ESL Reading Test Development and Analysis

Hyojung Lim
Michigan State University
hyojung@msu.edu
Young-Shin Kwon
Teachers College, Columbia University
yskwon79@gmail.com

The purpose of this paper is to present an overview of ESL reading test development and analysis in the context of a small-scale ESL classroom. We created 12 multiple-choice items for an ESL reading mid-term exam, administered the test in the Community English Program at Teachers College in fall 2008, and analyzed the test results to evaluate the reliability and validity of the test. We first describe the nature of reading ability by reviewing the literature on second language reading and reading assessment. Based on a widely shared definition of reading ability, we suggest a theoretical construct of reading ability and relevant observable variables. Following Bachman and Palmer (1996), this paper provides practical guidance for language teachers with regard to how to create reading test items and assess the test quality from describing the target language use (TLU) domain and task types, developing a test design statement, generating the blueprint for test operationalization, coding multiple choice items, to conducting item and distractor analyses. Issues revolving around L2 reading test development are further discussed.

We created this test as part of the mid-term exam designed for the ESL learners of the Advanced 2 (A2) class of the Community English Program (CEP) at Teachers College, Columbia University in New York City. The CEP is a lab-school where Teachers College students from the Applied Linguistics and TESOL program teach adults from the community as part of their practical training, apply various teaching methods based on linguistic and pedagogical theories, and collect data for empirical studies related to the instruction and assessment of second-language (L2) learners. The students of CEP are adult ESL learners, most of whom are either immigrants, international students who are planning to study or already in school, or family members of international students in the Columbia University community. The CEP curriculum consists of twelve levels in total, ranging from basic (B1 to B4), intermediate (I1 to I4), and to advanced (A1 to A4).

Participants of this study are advanced ESL students in the A2 evening class, who are still aiming for higher-levels of English proficiency, some seeking to advance to the levels of A3 and A4. About 67% of the students completed either graduate or post-graduate degrees and 50% are planning to stay in the United States for an academic or a vocational purpose. The A2 course focuses on further developing the four integrated skills in English. The course objectives are to improve students' skills with a focus on critical analysis and self-expression, and to help students with their knowledge and application of pragmatics. The mid-term exam accounts for 30% of their final

grade. By the time of data collection, the class theme was about the political or social issues as dealt with in the 2008 US presidential election. With regard to the language skills, students were taught lessons on (1) how to use contextual clues, analyze arguments, make inferences and generalizations and determine the purpose and function of a text for reading, (2) how to link paragraphs to essays, create an argument, organize information and use transitions for writing, (3) how to summarize, identify implications, and personalize the information for listening, and (4) how to continue a discussion, present ideas and debate a topic for speaking. In terms of reading, students were exposed to extensive reading of news articles both in and out of class.

The mid-term exam we designed for the A2 class can be classified as an achievement test or a progress test, given that it aims to measure the extent of learning or mastery within a specific instruction domain. The test result, as part of final grade, is used to make decisions about their advancement or competency. The mid-term may serve as a diagnostic test as well: the test result carries information about students' strengths and weaknesses, and thus can prescribe future teaching or learning directions for the rest of *SEM*ester. Since the purpose of the course is to improve integrated skills, listening and speaking are supposed to be assessed in the test. However, due to time constraints as well as test practicality, the mid-term includes only grammar, listening, reading and writing. In this paper, we focus on the reading test, since reading skills were more emphasized in the class during the first half of the semester than listening and speaking skills.

We will first describe the nature of reading ability based on the review of prior research on second language reading and assessment. Based on the prior literature on reading comprehension, we suggest a theoretical construct of L2 reading ability. The theoretical construct of reading ability provides a useful ground for the subsequent test construction: describing the target language use (TLU) domain and task types, writing test design statements, developing the blueprint for the test operationalization, coding multiple choice section, and finally administrating the test. Lastly, the test reliability and the construct validity will be assessed through item analyses.

Reading Ability

To measure learners' reading ability in the A2 class at CEP, essential is to first clarify what reading ability is and/or what reading components the test is to assess. Reading is a complex, multifaceted cognitive behavior that involves a number of linguistic and cognitive processes. Thus, it seems hardly possible to come up with one simple definition for it (Grabe & Stoller, 2002). Instead, many reading researchers have shed light on multiple aspects of the reading construct. Researchers have foraged for discrete factors that constitute L1 and L2 reading comprehension (Barnett, 1986; Devine, 1981), identified cognitive processes involved in different types of reading (Weir, Hawkey, Green, & Devi, 2009; Khalifa & Weir, 2009), and investigated strategies/skills that learners likely employ while reading L2 texts (Cohen & Upton, 2007; Savery, 2012; Sheorey & Mokhtari, 2001).

According to the information-processing approach, reading comprehension is considered as the

product of bottom-up and top-down reading skills. Grabe and Stoller (2002) characterize reading as a serial process consisting of two different levels: lower-level and higher-level processes. Lower-level processes include basic linguistic processes such as word recognition, syntactic parsing, and even simple sentence verification. Reading begins with decoding a string of letters in print, recognizing word meanings, parsing sentence structures, and finally to constructing clause-level, textual meaning units. To obtain a high level of comprehension, therefore, it is crucial for learners to be able to execute the lower-level processes automatically. Efficient processing frees up available mental resources, which eventually helps readers to hold more information in their memory (Daneman & Carpenter, 1980). Both L1 and L2 reading researchers have acknowledged the contribution of automatic bottom-up processing skills to the increased reading comprehension (Koda, 2005; Roberts, Christo, & Shefelbine, 2011). Conceivably, without processing lexical and syntactic information, readers cannot run any higher-level cognitive processes (e.g., inferences) where we believe ultimate comprehension takes place. In L2 reading, Alderson (1984) claims that foreign-language reading is a language problem rather than a reading problem; especially for those who are already literate in their L1, much of the difficulty in L2 reading comprehension could be mainly due to their language proficiency, not to their literacy skills. This is particularly true for educated adult language learners who already possess higher-order thinking ability in their native language but lack automatic processing skills in the L2. In terms of assessment, any reading tests are likely to assess lower-level linguistic processes in an implicit way; there is no reading test item that directly measures test takers' word recognition skills or sentence processing skills. Instead, bottom-up skills are often assumed to be tested in a rather unified or general way (Alderson, 2000).

The top-down approach to reading underscores the effects of higher-level reading processes on comprehension. This is where the schema theory comes into play. According to Grabe and Stoller (2002), the higher-level processes begin to play a role in the text model of comprehension, where readers draw main ideas and supporting details from a text at or beyond the clause-level meaning units. While reading, readers are likely to activate their content and formal schemata: content schemata means readers' background knowledge of the content area of the text, whereas formal schemata pertains to readers' knowledge of the rhetorical structures of different types of texts (Carrell & Eisterhold, 1983). The essential idea of the schema theory is that readers' familiarity with the discourse organization as well as with the topic facilitates their understanding of the text. Thus, reader variables, such as cultural background or topical knowledge, often become determining factors for the quality of comprehension. Finally, Grabe and Stoller explain that executive control (or metacognitive) processes are part of the higher-level reading processes. Previous empirical studies found that good readers have advanced synthesis and evaluation skills so that they can simultaneously monitor their comprehension and quickly adopt relevant reading strategies (Paris & Myers, 1981). In the context of L2 reading assessment that measures both language and reading ability, however, we believe that educated adult L2

learners should be forced to utilize their L2 linguistic knowledge and skills rather than their content knowledge, or general reasoning ability. Especially, international graduate students who usually have high-level literacy skills in their L1, meaning that they know how to approach a text and how to inspect their own understanding. As long as they meet the threshold of L2 language proficiency, if any, such learners should be able to transfer their cognitive and literacy skills to the second language (Cummins, 1991).

From a balanced perspective, the interactive model highlights that the bottom-up processing works in concert with the top-down processing, or vice versa. Interaction has been understood in many different ways. The "simple view of reading" proposed by Hoover and Gough (1990) views reading comprehension as the combination of word decoding and listening comprehension; lacking either decoding skills or listening ability can deteriorate the quality of reading comprehension. Rather, taking a "compensatory" approach, Stanovich (2000) points out the tendency of readers resorting to their higher-level processing skills to compensate for their deficiency in lower-level processing skills. For instance, readers often use context clues to guess the meaning of an unknown word and consequently improve their understanding of the text. In L2 reading, Bernhardt's (2005) compensatory model echoes Stanovich's view, thereby describing how L2 readers rely on their L1 literacy skills to improve L2 language-processing skills or how an increase in word knowledge helps to accelerate the processing of L2 sentences. Meanwhile, Grabe (1991) suggests a more general type of interaction; the interaction between a text and a reader. Readers form their reading comprehension by relating the given textual information to their background knowledge. Given that it is a reader who reconstructs the representation of a text, the way that the reader processes the text likely determines the type and level of comprehension. To us, the interaction discussed in Grabe seems to rather support the schema theory where high-level reading processes play a substantial role.

Another way to approximate the reading construct is to explore types of strategies that readers employ while reading. Researchers, in their examination of good and poor readers, have discovered that good readers are likely to adopt various effective reading strategies (Anderson, 1991; Ebrahimi, 2012; Paris, Limpson, & Wixson, 1983; Paris & Myers, 1981). In this regard, Grabe (2004) states that "a number of individual comprehension strategies have been shown to have a significant impact on reading comprehension abilities" (p.51). According to Fitzgerald (1995), reading strategies can be understood in two different ways: (a) psycholinguistic strategies that learners use to recognize and comprehend lexical items; and (b) metacognitive strategies that learners use to deal with a whole text and repair miscomprehension. The psycholinguistic strategies are similar to the compensatory strategies that L2 learners rest on to overcome linguistic limitations. In the context of assessment, Cohen and Upton (2007) documented the reading strategies based on international students' verbal reports. The observed strategies were categorized into three groups: (a) approaches to reading the passage (e.g., considering prior knowledge of the topic), (b) uses of the passage and the

main ideas to improve understanding (e.g., re-reading to clarify the ideas), and (c) identification of important information and the discourse structure of the passage (e.g., looking for sentences that convey the main ideas). Note that the reading strategies listed here are all language-independent, metacognitive strategies. According to Fitzgerald's (1995) collection of literatures on L2 reading strategies, the most common were: asking questions, rereading, imaging, using a dictionary, anticipating or predicting, reading fast or changing speed, associating, skipping, and summarizing. From learners' perspective, Judith (1995) discovered that scanning for specific information, skimming, re-reading, word-guessing skills and summarizing were valued most by students learning Spanish as a second language. Taken together, L2 readers use various types of strategies at all levels (e.g., lexical, sentential, and textual level) to maximize their comprehensio. They are likely to approach L2 reading as a problem-solving task, thereby evoking higher-order cognitive processes (e.g., monitoring), presumably in the same way that they would do in L1 reading. The reading strategies reviewed so far can be reduced to three major reading behaviors: reading to search for information, integrating pieces of information, and figuring out hidden meanings (e.g., an author's intention).

Lastly, but most importantly, the purposes of reading need to be taken into consideration, as reading itself is a purposeful behavior. According to Carver (1997), there are two types of reading: "rauding" and "reading to learn." The term "rauding" pertains to basic comprehension — reading a text to understand major points — while "reading to learn" involves the reconstruction of a text — figuring out main ideas and supporting details. Similarly, drawing from the cognitive processing model for reading comprehension, Khalifa and Weir (2009) propose two kinds of reading at two different levels: careful and expeditious reading at the local and global level, respectively. Careful reading is intended to extract complete meaning from a given text (Hoover & Tunmer, 1993). It is conceived as slow, careful, linear, and incremental reading. Conversely, expeditious reading is rapid, selective, and efficient reading, including scanning and skimming. Both readings can take place at the lexical or sentential (local) level, or at the paragraph or textual (global) level. Albeit using different terms, ETS (2000) suggests the purpose-driven framework for the iBT TOEFL reading test: reading to find information, reading for basic comprehension, reading to learn, and reading to integrate information. In light of item difficulty, reading to integrate information is thought to be more difficult than reading to find information, since the former requires relatively higher-order cognitive abilities. Taking learners' proficiency into account, we decided to include more inference-type questions. For advanced learners, such as those in the A2 class, reading should not be a language problem any longer. Rather, they are expected to read to synthesize and critique texts.

To sum up, the reading construct that we want to measure in the mid-term exam entails three variables: gist, details, and inference. The information-processing perspective on reading, the skill-and-strategy approach, and the reading-purpose perspective all provide strong rationale for the variables that we suggest. Given that lower-level processes are assessed in an implicit manner, we

expect our students to be able to make connections across sentences and paragraphs quickly and accurately so that they can correctly comprehend main ideas and supporting details in a given time. In terms of reading types to be tested, reading for gist and details may be associated with search reading, skimming, and reading to learn. More specifically, reading for gist can be involved in such items as summarizing a text, finding a main idea, or selecting a headline/title for the text. Reading for details can be induced by the items such as finding specific information, relating a pronoun to its referent, and rephrasing a given sentence. For the inference questions, learners have to make use of their content and formal schemata to answer the questions. Readers will be asked to derive both literal and implied meaning at lexical, sentential, and textual level, to guess an author's intention for using specific expressions in the flow of ideas, and to read an author's tone. Figure 1 summarizes the theoretical model of reading ability for the reading test in CEP A2 mid-term test.

Test Construction

The Target Language Use (TLU) Domain

The context of the target language use (TLU) domain is the CEP evening class of L2 learners at the Advanced 2 level, taught in a classroom at Teachers College. The class integrates all four language skills of reading, writing, listening, speaking, and includes grammatical contents, while following a weekly theme-based curriculum. The learners are all adults from various nationalities and cultural backgrounds, coming from different occupational backgrounds as well. They are generally enrolled in the CEP to advance their English proficiency overall, while some learners have specific purposes such as to enter a higher-education institution or an English-speaking workplace in the United States. While it would be difficult to pinpoint a specific TLU domain because of the broad background of the group of learners, we have decided that language instruction would be the most appropriate TLU domain for our subjects.

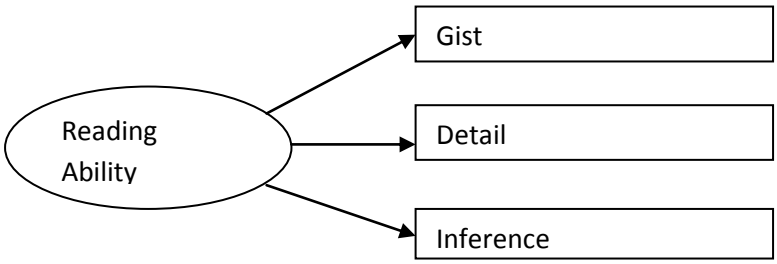

Figure 1. A theoretical model of reading ability.

We attended one class session to observe what themes the learners were specifically dealing with after looking over the syllabus and textbook for the class. At the time of observation, the class topic was the current presidential candidates and their campaigns on several socio-economic issues. The learners had been assigned to research information from the media and support one of the presidential candidates' campaigns on a specific issue (e.g., education, health care, and energy) based on their investigation. In pairs, they were to present a brief spoken debate on their ideas by supporting them with the information they found. According to the syllabus, the learners had also been instructed on writing academic essays on opinion-based subjects. We judged that after completing these instructional tasks, the learners should be able to understand and analyze fact-based information from the media. They should be able to read for specific details and infer further facts according to the given information. In both academia and the workplace, extracting information from the media or other informational sources, and making critical judgments of the information to form individual opinions are important abilities that are often required to competently perform a given duty, such as making decisions about a course of action or a direction that a business should take.

Taking these into consideration, for the reading task of our test we decided to use a news article on the subject of education and the differing views that the presidential candidates have on this issue. The skills needed for the reading tasks are (a) reading for gist (both at the passage level and paragraph level), (b) reading for detail such as for finding facts and correct word references, and (c) making correct inferences about the writer's purpose or rhetorical purpose.

Design Statement

Following Bachman and Palmer (1996), we developed a design statement for the current test (see Table 1). The design statement is essential for the subsequent procedure of test development, operationalization, trialing, and assessment use. Based on the design statement, the test structure and the task specifications are presented in the following section.

Operationalization

Test structure.

1. Number of tasks: The test consists of one task containing 12 items to measure the test takers' ability of reading in a language-instructional domain. The students must read a news article and answer twelve multiple-choice questions.

2. Salience of tasks: The reading task is clear, with clear labels and specific instructions provided.

3. Relative importance of tasks: All items within the task are of equal importance and worth the same amount of points in the mid-term exam.

4. Number of tasks per part: The reading part is one task consisting of a set of twelve multiple-choice questions. Table 2 provides a summary of the test structure.

Table 1
Design Statement

1. Test purposes	
A. Inferences	About test-takers' reading and writing ability in a language-instructional domain.
B. Decisions	
I. Stakes	Relatively high in the context of the course; results are used to determine advancement to the next level in the program (the mid-term examination counts for 30% of the final course grade).
II. Individuals affected	Test-takers (CEP students) and CEP class teacher
III. Specific decisions to be made	1. Achievement a. Progress: To determine if students have mastered the language skills covered up to the mid-term exam. b. Grading: Results are part of the mid-term grade, a component of the final grade, which determines advancement to the next level in the program. 2. Diagnosis a. For teachers: To evaluate each student's strengths and weaknesses in order to help students make further improvement. b. For students: To obtain information on their own strengths and weaknesses in order to identify and overcome weaknesses.
2. Description of TLU domain and task types	
A. Identification of tasks	
1. TLU domain	Language-instructional, but also possible to be real-life for some students.
2. Identification and selection of TLU tasks for consideration as test tasks	TLU tasks to be analyzed were identified based on the course syllabus and class handouts. The Reading Task (reading a news article and answering multiple-choice questions) is an instructional task similar to those performed in class. Reading a news article can also be a real-life task.
B. Description of TLU task types	Refer to Table 3 for the test task specifications.
3. Definition of constructs	
A. Language ability	The construct definition for this achievement test is based on both a theoretical model of language ability and the content of the class. The elements of language knowledge included in the construct definition are: - Reading ability a. Reading for gist (summary, main idea, title) b. Reading for details (fact finding, word reference)

	c. Inferencing (writer's purpose, rhetorical purpose, tone)
B. Strategic competence	Not included in the construct.
C. Topical knowledge	Not included in the construct. However, some degree of topic knowledge is assumed, as students are familiar with the topic dealt with in class (e.g., education, presidential candidate's debate).

Table 2
Test Structure

Construct	Task Type	Number of Tasks	Number of Items	Time	Scoring
Reading Ability • Gist • Detail • Inference Theme: Education issues in the US presidential debate	Selected-Response (Multiple Choice)	1	12	30 mins	Dichotomous Scoring 0/1 12 points available

Test task specifications.

1) Purpose: See the design statement (Table 1)
2) Definition of construct: See the design statement (Table 1)
3) Setting
   a) Physical characteristics: Classrooms (Horace Mann Hall Rm. 136) at Teachers College, Columbia University. See the test task specifications in Table 3.3 for a detailed description of classroom conditions.
   b) Participants: The CEP teacher and the test-takers (CEP students).
   c) Time of task: During class hours on Thursday, October 23, 2008.
4) Time allotment: Thirty minutes.
5) Instructions:
   a) Language: The target language (English) because test-takers have a variety of native languages. Separate instructions are provided for the reading part and the writing part, and the students are allowed to ask questions about instructions they are not sure of.
   b) Channel: Visual (writing).
   c) Instructions: See the copy of the test provided in Appendix C
6) Characteristics of input and expected response: See the test task specifications in TABLE 3 (Appendix A).
7) Scoring method:
   a) Criteria for correctness: The multiple-choice questions are scored dichotomously based on an objective answer key.
   b) Procedures for scoring the responses: The multiple-choice questions are scored dichotomously based on the objective answer key for the multiple-choice questions. One point is given for each correct answer and zero points are given for each incorrect answer, for a possible total of twelve points.

c) Explicitness of criteria and procedures: The test-takers are informed in general terms about the scoring criteria in the instructions. Table 3 summarizes the task specifications for each task. (See Appendix A)

Item Coding

The reading test consists of twelve multiple-choice items that are divided into three observable variables: understanding the gist, finding details, and making inferences. An inference is an overarching notion of guessing from the context, ranging from guessing meanings of new words to reading the author's tone. Table 4 illustrates the observable variables for each item and a brief description of their subordinate variables.

Administration Procedures

The test was administered to the CEP Level A2 evening class as part of their mid-term evaluation and took place in their original classroom during their usual class time. The students were given a separate listening and grammar test at the beginning (given by the CEP instructor) and afterwards handed out the test booklets on the reading and writing parts, which they could start immediately upon receiving it. As the test booklets were being handed out, the students were told how much time they had to complete the test, and were allowed to leave the classroom upon the completion of the test. The students were allowed to ask any questions that arose while taking the test. The entire test period lasted for two and a half

Table 4
Coding Multiple Choice Items for Reading

Observed Variable	Item Number	Description of the item
Gist	1	Giving a title to the entire reading passage.
	11	Understanding the main idea of a paragraph.
	12	Summarize the entire passage.
Inference	2	Reading the author's tone.
	3	Understanding a rhetorical purpose.
	4	Guessing an expression in context.
	6	Understanding a rhetorical purpose
	7	Understanding the metaphoric use of a word.
Detail	5	Comprehending specific information explicitly stated in the text.
	8	Comprehending specific information explicitly stated in the text.
	9	Finding the referent of a pronoun.
	10	Rephrasing a sentence.

hours, which included all four parts of the test (listening, grammar, reading, and writing). As the test-takers finished their test, they were given a brief post-test survey to fill in. The survey consisted of general-information questions about the test-taker (age, nationality, occupation, etc.) and some questions about self-perceived language proficiency and qualities of the test itself.

Test Takers

The number of the students was twelve from the evening A2 class of Community Language program at Teachers College in New York. Most of them were in their late twenties or early thirties, while one was in her late thirties and one in her early forties. Ten students completed their education at or beyond the graduate level, while only two students obtained up to a bachelor's degree. Their majors were as diverse as Social Work, Trading, Economics, Law Administration, English and American Literature, Electronic Engineering, Art Design, and Home Economics. Regarding nationality, the East-Asian students were dominant: seven from Japan, two from Korea and one from China. The two remaining participants were from Bolivia and Poland. Nine students were female and three were male. With regard to the length of stay in the United States, it varied from one and a half months to eight years. To be more specific, eight students (67%) lived in America for less than one year, while two students for more than five years. The post-test questionnaire was used to collect the participant information (Appendix B).

Test Instrument

The purpose of our test was to measure reading ability within a specific instructional domain. By reviewing various articles on reading ability, we have decided to include gist, inference and detail for the reading construct. The reading test consisted of twelve multiple-choice items; 3 for gist, 5 for inference and 4 for detail variable. The topic of the reading task was "American education" discussed in the 2008 US presidential election. A copy of the actual and the expected responses are attached in Appendix C.

Scoring Procedures

The multiple-choice reading task was scored objectively and dichotomously. One scorer rated every test paper using an objective answer key and assigned one point to correct answers and zero points for incorrect answers. The total score was the sum of the point that each item earned. The possible range of scores on this task was therefore 0 to 12.

Analyses and Results

Descriptive Statistics

The reading section had 12 multiple-choice questions, for a total possible score of 12 (k=12). One point was assigned to a correct answer and zero to an incorrect answer. In terms of the measures of central tendency, the mean was 6.75 (56.25%), the median was 6.50, and the mode was 6.00. The skewness value of the score distribution was −0.04. The kurtosis was 0.34. The kurtosis indicates the degree to which the distribution is peaked. Given that the skewness value and the kurtosis were close to zero, the test scores were normally distributed. In terms of the data dispersion, the range was 7.00, from a minimum score of 3.00 to a

maximum score of 10.00. The standard deviation was 1.91. The results are summarized in Table 5.

Table 5
Descriptive Statistics for the Reading Task

Statistics	Results
Number of participants (N)	12.00
Number of items (k)	12.00
Maximum possible score	12.00
Mean	6.75
Median	6.50
Mode	6.00
Skewness	– 0.04
Kurtosis	0.34
Range	7.00
Minimum	3.00
Maximum	10.00
Standard deviation	1.91

Considering that the test was an achievement test for the A2 class at CEP, we expected the distribution of scores to be negatively skewed, and ideally students were to answer 70% of the test correctly on average. However, our test results turned out to be undesirable for a criterion-referenced test: the skewness value of −0.041 and the kurtosis of 0.334 indicate that the test scores were normally distributed. Furthermore, the average of 6.75 means that only 56.25% of the test was answered correctly on average, which was somewhat lower that the cut-off line (70%) for the pass and fail standard at CEP.

From the statistical figures, we could infer that our test was somewhat difficult for the participants. Presumably, only a few students might have mastered the theme and the reading strategies previously taught in class. The larger proportion of inference questions might have raised the level of difficulty in that these questions usually require higher-order cognitive skills. Therefore, it could be that our test failed to correctly measure students' reading ability on the basis of the class objectives.

The standard deviation of 1.91, the kurtosis of 0.34 and the range 7.00 out of 12.00 suggest that the test scores are somewhat widely spread out. Thus, the group in the evening A2 class proved to be heterogeneous with regard to English reading ability. It may be that these students had not been correctly placed in the beginning, or has truly shown varying degrees of development in reading comprehension. In Table 6, the results are illustrated in the stem-and-leaf plot.

Table 6
Reading MC Stem-and-Leaf Plot

Frequency	Stem	Leaf
1	3.	0
1	5.	0
4	6.	0 0 0 0
3	7.	0 0 0
2	9.	0 0
1	10.	0

Internal Consistency Reliability and Standard Error of Measurement for the MC Task

This section evaluates the test reliability. Test reliability means the extent to which the results are consistent or stable. To be more specific, the reliability estimates are interpreted as the percent of systematic, consistent, or reliable variance in the scores of a test, including both true and random error variance. When it comes to the MC items, the internal consistency reliability across the 12 items was examined by calculating the reliability coefficient. The internal-consistency reliability informs us as to the degree to which each item

relates to all the other items. Subsequently, we calculated the standard error of measurement (*SEM*) to determine a confident interval of a student's score; the narrower *SEM* evidences the higher test reliability, meaning that test cores will less fluctuate if the test is repeated. We also calculated Cronbach's alpha as an alternative measure of the split-half reliability. The split-half reliability was not appropriate for this short test, because the number of test items was too small to separately score and compare the compare the odd-numbered and the even-numbered items. Table 7 presents the internal consistency for the 12 MC items.

Table 7
Internal Consistency Reliability Statistic for the Reading Task (K=12)

Cronbach Alpha Coefficient	Number of Items
0.343	12

Cronbach's alpha typically ranges from 0 to 1, with 1 being the most consistent. The coefficient 0.343 suggests that the internal consistency for our MC items were relatively low. With the reliability of 0 .343, the scores are around 34% consistent. That leaves 66% of measurement error or random variance in the scores. This implies that the degree to which the items relate to one another was somewhat low, so was the internal consistency of the test. There are several reasons for the unexpected results: First, the small number of items might be ascribed to the low consistency. The MC items were only twelve in total, consisting of three items for the gist, four for the detail, and five for the inference variable. Hence, every correlation between items should have a substantial impact on the reliability of the test. Second, the sample size of twelve students might have been too small to correctly calculate the reliability coefficient. Only a couple of students' mistakes in their responses could have affected the statistical analyses. In either case, the low internal consistency seems mainly due to the limited amount of data. All in all, we do not have sufficient evidence to say that our test is trustworthy. In addition to Cronbach's alpha, the *SEM* was calculated to determine the band around a student's score within which the student's score would probably fall, if the test were repeated. This gives an idea of how accurate an individual's true test score might be. The computation formula for *SEM* is given in Table 8, where the result for our test is summarized as well.

Table 8
Standard Error of Measurement for the Reading Task

$*SEM = S\sqrt{1-rxx}$.
$SEM = 1.913 \times 0.811 = 1.551$

* where S = standard deviation (retrieved from the descriptive statistics) and rxx = reliability estimate for the test, which is equal to the Cronbach's alpha coefficient

Based on the estimated *SEM* = 1.551, a 95% confidence (±2 *SEM*s) interval was calculated. According to the result, a student's score would consistently fall within a band of two *SEM*s higher and two *SEM*s lower than her raw score 95% of the time if s/he were to take the test multiple times. For instance, participant #3 received 7 out of 12, but it is 95% certain that the score would fall somewhere between 3.174 and 10.102 if the participant were to take the same test repeatedly. Since each item was scored dichotomously, we rounded up these values to 4 and 11, respectively.

Considering that the total reading score was 12, the *SEM* of 1.91 seems relatively large for the short test, and thus the 95% confidence interval for participant #3's score turned out to be too broad. This indicates that extra factors, other than one's reading ability, may have confounded the observed scores such as the degree of motivation, fatigue, and chance knowledge of item content.

Item Analysis

To search the causes for the low internal consistency, the 12 MC items were analyzed by calculating the item difficulty (or *p*-value), the item discrimination index (or *d*-value) and the "alpha if item deleted." To explain each term briefly, the item difficulty is an index that tells us the proportion of test takers who got the item correct in proportion to all the test takers who answered the item. The item discrimination indicates the degree to which the item discriminates between different groups. By convention, the high 27% of the students is compared with the low 27% in a norm-referenced test. Lastly, the "alpha if item deleted" shows a recalculated Cronbach's alpha if the item is deleted from the test. These statistical results were the bases for the decisions made on whether to delete or keep each item (see Table 9).

The p-values ranged from 0.167 (for item 6) to 0.917 (for item 11). In other words, item 6 was extremely difficult, therefore, only two participants got the answer correct, while item 11 was extremely easy, therefore, everyone except for one participant got it correct. The overall *p*-value of the twelve items was 0.576.

Given that an ideal achievement test aims for a *p*-value of 0.70, our test appeared to be somewhat difficult as a criterion-referenced test, which is consistent with the earlier report on the descriptive statistics. Except for items 8, 9 and 11 with *p*-values of 0.833, 0.833 and 0.917 respectively, the *p*-values of all the other items were lower than 0.70. Moreover, item 5, 6 and 10 were extremely difficult with *p*-values of 0.333, 0.167 and 0.250 respectively. By only looking at the estimated *p*-values,

Table 9
Item Analysis for the Reading Test

Item	Observed Variable	Difficulty (*p*-value)	Discrimination (*d*-value)	Alpha if item deleted	Decision
1	Gist	0.417	0.239	0.269	Keep
2	Inference	0.417	0.133	0.317	Keep
3	Inference	0.583	0.412	0.185	Keep
4	Inference	0.667	−0.417	0.516	Delete
5	Detail	0.333	−0.249	0.461	Delete
6	Inference	0.167	0.106	0.328	Keep
7	Inference	0.667	0.367	0.328	Keep
8	Detail	0.833	0.240	0.284	Keep
9	Detail	0.833	−0.021	0.367	Not sure
10	Detail	0.250	0.706	0.068	Keep
11	Gist	0.917	−0.189	0.396	Not sure
12	Gist	0.667	0.249	0.267	Keep

our test seems more like a placement or a proficiency test, rather than an achievement test. In terms of the difficulty level of each variable, the average *p*-value for gist items was 0.667, that of inference items 0.500, and that of detail items 0.562; inference items were relatively more difficult than the other two variables, as we expected.

To calculate the discrimination index, the point biserial correlation was utilized. The "corrected item-total correlation" was interpreted as the *d*-value. By convention, the items with a *d*-value of 0.40 and above are evaluated as very good items. Those with a *d*-value of 0.30 to 0.39 are considered as reasonably good items, but subject to improvement. On the other hand, items with a *d*-value of 0.20 to 0.29 do not effectively differentiate the high 27% from the low 27% of test-takers. Lastly, a *d*-value of 0.19 and below indicates that the item needs to be deleted or improved. Based on this standard, only three items (item 3, 5 and 10) were evaluated as the very good or relatively good items with the *d*-value of 0.412, 0.367 and 0.706, respectively. Five items were evaluated as either marginal (item 1, 8, and 12) or poor items (item 2 and 6) and thus presumably need to be deleted or revised. Lastly, four items (item 4, 5, 9 and 11) were almost non-discriminating or negatively discriminating with the *d*-value of −0.417, −0.249, −0.021 and −0.189 respectively. Overall, nine out of twelve items were labeled as marginal, poor and negatively discriminating items due to their low or negative *d*-values. Our conjecture is that the test might have been simply too difficult for all students. Both the high-scoring and the low-scoring group seem to have missed the same questions. Another possible scenario is that the low-scoring group might have scored some items correctly by chance, while the high-scoring group still missed the items.

To decide whether to delete or keep items, we referred to the "alpha if item deleted" and compared the recalculated alpha with the original alpha of 0.343. Although items 2 and 6 were evaluated as poor items with the *d*-value of 0.133 and 0.106 respectively, we decided to keep them in our test in that "the alpha if item deleted" rather decreased to 0.317 and 0.328 for item 2 and 6, respectively. These figures were slightly smaller than the original alpha of 0.343 and thus deleting these items would not help to increase the Cronbach alpha for the reading test. The same thing was true for the rest marginal items so we decided to keep item 1, 8, and 12.

When it comes to such questionable items as 4, 5, 9 and 11, more analyses are necessary to examine why the *d*-values turned out to be negative. In item 4, an inference question students had to infer a meaning of an expression in a context, it turned out that the lowest scorer got this question correct, while the highest missed the question. No consistent pattern was found among the middle group. The "alpha if the item deleted" went up to 0.516, which was much higher than the original alpha of 0.343. Since the item was considered to harm the test reliability with a negative discrimination index, we decided to eliminate item 4. Item 5 was a detail question that asked students to find information explicated in the text. Although searching for the explicit information was assumed to be an easy type of question, complex sentence structures of the text might have confused many students. The *p*-value of this item was 0.333, meaning that the question itself was too difficult so that only four students out of twelve scored

correctly. Since the "alpha if the item deleted" increased to 0.461, we decided to delete the item. Item 9 was another detail question that asked students to find the pronoun referent within a paragraph. Item 11 was a gist question that asked about the main idea of a paragraph. The negative *d*-value of these two items seemed to be due to their high *p*-values. In other words, the *p*-values of 0.8333 and 0.9167 for each item suggest that most of the students scored them correctly and thus the high and low groups were not properly distinguished. Given that the test was an achievement test and 70% of the students were expected to answer the questions correctly, we decided to keep the items despite the negative discrimination indices. Furthermore, the "alpha if the item deleted" for item 9 and 11 amounted only to 0.367 and 0.396, respectively. These figures were only a little larger than the original alpha of 0.343, compared to items 4 and 5 with the "alpha if item deleted" of 0.516 and 0.461. Based on these considerations, we decided to keep items 9 and 11.

All in all, we finally eliminated item 4 and 5 from the reading test and calculated the new Cronbach alpha (Table 10) and the new *SEM* (see Table 11). Consequently, the Cronbach alpha went up to 0.599 from 0.343. This implies the degree to which the items that relate to each other became higher, subsequently increasing the internal consistency. Likewise, the new *SEM* decreased to 1.324 from 1.551, which may also evidence the increased internal consistency. Taking participant #3 for example again, her raw score was 7, but the score was to vary between 4 and 11 (3.127 and 10.101 rounded due to the dichotomous scoring) when the *SEM* was 1.551 with a 95% confidence interval (±2 *SEM*s). Now, with the recalculated *SEM* of 1.324, her score would fall in between 5 and 10 (4.353 and 9.647 rounded due to the dichotomous scoring), if the test were repeated. Since the expected range between the lowest and the highest score with a 95% confidence interval (±2 *SEM*s) slightly decreased from 7 (11−4=7) to 5 (10−5=5) with the new *SEM*, it seems safe to say that the internal consistency of this test improved, though the range of 5 could be still large for this short test with the total score of 12.

Table 10
Internal Consistency Coefficient Revised

Cronbach Alpha	Number of Items
0.599	10

Table 11
Standard Error of Measurement for the Reading Test Revised

$*SEM = S\sqrt{1 - rxx}$
$SEM = 2.09 * 0.6332 = 1.324$

*where S = standard deviation (retrieved from the descriptive statistics) and rxx = reliability estimate for the test, which is equal to the Cronbach's alpha coefficient.

Distractor Analysis

We also performed a distractor analysis to evaluate the quality of the individual items and to see whether they correctly discriminated the high group from the low group. The discrimination index was calculated by comparing the high 27% group and the low 27% group in their responses to the key answers and other distractors. The three top-scoring students were separated from the three bottom scoring students, as three is approximately 27% of twelve. Those who scored 9 and 10 points were selected as the high group, while those who scored 3, 5 and 6 were treated as the low group. Since there were four students who received 6, one was

randomly selected among the four and consistently used for the distractor analysis across different items. To calculate the discrimination index, the number of the high-scoring students that answered the item correctly was subtracted by the number of the low-scoring students that answered the same item correctly and then divided by the number of the high group students.

The value of this index is scaled from −1 to 1; the value of 0 indicates that there is no discrimination. The ideal value for the key answer is 1 or positive at least, while the value for distractors should be −1 or negative. The formula to calculate the discrimination index is presented in Table 12.

Table 12
The Formula for The Discrimination Index for The Distractor Analysis

D represents the discrimination index: *Nch* stands for the number of the high-scoring students who got an item correct, *Ncl* means the number of the low-scoring students who got the item correct and *Nh* means the number of the high group students, the formula for the discrimination index is, *D= (Nch – Ncl)/Nh*

Item 4 (Inference) with a d-value of −0.417 and item 10 (Detail) with a d-value of 0.706 were chosen for the distractor analysis to investigate what led to the discrepancy. Based on the formula above, the distractor analysis for item 4 was summarized in Table 13.

Table 13
Distractor Analysis for Item 4

Question Type	Answer	High 27% *N*=3	Low 27% *N*=3	Total Count	Total %	Discrimination Index	Difficulty Factor
Key	a	0	0	0	0	0	0.667
Distractor	b	4	0	4	33	.33	
Distractor	c	2	3	8	67	−.33	
Distractor	d	0	0	0	0	0	

Item 4 was a negatively discriminating item, and thus we decided to delete it. The key answer was c and around 67% of the students answered this question correctly. The discrimination index shows that no one chose the distractor *a* and *d*, meaning that these distractors did not function well as intended. The entire low 27% group got this question correct, while one student from the high 27% group chose the distractor b. Consequently, the key answer turned out to be negatively discriminating, while the distractor *b* positively discriminating. This was a rather undesirable outcome in that ideally the discrimination index for the key answer should be a positive value or even 1 at the highest, while that of distractors should be a negative value or even −1 at the lowest. The undesirable function of the key answer and distractors in item 4 might have contributed to the negative *d*-value of −0.417 and the decreased internal consistency of the test. To improve the quality of the test, distractor *a* and *d* need to be replaced with more attractive

distractors; further revision is necessary to make the key answer *c* positively discriminating and make the distractor *b* negatively discriminating. Otherwise, it seems preferable to delete item 4 to increase the test reliability. Now we turn to item 10, which had a *d*-value of 0.706 and a *p*-value of 0.250. Table 14 summarizes the results.

Table 14
Distractor Analysis for Item 10

Question Type	Answer	High 27% *N*=3	Low 27% *N*=3	Total Count	Total %	Discrimination Index	Difficulty Factor
Key	A	3	0	3	25	1.00	
Distractor	b	0	0	1	8.3	0.00	
Distractor	c	0	2	5	42	−0.67	0.2500
Distractor	d	0	1	2	17	−0.33	
Other				1	8.3		

Although this item was evaluated as an extremely difficult item with the *p*-value of 0.250, the distractor analysis revealed that it properly discriminated the high group from the low group, with well-devised distractors. The entire high-scoring group chose the key answer, while the low and the medium group selected other distractors. Consequently, the discrimination index of the key answer turned out to be 1, meaning that the item perfectly distinguished the high group from the low group. Distractor c was the most attractive, in that 42% of the students responded to it, and 40% of the respondents were from the low group. Distractor b and d also appealed to around 8% and 17% of the students respectively, but not to any in the high group. That being said, all distractors seem to have reasonably served their purpose.

Evidence for Construct Validity with the MC Task

Finally, the correlations among reading variables were examined to assess the construct validity of the MC items. Construct validity pertains to the question of the extent to which a test measures the underlying psychological constructs of the test. Earlier in the paper, we decided to have gist, detail and inference variables to estimate the reading construct. That is, the three variables should be correlated with one another, as they all measure the same underlying construct. The Pearson product-moment correlation was computed; the range of the Pearson correlation coefficient ranges from +1 to −1. A positive value indicates a direct, linear relationship between the variables while a negative value indicates an inverse relationship. According to Brown (2005), there is a high correlation between the two variables when the coefficient equals to 0.75 or above, a moderate correlation when it falls between 0.5 and 0.74, a low correlation when it comes between 0.25 and 0.49. If the coefficient is below 0.25, it is safe to say that the variables are uncorrelated. When the correlation coefficient is close to 0, in either a positive or a negative figure, it indicates little or no correlation between the variables. Using these standards, we summarized the correlation analyses in Table 15.

Table 15
*Correlation Matrix between Variables for the Reading Test (*K=*12,* N=*12)*

Scale	Gist	Detail	Inference
Gist	1		
Detail	0.142	1	
Inference	0.369	0.367	1

Note: * indicates significance at the α=0.05 level (2-tailed)

A low correlation of 0.367 was found between inference and detail; another low correlation of 0.369 between inference and gist; near-zero correlation of 0.142 between detail and gist. Such a low or no correlation among reading variables refutes the sound construct validity of the test. In light of the generalizability of the correlation coefficient, each correlation coefficient turned out to be statistically insignificant. In considering that the observed correlations were smaller than the critical value of 0.576 (*df*=10) at the 0.05 level, we cannot rule out the possibility that the results presented in Table 15 were possibly due to chance.

Since items 4 and 5 were judged to have depressed the test reliability with the lowest d-value, we took out the two items from the analyses. After the deletion of the two items, the correlations among variables of the revised test were recalculated as shown in Table 16.

Table 16
*Correlation Matrix Between Variables for the Revised Test (*K=*10,* N=*12)*

Scale	Gist	Detail	Inference
Gist	1		
Detail	0.310	1	
Inference	0.442	0.497	1

Note: * indicates that correlation is significant at the α=0.05 level (2-tailed).

As a consequence, the magnitude of correlations among the variables slightly increased. A somewhat moderate correlation was found between inference and detail with the coefficient of 0.497. Still, a low correlation was estimated between gist and inference with that of 0.442 and between gist and detail with that of 0.310. Again, the correlational evidence among reading variables from the revised test was not sufficient to verify the construct validity of the reading test. We cannot guarantee the generalizability of the test result, since the observed correlations were not statistically significant at the 0.05 level.

Discussion and Conclusions

The purpose of this study was to demonstrate how to design the reading test, analyze the results, and evaluate the quality of the test. Given that it was an achievement test, its purpose was to measure the extent of learning or mastery within a specific instructional domain. Based on the theoretical model of the reading construct and the course syllabus specific to the A2 evening class, we developed 12 MC items for the reading test. We expected our test to correctly measure the underlying construct of reading ability. Although the topic of the passage was intended to correspond to the class theme, we tried to make the items not susceptible to their topical knowledge; to answer

questions, test-takers needed to closely read the passage.

Overall, the reading test turned out to be somewhat difficult for the students, in that the means, medians and modes of the construct did not meet the general standards of those of an achievement test. Even so, the test scores were rather normally distributed, indicating that the participant group was not as homogenous as we expected in terms of their reading ability. The results might suggest that the CEP placement test failed to place them according to their true language abilities, thereby calling for test improvement. Otherwise, it may also be that the participants were not motivated enough to do their best on the examination. All of the participants were adult ESL learners with a high level of general education, having at least a bachelor's degree. They voluntarily attended the CEP program to develop their general English ability, and hence they might not have felt much pressure about taking the test.

When it comes to evaluating the reliability and the construct validity of our reading test, by performing the item analysis, nine out of twelve items were evaluated as either marginal, poor, or even negatively discriminating items in our pilot test. Taking into account the "alpha if item deleted" and the p-values, we decided to delete two items. As a result, the Cronbach's alpha for the reading test increased, but still no statistically sufficient evidence was found for the construct validity. Thus, the reading test might not have been as successful in correctly measuring the underlying reading construct.

The undesirable outcome of the reading test seemed partially due to its elicitation method. While a writing or speaking task is a relatively direct test task, where test-takers are required to do the actual skill, the MC items are devised to indirectly assess the intangible construct, reading ability. Thus, it is questionable if such items can actually tap into test-takers' true reading ability. Murphy et al. (1998) also indicates the fragility of the evidence surrounding reading assessment. Given that reading itself is a "complex and multifaceted process (p. 6)," it must be extremely challenging to access the abstract construct precisely.

Limitations

One of the main limitations, as mentioned several times earlier, was the small number of participants in the study (N=12) and limited number of items given on the test (K=10). These limited numbers could have been a factor that restricted evidence for the validity and generalizability of the test. Moreover, the range of ability among the participants was presumably rather narrow in that they were in the same level of CEP classes, limiting the variability of possible scores. A small range of variability can depress the correlation coefficients, and as a result, bring down test validity and generalizability.

Another limitation of the test is that, although the test items were created based on the CEP course syllabus to measure the participants' level of achievement, their scores did not reach a level that is generally expected in an achievement test. More specifically, an achievement test generally brings about an average score of 70% (which is also the cut-off score for CEP students when they advance to the next level), whereas our reading test average was only 56.25%. This figure could mean that the participants performed poorly overall, but on the other hand, it can also imply that the difficulty level of the overall test was rather high for the participants, or

even that the test was not an adequate representation of what they learned up to the mid-term exam.

There are some possible improvements we would make to the process of this project were we to administer it again. First of all, we would try to adjust the difficulty level of test, double-checking whether the items accurately reflect the course contents so that it would better serve as an achievement test. Closer communication with the teacher during the process of the test creation could help in carrying out this goal. It would also be helpful to administer a trial test with the items or have peers review the items to receive specific feedback before using the test. Finally, having a larger pool of participants would definitely help to obtain more reliable statistics when analyzing the test results.

References

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. Charles Alderson & A. H. Urquhart (Eds.), *Reading in a Foreign Language* (pp. 1–27). London: Longman.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Anderson, N. J. (1999). *Exploring second language reading: Issues and strategies*. Boston, MA: Heinle & Heinle Publishers.

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*(4), 460–472.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133–150. doi:10.1017/S0267190505000073

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

Carrell, P. L., & Eisterhold, J. C. (1983). Schema Theory and ESL Reading Pedagogy. *TESOL Quarterly, 17*(4), 553–573.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19*(4), 727–752.

Carver, R. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Scientific Studies on Reading, 1*, 3–43.

Coady, J. M. (1979). A psycholinguistic model of the ESL reader. In R. McKay et al. (Eds.), *Reading in a second language* (pp. 5–12). Rowley, MA: Newbury House.

Cohen, A., & Upton, T. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL. *Language Testing, 24*(2), 209–250.

Cummins, J. (1991). Conversational and academic language proficiency in bilingual contexts. *AILA Review, 8*, 75–89.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450–466.

Ebrahimi, S. S. (2012). Reading strategies of Iranian postgraduate English students living at ESL context in the first and second language. *International Conference on Education and Management Innovation, 30*, 195–199.

Educational Testing Services. (2000). Test of English as a Foreign Language. Retrieved from

http://www.ets.org/Media/Research/pdf/RM-00-04.pdf

Fitzgerald, J. (1995). English-as-a-second-language learners' cognitive reading processes: A review of research in the United States. *Review of Educational Research, 65*(2), 145–190. doi:10.3102/00346543065002145

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1–13). Mahwah, NJ: Erlbaum.

Goodman, K. S. (1973). *Theoretically based studies of pattern of miscues in oral reading performance (Final Report Project No. 9-0375).* Washington, DC: US Department of Health, Education and Welfare, Office of Education, Bureau of Research.

Goodman, K. S. (1994). Reading, writing, and written texts: A transactional socio-psycholinguistics view. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (pp. 1093–1130). Newark, DE: International Reading Association.

Grabe, W. (1991). Current development in second language reading research. *TESOL Quarterly, 25*(3), 375–406.

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading*. Harlow, England: Pearson Education.

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics, 24*, 44–69.

Hoover, W. A., & Gough, P.B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127–160.

Judith, R. (1995). Student responses to reading strategies instruction. *Foreign Language Annals, 28*(2), 262–273.

Khalifa, H., & Weir, C. (2009). Examining Reading: Research and practice in assessing second language reading. *Studies in Language Testing (Vol. 29)*. Cambridge, England: UCLES/Cambridge University Press.

Kitao, S. K. (1989). *Reading, schema theory and second language learners*. Tokyo: Eichosha Shinsha.

McGinley, M. (1992). The role of reading and writing while composing from sources. *Reading Research Quarterly, 27*(3), 226–248.

Murphy, S., Shannon, P., Johnston, P., & Hansen, J. (1998). *Fragile evidence: A critique of reading assessment*. Mahwah, NJ: Erlbaum.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology, 8*, 293–316.

Paris, S., & Myers, M. (1981). Comprehension monitoring, memory, and study strategies of good and poor readers. *Journal of Literacy Research, 13*(1), 5–22. doi:10.1080/10862968109547390

Savery, N. (2012). Targeted reading comprehension strategies instruction for raising reading levels in tertiary contexts. *Journal of Academic Language & Learning, 6*(1), 32–47.

Sheorey, R., & Mokhtari, K. (2001). Differences in the metacognitive awareness of reading strategies among native and non-native readers. *System, 29*(4), 431–449. doi:10.1016/S0346-251X(01)00039-2

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford Press.

Appendix A
Test Task Specifications

	Task 1 Multiple choice
SETTING	
Physical characteristics	Location: room HM 136 at Teachers College. Noise level: low to moderate depending if the door is open. Temperature and humidity: cool and moderate in humidity. Seating conditions: each test taker has his/her own seat in an auditorium type classroom. Lighting: well lit. Materials and equipment an degree of familiarity: pens or pencils, paper provided, students refer to clock on front wall to keep time.
Participants	The CEP teacher and the students
Time of task	During class hours at 7 PM on Thursday, October 23, 2008.
INPUT	
Format	
Channel	Visual
Form	Language
Language	Target: English as a second language
Length	Instructions: one to two sentences, Reading passage: ten paragraphs
Type	Item: elicit selected response
Speededness	Unspeeded
Vehicle	Live
Language characteristics	
Organizational characteristics	
Grammatical	Vocabulary: general Morphology and syntax: standard English Graphology: typewritten
Textual	Cohesion: cohesive Organization: focused discussion and analysis
Pragmatic characteristics	
Functional	Ideational, manipulative, and heuristic
Sociolinguistic	Dialect/variety: standard Register: formal Naturalness: natural Cultural references and figurative language: related to topic
Topical characteristics	Restricted: education issues in the US presidential candidate debate

EXPECTED RESPONSE	
Format	
Channel	Visual
Form	Non-language; circling the correct letter
Language	Target: English as a second language
Length	Short: 12 MC items
Type	Selected response
Speededness	Generally unspeeded
Language characteristics	
Organizational characteristics	
Grammatical	Vocabulary: general Morphology and syntax: standard English Graphology: circled responses
Textual	Cohesion: cohesive Organization: extended discussion and analysis
Pragmatic characteristics	
Functional	Ideational and heuristic
Sociolinguistic	Dialect/variety: standard Register: formal. Naturalness: natural Cultural references and figurative language: related to topic
Topical characteristics	Restricted: education issues in US presidential candidate debate
RELATIONSHIP BETWEEN INPUT AND RESPONSE	
Reactivity	Non-reciprocal
Scope of relationship	Broad to work with the general gist and inference questions. Narrow to work with the vocabulary and grammar in context questions
Directness of relationship	Direct

Appendix B

Test-taker Survey

A. Personal characteristics

1. Age:
2. Gender: M / F
3. Nationality:
4. Native language:
5. How long have you been in the United States? _____
6. What is the level of education that you completed at the most recent years?
   a. elementary b. secondary c. undergraduate d. graduate e. post-graduate
7. Are you planning to go to college or graduate school, or find a job in the United States? Y / N

B. Topical knowledge

1. How often do you read an American newspaper?
   a. everyday b. every either day c. once a week d. once a month e. never
2. How many hours do you spend in reading a newspaper?
   a. less than half an hour b. half an hour c. one hour d. two hours e. more than two hours
3. Are you interested in 2008 US presidential election?
   a. very much b. interested c. only a little interested d. not interested
4. What did you major in, if you have a bachelor degree? _____

C. Levels and profiles of language knowledge

1. How much time have you spent studying English (in a secondary or post secondary school)? _____
2. Have you ever taken any standardized English exam (e.g., TOEFL, TOEIC) before? If so, which test was it and what was your score? _____

D. Possible affective responses to taking the test

(5 = strongly agree , 1 = strongly disagree)

I was nervous while taking the mid-term exam ……. 5 4 3 2 1

I am familiar with the types of questions in the reading section. ……. 5 4 3 2 1

I am familiar with the type of writing question ….…….. 5 4 3 2 1

I felt the level of the reading questions was difficult …….. 5 4 3 2 1

I felt the level of the writing question was difficult …….. 5 4 3 2 1

E. Reflecting on the test questions.

1. How did you solve the following question in the reading section?
   a. I already knew this information from the media (e.g., newspaper or TV).
   b. I skimmed the reading passage to find the information.
   c. I just guessed randomly.
   d. Other ways _____

9. In his presidential campaign on the issue of education, John McCain: a. disagrees with the idea of NCLB. b. suggests more grants for preschool programs. c. wants to reward high-achieving teachers with federal money. d. plans to increase government funding for independent schools.

2. How did you solve the following questions in the reading section?
   a. I already knew the meaning of the word before taking this test.
   b. I inferred the meaning from the content of the passage.
   c. There is a similar word in my first language.
   d. I just guessed randomly.
   e. Other ways _____.

6. In line 12, what does “glum” mean? a. puzzling b. convincing c. discouraging d. self-explaining	12. What is the meaning of “diluted” in line 46? a. less effective b. risky to carry out c. more troublesome d. difficult to clean up

3. How do you rate your reading ability in your first language?
   a. Advanced
   b. High-intermediate
   c. Low –intermediate
   d. Beginner

4. How do you rate your writing ability in your first language?
   a. Advanced
   b. High-intermediate
   c. Low –intermediate
   d. Beginner

5. In the writing section, was it helpful to have the planning chart before writing the essay?
   a. Very helpful
   b. Somewhat helpful
   c. Only a little helpful
   d. Not helpful

Appendix C

Mid-term Evaluation for CEP A2 Evening Class

Name: _____

Instructor: Abbi Leman (A2 Evening)

Date: Oct. 23, 2008

READING SECTION

You have 30 minutes to complete the following reading tasks.

Directions: Read the passage. Circle the correct letter.

"OUR nation is at risk. Our once unchallenged pre-eminence in commerce, industry, science and technological innovation is being overtaken by competitors throughout the world." So reported an education commission in 1983. That report was a turning point for American schools, helping spur a wave of reform. But 25 years later the state of American education is in a muddle.

In some ways its public schools have improved. America's nine-year-olds scored 22 points higher on a national maths test in 2004 than they had in 1982. But in many areas America still languishes, as described in a recent report by Ed in '08, an advocacy group. The percentage of 17-year-olds with basic reading skills has dropped, from 80% in 1992, when the current test was introduced, to 73% in 2005. On the international stage, American students are doodling while others scribble ahead. The Organization for Economic Co-operation and Development has a glum statistic: in the most recent ranking of 15-year-olds' skill in maths, America ranked 25th out of 30. Though America's universities remain pre-eminent in the world, they have grown increasingly unaffordable. Barack Obama notes that between 2001 and 2010, two million qualified students will not go to university because they cannot afford it.

Efforts to move America forward have proceeded inconsistently. A federal bill, No Child Left Behind (NCLB) was passed with broad support in 2002, the culmination of a long push to set high standards and hold schools accountable for meeting them. It requires states to test students on maths and reading; science is being added. Schools that do not progress towards meeting state standards face financial sanctions.

But the law is hotly debated. George Miller, a Democratic congressman, calls NCLB "the most negative brand in America"—and he was one of the law's architects. Teachers' unions utter no four-letter word with more anger than NCLB. They say the law forces "teaching to the test", that the sanctions are too strong and the carrots too small. Even those who still support the law find problems with it. NCLB, for example, does not chart a student's progress.

Some states have set their standards very low. Some 90% of Mississippi's fourth-graders were labeled "proficient" or better on a state reading test in 2007; only 22% were so described after a national test.

Unsurprisingly, advocates from all corners are trying to make education a main campaign issue. Ed

in ’08 points out that many of the proposals from “A Nation at Risk” have been ignored: standards remain weak, few districts pay teachers by results and calls for a longer school year have gone disregarded. But despite a budget of $60 million, Ed in ’08’s campaign has had little impact.

Mr. Obama is at least taking the problem seriously. His plans run the gamut, from grants for preschool programs to a $4,000 tax credit for university fees. He is vague about NCLB, but has resisted calls to throw out the law. He suggests improving it through more sophisticated tests, measuring students’ progress over time and giving schools more resources. In September he announced new plans to double federal funding for independent or “charter” schools. A separate “innovative schools fund” would help districts to create a portfolio of successful school types, including charters.

Perhaps most interesting are his plans for teachers. He would give extra money to districts that work with their unions to form “career ladders”. These could include pay increases for a list of achievements, from teaching in hard-to-staff schools to lifting students’ performance.

But a good scheme on paper may be diluted in practice. Negotiations over pay are messy at best.

For his part, Mr. McCain offers promising opinions but few details. He supports NCLB but has said little about how to strengthen its main tenets. He supports charter schools (like Mr. Obama) and voucher programs (unlike Mr. Obama, who is dead-set against them), but has said little about how he might expand them. His boldest ideas center around using federal money to let parents choose tutors and principals reward good teachers.

In the debate over how a president might help America’s schools, a main obstacle is that, traditionally, it has not been his job to help them much at all. The national government provides less than 10% of total spending on schools. Indeed, states and cities continue to be the boldest innovators. Chicago is opening dozens of new schools, including charter schools, in its poorest areas. Cities such as Denver and New York now have schemes to reward teachers for their skill. The results there are mildly encouraging.

The two candidates offer different plans for how they might push these reforms along. Both, however, have largely overlooked the most obvious role. At the very least, the next president could help to create a better benchmark for student achievement. As Mississippi proves all too well, a state standard can be an elastic ruler.

What is the best title for the passage? - GIST

a. Under NCLB, even strong schools falter
b. Can the candidates fix America's decidedly mediocre schools?
c. Can school equity be achieved with a larger education budget?
d. Discrepancies between McCain and Obama over education policies

1. What is the author's overall tone in the passage? - INFERENCE

a. Ironic.
b. Neutral.
c. Critical.
d. Hopeful.

2. Why does the author mention the education commission in 1983 in the beginning? -INFERENCE

a. To point out the effects of American education on other social areas since 1983
b. To emphasize that American education has been a problem for the past 25 years
c. To give an example of the efforts that a government made to improve education
d. To relate the event to the education policies that two presidential candidates suggest

3. What does it mean to "be in a muddle" in line 5? - INFEERENCE

a. be in mental stress
b. lack attention to details
c. be in a disorderly condition
d. have no sense of responsibility

4. What is true according to the 2nd paragraph (lines 6–16)? - DETAIL

a. Public schools in America have made overall improvement.
b. America still has a relatively good international ranking in math skills.
c. Math abilities of nine-year-old children enhanced significantly by 2004.
d. Even American universities are falling behind in terms of academic competence.

5. Why does the author mention George Miller in line 22? - INFERENCE

a. To reveal the controversy of the NCLB debate
b. To give a specific example of one limitation of NCLB
c. To provide evidence of how strongly NCLB is opposed
d. To suggest that Miller would be able to improve NCLB

6. Which of the following would best replace the word "carrots" in line 25? - INFERENCE

   a. support
   b. rewards
   c. challenge
   d. standards

7. In his presidential campaign on the issue of education, John McCain: - DETAIL

   a. disagrees with the idea of NCLB.
   b. suggests more grants for preschool programs.
   c. wants to reward high-achieving teachers with federal money.
   d. plans to increase government funding for independent schools.

8. In line 39, what does "it" refer to? - DETAIL

   a. NCLB
   b. the law
   c. the gamut
   d. the problem

9. Based on the facts in the 5th paragraph (lines 28-30), which of the following correctly rephrases "a state standard can be an elastic ruler" in line 62? - DETAIL

   a. Standards set within the state can be misleading
   b. The state legislators can be flexible in law making
   c. States can set standards that increase student performance
   d. Sometimes states can measure students upon a rigorous standard

10. The main point of the 10th paragraph (lines 53-58) is: - GIST

    a. Future American president needs to allow more budgets for schools.
    b. Efforts for better education have been made mostly at the state level.
    c. The national government is planning to give teachers more incentives.
    d. Many states are against education policies that the federal government suggests.

11. What is the best conclusion of the passage? - GIST

    a. Teachers and government officials must all cooperate toward improving the effects of NCLB.
    b. Enhancing the quality of education in America will be a major job for the next president.
    c. States and cities should take the more initiative role to improve schools and student performance.
    d. The first step towards reform can be made if the national government increases the funds for education.