
<td>0.31</td>
<td>0.99</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Rad.3.6</td>
<td>.84</td>
<td>.37</td>
<td>0.40</td>
<td>-0.64</td>
<td>0.31</td>
<td>0.93</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>Rad.3.7</td>
<td>.68</td>
<td>.47</td>
<td>0.35</td>
<td>0.41</td>
<td>0.25</td>
<td>1.03</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Rad.3.8</td>
<td>.74</td>
<td>.44</td>
<td>0.41</td>
<td>0.08</td>
<td>0.26</td>
<td>0.96</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>Rad.4.1</td>
<td>.93</td>
<td>.25</td>
<td>0.45</td>
<td>-1.67</td>
<td>0.43</td>
<td>0.84</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>Rad.4.2</td>
<td>.93</td>
<td>.25</td>
<td>0.11</td>
<td>-1.67</td>
<td>0.43</td>
<td>1.06</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Rad.4.3</td>
<td>.70</td>
<td>.46</td>
<td>0.31</td>
<td>0.28</td>
<td>0.26</td>
<td>1.06</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Rad.4.4</td>
<td>.76</td>
<td>.43</td>
<td>0.37</td>
<td>-0.06</td>
<td>0.27</td>
<td>0.98</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Rad.4.5</td>
<td>.89</td>
<td>.32</td>
<td>0.38</td>
<td>-1.07</td>
<td>0.35</td>
<td>0.91</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>Rad.4.6</td>
<td>.86</td>
<td>.35</td>
<td>0.36</td>
<td>-0.84</td>
<td>0.33</td>
<td>0.95</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td>Mul.5.1</td>
<td>.92</td>
<td>.27</td>
<td>0.26</td>
<td>-1.49</td>
<td>0.41</td>
<td>0.98</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Mul.5.2</td>
<td>.65</td>
<td>.48</td>
<td>0.20</td>
<td>0.6</td>
<td>0.25</td>
<td>1.2</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Mul.5.3</td>
<td>.90</td>
<td>.30</td>
<td>0.35</td>
<td>-1.2</td>
<td>0.37</td>
<td>0.94</td>
<td>-0.1</td>
<td></td>
</tr>
<tr>
<td>Mul.5.4</td>
<td>.85</td>
<td>.36</td>
<td>0.50</td>
<td>-0.74</td>
<td>0.32</td>
<td>0.84</td>
<td>-0.8</td>
<td></td>
</tr>
<tr>
<td>Mul.5.5</td>
<td>.85</td>
<td>.36</td>
<td>0.36</td>
<td>-0.74</td>
<td>0.32</td>
<td>0.94</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>Mul.5.6</td>
<td>.76</td>
<td>.43</td>
<td>0.44</td>
<td>-0.06</td>
<td>0.27</td>
<td>0.92</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td>Mul.5.7</td>
<td>.70</td>
<td>.46</td>
<td>0.12</td>
<td>0.28</td>
<td>0.26</td>
<td>1.24</td>
<td>2.1</td>
<td></td>
</tr>
</tbody>
</table>
Table 10.2 also shows the Rasch calibrated item statistics. Individual item difficulty measures spread over 3.73 logits (from -1.67 to 2.06 logits). The item separation index is 3.14 (or 4.52 strata) with a reliability coefficient of 0.91. This means that the 54 items can be reliably divided into more than four statistically significant difficulty levels. The item infit MnSq statistics range from 0.55 to 1.78. Among the 54 items, two items (or 3.70%) have infit MnSq statistics over 1.50: Ord.7.4 (infit MnSq = 1.61) and Con.9.1 (infit MnSq = 1.78). These two items need further inspection for their
functioning. Moreover, item discrimination statistics (point biserial correlation) spread from -0.01 to 0.74 with an average of 0.40.

The Rasch calibrated person measures spread over 5.18 logits (from -0.65 to 4.52 logits) with a mean of 1.51 logits (SD = 1.42). The person infit MnSq statistics range from 0.58 to 2.03, with two examinees’ infit MnSq statistics above the 1.5 threshold. The person separation index is 2.68 (or 3.90 strata) with a reliability coefficient of 0.88. This means that the examinees can be differentiated for at least three statistically significant ability levels (i.e., high, mid, and low) with a good level of reliability.

Figure 10.1 presents the item-examinee map. The leftmost column is the logit scale, based on which examinee ability and item difficulty can be compared. The second column to the left shows the ability distribution of examinees with each asterisk (*) representing one examinee. Higher ability examinees appear in higher positions than lower-ability examinees. The third column to the left shows the difficulty distribution of the 54 test items. More difficult items occupy higher positions on the scale than easier items. As Figure 10.1 shows, the difficulty of these test items generally matches the ability of the examinees on the lower stretches of the logit scale, suggesting that the test overall is relatively easy. There are also cases of measurement redundancy. For example, three items in Section #6 (i.e., Bla.6.3, Bla.6.6, and Bla.6.7) show almost identical levels of difficulty. The two rightmost columns show the structures of the two rating scales for the word order section (Section #7, column S.2), and for the constructed responses sections (Sections #8 and 9, column S.3). These two columns visualize how the two rating scales have been used in evaluating the examinees’ responses: Within each rating scale, the score category of 3 covers a wider range of the underlying ability being assessed than the other score categories; on the other hand, the score categories of 1 and 2 for Sections #8 and 9 (column S.3) cover a fairly narrow range of the underlying ability.
\begin{figure}
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\textbf{Measr} & \textbf{Examinees} & \textbf{-Items} & S. 2 & S. 3 \\
\hline
5 & *** & + & (3) & (3) \\
\hline
4 & * & + & + & \\
\hline
3 & **** & + & + & \\
\hline
2 & * & Bla. 6.2 & + & + \\
\hline
1 & ** & Con. 9.2 Con. 9.3 Ord. 7.5 & + & + \\
\hline
0 & ** & Bla. 6.5 Con. 9.4 Lis. 2.1 Mul. 5.2 & --- & \\
\hline
-1 & * & Ord. 7.7 Rad. 3.2 & 1 & \\
\hline
-2 & * & Con. 8.1 Ord. 7.4 Rad. 3.4 & 1 & \\
\hline
\hline
\end{tabular}
\caption{Figure 10.1 Item-examinee Map Based on the Test}
\end{figure}
RQ2 investigates how the rating scales function. Table 10.3 displays the data of the rating scale for Section #7 (word order). It shows that each score category is used at least 10 times (see counts used column), that the frequency distribution of score usage (see % column) follows a regular upward pattern that peaks at the score of four. Moreover, the average observed measures of the four score categories increase monotonically (see avge. meas. column), and the difference between the observed measures and the expected measures (see exp. meas. column) is small (i.e., ranging from 0.01 to 0.24 logit). The outfit MnSq values of all score categories are under 2.0, indicating satisfactory model fit. Step thresholds, however, indicate an instance of disordereding (i.e., step thresholds do not increase monotonically): whereas the step threshold for score category 1 is at -.38 logit, the step threshold for score category 2 is at -.54 logit. This step disordereding is further illustrated in Figure 10.2, which shows the structure of the rating scale. The horizontal axis in Figure 10.2 indicates the difference between examinee ability and item difficulty, and the vertical axis indicates the probability (0%-100%) of assigning a particular score. The four curved lines are the probability curves for the four score categories. These probably curves indicate which score category is the most likely to be awarded given a specific combination of item difficulty and examinee ability. For example, if an examinee’s ability is 0.5 logit above the difficulty of an item, he/she is most likely to receive a score of 2. The step disordereding for score category 1 is shown by the fact that its probability curve is submerged under the probability curves of score categories 0 and 2 and thus never forms a distinct “hill” (as in the case of score category 2). This result suggests that score category 1 is poorly defined or too narrowly defined to be effectively separated from neighboring score categories. Finally, the distance between step thresholds are 0.16 logit (between score categories 1 and 2) and 1.46 logits (between score categories 2 and 3). In Figure 10.2, the distinct and wide “hill” in which score category 2 is most probable reflects the relatively large distance between the step thresholds of score categories 2 and 3 (i.e., 1.46 logit).
Table 10.3 Statistics for the Rating Scale Assessing Responses for Section #7 (Word Order)

<table>
<thead>
<tr>
<th>Score</th>
<th>Counts used</th>
<th>%</th>
<th>Avge. meas.</th>
<th>Exp. meas.</th>
<th>Outfit MnSq</th>
<th>Meas.</th>
<th>S.E.</th>
<th>Category</th>
<th>Most probable from</th>
<th>Category peak probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>79</td>
<td>12%</td>
<td>-.52</td>
<td>-.53</td>
<td>.9</td>
<td>(-1.94)</td>
<td></td>
<td>low</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>88</td>
<td>13%</td>
<td>-.23</td>
<td>.01</td>
<td>.6</td>
<td>-.38</td>
<td>.14</td>
<td>-.64</td>
<td>1.33</td>
<td>30%</td>
</tr>
<tr>
<td>2</td>
<td>215</td>
<td>32%</td>
<td>.74</td>
<td>.73</td>
<td>.9</td>
<td>-.54</td>
<td>.11</td>
<td>.50</td>
<td>-.10</td>
<td>-.46</td>
</tr>
<tr>
<td>3</td>
<td>322</td>
<td>44%</td>
<td>1.96</td>
<td>1.90</td>
<td>.9</td>
<td>.92</td>
<td>.10</td>
<td>(2.17)</td>
<td>1.37</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>(Mean)</td>
<td></td>
<td>(Mean)</td>
<td>(Modal)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Regarding the functioning of the rating scale used for assessing constructed responses in Sections #8 and #9, Table 10.4 shows that each score category is used at least 10 times; however, the frequency distribution of score usage shows a somewhat irregular pattern: score categories 0 and 3 are used more frequently than score categories 1 and 2. The observed average measures of each score category increase steadily (see aveg. meas. column) and the difference between observed and expected measures is small (i.e., ranging from 0.05 to 0.27 logit). In examining the step thresholds, there are clear instances of step disordering: the step thresholds for score categories 1, 2, and 3 actually steadily decrease (i.e., at .30, .29, and -.59 logit, respectively). This finding, in conjunction with the small distances between step thresholds (i.e., 0.01 logit between score categories 1
and 2, and 0.88 logit between score categories 2 and 3), indicates that score categories 1 and 2 are likely to be poorly or narrowly defined. The step disordering is further illustrated in Figure 10.3, in which the probability curves for score categories 1 and 2 never form distinct “hills.”

Figure 10.3 Rating Scale Structure for Sections #8 and 9 (Constructed Responses)
Table 10.4 Statistics for the Rating Scale Assessing Responses for Sections #8 and #9 (Constructed Responses)

<table>
<thead>
<tr>
<th>Score</th>
<th>Counts used</th>
<th>%</th>
<th>Avge. meas.</th>
<th>Exp. meas.</th>
<th>Outfit MnSq</th>
<th>Meas.</th>
<th>S.E.</th>
<th>Category</th>
<th>Most probable from</th>
<th>Category peak probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>95</td>
<td>12%</td>
<td>-.16</td>
<td>-.23</td>
<td>1.4</td>
<td>(-1.35)</td>
<td></td>
<td>low</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>9%</td>
<td>.11</td>
<td>.21</td>
<td>1.1</td>
<td>.30</td>
<td>.13</td>
<td>-.34</td>
<td>-.86</td>
<td>24%</td>
</tr>
<tr>
<td>2</td>
<td>84</td>
<td>11%</td>
<td>1.06</td>
<td>.79</td>
<td>1.9</td>
<td>.29</td>
<td>.12</td>
<td>.37</td>
<td>.02</td>
<td>19%</td>
</tr>
<tr>
<td>3</td>
<td>517</td>
<td>68%</td>
<td>1.76</td>
<td>1.81</td>
<td>1.1</td>
<td>-.59</td>
<td>.10</td>
<td>(1.33)</td>
<td>.87</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Mean)</td>
<td></td>
<td></td>
<td></td>
<td>(Modal)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

**RQ1. Test Difficulty, Discrimination, Reliability, and Rasch Model Fit**

RQ1 examines the measurement quality of the 54-item test in terms of difficulty, discrimination, reliability, and Rasch Model fit. In terms of difficulty, although there is a relatively wide range for the binary-scored and rated items, the mean test score of 74.63% suggests that the test overall is relatively easy for the examinee group. This observation is corroborated by the distribution of items along the logit scale in Figure 10.1: while the items span from -2.0 to 2.0 logits, they generally correspond to the lower-ability examinees. Because the test was developed as a summative assessment tool for evaluating the linguistic knowledge covered in the course, the examinees can be expected to do well on the test, an observation also made by Li (2019) for a similar assessment situation. Hence, the difficulty level of the test is presumably appropriate for the course. Admittedly, the desirable level of difficulty for a classroom test like this one is likely open to discussions and varies across different Chinese programs. It would be an interesting project if data can be collected from other comparable institutions in the U.S. to examine the national trend in setting the difficulty benchmark for summative tests used in CSL classrooms.

There are two additional observations regarding test difficulty across sections. The first is that the binary-scored part (Sections #1 to #6) and the rated part (Sections #7 to #9) are comparable in difficulty. Admittedly, except for a few items in Sections #8 and #9, the rated part of the test does not require examinees to write characters from memory (which may have reduced the difficulty of the items), although it does require their ability to recognize characters as in the binary-scored part of the test. This finding may inform the ongoing debate regarding whether to delay the introduction of characters in a Chinese curriculum (e.g., Knell & West, 2017; Krashen, 2017; Ye, 2013). Because the Chinese program takes a no-delay approach, the implication is that, after one year of classroom instruction, students can perform satisfactorily (based on the mean test score) on a test that assess character recognition ability, and that character writing by itself does not necessarily contribute to test difficulty.

The second observation is that question format may lead to different
levels of difficulty. For example, although both Sections #5 and #6 target vocabulary knowledge, Section #6 (multiple choice with a word bank) appears to be more difficult than Section #5 (regular multiple choice with four options) (see Tables 10.1 and 10.2). Section #6 is more difficult than Section #5 for two possible reasons: (1) examinees have a larger number of options to choose from, and (2) although there can be more than one correct answer for certain blanks (e.g., both E and I can fit the first blank of the second item in Section #6), examinees need to make sure that all other blanks are filled with appropriate words. In other words, examinees need to simultaneously attend to multiple items and potentially appropriate options in completing Section #6, which likely adds to the difficulty of this section.

Concerning test and item discrimination (i.e., to what extent a test or an item can differentiate higher ability examinees from their lower-ability counterparts), the point biserial correlation coefficients in Table 10.2 show a relatively wide range between -0.01 and 0.74 for individual items with an average of 0.40 for the entire test. Fulcher and Davidson (2007, p. 104) suggests a conservative cutoff point of 0.25 for evaluating test discrimination. With this criterion, although the entire test appears to have a sound level of discrimination, seven out of the 54 items (12.96%) exhibit relatively low discriminatory power. The negative and very low point biserial correlation coefficients (i.e., less than 0.10) for items Lis.1.2 and Lis.1.3 are particularly problematic these statistics indicate that examinees’ performance on these two items are barely related to their overall performance on the test. In hindsight, these two items do appear to be somewhat convoluted in terms of content; hence, revisions would be necessary for future administrations of the test.

In terms of test reliability, the Rasch person (examinee) reliability coefficient (similar to Cronbach’s α) is 0.88. For a classroom test developed by course instructors, this number indicates a fairly good level of test reliability (Kline, 2000, p. 13)

Finally, in terms of Rasch model fit, all 54 items (except for two) show acceptable infit MnSq statistics (i.e., between 0.5 and 1.5); meanwhile, the infit MnSq statistics of all 88 examinees (except for two) fall into the 0.5–1.5 range. The misfit ratios for items and persons (examinees) are 3.70% and
2.27%, which are below the 5.00% threshold (Boone et al., 2014). The relatively low item misfit ratio indicates that the test exhibits a satisfactory level of model fit; likewise, the relatively low person misfit ratio suggests that the test response behavior of the majority of the examinees conform to the expectations of the Rasch model. Hence, it is justified to add up the scores of individual test items and use the overall test scores as an indicator of the examinees’ linguistic knowledge.

In summary, the statistics regarding test difficulty, discrimination, reliability, and Rasch model fit demonstrate that the 54-item test exhibit overall satisfactory measurement quality for a summative assessment of the mastery of linguistic knowledge taught in an elementary Chinese language course. The seven items with relatively low discriminatory power (i.e., below 0.25) and the two misfit items (i.e., infit MnSq value larger than 1.5) should be revised for future applications of the test to comparable examinees.

**RQ2. Rating Scale Functioning**

RQ2 investigates the functioning of the rating scales developed for evaluating responses to Sections #7, #8, and #9. The results clearly show that both rating scales need to be revised to achieve better measurement quality. A shared issue of both rating scales is step disordering (see the Criteria for Evaluating Rating Scale Quality section). According to Linacre (2002), step disordering occurs when a rating category is poorly defined or too narrowly defined in terms of the underlying construct. In other words, the problematic score category cannot be distinctively separated from its neighboring score categories, which suggests an overlap in the underlying ability being assessed.

Specifically concerning the rating scale for Section #7 (word order), Figure 10.2 and the distance (see the results section under RQ2) between step thresholds both suggest that score category 1 is probably too narrowly defined for its substantive meaning. Because Section #7 expects examinees to construct sentences with intended structure and meaning, in hindsight it can be argued that a score of zero and a score of one, as defined in the current version of the rating scale, may not be substantively different from
each other because neither indicates the ability to form complete sentences in terms of structure and meaning. Therefore, it is advisable to revise the rating scale by combining these two score categories in future scoring of examinee responses.

The rating scale developed for Sections #8 and #9 also show step disordering. Specifically, score categories 1 and 2 both appear to be too narrowly defined based on Figure 10.3 and the distance (see the results section under RQ2) between step thresholds. Essentially, Sections #8 and #9 assess the ability to comprehend individual questions and short paragraphs by asking examinees to construct short responses. In other words, reading comprehension should be the main focus of evaluation, as long as the constructed responses can clearly express the intended meaning. With such an understanding, minor grammatical errors that do not interfere with meaning expression should not be factored into the rating scale for assessing reading comprehension. In hindsight, therefore, score category 2 may not be necessary because it was created to capture the nuances between a response that is grammatically perfect and a response that contains minor grammatical error(s). Combining score categories 2 and 3 can be an option to enhance proper rating scale functioning. On the other hand, although the statistics indicate that score category 1 is also narrowly defined, conceptually it does represent a level of reading comprehension ability that is distinct from score category 0 (no comprehension at all) and score categories 2 and 3 (complete reading comprehension). Therefore, it would be advisable to revise the current rating scale to include three levels (i.e., 0, 1, and 2).

In reflecting upon the observed instances of step disordering for both rating scales, the crux of the issue seems to be an over-emphasis on the descriptive function of the score categories and an unintended sacrifice of the interpretive value of the score categories. As described in the method section, the development of the rating scales followed a bottom-up approach, that is, the researchers read through all responses and identified benchmark responses before setting rating score categories. This process tends to capture as much nuances among the responses as possible. For example, score categories 2 and 3 for Sections #8 and #9 differ only in terms of grammaticality of responses. Although this distinction is useful in describing
the differences in responses to these two sections, it may not be very helpful for capturing the differences in the examinees’ reading comprehension ability, which is the focal underlying ability of these two sections. Hence, the implication for developing rating scales for classroom assessment is to focus on the interpretive value of the rating scale based on the focal construct. Pedagogically, these findings highlight the importance of setting clear objectives for assessment in the process of developing scoring rubrics (e.g., a rubric for assessing reading comprehension should primarily focus on evaluating reading comprehension), rather than attempting to attend to all details in examinee responses.

**Conclusion and Future Research**

This study examines the measurement quality of a summative assessment instrument used in a CSL classroom. The results regarding overall test difficulty, discrimination, and reliability appear to be satisfactory for the intended purpose of the test. Meanwhile, several items showing below-threshold quality in terms of item discrimination and Rasch Model fit need to be revised or replaced for future use of the test. This study is the first empirical effort to investigate the functioning of rating scales developed for CSL classroom assessment. Somewhat surprisingly, as discussed earlier, the instructor-developed ratings scales did not function fully as intended according to the criteria set in this study.

Although the three rated sections (with 17 items) examined in this study constitute a relatively small part of the test (with 54 items), and although the test by itself only accounts for 16% of the students’ final course grade, the issues raised here are likely to have wider implications for CSL classroom teaching. Given the ubiquity of instructor-developed rating scales in CSL classrooms, and because the scores based on such rating scales contribute to students’ course grades (and their GPAs), it is important to ensure the measurement quality of such scales. Therefore, CSL instructors should not only focus on developing their rubrics, but also be aware of the potential limitations of their self-developed instruments for assessment. Ideally, CSL instructors and researchers can work together to check the proper functioning of such instruments, which contributes to the quality of CSL
classroom assessment. Although the content of this study may be technical and although the findings were based on a particular Chinese program, the issues discussed here may be generalizable to similar assessment contexts in other Chinese programs.

There are two directions for future research. First, it would be meaningful to investigate the quality of other types assessment instruments commonly used in CSL classrooms. This study focused on one summative assessment instrument, but there are also formative assessment tools based on which instructors evaluate students’ progress in order to inform subsequent teaching. The quality of formative assessment would be an interesting topic to pursue in future research. Second, this study focuses on the measurement characteristics of a CSL classroom test. The results reported here constitute partial (albeit important) evidence of a validity argument for the intended use of the test; hence, this study does not represent a comprehensive effort on test validation. For the purpose of enhancing assessment quality and accountability of CSL teaching, it would be desirable to design and implement a full-fledged validation research project.

References
assessment. TESOL Quarterly, Special Issue 43(3), 393–415.


Appendix 1

The Final Written Test

1. Listen to a dialogue and answer the following T/F questions. The dialogue will be read twice for you.

男：哎，你看，我今天新买了一件衣服。你觉得怎么样？
女：大小很合适。。。你在哪儿买的？
男：就在学校的书店买的，不贵，就200块钱。
女：200块钱！这还不贵吗？我觉得颜色也不好看。我知道你喜欢咖啡色，为什么买了这个颜色的？
男：我想买咖啡色的，可是不是太大，就是太小。
女：那你为什么买这件呢？
男：因为，嗯，那个售货员。。。
女：那个售货员很漂亮，是吗？
男：哪里！她不认识我，不过我知道你和她是朋友。
女：是吗？我怎么不知道我的朋友在学校的书店工作呢？
男：嗯，那你明天自己去看看吧。

(T) The woman was a little surprised that the man bought the coat.
(F) The man bought the coat because the price was appropriate.
(T) We can assume that many of the man’s clothes were brown.
(F) The salesperson turned out to be the man’s friend.
2. Listen to a monologue and complete the following T/F questions. The monologue will be read twice for you.

王朋，你好！我是白英爱。我今天早上才知道你今天晚上要去机场。你告诉李友了，可是你怎么没有告诉我呢？你别坐公共汽车去机场。虽然坐公共汽车去机场很便宜，可是也很麻烦：你得先坐公共汽车坐十二站，然后换地铁，一共得花两个小时。我今天晚上有空儿。我可以开车送你去机场。你回宿舍以后给我打个电话，好吗？对了，你的飞机是七点的还是八点的？我们晚上见！

(T) Li You learned about Wang Peng’s travel plan earlier than Bai Ying’ai.
(F) Bai Ying’ai will call Wang Peng again this evening after returning to her dorm.
(F) According to Bai Ying’ai, the public transportation to the airport is not only convenient but also inexpensive.
(F) Bai Ying’ai already knows the departure time of Wang Peng’s flight.

3. Provide the meaning for each radical.

线 Radical Meaning ___________________
懂 Radical Meaning ___________________
澡 Radical Meaning ___________________
衬 Radical Meaning ___________________
钱 Radical Meaning ___________________
换 Radical Meaning ___________________
城 Radical Meaning ___________________
起 Radical Meaning ___________________

4. Radical knowledge application.

Circle the character that means “to throw” 殆 投 落
Circle the character that means “copper” 铜 镜 洞
Circle the character that means “(book) shelf” 驾 茄 架
Circle the character that means “tomb” 坟 纹 蚊
Circle the character that means “to worry” 虎 虑 虚
Circle the character that means “to stare at” 汀 订 盯
5. Multiple choice.

(1) A: 今天晚上你打车回家________开车回家？
   B: 我没有钱，所以我坐地铁回家。
   A. 不是  B. 还是  C. 也是  D. 或者

(2) A: 你怎么去机场？
   B: 我先坐公共汽车到 five points, ________ 坐地铁去机场。
   A. 最后  B. 后来  C. 以后  D. 然后

(3) A: 请问你想买什么？
   B: 我想买一________鞋子和一________裤子。
   A. 条、条  B. 双、件  C. 件、条  D. 双、条

(4) 昨天我去了一________音乐会。回家以后我写了一________日记。
   A. 个、个  B. 个、篇  C. 篇、篇  D. 个、张

(5) Which is the correct way to say RMB 14.95?
   A. 十四块九十五分  B. 十四块五毛九分  C. 四十块九毛五分  D. 十四块九毛五分

(6) 昨天早上我吃完饭________去学校了。
   A. 就  B. 没  C. 在  D. 不

(7) A: 王朋，我不________开车，你________教我吗？
   B: 没问题！你什么时候有空儿？
   A. 能、能  B. 会、能  C. 可以、能  D. 能、会

(8) Which of the following is a grammatically correct sentence?
   A. 今天晚上我得给我的老师打一个电话。
   B. 我得给我的老师打一个电话今天晚上。
   C. 今天晚上我得给一个电话打我的老师。
   D. 我得给打我的老师一个电话今天晚上。
6. Choose the appropriate words from the word bank to fill in the blanks.
A. 别  B. 容易  C. 就  D. 得  E. 已经  
F. 一共  G. 在  H. 特别  I. 最近  J. 先
(1) 今天的中文考试很____B____，所以我考得很不错。
(2) A: 我____E____两个月没有看到你了，你____I____怎么样？
     B: 我____H____忙，有很多功课。我觉得很累。
(3) 先生，你买了两张飞机票和一张火车票，____F____是 2300 块。
(4) A: 文中，我们现在一起回家吧。
     B: 我还得去见常老师。你____J____回家，我见了常老师以后____C____回家。

7. Re-arrange the following words to form grammatical sentences.
(1) 发短信你我打或者的手机给有空儿的话
   ________________________________________________。
(2) 复习中文我晚上都每天
   ________________________________________________。
(3) 可是太贵虽然这书有意思了本
   ________________________________________________。
(4) 一样这和那件件衣服好看衣服
   ________________________________________________。
(5) 我就合适如果的话买长短
   ________________________________________________。
(6) 得学以外中文专业课还我除了我很忙，
   ________________________________________________。
(7) 今天我要去看书下午到图书馆
   ________________________________________________。
(8) 的吗你们课文了都预习今天
   ________________________________________________。

8. Answer the following questions based on your own circumstances. You should use full sentences written in Chinese characters to answer these questions.
(1) 你常常去你的中文老师的办公室问问题吗？
你觉得你写汉字写得怎么样？
你每个星期有几节中文课？
你喜欢什么颜色，红色、黄色、还是咖啡色？
你喜欢一边吃饭一边看电视吗？

9. 读 comprehension. Answer the following questions in Chinese.
中国的大小孩子们都特别喜欢中国新年。因为除了可以穿新衣服和新鞋子以外，他们的爸爸妈妈还会给他们钱。可是李红不喜欢中国的新年。李红今年二十五岁了，有先生，可是没有孩子，所以别人不给李红钱，她还得给别的小孩很多钱。李红还说新年的时候公共汽车和地铁都特别少。因为她和她先生都没有车，所以出去玩儿很不方便。她觉得中国新年很没有意思。
(1) 李红是 男的 (male) 还是 女的 (female)？(2) 你怎么知道？
(3) 李红为什么不喜欢中国新年？(List two reasons)
(4) 中国的大小孩子为什么喜欢中国新年？