Measurement Quality of an Instructor-Developed CFL Summative Assessment Instrument: An Exploratory Pilot Study

Shuai Li
Georgia State University, sli12@gsu.edu

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, "Measurement Quality of an Instructor-Developed CFL Summative Assessment Instrument: An Exploratory Pilot Study" (2018). World Languages and Cultures Faculty Publications. 68.
https://scholarworks.gsu.edu/mcl_facpub/68

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 of an Instructor-developed CFL Summative Assessment

Instrument: An Exploratory Pilot Study

Shuai Li

Georgia State University

Cite as:
QUALITY OF CFL SUMMATIVE ASSESSMENT

Abstract
This study examined the measurement quality of an instructor-developed classroom assessment instrument (i.e., a final written test) used to assess linguistic knowledge covered in an undergraduate elementary Chinese course at a U.S. university. Participants were 222 learners enrolled in the Chinese course from Fall 2011 to Fall 2013. Analyses were performed on a subset of 64 binary-scored (0/1) test items. The 64 items showed acceptable overall test reliability, test discrimination, and Rasch model fit. Meanwhile, insufficient number of difficult items, below-threshold discriminatory power of certain items, and existence of measurement redundancy were also found. Strategies for improving the measurement quality of the test were discussed.

Keywords: Measurement quality, Classroom assessment, Chinese, linguistic knowledge

摘要
本文考察了一套用于美国一所大学的本科初级汉语课程的总结性评价工具（即期末考试试卷）的测量质量。这是一套授课教师自行编制的、用来测试学生掌握初级汉语课程所教授的语言知识程度的试卷。被试为 222 名在 2011 年秋季学期至 2013 年秋季学期期间注册初级汉语课程的学生。对该试卷所包含的 64 道 0/1 计分的试题的测试数据进行分析，结果表明，64 道试题总体信度、区分度及 Rasch 模型拟合度达到了可接受的水平。同时，研究也揭示了该试卷存在的问题，即难度较高的试题数量不足，部分试题区分度未达到临界水平，以及存在测量冗余的现象。本文最后讨论了提高该试卷测量质量的方法。

关键词：测量质量、课堂测试、总结性评价、中文、语言知识
QUALITY OF CFL SUMMATIVE ASSESSMENT

Introduction

Since the beginning of the 21st century, the field of tertiary level CFL (Chinese as a foreign language) education in the U.S. has witnessed rapid development (Wang & Ruan, 2016), as reflected by the increasing enrollment in college Chinese classes (Goldberg, Looney, & Lusin, 2015). This upward trend has fueled a growing interest in CFL curriculum development (e.g., Chen, 2012; Everson, 2011; Li, Wang, & Wang, 2013; Li & Zhang, 2016; Wen, 2015). The field’s passion for curriculum development, however, contrasts sharply with a lack of research on assessment at both program/curriculum and course levels, which led Ke (2012, p. 87) to conclude that “empirical research in CFL assessment has largely been ignored”. Because assessment is an integral component of language curriculum development (Brown, 1995), it merits attention from CFL educators. Classroom assessments (e.g., instructor-developed tests and quizzes, portfolios, interviews) may particularly require research effort due to their ubiquitous use in virtually all Chinese classes. This study contributes to CFL assessment research by exploring the measurement quality of an instructor-developed summative assessment instrument (i.e., a final written test) adopted in an undergraduate elementary CFL course in the U.S.

Classroom-Based Language Assessment

Language assessment denotes a process of gathering test and non-test data to infer an individual’s language-related characteristics (Purpura, 2016). The notion incorporates large-scale standardized tests (e.g., Hanyu Shuiping Kaoshi) developed by language assessment experts and classroom assessments created by instructors (Cheng, 2013; Cockey, 2014; Purpura, 2016). Classroom-based language assessment may take various forms, such as presentations, compositions, portfolios, interviews, self-assessment, peer assessment, and paper-and-pencil tests. Classroom-based
QUALITY OF CFL SUMMATIVE ASSESSMENT

language assessment can be formative or summative (Cheng, 2013; Cockey, 2014), reflecting the distinction of assessment for learning and assessment of learning (Rea-Dickins, 2008). Formative assessment is administered during instruction; the results are interpreted as students’ learning progress and are used as feedback to students and instructors to guide subsequent learning and teaching. Summative assessment is typically administered after instruction to evaluate learning outcomes. The results of summative assessment are often used for making decisions on course grades, placement, and program completion, among others.

A Case for Quality Check of CFL Classroom-Based Assessment

Regardless of assessment type (formative or summative), classroom-based language assessment should be developed based on the curriculum in specific teaching contexts (Cockey, 2014). This entails articulating instructional goals and learning objectives at various levels, and developing assessment instruments accordingly. Bachman and Palmer (2010, pp. 477-485), for example, described the rationale behind planning and designing a speaking test (e.g., the needs and consequences of the test) for an elementary CFL course offered at a U.S. university. The authors, however, did not show the actual test, nor did they provide validity evidence to support the use of the test. In another study, Ke (2006) proposed a model of formative assessment for the Chinese program at The University of Iowa. The Chinese program’s curriculum goals were developed according to mainstream proficiency standards and frameworks (e.g., The Common European Framework). Based on the articulated curriculum goals, Ke’s model adopted a task-based approach and was characterized by criterion-referenced testing, skill integrated assessment tasks, componential scoring, and simulation of classroom activities. Ke, however, did not provide data to demonstrate the functioning of his model; he instead called for research to examine the extent to which the model
QUALITY OF CFL SUMMATIVE ASSESSMENT

helps achieve the curriculum goals of the Chinese program.

Both aforementioned studies show how theoretically informed and practically motivated program-wide CFL assessment can be conceptualized and developed. However, many Chinese programs in North America may not be ready for the construction of such program-level assessments due to a lack of well-articulated curriculum goals: A recent survey of tertiary level CFL programs in North America (Li, Wen, & Xie, 2014) showed that only 55.1% and 45.5% of the respondent programs articulated curriculum-level and program/department-level learning outcomes, respectively; on the other hand, 62.9% of the programs developed grade-level learning objectives.

Although developing program-level assessment may not be an immediate option for many Chinese programs, classroom assessments, particularly those developed by instructors, are widely and regularly used in all Chinese classes. Because numerous pedagogical decisions (e.g., grade assignment) are dependent on the results of those instructor-developed assessment instruments, the quality of such instruments is of critical importance to all parties involved. However, little empirical research is available on the quality of instructor-developed CFL classroom assessment (Ke, 2012), which has led to calls for studies on the effectiveness of assessment methods based on classroom-generated data (Xiang, 2016).

This study thus explores the measurement quality of an instructor-created summative assessment tool. Following Bachman (2004), indicators of measurement quality include reliability, difficulty, discrimination, and IRT-based (item response theory) measures. Although different IRT models exist, this author adopted the Rasch Model because it enables the construction of a “measurement scale with interval scale properties” (Bond & Fox, 2007, p. 265). Aspiring for an interval scale (which allows
addition and subtraction) is particularly important in the context of evaluating classroom-based summative assessments, because the default method of score reporting in CFL classes (i.e., providing a summed test score) relies on the construction of an interval scale in the first place. The Rasch Model is introduced below.

The Rasch Model for Quality Check

The Rasch Model is a probability-based psychometric model for measuring latent psychological traits. A latent psychological trait (e.g., language ability) is an personal attribute inferable from observations of one’s behaviors. In applied linguistics, the Rasch Model has been widely applied to performance assessment (e.g., Eckes, 2015; Liu, 2006; McNamara, 1996) for examining the measurement quality of assessment instruments, rater behaviors, and examinee responses.

The Rasch Model has two fundamental principles (Rasch, 1960, p. 117, cited in Bond & Fox, 2007, p. 10). First, a more difficult item should have a lower probability of being answered correctly than an easier item by any given person; second, a higher achiever should have a higher probability of solving any item than a lower achiever. Essentially, the Model postulates that, in an idealized situation (e.g., no time pressure), the probability of a given person who correctly answers a given test item is contingent upon the difference between the person’s ability and the item’s difficulty. For example, when a person’s ability level is on par with an item’s difficulty level, he/she should have a 50% chance to correctly answer the item; the chance should be less than 50% if a person’s ability level is lower than an item’s difficulty level, and vice versa.

To enable direct comparisons between personal abilities and item difficulties, the Rasch Model transforms raw test scores, which typically belong to the category of
QUALITY OF CFL SUMMATIVE ASSESSMENT

ordinal scale, to log odds ratios on an interval scale called logit scale. The logit scale has a zero point and extends to positive and negative infinity. The measurement unit of the logit scale is called logit. A larger logit value denotes a higher level of personal ability (called person measure) and item difficulty (called item measure), and a smaller logit value denotes a lower level of personal ability and item difficulty. The value of zero (0) logit is conventionally set as the mean of the item difficulty measure.

To assist visual comparison between personal ability and item difficulty, the Rasch Model generates an item-person map (Fig.1, detailed in the Results section).

As a means of quality check, the Rasch Model requires individual item functioning and a person’s item response behaviors to conform to the two aforementioned principles before constructing the logit scale. In other words, a prerequisite of interpreting the results of Rasch analysis is that a given data set has to fit the Model’s a priori requirements. The extent to which a data set conforms to model expectation is expressed through fit statistics, which the Rasch Model calculates for each item and for each person. Applied linguists typically review the infit MNSQ (Mean Square) statistics (which are not sensitive to outliers) to evaluate model fit. A value of one (1) indicates expected model fit, and the acceptable range of infit MNSQ values is 0.5 to 1.5 (Boone, Staver, & Yale, 2014, p.166; Wright & Linacre, 1994). An infit MNSQ value larger than 1.5 means excessive amount of randomness than the Model predicts (i.e., the individual’s test behavior and/or item functioning are too unpredictable); an infit MNSQ value below 0.5 indicates far less randomness than the Model expects (i.e., the individual’s test behavior and/or item functioning are too predictable).

There is a practical benefit of aspiring for model fit in the context of classroom-based language assessment. Typically, instructors develop a test and use the total test score as the basis for decision-making (e.g., a student may receive an “A” because his
total test score is above 90%). Procedures such as these are too common to prompt us to reflect upon the extent to which summing up the scores of individual test items is justified in the first place. To illustrate, when certain items are more likely to be answered correctly by lower achievers than by higher achievers, such items should be scrutinized before being included in calculating total test scores. Likewise, when a higher achiever answers easy items incorrectly, his/her performance on these items would call for examination before those responses can be used to infer his/her ability. By requiring items and examinees to conform to the model’s expectation, Rasch analysis allows identification of potentially problematic items and persons in specific assessment contexts. As such, it paves the way for constructing an interval scale of measurement, which helps justify the practice of calculating total test scores based on individual items.

The Present Study

Although instructor-developed classroom-based language assessment can take various forms, this study focused on a summative assessment (i.e., a final written test) because of its relatively higher stakes in its instructional context: The test carries the largest weight (16%) of a student’s overall course grade. In addition, based on this author’s experience and observation of CFL teaching practices at multiple U.S. universities, written tests assessing linguistic knowledge are a widely adopted format of assessment. The research question was:

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

Method

The Research Context
The research site is an undergraduate Chinese program within a foreign languages department at a four-year public university in the U.S. In many ways, this program demonstrates the characteristics of young and developing undergraduate CFL programs in U.S. colleges. To begin with, while it has yet to offer a B.A. degree in Chinese, the program collaborates with several academic departments for interdisciplinary B.A. programs and a Chinese Studies Minor. Consequently, the curriculum of the program has undergone refinement. Concerning course offerings, the Chinese program regularly offers first-year to third-year Chinese language courses and two content courses taught in English. Fourth-year Chinese courses have yet to be taught due to insufficient enrollment. Although the program typically attracts 120 to 150 individual registrations each semester, over 70% of those enrolled are in the first-year classes. Finally, in terms of course staffing, there are two full-time faculty members with a few need-based part-time instructors.

This study focuses on a selection of items of a final written test for the CHN 101 course (the lowest level of Chinese instruction, 3 credits) of the Chinese program. While there is no specific proficiency goal set for this course due to the program’s developing nature, the overall learning objectives are four-fold: Linguistic knowledge (e.g., Pinyin system, radicals, characters, sentence patterns), communicative functions (e.g., discussing family), cultural understanding (e.g., appropriate greetings and responses), and independent learning skills (e.g., using online resources). Among these objectives, the instructors consider linguistic knowledge and communicative functions as the most important. The scope of the four learning objectives is primarily based on the textbook, *Integrated Chinese* (Level 1, Part 1, 3rd edition) and supplemented by instructor-developed materials. The CHN 101 course covers the introductory chapter (Pinyin and characters/radicals) and the first five chapters.
Both formative and summative assessments are used in the course, including chapter tests and quizzes (formative and summative), oral and written homework assignments (formative), tutoring sessions (formative), presentations (formative), and final oral and written exams (summative). These assessment methods primarily evaluate students’ mastery of linguistic knowledge and communicative functions. For example, the curriculum sets up its final oral test in the form of a skit performance to focus mainly on communicative functions and, to a lesser extent, linguistic knowledge. In contrast, its final written test (detailed below) mostly focuses on assessing linguistic knowledge and its applications. The course’s final exam (including oral and written tests) accounts for 20% of one’s final course grade. The final exam’s oral portion constitutes 20% of and the written test 80% of the final exam’s score. As such, the final oral test accounts for 4% (20% × 20%) of one’s overall course grade, whereas the final written test accounts for 16% (20% × 80%).

Test Development and Structure

Assessing the linguistic knowledge covered in the CHN 101 course is the primary function of the final written test. As determined by the course syllabus and the textbook, such linguistic knowledge covers the Pinyin system (i.e., initials, finals, and tones), radicals, characters and words, sentence structures, and conventionalized discourse patterns (e.g., question-response pairs in typical communicative situations).

In developing the test, the instructor relied mostly on his teaching experience. For example, phonemic pairs such as “uo – ou” and “x – sh” that were found to be difficult for students to differentiate were included. Items like these also reflected classroom activities, during which students were asked to differentiate pairs of initials, finals and tones.

The administered final written test observed for this study contained 11 sections.
Pinyin knowledge was assessed in Section #1 (five items). In responding to each item, students were asked to choose the correct Pinyin syllable(s) (out of two choices) that they hear.

Radical knowledge was targeted in Sections #4 and #5. In Section #4, students were tasked with first identifying the radical of each character and then providing the (English) meaning of each radical. In Section #5, students needed to use their radical knowledge to infer the meaning of an unknown character. For example, students were asked which of the following three characters mean “river” (i.e., 柯, 舶, 河) by utilizing their knowledge of the water radical (氵).

In assessing vocabulary and grammar knowledge, Section #6 (eight items) and Section #7 (seven items) represented two types of multiple-choice questions. While each item in Section #6 asked students to choose a correct answer out of four options, the items in Section #7 had a shared word bank from which students needed to choose one that best fitted each item. Additionally, Section #10 (eight items) assessed word order: Students needed to string words together to form grammatically correct sentences and then provide English translations of these sentences.

To assess the knowledge of discourse patterns (e.g., question–answer pairs), Section #2 (eight items) was based on both aural and written prompts, whereas Section #8 (five items) was based on written prompts only. In Section #2, students would first hear an utterance and then choose the best reply out of three options; Section #8, on the other hand, required students to match questions to appropriate responses (both in written format). These two types of questions were included because they resembled what the students did both in the classroom and on their homework assignments.

The aforementioned sections all targeted specific domains of linguistic knowledge.
QUALITY OF CFL SUMMATIVE ASSESSMENT

The instructor also felt a need to assess the ability to apply linguistic knowledge (e.g., vocabulary, grammar, and discourse knowledge) to meaning comprehension and production. Hence, Section #3 (three items) was a listening comprehension section in which students would listen to a short dialogue (176 characters) and make true/false judgments. Section #9 (five items) was a reading comprehension section with a short paragraph (221 characters) and five multiple choice questions. Finally, Section #11 (five questions) assessed writing skills by asking students to provide constructed responses to prompt questions in Chinese characters.

Appendix One shows the nine binary-scored sections of the test; Sections #10 and #11 were not included in the Appendix because they were not binary-scored and therefore were not analyzed here.

Examinees

A total of 264 students enrolled in the CHN 101 course from Fall 2011 to Fall 2013. After excluding those who withdrew from the course or did not show up for the final exam, the remaining 222 students became the examinees of this study.

Data Analysis

A correct response was coded as 1 and an incorrect response 0. Each item in Section #4 was coded twice for identification of radicals and radical meanings, respectively. Each item in Section #4 was thus coded as two separate items for analysis. The total number of items analyzed was 64 and the score range of each examinee was 0–64. The data was analyzed under the Rasch Dichotomous Model with the software Winsteps (Version 3.80.1). Additional statistical analysis was conducted with SPSS Version 16.0.

Results

Table 1 shows difficulty, reliability, and discrimination of the 64 items. The
individual item means ranged between 0.45 and 1.00 with an average of 0.86 (out of 1.00). Item discrimination statistics (item-total correlation coefficients) spread from 0.06 to 0.50 with an average of 0.28. The overall reliability of the 64 items (Cronbach’s alpha) was 0.86. Table 2 shows the means and standard deviations of each section: The most difficult section was listening comprehension (Section #3) and the easiest section was discourse patterns with matching questions (Section #8). Individual examinees’ test scores ranged between 28 (out of 64, or 43.75%) and 64 (or 100%) with a mean of 55.04 (out of 64, or 86.01%).

Table 1 also shows the Rasch calibrated item statistics. Individual item difficulty measures spread over 7.49 logits (from -4.43 to 3.06 logit). The item separation index was 3.50 (or 5 strata) with a reliability coefficient of 0.92. This means that the 64 items can be reliably divided into five statistically significant difficulty levels. The item infit MNSQ statistics, except for the item Dis8.1, ranged from 0.83 to 1.18. The Rasch calibrated person measures spread over 6.75 logits (from -0.40 to 6.35 logit) with a mean of 2.83 logits (mean = 1.24). The person infit MNSQ statistics ranged from 0.63 to 1.82, with two examinees’ infit MNSQ statistics falling above the upper threshold of 1.5. The person separation index was 1.80 (or 2.73 strata) with a reliability coefficient of 0.76. This means that the examinees could be differentiated for at least two ability levels (i.e., high and low) with an acceptable level of reliability.

Fig. 1 is the item-person map. It compares examinee ability and item difficulty on the same logit scale. The left side of the map shows the distribution of examinee ability. Each pound sign (#) represents two examinees and each dot (.) one. Higher achievers occupy higher positions on the scale than lower achievers. The right side of
the map shows the difficulty level of the 64 items. More difficult items are positioned higher on the scale than less difficult ones. The two M letters on the axis in the middle of the map respectively denote the means of examinee ability and item difficulty. Letters S and T on each side of the axis indicate one and two standard deviations. As Fig. 1 shows, the 64 items were, in general, relatively easy for the examinees as a group, because the difficulty levels of the majority of the items were below most of the examinees’ ability level. This observation is corroborated by the fact that the mean person ability measure was 2.83 logits above the mean item difficulty measure (set to be at 0 logit).

Fig. 1, along with the corresponding item statistics in Table 1, offers additional insights into item function. First, different sections appeared to vary in terms of the homogeneity of item difficulty. To illustrate, the items of Section #8 (i.e., Dis8.1, Dis8.2, Dis8.3, Dis8.4 and Dis8.5) were clustered at the lower end of the logit scale, suggesting that these items are of a similar level of difficulty. In contrast, Section #6 and Section #7 (targeting vocabulary and grammar) showed a relatively wide range of difficulty level among items. For example, item Gra6.8 was the most difficult item (3.06 logits), but item Gra6.3 was far less difficult (-1.55 logits).

The second observation is that item format appeared to affect item difficulty. This is most obvious when comparing sections measuring the same linguistic knowledge. For instance, Sections #2 and #8 were both designed to assess discourse patterns, yet the multiple-choice items involving listening and reading in Section #2 were more difficult than the reading-based matching items in Section #8.

The third observation is that there were instances of measurement redundancy, especially among items within the same section. For example, items Rad4.1m and Rad4.6m (Section #4), which asked examinees to write out the English meanings of
certain radicals, had similar difficulty measures of 1.96 and 1.94 logits respectively. The most obvious case was Section #8, where four of the five items shared the same measure (-3.21).

**Discussion**

This study examined the measurement quality of the binary scored sections of the test (Sections #1 to #9) regarding reliability, difficulty, discrimination, and Rasch model fit. The Cronbach’s alpha of 0.86 and the Rasch person reliability coefficient of 0.76 indicate an acceptable to good level of reliability (Kline, 2000, p. 13). Because the test is not a standardized test, but rather a classroom assessment, the observed levels of reliability are appropriate, although a higher level of reliability would always be desired.

In interpreting the findings regarding the difficulty measure, one should note that the test was not intended to be a proficiency test with the purpose of differentiating levels of performance, because the CHN 101 course had yet to set its proficiency goals by the time of data collection. Rather, the test was created as a summative assessment for evaluating the linguistic knowledge covered in the Chinese course. The students could be expected to do well on the test because the targeted linguistic knowledge was already taught and practiced throughout the semester. Therefore, the mean test score of 86.01% (range: 43.75% – 100.00%) may be acceptable to this Chinese course. Admittedly, the appropriate difficulty level of a test is open to discussion, because different instructors and programs are likely to set different benchmarks in their respective assessment contexts.

While the 64 binary scored items generally leaned towards the easy end on the difficulty measure, individual item difficulty (Table 1) showed a wide range from 0.45 to 1.00. The distribution of item difficulty along the logit scale (Fig. 1) largely
shows equal (vertical) intervals among items, except between those towards the difficult and the easy ends (e.g., there is a relatively large distance between Gra6.8 and Lis3.3). This distributional pattern suggests that, although the test as a whole allowed for relatively precise measurement of the targeted linguistic knowledge, it could be further improved by developing items to fill the relatively wide “gaps” along the logit scale.

There was also noticeable variation in the difficulty of the individual sections. As Table 2 shows, sections assessing the ability to use multiple aspects of linguistic knowledge for meaning comprehension (e.g., Section #3 on listening comprehension) appeared to be more difficult than sections measuring specific aspects of linguistic knowledge (e.g., Section #8 on discourse patterns). Meanwhile, in measuring the same aspect of linguistic knowledge (e.g., knowledge of discourse patterns), sections involving multiple modalities and skills (e.g., Section#2 involved aural and written modalities as well as listening and reading skills) were more difficult than sections involving one single modality and skill (e.g., Section #8 involved written modality and reading skill only). These findings indicate that items assessing the integration of different linguistic knowledge and/or skills likely require a higher level of mastery of linguistic knowledge as assessed by the test.

With regard to test/item discrimination (i.e., how well a test/item can differentiate high and low achievers), individual item discrimination showed a relatively wide range (i.e., 0.06 – 0.50) with an average of 0.28. Adopting a somewhat stringent threshold value of 0.25 (Fulcher & Davidson, 2007, p. 104), twenty-two (or 34.38%) of the 64 items were below that threshold. When using a commonly accepted cut-off value of 0.20, 16 items (or 25%) failed to reach the threshold. Because many items were fairly easy for the majority of examinees, the ceiling effect might have
contributed to the low discrimination of some items. Items with excessively low level of discrimination (i.e., close to zero) can be dropped because they may measure a construct different from what the test is supposed to measure.

The last indicator of the measurement quality of the test is Rasch model fit. The infit MNSQ statistics of 63 out of the 64 items were within the range of 0.5 to 1.5. The only problematic item was Dis8.1, for which the Rasch Model did not output a fit statistic due to a ceiling effect (i.e., the item was answered correctly by all examinees). This item should be revised or removed to enhance measurement quality of the test. Overall, because the majority of the test items exhibited a satisfactory level of model fit, it is largely justifiable to use the sum score of the test as an indicator of masteries of linguistic knowledge as operationalized in this study.

With respect to model fit of individual examinees, results indicated that 220 (or 99.10%) of the 222 examinees’ infit MNSQ statistics were within the range of 0.5–1.5. The misfit rate was 0.90% (two out of 222), which is below the threshold of 5% (Boone, Staver, & Yale, 2014). These findings indicate that the item response patterns of the majority (220) of the examinees fit the expectation of the Rasch Model, suggesting that it is defensible to use the test scores as an indicator of their linguistic knowledge. For the two misfit examinees, one needs to review their unexpected item responses (e.g., a lower-ability examinee correctly answers a very difficult item) before the test scores can be interpreted as a measure of their linguistic knowledge.

The test can be further improved with regard to item discrimination and item difficulty. There are several options to cope with low item discrimination. Because excessively easy or difficult items tend to have relatively low discriminatory power, such items (e.g., item Pin1.2) can be revised or replaced. Another option is to examine whether an item was poorly constructed. One example is Pin1.5, an item designed to
assess whether or not students are able to differentiate the three initials (i.e., $j$, $q$, and $x$). Although the item possessed a satisfactory level of difficulty (i.e., mean = 0.76, $SD = 0.43$), its discrimination statistic was 0.16. A potential issue might be the fact that this item targeted two pairs of initials (i.e., $q$ vs. $x$, and $j$ vs. $q$), which might have confused high and low achievers alike (this hypothesis, however, should be verified through introspective verbal reports from examinees). A possible strategy to revise this item, therefore, is to make it focus on one pair of initials (e.g., $q$ vs. $x$).

To improve the difficulty distribution of the items, more difficult items may need to be developed if the course instructor is to aim for a higher benchmark of overall difficulty. Again, there are several options. First, because items requiring the integration of multiple aspects of linguistic knowledge for meaning interpretation (e.g., Sections #3 and #9) were found to be more challenging than those focusing on one single aspect of linguistic knowledge (e.g., Sections #1 and #8), one could increase the number of items in the former category. Second, items involving multiple skills (i.e., Section #2) were found to be more difficult than those based on one single skill (i.e., Section #8), thus one could develop items that require an integrated application of skills in assessing linguistic knowledge. Finally, additional item analysis can be done to see whether the different components of an item functioned adequately. For example, Gra6.3 (i.e., mean = 0.98, $SD = 0.15$) was the easiest item in Section #6 with below-threshold discriminatory power (i.e., 0.17). Because this was a multiple-choice item, its easiness might have been due to the fact that some of its distractors did not function as intended. A distractor analysis showing the percentage of incorrect responses for each of the three distractors would allow detection of problematic distractors (e.g., no examinee chose that option) and one can make revisions accordingly.
Finally, to address the issue of measurement redundancy (e.g., Rad4.4i and Rad4.5i), one could review item statistics of each test section and drop items of comparable difficulty index and discriminatory power. However, caution is needed here – even though some items pose risks of redundancy from a measurement perspective, there may be pedagogical reasons to justify their inclusion in the test. A good example to illustrate this point is Section #4, which asked students to identify the radicals of each character (e.g., Rad4.1i in Table 1). The entire section was found to be quite easy and several items had comparable discriminatory power. Before proceeding to remove them, a question to ask is whether it is pedagogically meaningful to ensure that students are actually able to identify certain radicals (e.g., 氵) that are covered in the CHN 101 course curriculum. If the answer is a “yes”, then one should still keep such items. In other words, because the test was designed to assess students’ masteries of instructional content, ensuring content coverage should be a more important consideration than reducing measurement redundancy. Although item statistics can inform researchers about the potential measurement issues of a test, whether and how to revise the test should be based on pedagogical considerations.

Limitations and Implications

This study explored the measurement quality (i.e., reliability, difficulty, discrimination, and Rasch model fit) of 64 binary scored items developed by a Chinese language instructor as part of a summative assessment instrument of an elementary CFL course. The results showed acceptable level of overall discrimination, reliability, and Rasch model fit; meanwhile, instances of measurement redundancy, unsatisfactory level of individual item difficulty and discrimination, and misfit of persons were also found.

There are several limitations of this study. First, as mentioned in the Method
QUALITY OF CFL SUMMATIVE ASSESSMENT

section, this study focused on the 64 binary scored items of the test, while leaving out the remaining 13 items (in Sections #10 and 11) that were scored based on rating scales. Hence, the results reported here constitute a partial revelation of the measurement quality of the summative assessment instrument. A project is currently underway to examine the functioning of the rating scales, raters, and items for the two sections. The second limitation is that an examination of the measurement quality does not constitute a full-fledged validation of the test in relation to its intended purposes. To this end, additional evidence will be needed, including, for example, the strength of correlation between the examinees’ test performance and their course grades, and the cognitive processes involved in responding to the different test items. Finally, test development and validation is an iterative process, involving several rounds of piloting and revision. Hence, this study served as a pilot project that provided empirical grounds for subsequent test revisions. It would be interesting to revise the test, re-administer it, and examine the measurement quality of the said test.

There are two ways that CFL educators can benefit from this study. Practically, for programs adopting the same textbook (*Integrated Chinese*, Level 1, Part 1, 3rd edition), the test items with acceptable measurement quality in terms of difficulty, discrimination, and *Rasch* model fit may be included in their respective assessment instruments. However, one should note that this test as general course assessment focuses on assessing linguistic knowledge and that it is not intended to assess proficiency. Moreover, in delineating the rationale and procedures for creating a classroom-based summative assessment instrument, and in examining the measurement quality of the instrument, this study hopes to demonstrate what CFL teacher-researchers can do to check and improve the quality of their assessment tools. Admittedly, effective assessment instruments are developed in specific contexts to
serve specific purposes, therefore the results of one study may not be readily transferrable to other assessment environments. However, by investigating the quality of one instructor-developed assessment tool in one specific CFL instructional context, we are beginning a process of accumulating empirical evidence that can reveal issues associated with instructor-developed assessment instruments.

References


QUALITY OF CFL SUMMATIVE ASSESSMENT 22


## QUALITY OF CFL SUMMATIVE ASSESSMENT

### Table 1

*Item Statistics*

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>Corrected item-total correlation</th>
<th>Rasch analysis</th>
<th>Measure</th>
<th>Model S.E.</th>
<th>Infit MNSQ</th>
<th>Infit ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pin1.1</td>
<td>.96</td>
<td>.20</td>
<td>.21</td>
<td>-.91</td>
<td>.35</td>
<td>.99</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Pin1.2</td>
<td>.97</td>
<td>.18</td>
<td>.06</td>
<td>-1.19</td>
<td>.39</td>
<td>1.05</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td>Pin1.3</td>
<td>.92</td>
<td>.27</td>
<td>.08</td>
<td>-.18</td>
<td>.27</td>
<td>1.13</td>
<td>.7</td>
<td></td>
</tr>
<tr>
<td>Pin1.4</td>
<td>.94</td>
<td>.24</td>
<td>.15</td>
<td>-.50</td>
<td>.30</td>
<td>1.05</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td>Pin1.5</td>
<td>.76</td>
<td>.43</td>
<td>.16</td>
<td>1.37</td>
<td>.17</td>
<td>1.18</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Pin1.6</td>
<td>.97</td>
<td>.16</td>
<td>.23</td>
<td>-1.36</td>
<td>.42</td>
<td>.98</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Dis2.1</td>
<td>.96</td>
<td>.20</td>
<td>.39</td>
<td>-.91</td>
<td>.35</td>
<td>.87</td>
<td>-.3</td>
<td></td>
</tr>
<tr>
<td>Dis2.2</td>
<td>.75</td>
<td>.43</td>
<td>.29</td>
<td>1.40</td>
<td>.17</td>
<td>1.06</td>
<td>.7</td>
<td></td>
</tr>
<tr>
<td>Dis2.3</td>
<td>.47</td>
<td>.50</td>
<td>.24</td>
<td>2.94</td>
<td>.15</td>
<td>1.09</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>Dis2.4</td>
<td>.81</td>
<td>.39</td>
<td>.26</td>
<td>.99</td>
<td>.19</td>
<td>1.07</td>
<td>.7</td>
<td></td>
</tr>
<tr>
<td>Dis2.5</td>
<td>.93</td>
<td>.25</td>
<td>.39</td>
<td>-.33</td>
<td>.28</td>
<td>.88</td>
<td>-.5</td>
<td></td>
</tr>
<tr>
<td>Dis2.6</td>
<td>.59</td>
<td>.49</td>
<td>.46</td>
<td>2.32</td>
<td>.15</td>
<td>.90</td>
<td>-1.9</td>
<td></td>
</tr>
<tr>
<td>Dis2.7</td>
<td>.85</td>
<td>.36</td>
<td>.32</td>
<td>.69</td>
<td>.20</td>
<td>.99</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Dis2.8</td>
<td>.76</td>
<td>.43</td>
<td>.44</td>
<td>1.34</td>
<td>.17</td>
<td>.91</td>
<td>-1.1</td>
<td></td>
</tr>
<tr>
<td>Lis3.1</td>
<td>.77</td>
<td>.42</td>
<td>.16</td>
<td>1.28</td>
<td>.17</td>
<td>1.17</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>Lis3.2</td>
<td>.81</td>
<td>.39</td>
<td>.30</td>
<td>.99</td>
<td>.19</td>
<td>1.02</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Lis3.3</td>
<td>.57</td>
<td>.50</td>
<td>.26</td>
<td>2.44</td>
<td>.15</td>
<td>1.11</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>Rad4.1i</td>
<td>.93</td>
<td>.25</td>
<td>.19</td>
<td>-.33</td>
<td>.28</td>
<td>1.06</td>
<td>.4</td>
<td></td>
</tr>
<tr>
<td>Rad4.2i</td>
<td>.95</td>
<td>.22</td>
<td>.18</td>
<td>-.69</td>
<td>.32</td>
<td>1.04</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Rad4.3i</td>
<td>.99</td>
<td>.12</td>
<td>.17</td>
<td>-2.09</td>
<td>.59</td>
<td>1.00</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Rad4.4i</td>
<td>.96</td>
<td>.19</td>
<td>.21</td>
<td>-1.04</td>
<td>.37</td>
<td>1.01</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Rad4.5i</td>
<td>.96</td>
<td>.19</td>
<td>.26</td>
<td>-1.04</td>
<td>.37</td>
<td>.95</td>
<td>.0</td>
<td></td>
</tr>
<tr>
<td>Rad4.6i</td>
<td>.98</td>
<td>.15</td>
<td>.24</td>
<td>-1.55</td>
<td>.46</td>
<td>.98</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Rad4.7i</td>
<td>.98</td>
<td>.15</td>
<td>.14</td>
<td>-1.55</td>
<td>.46</td>
<td>1.03</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Rad4.8i</td>
<td>.97</td>
<td>.18</td>
<td>.19</td>
<td>-1.19</td>
<td>.39</td>
<td>1.02</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Rad4.1m</td>
<td>.66</td>
<td>.48</td>
<td>.33</td>
<td>1.96</td>
<td>.16</td>
<td>1.03</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>Rad4.2m</td>
<td>.82</td>
<td>.38</td>
<td>.28</td>
<td>.89</td>
<td>.19</td>
<td>1.05</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>Rad4.3m</td>
<td>.94</td>
<td>.24</td>
<td>.24</td>
<td>-.50</td>
<td>.30</td>
<td>.99</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Rad4.4m</td>
<td>.89</td>
<td>.32</td>
<td>.35</td>
<td>.29</td>
<td>.23</td>
<td>.95</td>
<td>-.3</td>
<td></td>
</tr>
<tr>
<td>Rad4.5m</td>
<td>.93</td>
<td>.25</td>
<td>.35</td>
<td>-.33</td>
<td>.28</td>
<td>.91</td>
<td>-.4</td>
<td></td>
</tr>
<tr>
<td>Rad4.6m</td>
<td>.66</td>
<td>.47</td>
<td>.34</td>
<td>1.94</td>
<td>.16</td>
<td>1.02</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td>Rad4.7m</td>
<td>.91</td>
<td>.28</td>
<td>.35</td>
<td>-.05</td>
<td>.25</td>
<td>.92</td>
<td>-.4</td>
<td></td>
</tr>
<tr>
<td>Rad4.8m</td>
<td>.85</td>
<td>.36</td>
<td>.35</td>
<td>.65</td>
<td>.20</td>
<td>.97</td>
<td>-.2</td>
<td></td>
</tr>
<tr>
<td>Rad5.1</td>
<td>.99</td>
<td>.09</td>
<td>.13</td>
<td>-2.51</td>
<td>.72</td>
<td>.99</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Rad5.2</td>
<td>.95</td>
<td>.22</td>
<td>.28</td>
<td>-.69</td>
<td>.32</td>
<td>.98</td>
<td>.0</td>
<td></td>
</tr>
<tr>
<td>Rad5.3</td>
<td>.96</td>
<td>.21</td>
<td>.29</td>
<td>-.80</td>
<td>.33</td>
<td>.97</td>
<td>.0</td>
<td></td>
</tr>
<tr>
<td>Rad5.4</td>
<td>.98</td>
<td>.13</td>
<td>.17</td>
<td>-1.79</td>
<td>.51</td>
<td>1.00</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Rad5.5</td>
<td>.89</td>
<td>.31</td>
<td>.20</td>
<td>.24</td>
<td>.23</td>
<td>1.07</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>Rad5.6</td>
<td>.89</td>
<td>.31</td>
<td>.37</td>
<td>.24</td>
<td>.23</td>
<td>.93</td>
<td>-.4</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>Mean</td>
<td>SD</td>
<td>Corrected item-total correlation</td>
<td>Rasch analysis</td>
<td>Measure</td>
<td>Model S.E.</td>
<td>Infit MNS Q</td>
<td>Infit ZSTD</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>----</td>
<td>----------------------------------</td>
<td>----------------</td>
<td>---------</td>
<td>------------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gra6.1</td>
<td>.90</td>
<td>.31</td>
<td>.25</td>
<td>.18</td>
<td>.23</td>
<td>1.01</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Gra6.2</td>
<td>.88</td>
<td>.33</td>
<td>.42</td>
<td>.39</td>
<td>.22</td>
<td>.91</td>
<td>-.6</td>
<td></td>
</tr>
<tr>
<td>Gra6.3</td>
<td>.98</td>
<td>.15</td>
<td>.17</td>
<td>-.155</td>
<td>.46</td>
<td>.98</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Gra6.4</td>
<td>.94</td>
<td>.24</td>
<td>.26</td>
<td>-.41</td>
<td>.29</td>
<td>1.01</td>
<td>.1</td>
<td></td>
</tr>
<tr>
<td>Gra6.5</td>
<td>.94</td>
<td>.24</td>
<td>.40</td>
<td>-.41</td>
<td>.29</td>
<td>.88</td>
<td>-.5</td>
<td></td>
</tr>
<tr>
<td>Gra6.6</td>
<td>.87</td>
<td>.33</td>
<td>.37</td>
<td>.44</td>
<td>.22</td>
<td>.95</td>
<td>-.3</td>
<td></td>
</tr>
<tr>
<td>Gra6.7</td>
<td>.79</td>
<td>.41</td>
<td>.15</td>
<td>1.13</td>
<td>.18</td>
<td>1.17</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>Gra6.8</td>
<td>.45</td>
<td>.50</td>
<td>.18</td>
<td>3.06</td>
<td>.15</td>
<td>1.14</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>Gra7.1</td>
<td>.57</td>
<td>.50</td>
<td>.41</td>
<td>2.41</td>
<td>.15</td>
<td>.94</td>
<td>-1.0</td>
<td></td>
</tr>
<tr>
<td>Gra7.2</td>
<td>.99</td>
<td>.12</td>
<td>.12</td>
<td>-2.09</td>
<td>.59</td>
<td>1.02</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Gra7.3</td>
<td>.81</td>
<td>.40</td>
<td>.47</td>
<td>1.03</td>
<td>.18</td>
<td>.88</td>
<td>-1.2</td>
<td></td>
</tr>
<tr>
<td>Gra7.4</td>
<td>.80</td>
<td>.40</td>
<td>.48</td>
<td>1.06</td>
<td>.18</td>
<td>.87</td>
<td>-1.3</td>
<td></td>
</tr>
<tr>
<td>Gra7.5</td>
<td>.69</td>
<td>.46</td>
<td>.41</td>
<td>1.78</td>
<td>.16</td>
<td>.94</td>
<td>-.8</td>
<td></td>
</tr>
<tr>
<td>Gra7.6</td>
<td>.96</td>
<td>.21</td>
<td>.40</td>
<td>-.80</td>
<td>.33</td>
<td>.88</td>
<td>-.4</td>
<td></td>
</tr>
<tr>
<td>Gra7.7</td>
<td>.81</td>
<td>.39</td>
<td>.36</td>
<td>.99</td>
<td>.19</td>
<td>.98</td>
<td>-.2</td>
<td></td>
</tr>
<tr>
<td>Dis8.1</td>
<td>1.00</td>
<td>.00</td>
<td>n/a</td>
<td>-4.43</td>
<td>1.83</td>
<td>minimum measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dis8.2</td>
<td>.99</td>
<td>.07</td>
<td>.28</td>
<td>-3.21</td>
<td>1.01</td>
<td>.91</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Dis8.3</td>
<td>.99</td>
<td>.07</td>
<td>.28</td>
<td>-3.21</td>
<td>1.01</td>
<td>.91</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Dis8.4</td>
<td>.99</td>
<td>.07</td>
<td>.28</td>
<td>-3.21</td>
<td>1.01</td>
<td>.91</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Dis8.5</td>
<td>.99</td>
<td>.07</td>
<td>.28</td>
<td>-3.21</td>
<td>1.01</td>
<td>.91</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td>Read9.1</td>
<td>.70</td>
<td>.46</td>
<td>.34</td>
<td>1.73</td>
<td>.16</td>
<td>1.02</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td>Read9.2</td>
<td>.69</td>
<td>.46</td>
<td>.30</td>
<td>1.76</td>
<td>.16</td>
<td>1.07</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Read9.3</td>
<td>.95</td>
<td>.22</td>
<td>.25</td>
<td>-.69</td>
<td>.32</td>
<td>.98</td>
<td>.0</td>
<td></td>
</tr>
<tr>
<td>Read9.4</td>
<td>.89</td>
<td>.31</td>
<td>.50</td>
<td>.24</td>
<td>.23</td>
<td>.83</td>
<td>-1.1</td>
<td></td>
</tr>
<tr>
<td>Read9.5</td>
<td>.62</td>
<td>.49</td>
<td>.25</td>
<td>2.16</td>
<td>.15</td>
<td>1.09</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>.86</td>
<td></td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* *The item labeled as Dis.8.1 was excluded from calculation of reliability statistics due to lack of variance.*
Table 2
*Means and Standardized Deviations of Each Section*

<table>
<thead>
<tr>
<th>Section in test</th>
<th>Section</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinyin ($k = 6$)</td>
<td>#1</td>
<td>0.92</td>
<td>0.25</td>
</tr>
<tr>
<td>Radical identification ($k = 8$)</td>
<td>#4</td>
<td>0.96</td>
<td>0.18</td>
</tr>
<tr>
<td>Radical meaning ($k = 8$)</td>
<td>#4</td>
<td>0.83</td>
<td>0.34</td>
</tr>
<tr>
<td>Radical application ($k = 6$)</td>
<td>#5</td>
<td>0.94</td>
<td>0.21</td>
</tr>
<tr>
<td>Vocab &amp; grammar (multiple choice) ($k = 8$)</td>
<td>#6</td>
<td>0.84</td>
<td>0.31</td>
</tr>
<tr>
<td>Vocab &amp; grammar (word bank) ($k = 7$)</td>
<td>#7</td>
<td>0.80</td>
<td>0.35</td>
</tr>
<tr>
<td>Discourse patterns (multiple choice) ($k = 8$)</td>
<td>#2</td>
<td>0.76</td>
<td>0.38</td>
</tr>
<tr>
<td>Discourse patterns (matching) ($k = 5$)</td>
<td>#8</td>
<td>0.99</td>
<td>0.06</td>
</tr>
<tr>
<td>Listening comprehension ($k = 3$)</td>
<td>#3</td>
<td>0.71</td>
<td>0.43</td>
</tr>
<tr>
<td>Reading comprehension ($k = 5$)</td>
<td>#9</td>
<td>0.77</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Figure 1. Item-person map of the test.
Appendix One
The Final Written Test (Section #1 to Section #9)

1. Choose the word that you hear.

( ) A. nǐ pái  
B. nǚ pái

( ) A. kāi xǐ  
B. kāi shǐ

( ) A. tóu tuó  
B. tuó tóu

( ) A. shéng gōng  
B. sèng gōng

( ) A. qí jì  
B. xí qì

( ) A. bái shì  
B. pái chì

2. Choose the most appropriate reply turn for the sentence that you hear.

Prompt: 我们明天晚上一起去吃中国饭，好吗？
( A. 好不好？  
B. 好久不见！  
C. 太好了！

Prompt: 李友，好久不见，你怎么样？
( A. 你好，我是王朋。  
B. 我也不错。  
C. 我很好，你呢？

**Prompt: 明天是星期六，我们去看球，怎么样？
( A. 我觉得打球很有意思。  
B. 我周末只想睡觉。  
C. 我也常常看电视。

Prompt: 现在几点？
( A. 四月三十号。  
B. 四点三十分  
C. 一九四九年。

Prompt: 请问你做什么工作？
( A. 我是他的朋友。  
B. 我是律师。  
C. 我是中国人。

Prompt: 你想喝茶还是喝可乐？
( A. 那是一瓶可乐。  
B. 这是一杯茶。  
C. 来一杯茶吧。

Prompt: 认识你我很高兴！
( A. 我不认识你。  
B. 我是王朋的朋友。  
C. 认识你我也很高兴。

**Prompt: 我要一瓶可乐，可以吗？
( A. 对不起，我家没有可乐。
    B. 我也喜欢喝可乐。
    C. 我常常喝可乐。

** This item appears to have more than one acceptable answer. Revision is in order.

3. Listen to a dialogue and complete the following T/F questions.

(The context of the story is set in the U.S.)
小王（男）：小李，这个周末你有空吗？
小李（女）：这个周末我没事。怎么了，小王？
小王（男）：你喜欢看电影，我也喜欢看电影，所以周末我想请你去看一个电影。
小李（女）：是外国电影吗？
小王（男）：不是，是一个美国电影。
小李（女）：嗯，我觉得美国电影没有意思。我觉得中国电影很有意思。
小王（男）：是吗？那我们去看一个 Jacky Cheng 的电影，怎么样？
小李（女）：太好了？几点？
小王（男）：星期天晚上 6 点，可以吗？
小李（女）：星期天晚上我爸爸妈妈和我要去吃中国菜呢。
小王（男）：那算了吧。

( ) At the beginning of the conversation, Xiao Wang invited Xiao Li to a Chinese movie.
( ) Xiao Li does not like domestic movies.
( ) Xiao Wang and Xiao Li will watch a Chinese movie together this weekend.

4. Circle the radical of the following characters and write out the meaning for each radical.
视 Radical Meaning ___________________
晨 Radical Meaning ___________________
嘴 Radical Meaning ___________________
刻 Radical Meaning ___________________
期 Radical Meaning ___________________
饱 Radical Meaning ___________________
侣 Radical Meaning ___________________
贵 Radical Meaning ___________________

5. Radical knowledge application.
Circle the character that means “grandma” 奶 仍 禧
Circle the character that means “river” 柯 阝 河
Circle the character that means “sad” 翡 悝 悲
Circle the character that means “finger” 指 赤 脂 醋
Circle the character that means “toe” 踝 趾 祇
Circle the character that means “vocabulary” 词 饲 嗦

6. Multiple choice.
(1) Which of the following is the correct way to say “September 10th, 2012”?

A. 十号九月二零一二年
B. 九月十日二零一二年
C. 二零一二年九月十日
D. 二零一二年十日九月
QUALITY OF CFL SUMMATIVE ASSESSMENT 30

(2) 昨天晚上我 6 点吃饭，可是李友 11 点 _______ 吃饭。
   A. 没    B. 还    C. 不    D. 才
(3) 王朋家有 _______ 口人？
   A. 什么    B. 吗    C. 谁    D. 几
(4) 白英爱：你为什么认识王朋？
   高小音：________ 王朋是我的同学。
   A. 可是    B. 因为    C. 还    D. 还是
(5) 昨天你看电视 _______ 吗？
   A. 了    B. 没    C. 吧    D. 呢
(6) 我介绍 _______，这是我的爸爸。
   A. 两点    B. 一点    C. 一下    D. 两下
(7) 王朋的姐姐漂亮 _______ 王朋的妹妹漂亮？
   A. 还    B. 还是    C. 也是    D. 不是
(8) 你爸爸在 _______ 工作？
   A. 哪儿    B. 一点儿    C. 什么    D. 那儿

7. Choose the appropriate words from the word bank to fill in the blanks.
   A. 有的时候    B. 常常    C. 外国    D. 那    E. 只
   F. 不错    G. 没有意思    H. 岁    I. 太    J. 不

(1) 我周末常常打球，_______ 我也唱歌和跳舞。
(2) A: 今天晚上我们去看一个 _______ 电影，怎么样？
   B: 今天晚上我 _______ 想睡觉。
   A: _______ 我去找别人吧。
(3) 你的中文很 _______。
(4) 我的姐姐今年二十三 _______。
(5) 你的中文 _______ 好了。

8. Match the questions on the left with the appropriate responses on the right.

   ( ) 你弟弟是他吗？     a. 我不认识。
   ( ) 你妈妈做什么工作？  b. 我弟弟不是大学生。
   ( ) 你妈妈喜欢听音乐吗？ c. 我妈妈觉得听音乐没有意思。
   ( ) 你弟弟也是大学生吗？ d. 我弟弟不是他。
   ( ) 你认识王朋的弟弟吗？ e. 我妈妈是老师。

9. Reading comprehension.

昨天晚上是白英爱的生日，所以白英爱请王朋和李友晚上六点去她家玩儿。白英爱还请了她的同学高文和中和高小音。可是因为高小音昨天晚上没有空，所以没有去。王朋和李友五点半才下课，所以他们六点半才到白英爱的家。高文中昨天没事儿，所以五点半就到白英爱家了。

在白英爱家，白英爱问她的朋友们想喝点儿什么。王朋要了一杯水。李友要了一瓶可乐。高文中只想喝茶。可是白英爱家没有茶，所以高文中什么也没有喝。他们一起吃中国饭，一起唱歌、跳舞和聊天儿。他们也看了
一个中国电影。王朋、李友和高文中晚上十二点才回家。

(1) 昨天晚上几个朋友去了白英爱的生日晚会(party)？__________
   A. 两个人       B. 三个人   C. 四个人     D. 五个人

(2) How much time did Wang Peng spent on his way to Bai Ying’ai’s home?
   _________
   A. 1 hour       B. 2 hours   C. 0.5 hour   D. 1.5 hours

(3) 高文中喜欢喝什么？__________
    A. 酸       B. 茶       C. 水       D. 咖啡

(4) 白英爱和她的朋友们没做什么？__________
    A. 聊天儿     B. 跳舞     C. 打球     D. 唱歌

(5) How much time did Gao Wenzhong spent at Bai Ying’ai’s home?
    _________
    A. 5 hours     B. 5.5 hours  C. 6 hours    D. 6.5 hours