A Reading and Writing Placement Test: Design, Evaluation, and Analysis

Hyun Jung Kim[1] and Hye Won Shin[2]
Teachers College, Columbia University

ABSTRACT

Placement tests, along with the growing interest in their validation, have become increasingly important in English as a Second Language programs. To this end, the present paper illustrates procedures in designing a placement test and using it to evaluate students' language ability by means of statistical analysis. 29 participants from three proficiency levels (beginning, intermediate, and advanced) took reading and writing placement test sections. Students' performance at each proficiency level was analyzed separately and compared across proficiency levels. In addition, analyses of responses on a survey revealed a relationship between language learning attitudes and behaviors exhibited by the participants and their performance on the placement test. Through close examination of the process of placement test design, evaluation, and analysis, this paper provides practical guidelines for reading and writing placement testing.

INTRODUCTION

As Brown (2005) emphasized, a placement test should efficiently separate students into appropriate levels with high reliability and validity. Unfortunately, teachers are often unaware of placement assessment processes. As a result, they often fail to balance their focus on assessment and teaching itself, which should be always considered together. The current study provides a practical example of the whole process of placement testing including design, evaluation, and analysis, an area often overlooked by teachers.

Thus, the purpose of our study is threefold: (a) to develop an ESL placement test that measures reading and writing abilities effectively, (b) to evaluate the test results to determine whether the test achieves its objective in grouping students into appropriate levels, and (c) to investigate the determinants of reading and writing abilities based on a questionnaire by employing ANOVA and multiple regression analysis. In the following sections, we present a conceptual framework for designing reading and writing sections of a placement test that is based on prior research. We then discuss the planning and development of the reading and

[1] Hyun Jung Kim is a doctoral student in Applied Linguistics at Teachers College, Columbia University. Her research interests include second language assessment and applied psychometrics. Correspondence should be sent to Hyun Jung Kim, 37 Bergen Blvd. Apt #1, Fairview, NJ 07022. Email: hjk2104@columbia.edu.
[2] Hye Won Shin is a doctoral student in Applied Linguistics at Teachers College, Columbia University. Her research interests include second language acquisition and assessment. Correspondence should be sent to Hye Won Shin, 3241 S. Sepulveda Blvd. Apt. 202, Los Angeles, CA 90034. Email: hwyu73@hotmail.com.

writing placement tests, including the target language use domain, design statement, item coding for the multiple choice (MC) section, and administration procedures. Finally, we present statistical analyses and discussion of the study results.[3]

LITERATURE REVIEW

Reading Ability

Reading can be defined as the interaction between the reader and the text (Aebersold & Field, 1997). This dynamic relationship portrays the reader as creating meaning of the text in relation to his or her prior knowledge (Anderson, 1999).

Much research has been done on how to assess L2 learners' reading ability. Weir (1997) introduced two distinctive views of reading for assessment: the unitary and multidivisible views of reading. In the unitary approach, expeditious, quick, purposeful, and efficient reading ability as a whole is evaluated. In the multidivisible approach, on the other hand, "if specific skills, components or strategies could be clearly identified as making an important contribution to the reading process, then it would of course be at least possible, if not necessary, to test these and to use the composite results for reporting on the reading proficiency revealed" (p. 44). Based on this notion, microlinguistic test items are used to measure different reading skills.

A skills approach, which is compatible with the multidivisible view of reading, has been influential in L2 reading assessment, although the presence of separate subskills is still debatable (Alderson, 2000; Weir, 1997). Numerous subskills identified under reading ability may serve as operational definitions of reading ability for a given testing context. For example, Diagnostic Language Testing (DIALANG) and the First Certificate in English (FCE) include four main reading features: (1) identifying main idea(s), (2) understanding detailed information, (3) inferring meaning, and (4) lexical inferencing from context, while the International English Language Testing System (IELTS) includes only the first three features (Alderson, 2000). In addition, the English for Academic Purposes (EAP) test, developed to provide diagnostic information about non-native speaking students at the University of Melbourne, also adopted the four features of the reading construct specified in DIALANG and FCE (Lumley, 1993).

Along with operational definitions, researchers must decide upon the testing method at the test design stage to collect relevant information about test-takers' reading ability. Alderson (2000) listed a number of test techniques or formats often used in reading assessments, such as cloze tests, multiple-choice techniques, alternative objective techniques (e.g., matching techniques, ordering tasks, dichotomous items), editing tests, alternative integrated approaches (e.g., the C-test, the cloze elide test), short-answer tests (e.g., the free-recall test, the summary test, the gapped summary), and information-transfer techniques. Among the many approaches to testing reading comprehension, the three principal methods have been the cloze procedure, multiple-choice questions, and short answer questions (Weir, 1997). In an attempt to identify the effectiveness of these three different test methods, Wolf (1993) found that test-takers performed better on multiple-choice items than on the other two methods. The following explanations have

[3] Procedures used in this paper are based on Dr. James Purpura's Second Language Assessment course (A&HL 4088).

been put forth to account for the difference: (a) open-ended and cloze tasks require the presence of language production skills, (b) open-ended and cloze tasks require added memory skills, and (c) multiple-choice tasks allow for guessing. Wolf concluded that test methods differentiate test-takers' ability to demonstrate their comprehension and that different methods measure different abilities.

The existence of testing method effects demonstrates that a single test technique is limited in eliciting all aspects of a test-taker's reading ability. Consideration must be given to the individual, private nature of the reading process, and the various purposes for which the test is used. In other words, "there is no one 'best method' for testing reading" (Alderson, 2000, p. 203). One single method is frequently used for practical reasons, but often at the expense of test validity. To elicit better information about a test-taker's ability, multiple methods which conform to the construct being measured are necessary (Alderson, 2000; Alderson & Banerjee, 2002). In any case, as Alderson (2000) writes, "we should always be aware that the techniques we use will be imperfect" (p. 270).

Writing Ability

One way to understand writing is to examine its process as a cognitive skill that draws equally on language and cognitive resources. In an effort to assess L2 writing ability, researchers have presented various models that describe the writing process (for a detailed review, see Raimes, 1991). First, the form approach addressed a single concern: *grammatical form*. Early second language composition pedagogy mirrored the audiolingual method of instruction in second language teaching where writing served to reinforce oral patterns of the language and test the application of grammatical rules (Matsuda, 2003; Raimes, 1991). In other words, this particular teaching paradigm supplemented learners' use of forms by means of oral communication.

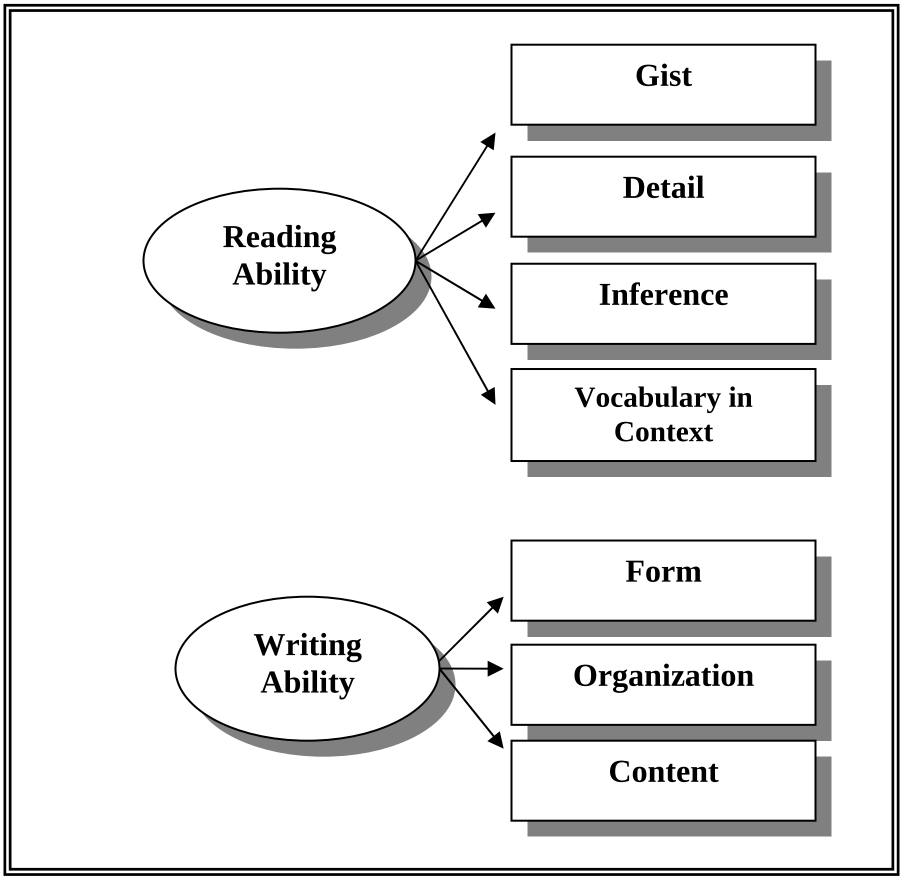
Second, the process approach was adopted by ESL teachers and researchers in reaction to the form approach to "develop an interest in what L2 writers actually do as they write" (Raimes, 1991, p. 409). Rather than concentrating on the finished product, the definition of writing expanded to include many of the recursive steps taken in order to complete a writing task while promoting conscious thought about writing and incorporating and evaluating different cognitive strategies and metacognitive processes used by writers. These steps included the stages of planning, writing, revision, and editing (Cumming, 2001). Hence, the process approach emphasized *organization* rather than grammatical form.

Another approach, the content approach, was put forth by the researchers and practitioners who argued that the process approach failed to meet the needs of non-native students (Raimes, 1991, Horowitz, 1986). For example, Horowitz claimed that the process approach fell short in preparing students for academic writing tasks such as laboratory reports. As a result, students were essentially left unprepared and unable to meet the expectations of the target academic culture. Hence, experts proposed a shift towards teaching and researching academic discourse genres, this time focusing their attention on *content*.

In addition to difficulties with choosing the appropriate construct for L2 writing ability, researchers have faced problems with the design and validation of scoring schemes (Garrett, Griffiths, James, & Scholfield, 1995). For instance, holistic scoring schemes are generally considered to result in less specific information about a student's test performance than analytic scoring schemes. Raters themselves, experienced and inexperienced, have been the subject of investigation in the last decade (Cumming, 1990). For example, studies have shown that L2 speakers tend to interpret test scores much more severely than native speakers (Alderson & Banerjee, 2002). Examinations of these rating behaviors show the consequences of raters' severity and leniency for performance scores.

Based on the literature review regarding reading and writing ability in an L2 setting, we constructed a framework, depicted in Figure 1, which can be used to develop an ESL placement test. Hence, we decided to measure and define reading ability as gist, detail, inference, and vocabulary in context, while writing ability as form, organization, and content.

FIGURE 1
Conceptual Framework of L2 Reading and Writing Ability



PLANNING AND DEVELOPMENT OF THE TEST[4]

Target Language Use Domain

The Community English Program (CEP) at Teachers College, Columbia University in New York City is open to a wide range of individuals interested in learning English for communicative purposes. Since there is a substantial degree of heterogeneity across students in terms of their English ability, placement testing is important for the program; identifying learners' proficiency levels and placing them in appropriate classes are imperative tasks for the efficiency and effectiveness of the CEP.[5] Through this project, we developed new items for the CEP placement test.

Given the diversity of backgrounds and language needs among the CEP students, it was difficult to define a single context of the target language use domain. Likewise, it was not possible to list the numerous target language use settings. These two factors prompted us to select the history of the state of New York as the content domain for the reading section. We speculated that CEP students would have a relatively homogenous level of prior knowledge of the content, regardless of the number of years they had been living in New York City or in the United States. By targeting this content domain, we controlled, to some extent, for the influence of each student's prior background knowledge on performance.[6] On the other hand, the content of the writing section was chosen to reflect an everyday situation: writing an informal postcard.

As both the reading and writing sections were part of the placement test, no target proficiency level was specified. However, as the purpose of this study was to differentiate students with respect to their L2 reading and writing ability, an intermediate proficiency level was the target in selecting and developing the reading and writing items.[7]

Test Structure

The purpose of the test was to gain information on the test-takers' English communicative ability in the area of reading and writing. To that end, we presented two reading tasks comprised of 12 multiple-choice items in total. Students were given 20 minutes to complete the tasks. Each item was scored as 0 or 1; students who answered all questions correctly received an overall reading score of 12. In addition, students were given another 20 minutes to complete an informal, descriptive writing task. The writing task was scored based on the sum of three criteria, each scored on a 5-point rating scale. Table 1 outlines the structure of the two test sections.

[4] For the overview of our placement test, see the Design Statement (Appendix 1).

[5] The CEP administers its placement test upon enrollment. Based on the scores, learners are placed into 12 levels, ranging from basic (levels B1 to B4) to intermediate (levels I1 to I4) to advanced (levels A1 to A4). The current CEP placement test is composed of five sections: grammar, reading, writing, listening, and speaking.

[6] We admit that this content might not be most representative of what CEP students would be required to read in everyday life.

[7] In practice, a placement test should be augmented by including questions developed for all proficiency levels (i.e. beginning, intermediate, and advanced) in order to maximize the discriminatory power of the test.

TABLE 1
Description of Test Structure

Construct	Task type	No. of tasks	No. of items	Time	Scoring
Reading ability (Gist, Detail, Inference, Vocabulary in Context)	Selected response: multiple-choice task	2	6 items for each task (12 in total)	20 min.	-Dichotomous (0/1) -12 points available
Writing ability (Form, Organization, Content)	Extended production: informal, descriptive writing task	1	1 item	20 min.	-Rating scale (3 criteria) -Analytic scoring (5 points for each criterion/ 15 points available)

METHOD

Participants

Our test was administered to four intact CEP class levels: one beginning (B4), two intermediate (I4), and one advanced (A4). In total, 29 participants (23 females, 6 males) from these three proficiency levels participated in the study (*n*=8 for B-4, *n*=22 for I-4, and *n*=3 for A-4).

In addition to the writing and reading tasks, participants completed a student survey (see Appendix 2) that provided demographic information about them. Ages ranged from the teens to the early eighties, with the majority of participants in their thirties. Education level varied as well—from middle school to post-graduate—including 7 two-year college graduates, 5 four-year college graduates, and 11 graduate degree holders. Native languages were also diverse. Respondents included one Russian, one Brazilian, one Danish, two French speakers, three Polish, three Ukrainian, three Korean, six Japanese, and seven Spanish speakers. Finally, participants' length of English study ranged from five months to ten years.

Administrative Procedures

Before administering the test, the proctor explained the purpose of the study and the test structure. First, participants completed the background survey. Next, the proctor distributed the test booklets and specified the duration of the reading section. Students were not allowed to turn to the writing section until the 20-minute time limit was up. The writing section followed the reading section and students were given another 20 minutes to complete it. Test booklets were collected after the 40-minute test period was over.

Instruments

Participants completed three tasks**—**two reading tasks and one writing task. First, participants completed the reading tasks after reading articles on the history of New York State.

They were given 20 minutes to read two passages, *Growth of New York State* and *History of West Point Military Academy*, and answer 12 multiple-choice questions (see Appendix 3 for the instruments used for the reading tasks). The 12 multiple-choice items were designed to measure reading ability focusing on gist, detail, inference, and vocabulary in context. Table 2 shows the observed variables in each item.

TABLE 2
Multiple-Choice Item Coding for the Reading Tasks

Task	Item No.	Observed variables
1	1	Vocabulary in context
	2	Detail
	3	Vocabulary in context
	4	Inference
	5	Gist
	6	Gist
2	7	Gist
	8	Vocabulary in context
	9	Detail
	10	Detail
	11	Inference
	12	Inference

Next, participants were given 20 minutes to complete the writing task which asked them to describe the three most interesting places to visit in their home country. For the task, they were told to imagine they were writing a postcard to a classmate (see Appendix 4 and Appendix 5 for the instrument used in the writing task and a sample response written by an examinee who received perfect scores across the three domains of form, organization, and content).

Scoring Procedures

The multiple-choice reading test was scored dichotomously. Each item response was scored as either correct (1 point) or incorrect (0 points). The total score was obtained by adding the number of correct answers. A maximum score of 12 was possible.

For the writing test, an analytic scoring rubric covering form, organization, and content was used (see Appendix 6). It was adapted from the booklet for *On Target 1* (Purpura & Pinkley, 2000). The rubric consisted of five scales from "5=complete control" to "1=little or no control" for each of the three writing components. To reduce subjectivity in scoring the writing test, two experienced raters first established a norm in an effort to maintain consistency between themselves,[8] and then independently read and scored all 28 writing tasks. For each test-taker, scores on each of the three writing components were averaged across the two raters to obtain a composite score. The three averaged composite scores were then added together resulting in an

[8] Any future study should consider increasing the number of raters within the budget constraint.

overall score for the writing test. The highest possible score for the writing task was 15 with a maximum rating of 5 for each component (form, organization, and content).

RESULTS AND ANALYSES

Results for the Multiple-Choice Task: Reading Ability

Descriptive Statistics

In order to better understand the nature of the test and the comparative abilities of the test-takers, descriptive statistics were calculated by proficiency level and for the three levels combined. The following section is an analysis of the three different levels (beginning, intermediate, and advanced) in addition to an analysis of the performance of all test-takers on the placement test.

Table 3 shows the descriptive statistics of the reading test for all three groups as well as all levels combined. Overall, the mean score was 6.17 (out of a total possible of 12). By proficiency level, participants in the advanced group scored highest (*M* = 9.00) while the beginning group scored lowest (*M* = 3.75). Beginning and intermediate level test-takers had a larger spread of scores than advanced test-takers (as indicated by the range of scores and standard deviations), where distribution of scores was at the high end for the advanced group and at the low end for the beginning level group. It should be noted that the advanced group had only three participants.

TABLE 3
Descriptive Statistics of the Reading Ability Test by Proficiency Level

	Beginning 4	Intermediate 4	Advanced 4	All Levels
Number of Test-takers (*N*)	8	18	3	29
Total Items (*K*)	12	12	12	12
Mean	3.75	6.78	9.00	6.17
Mode	1.00	5.00	10.00	5.00
Median	4.00	6.50	10.00	6.00
Minimum	1.00	4.00	7.00	1.00
Maximum	6.00	10.00	10.00	10.00
Range	5.00	6.00	3.00	9.00
Standard Deviation (*SD*)	2.05	1.86	1.73	2.48
Kurtosis	-1.60	-1.46	N/A	-0.36
Skewness	-0.35	0.18	-1.73	-0.26

Negative kurtosis overall indicates a flat distribution of scores. This distribution demonstrates a considerable amount of variability or heterogeneity in terms of reading proficiency of test-takers because the scores are spread widely.

Overall, scores were negatively skewed, meaning that there were more high scores than low scores. Negative skew is not desired on a placement test, where the beginning level respondents scored relatively well and the advanced level test-takers did not find the test demanding enough. A slight positive skew was found for intermediate level participants, who found the reading section somewhat difficult.

Internal Consistency Reliability

Internal consistency informs us about the degree of relatedness of the items on a test. To estimate the internal consistency reliability of the reading MC items, *Cronbach's Alpha*, which is widely used with dichotomously scored items, was calculated.

The internal consistency reliability of the 12 reading items was found as neither very good nor very bad (see Table 4). Cronbach's Alpha was calculated as 0.631 for the 12 reading items. Reliability values ranged from 0 to 1, with 0 representing no reliability and 1 as perfect reliability. The estimated reliability value, 0.631, is not large enough to report that the items on the reading subtest showed a high degree of homogeneity. Instead, it is modestly reliable, meaning that the 12 items measured the construct with a moderate degree of consistency.

TABLE 4
Reliability Statistic for the Reading Test (*N*=29)

Cronbach's Alpha	N of Items
0.631	12

Two possible reasons might explain the moderate degree of the reliability estimate. First, it seems that this degree of reliability resulted from the limited number of items. As there were only 12 items on the test, each item influenced the reported estimate greatly. Another possible reason is the limited number of examinees. A sample size of 29 is rather small for statistical analyses. The number of items and the number of examinees do affect the degree of reliability. When the number of items and the number of examinees increase, the reliability estimate also increases. Thus, the reported moderate Cronbach's Alpha is likely due, in part, to the limited numbers of items and examinees.

Item Analyses for the MC Section

The means of the 12 items were interpreted as a measure of item difficulty. Expressed as *p*, the item difficulty provides information about the proportion of examinees who answered an item correctly. Item discrimination, the degree to which the items discriminate among examinees, was also calculated since item difficulty alone does not provide enough information in terms of which test-takers get an item correct. In addition, in order to decide whether an item should be revised, kept as is, or deleted, the reliability coefficients, if an item is deleted, were calculated. The results of the item analyses for the 12 reading items are summarized in Table 5.

TABLE 5
Item Analyses for the Reading Test (*N*=29)

Item	Variable	Difficulty (*p*)	Discrimination (pt-biserial correlation)	Alpha if item deleted	Decision
1	ReadVoc1	0.43	0.219	0.623	Revise
2	ReadDet1	0.93	0.382	0.605	Keep
3	ReadVoc2	0.50	0.078	0.651	Delete
4	ReadInf1	0.79	0.213	0.622	Revise
5	ReadGist1	0.25	0.442	0.580	Keep
6	ReadGist2	0.43	0.289	0.609	Keep
7	ReadGist3	0.25	0.280	0.610	Keep
8	ReadVoc3	0.21	0.078	0.644	Delete
9	ReadDet2	0.61	0.365	0.593	Keep
10	ReadDet3	0.75	0.549	0.559	Keep
11	ReadInf2	0.64	0.457	0.573	Keep
12	ReadInf3	0.25	0.090	0.644	Delete

Note. Voc=Vocabulary in context; Det=Detail; Inf= Inference

Since the 12 reading items were designed as part of a placement test, a *p*-value range from 0.3 to 0.9 was desirable. According to this standard, Item 2 was too easy with a *p*-value of 0.93, meaning that the 93 percent of examinees correctly answered this item. On the other hand, four items were too difficult—Items 5, 7, and 12 with *p*-values of 0.25 and Item 8 with a p-value of 0.21. Only 21 percent of examinees responded correctly to Item 8, which was of greatest difficulty. Among the items within the range of 0.3 to 0.9, Items 4 and 10 were relatively easy items, with *p*-values of 0.79 and 0.75, respectively, while Items 1 and 6 were difficult items, with *p*-values of 0.43. Items 3, 9, and 11 represented moderate difficulty with *p*-values of 0.5, 0.61, and 0.64, respectively. These results suggest that in order for these 12 items to function as placement test items, the ones outside of the standard brackets (*p*-value between 0.3 and 0.9) need to be revised. Specifically, Item 2 should be made more difficult, while Items 5, 7, 8, and 12 should be revised to be easier. One limitation of this item analysis that needs to be pointed out is the dependence on the sample. Since item difficulty was calculated only with a limited number of examinees, the difficulty levels of items presented above may not be generalized. A different sample of examinees might generate different item difficulty indices.

The adjusted item-total correlation (i.e., the point biserial correlation) is interpreted as item discrimination. The following is the item discrimination index interpretation: a very good item has a *D* index of 0.4 and above, a reasonably good item has a *D* of 0.30 to 0.39, a marginal item has a *D* of 0.20 to 0.29, and a poor item has a *D* of 0.19 and below. According to Table 5, Items 5, 10, and 11 discriminated between high and low level examinees very well with *D* of 0.442, 0.549, and 0.457, respectively. Items 2 and 9 discriminated reasonably well. Items 1, 4, 6, and 7 were marginal items and should be revised. Items 3, 8, and 12 showed very low discrimination indices of 0.078, 0.078, and 0.090, respectively. These three items did not function appropriately in terms of discriminating examinees. Thus, they should be rejected or improved.

The "alpha if item is deleted" column indicates the change in reliability (i.e., Cronbach's Alpha) of the test if only that particular item is deleted. In order to keep an item, the alpha in this column should remain the same or decrease. This indicates that if the item is deleted, the internal consistency of the test overall will decrease. On the other hand, if the alpha increases after an item is deleted, the item should be deleted. This implies that if the item is deleted, the internal consistency of the test will be improved. Table 5 suggests that the test's reliability can be improved by deleting Items 3, 8, and 12 after comparing the "alpha if item deleted" value with the initial alpha value (0.631 from Table 4). Indeed, these three items were also problematic in terms of item difficulty and discrimination as presented above; all of them showed very low *D*-indices, and Items 8 and 12 were outside of the desirable difficulty range for a placement test. Thus, these items were deleted. In addition, Items 1 and 4 were revised since they showed only a slight decrease in alpha after they were deleted. These results were confirmed by the item discrimination analysis; that is, Items 1 and 4 were marginal items that required improvement. Other items showed a decrease in alpha when they were deleted, and therefore, the decision was made to keep them. After deleting Items 3, 8, 12, Cronbach's Alpha increased to 0.703 (see Table 6) from the initial value of 0.631.

TABLE 6
Reliability Statistics for the Reading Test (*N*=29)

Cronbach's Alpha	N of Items
0.703	9

Among the three deleted items, Items 3 and 8 tested vocabulary in context and Item 12 tested inference. The inappropriateness of the vocabulary items might be explained by students' misuse of vocabulary knowledge. The items were designed to elicit examinees' ability to deduce the meaning of a word from the given text. Since some of the distractors had possible meanings of the word outside of the text, it is likely the examinees chose an answer without considering the text. On the other hand, Item 12 was far too difficult according to its *p*-value. This might have caused the item to function inappropriately.

Evidence of Construct Validity within the MC Task

In this section, the construct validity of the multiple-choice reading task is assessed by a Pearson product-moment procedure to determine the correlation between the four domains—gist, vocabulary, inference, and detail—of the reading task. The correlation between two variables represents the degree to which they are related. Relationships between the variables, however, vary in terms of strength and direction. The correlation coefficient calculates the strength and direction of the relationship between the two variables. Since the scores are interval in nature and normally distributed, the Pearson product-moment correlation coefficient, *r*, was calculated. The values of the correlation coefficients can range from -1.00 to +1.00.

Correlation coefficients among items can be high (r = 0.75 or above), moderate (r = 0.5 to 0.74), low (r = 0.25 to 0.49), uncorrelated ($r < 0.25$), or not correlated at all ($r = 0$). A correlation coefficient less than 0 is a negative correlation. An example of this relationship is when participants' scores are high in one domain but low in another domain. Thus, negative correlation coefficients indicate inverse relationships and are identified by the presence of a minus sign, while positive correlation coefficients indicate direct relationships. According to the theoretical model of reading ability, one would expect to see all the variables positively correlated with one another because they are all components of reading ability.

Results showed mixed findings. On the one hand, strong correlation between components of the reading construct existed; on the other hand, mastery of a component within the reading ability construct did not correlate significantly with others within the reading ability construct. Table 7 illustrates the correlation matrix for the original test (*K* = 12) and Table 8 shows the correlation matrix for the revised test (*K* = 9). Analyses of both were conducted in order to compare how the correlations would change, if at all, after the three items were deleted from the original test.

TABLE 7
Correlation Matrix for the Reading Test (*K*=12, *N*=29)

Scale	Gist	Vocabulary	Inference	Detail
Gist	1.000			
Vocabulary	0.276	1.000		
Inference	0.420*	0.016	1.000	
Detail	0.321	0.484**	0.368*	1.000

**p*<.05. ***p*<.01.

The findings summarized in Table 7 suggest that the reading test may have indeed served as a good tool for measuring participants' reading ability. In other words, the significant correlations appear to be due to factors other than chance. Correlations between inference and gist items ($r = 0.420$) and between detail and inference items ($r = 0.368$) were statistically significant at the $\alpha = 0.05$ level, meaning that there is a 95% chance that the correlation between the variables was not due to chance. The correlation between detail and vocabulary ($r = 0.484$) was statistically significant at the $\alpha = 0.01$ level, indicating that one can be 99% confident that this correlation was not a chance phenomenon. Moreover, the correlation coefficients themselves ($r = 0.420$, $r = 0.368$, and $r = 0.484$) were relatively low and positive. This is desirable because each component was intended to test a different aspect of the skill of reading, therefore adding to the overall picture of reading ability.

Correlations that were not found to be statistically significant might be due to chance phenomena. However, even though the correlation coefficients between, for example, detail and gist were nonsignificant—a low correlation of 0.321—the result suggests that these two variables were measuring two related constructs. In fact, one can assert that the test may have yielded a good measure of reading ability. Yet, the lowest correlation is shown between inference and vocabulary ($r = 0.016$) and a comparatively low correlation is exemplified between vocabulary and gist ($r = 0.276$). These lower correlations imply that the items were less homogeneous. This accounts for the lower internal consistency of the test as a whole.

After deleting the three problematic items from the original test, some interesting differences occurred in the relationships among the scores associated with each of the four constructs (see Table 8). For instance, the only statistically significant correlation was displayed between inference and gist ($r = 0.469$) at the $\alpha = 0.05$ level. This implies that there is a 95% chance that the correlation observed between the inference and gist variables was not due to chance. All other correlations were not statistically significant. Yet, the correlation coefficients were relatively low ranging from 0.223 to 0.469, implying the variables were probably testing different skills.

TABLE 8
Correlation Matrix between Variables for the Reading Test (*K*=9, *N*=29)

Scale	Gist	Vocabulary	Inference	Detail
Gist	1.000			
Vocabulary	0.353	1.000		
Inference	0.469*	0.223	1.000	
Detail	0.321	0.235	0.349	1.000

*p<.05. **p<.01.

.

The results summarized in Table 8 appear quite surprising. One would expect that after deletion of the three items, more correlation coefficients would be found to be significant. Unfortunately, as shown by the correlation matrix, fewer correlations were actually found after these deletions. Three factors may have contributed to this unexpected finding. First, the grouping together of the three proficiency levels (beginning, intermediate, and advanced) may have affected the value of the correlation coefficient. Second, upon a closer examination of the correlation coefficient formula it becomes apparent that as the variation decreases the correlation varies. Therefore, variability of correlation is somewhat dependent on variation being low. Hence, one cannot assume that when Cronbach's Alpha increases, the correlation, too, will increase. Lastly, the significance of the correlations is dependent on the sample size. When the total sample size is only 29, it is difficult to assert that the correlation is due to chance alone. Furthermore, the lack of a significant correlation between detail and vocabulary and between detail and inference may be due to the noticeable reduction of items (two vocabulary items and one inference item were deleted).

Results for the Extended Production Task: Writing Ability

Descriptive Statistics

The descriptive statistics presented in Table 9 summarize participants' performance on the writing section by level (beginning, intermediate, and advanced) and by the three levels combined. The scores are averaged across rater 1 and rater 2 for each of the three components (content, organization, and form). Overall, the mean total writing score was 2.63 (*SD* = 1.14) out of a possible 5. By proficiency level, the advanced group scored highest and the beginning group scored lowest. Mean writing scores clustered together closely. For instance, the three advanced test-takers each scored a 5 as indicated by their total writing score. For beginning and intermediate participants, the range in scores was only 2.

Overall, the distribution of writing scores was negative (kurtosis = -0.613). This distribution suggests heterogeneous participants with a sizeable amount of variability in writing ability. The sizable positive kurtosis observed among beginning level test-takers indicates the distribution was peaked, denoting homogeneity among participants. The scores were clustered over a narrower distribution such that beginning level participants were not normally distributed in terms of their reading ability.

The positively-skewed distribution of all participants' scores (where there are more low scores than high scores) was expected in a placement test. This can be a result of limited English mastery in beginning level students. The positive skew of beginning participants' scores may be due to lack of knowledge of organization, content, and language components that encompass writing ability. Similarly, intermediate level test-takers also found the test a little difficult, but to a lesser degree than their beginning level counterparts.

TABLE 9
Descriptive Statistics, Extended Production Tasks and Total

	Beginning (B4)				Intermediate (I4)				Advanced (A4)				All Levels			
	Content	Organization	Form	Writing Total	Content	Organization	Form	Writing Total	Content	Organization	Form	Writing Total	Content	Organization	Form	Writing Total
Number of Test-takers (*N*)	8	8	8	8	18	18	18	18	3	3	3	3	29	29	29	29
Total Items (*K*)	1	1	1	3	1	1	1	3	1	1	1	3	1	1	1	3
Mean	1.63	1.25	1.25	1.25	3.28	3.06	2.94	2.89	4.67	4.67	5.00	5.00	2.78	2.60	2.52	2.63
Mode	1	1	1	1	3	3	3	3	5	5	5	5	1(a)	1(a)	1(a)	1
Median	1.00	1.00	1.00	1.00	3.00	3.00	3.00	3.00	5.00	5.00	5.00	5.00	3.00	3.00	2.50	2.67
Minimum	4	3	3	3	2	2	2	2	4	4	5	5	1	1	1	1
Maximum	4	3	3	3	4	4	4	4	5	5	5	5	5	5	5	5
Range	3	2	2	2	2	2	2	2	1	1	0	0	4	4	4	4
Standard Deviation (SD)	1.061	.707	.707	.707	.669	.725	.639	.676	.577	.577	.000	.000	1.123	1.213	1.176	1.139
Kurtosis	3.937	8.000	8.000	8.000	-.564	-.904	-.143	-.531	N/A	N/A	N/A	N/A	-.707	-.740	-.211	-.613
Skewness	1.960	2.828	2.828	2.828	-.382	-.086	.041	.132	-1.732	-1.732	N/A	N/A	-.124	.160	.451	.143

Note. (a) = Multiple modes exist. The smallest value is shown.

Figures 2, 3, and 4 show the distribution of scores of the beginning and intermediate levels separately as well as the scores for all levels combined. A graphic representation of the advanced level was omitted as the small sample size misrepresented the scores. For Figures 2, 3, and 4, larger frequency of occurrences is presented by taller bars. Beginning level test-takers most often scored 1, while intermediate level test-takers most often scored 3. As presented in Figure 4, the most frequent scores achieved by the group as a whole were 1 and 3. The positively skewed distribution of both individual levels and all participants shows that there was a higher frequency of low scores and a lower frequency of high scores. Considering the kurtosis of 8.000 and -0.531 for beginning and intermediate level test-takers respectively, it can be inferred that the beginning level group was relatively homogeneous while the intermediate level group was relatively heterogeneous in terms of their writing ability. Results of the writing test for all levels combined yielded a negative kurtosis of -0.613 with a standard deviation of 1.139. In a placement test, a very easy or difficult item would not be appropriate. Therefore, a wide spread of scores with only a small number of students obtaining any one score is desirable.

FIGURE 2
Histogram of the Writing Subtest Scores for Beginning Level (*N*=8)

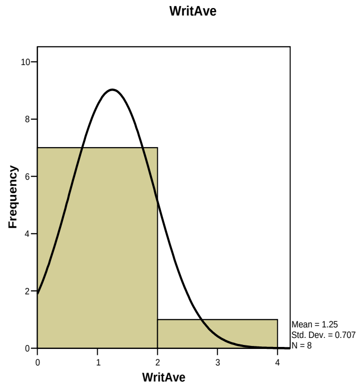


FIGURE 3
Histogram of the Writing Subtest Scores for Intermediate Level (*N*=18)

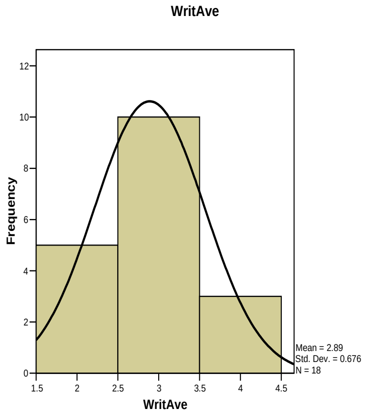
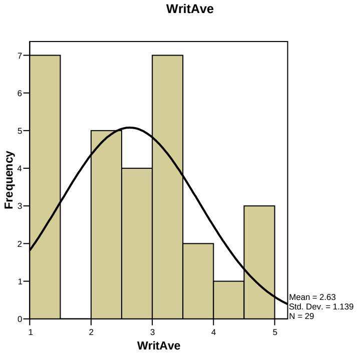


FIGURE 4
Histogram of the Writing Subtest Scores for All Levels (*N*=29)



Internal Consistency Reliability

In order to check internal consistency reliability of the extended production task, Cronbach's Alpha was calculated. Cronbach's Alpha was used for the writing test since it applies to items scored on an ordinal scale as well as to dichotomously scored items. For the calculation, the average scores from the two raters were used for the three domains of the writing scoring rubric. These three domains of form, content, and organization were treated as items.

The estimate of internal consistency reliability was 0.971 for the three domains (see Table 10). With a value close to 1, this number indicates a near-perfect internal consistency. Thus, one can argue that the three domains (form average, content average, and organization average) consistently measured the same construct of writing ability.

TABLE 10
Reliability Statistic for the Writing Test (*N*=29)

Cronbach's Alpha	N of Items (N of domains)
0.971	3

Inter-Rater Reliability

Using a correlation procedure, inter-rater reliability was calculated. Inter-rater reliability represents the degree of agreement in scoring between raters, and correlation refers to the degree to which one variable varies with another. First, the correlation between rater 1 and rater 2 was calculated across the three domains of form, content, and organization. Since the scores were ordinal, a Spearman Rank-Order correlation procedure was used.

As shown in Table 11, the correlation coefficients between raters 1 and 2 were 0.829 for the content domain, 0.904 for organization, and 0.854 for form, respectively. They were all statistically significant at the $\alpha = 0.01$ level, indicating that the first rater's score on each domain significantly correlated with the second rater's score on the same domain. As the two raters' scores were positively correlated at the 99% significance level, it can be inferred that the two raters showed a high degree of agreement in scoring the three variables. These findings also imply that the two raters interpreted the scoring rubric in the same way and shared an understanding of the elements that should be included for each domain.

TABLE 11
Correlation Matrix for the Writing Test, Spearman Rank-Order (*N*=29)

	ContR1	ContR2	OrgR1	OrgR2	FormR1	FormR2
ContR1	1.00					
ContR2	0.829**	1.00				
OrgR1			1.00			
OrgR2			0.904**	1.00		
FormR1					1.00	
FormR2					0.854**	1.00

**p*<.05. ***p*<.01.

Next, the correlation between the writing total score for rater 1 and the writing total score for rater 2 was calculated. For the total score correlation, a Pearson product-moment correlation procedure was used since the total scores are continuous in nature.

It is evident that the total writing scores from rater 1 and 2 were highly correlated (see Table 12). The correlation coefficient between raters was 0.936, showing a strong positive correlation. Moreover, it was statistically significant at the $\alpha = 0.01$ level. As a result, it can be assumed that the two raters scored the examinees' writings with the same criteria in mind.

TABLE 12
Correlation Matrix for the Writing Test, Pearson Product-Moment (*N*=29)

	Rater 1 (WritTotR1)	Rater 2 (WritTotR2)
Rater 1 (WritTotR1)	1.00	0.936**
Rater 2 (WritTotR2)	0.936**	1.00

**p*<.05. ***p*<.01.

Internal consistency reliability (expressed by Cronbach's Alpha) can be called *internal reliability*. This is because items are a part of the actual test. On the other hand, since raters are not a part of the actual test, inter-rater reliability is considered *external reliability*. For the writing portion of the test, the former was at the level of 0.971 and the latter was calculated as being above 0.936. Analyses therefore revealed that external reliability (i.e., inter-rater reliability) was a more conservative estimate of the writing test reliability than its internal reliability.

Evidence of Construct Validity within the Extended Production Task

The construct validity of the extended production task was determined by the Pearson product-moment procedure used to obtain the correlations among content, organization, and form for the writing task. Once again, Pearson product-moment correlations were used because the average scores are interval in nature. The degree to which these variables were correlated is illustrated in Table 13.

TABLE 13
Correlation Matrix for the Writing Test (*N*=29)

Scale	Content	Organization	Form
Content	1.00		
Organization	0.922**	1.00	
Form	0.895**	0.937**	1.00

**p*<.05. ***p*<.01.

In interpreting the relationships between the variables in the extended production task, the results generally exhibited high correlations at the $\alpha = 0.01$ level. Organization was highly correlated with content ($r = 0.922$) and form ($r = 0.937$), both statistically significant at the $\alpha = 0.01$ level. In other words, there is a 99% chance that these correlations were not due to chance alone. Similarly, correlation between content and form ($r = 0.895$) was significant at the $\alpha = 0.01$ level.

The strong relationships between the variables provide some evidence for a claim that all three domains were measuring the same underlying construct—namely, writing ability. In other words, these writing components were indeed measuring what they purported to measure. Accordingly, the results support our earlier assertion based on our literature review that writing ability consists of content, organization, and form control. However, another interpretation of the high correlations is that these three variables are not separable. If this is the case, dropping one of the three domains may be considered.

Determinants of Reading and Writing Abilities

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a method used to investigate the similarities and/or differences among two or more groups (i.e., between-group comparison). More specifically, if employed, it allows researchers to examine whether the means and standard deviations of two or more groups are the same or different based on *F*-statistics. In our study, we tested the null hypothesis, that is, that the same mean exists across the three proficiency levels (beginning, intermediate, and advanced). The null hypothesis is defined as a hypothesis of no difference or as Bachman (2004) stated, “if the observed difference is entirely due to chance, then this implies that there is no real difference between the two means” (p. 214). Here, the null hypothesis can be written as follows: Mean (beg) = Mean (inter) = Mean (adv).

The averages of reading and writing scores of the three proficiency levels were compared (see Table 14). This analysis offers information on whether there were statistically significant differences in reading and writing performances or abilities across the three groups.

TABLE 14
Analysis of Variance in Test Scores

Skill	Beginning (B4)[9]	Intermediate (I4)	Advanced (A4)	F
Reading Score	3.43	6.74	9.00	12.03**
Writing Score	6.57	17.26	28.00	63.02**
Total Score	10.00	24.00	37.00	73.09**

**p*<.05. ***p*<.01.

The average reading scores of the three groups (beginning, intermediate, and advanced) were 3.43, 6.74, and 9.00, respectively. To determine if these means were different from each other statistically, we calculated an *F*-ratio as follows:

F-ratio calculation:
$F\ [k\text{-}1, n\text{-}k] = \{R^2/(k\text{-}1)\}/\{(1\text{-}R^2)/(n\text{-}k)\}$
where R^2 = R-squared, *n*= number of samples, and *k*= number of groups

As shown above, $\{(1\text{-}R^2)/(n\text{-}k)\}$ is the amount of discrepancy we would expect due to chance while the $\{(R^2/(k\text{-}1)\}$is the amount of discrepancy that is due to any differences between the groups (Bachman, 2004). The obtained *F*-ratio reading score is 12.03, which is much larger than the critical value at 95% ($F_{95\%}$*[3-1, 29-3] = $F_{95\%}$*[2, 26] = 3.39) or the critical value at 99% ($F_{95\%}$*[2, 26] = 5.57). Thus, we can conclude that the null hypothesis of [Mean (beginning) = Mean (intermediate) = Mean (advanced)] can be rejected at both the 95% and 99% significance levels. This result shows that: (1) the placement test was sufficient in distinguishing across students' different reading ability and (2) the test successfully separated the three proficiency level students.

Similar results were obtained for both the writing score and the total score (the sum of the reading score and writing scores). Specifically, the *F*-ratio for the writing score was 63.02 and for the total score, 73.09. These results allow us to conclude that each group's average writing and total scores were statistically different from each other at the 99% significance level. In sum, ANOVA analysis on the reading, writing, and total scores suggests that the placement test was effective in identifying the three proficiency groups in terms of reading, writing, and overall language abilities.

Second, differences between groups in terms of learners' desire to improve specific language skills were examined (see Table 15). Students were surveyed before the placement test on how interested they were in improving their language skills in reading, writing, listening, and speaking. Items marked were scored as either 1 or 0. A "1" signifies that the participant is

[9] After reviewing statistical properties of each group, we found that one student (ID #2) in the "beginning" group performed exceptionally well, which may have distorted the ANOVA/regression analysis. Choosing between excluding the student from further analyses and reclassifying this participant, we chose to reclassify the student into the "intermediate" group. The decision was made after considering two possibilities: (1) this student's skill had improved significantly during the last few months, or (2) this student had not been able to show his/her ability fully during the placement test due to reasons such as health problems or jet-lag. We will investigate this matter by interviewing the students in-depth in the future.

interested in improving, for example, his/her reading skills, while a "0" suggests no interest in improving that particular skill.

TABLE 15
Analysis of Variance in Language Skill Improvement Preference

Intention	Beginning (B4)	Intermediate (I4)	Advanced (A4)	F-statistic
Reading	0.57	0.53	1.00[10]	1.18
Writing	0.43	0.68	0.67	0.68
Listening	0.71	0.68	0.67	0.01
Speaking	0.86	0.89	0.67	0.53

**p*<.05. ***p*<.01.

Table 15 illustrates the percent of participants who desired to improve their skills in reading, writing, listening and speaking. For instance, 57% of beginning level group members wanted to improve their reading skills. ANOVA tests were performed across the three groups on each of the skills. There were no significant differences in any of the skills among the three groups since the *F*-statistics are all below the critical value. In other words, the three groups had similar degrees of preference across the four language skills.

Interestingly, the beginning level group showed more interest in improving listening (71%) and speaking skills (86%) than reading (57%) and writing (43%) skills, while the majority of learners (89%) in the intermediate level group hoped to improve their speaking skills. This finding is intriguing and warrants further investigation.

Third, we examined the data for any behavioral differences among the groups in terms of time spent on activities related to reading skills (see Table 16). Specifically, participants were asked to indicate how many hours per week they engaged in reading-related activities such as reading books, Internet surfing, and reading magazines and newspapers (see Appendix 2). It might be argued that students with better reading skills spend more time reading books since they are comfortable with the reading activity, and that this could improve their reading skills considerably. To examine the reasons for this relationship, a regression analysis (with applicable control variables) would be more appropriate than an ANOVA analysis.

Table 16 suggests that the groups indeed had different reading habits regarding books; beginning- and intermediate-level groups spent a similar number of hours per week (1.43 hours and 1.37 hours, respectively) on reading books, while the advanced level group spent 3.33 hours per week. The *F*-ratio for reading books was 3.82, which is significant at the 95% confidence level. As for Internet surfing hours, the advanced level group spent more time than any other group but the difference was not significant.

Little difference was found in terms of hours spent reading magazines or newspapers; however, it is interesting to note that the intermediate level group devoted relatively more time to reading magazines and newspapers than the other two groups. Even though these group differences are not significant, the behavioral patterns can be explored further using multiple regression analysis. Specifically, we can relate the different reading habits or reading input

[10] Here, "1.00" means that all three students in the advanced group answered "yes" to the question.

measured as hours spent (independent variable) to reading performance or reading ability (dependent variable) at the time of the placement test.

TABLE 16
Analysis of Variance in Reading Habits

Activity	Beginning (B4)	Intermediate (I4)	Advanced (A4)	F-statistic
Books	1.43	1.37	3.33	3.82*
Internet Surfing	1.00	1.74	3.67	2.42
Magazine	1.14	1.63	1.33	0.24
Newspaper	1.43	2.16	1.67	0.61

***p*<.05. ***p*<.01.

Fourth, we investigated whether there were differences among the three groups with respect to writing habits (see Table 17). Students were asked to complete a survey regarding the time they spent per week as measured in hours on writing text-messages, writing emails, filling in forms, Internet chatting, writing shopping lists, and writing notes.

TABLE 17
Analysis of Variance in Writing Habits

Activity	Beginning (B4)	Intermediate (I4)	Advanced (A4)	F-statistic
Text Messaging	1.00	0.58	1.00	0.67
Emails	1.00	1.37	1.67	0.29
Forms	0.57	1.05	1.00	0.48
Internet Chatting	1.14	0.58	1.00	1.76
Shopping Lists	1.29	0.58	1.00	1.25
Notes	0.86	1.58	3.00	1.76

**p*<.05. ***p*<.01.

All three groups spent comparable amounts of time on these six activities as no significant differences were found. However, we observed interesting patterns among the groups: (a) the beginning level group spent more time writing text messages (1.00 hour), chatting over the Internet (1.14), and writing shopping lists (1.29), (b) the intermediate level group spent more time writing emails (1.37) and filling in forms (1.05), and (c) the advanced level group spent far more time on writing notes (3.00) than any other group, but less time on writing emails (1.67), filling in forms (1.00), or text messaging (1.00).

Fifth, we examined the effects of prior English education, which was measured as time (in years) students had spent studying English before joining the CEP at Teachers College.

TABLE 18
Analysis of Variance in Prior English Study

Prior Study	Beginning (B4)	Intermediate (I4)	Advanced (A4)	F
Years	1.60	5.42	8.33	6.26**

*p<.05. **p<.01.

We found that there were significant differences among groups in terms of prior English study (see Table 18). The beginning level group had studied English 1.6 years on average. For the intermediate level group the average was 5.42 years, while for the advanced level group it was 8.33 years. The *F*-ratio was 6.26, which was significant at the 99% confidence level. In sum, a large portion of variation in reading/writing scores can be explained by this variable.

In conclusion, the findings show that there are significant differences among groups in terms of reading/writing performances or abilities. Not surprisingly, prior study time was found to be an important determinant of the score differences between groups. No significant differences were seen in terms of intention to improve specific language skills (i.e., reading, writing, listening, speaking) and time spent on reading/writing-related activities. Nonetheless, some interesting behavioral patterns across groups were discovered, which will be explored further via multiple regression analysis in the following section.

Multiple Regression Analysis

The goal of multiple regression analysis is to identify possible determinants of the reading and writing scores.[11] First, we estimated a simple multiple regression model for the reading score as a dependent variable. Specifically, we model the relationship as follows:

Reading performance regression:
SR = f [RB, RS, RM, RN, GEND, AGE, EDUC, TIME]
where SR: Reading Score, RB: Reading Books, RS: Reading via Surfing, RM: Reading Magazine, RN: Reading News, GEND: Gender, AGE: Age, EDUC: Education, TIME: Time to learn English before joining the Columbia ESL program

The estimation results are summarized in Table 19:

[11] For the estimation of both reading performance as well as writing performance regression models, we entered all specified independent variables at a time, that is, we used Enter method.

TABLE 19
Regression Analysis of Reading Performance

Variable	Unstandardized Coefficient (B)	Std. Error	T-statistic	Significance of T-statistic
Constant	5.60	4.18	1.34	0.20
RB	0.50	0.52	0.96	0.35
RM	-0.01	0.40	-0.01	0.99
RN	0.09	0.44	0.21	0.84
RS	-0.23	0.41	-0.56	0.59
EDUC	0.27	0.46	0.59	0.56
AGE	-0.34	0.62	-0.54	0.59
GEND	-0.78	1.34	-0.58	0.57
TIME	0.24	0.15	1.67	0.11
R^2: 0.25				

**p*<.05. ***p*<.01.

As evident from Table 19, no significant determinants of reading ability were found. R^2 of the estimation was not high (0.25), indicating that only 25% of variance in reading scores was accounted for by the eight independent variables. These results imply that an important factor may have been overlooked in the survey with respect to test-takers' reading performance.

Second, a multiple regression model was estimated for the writing score as a dependent variable. Specifically, we model the relationship as follows:

Writing performance regression:
SW = f [WT, WE, WF, WC, WL, WN, AGE, EDUC, TIME]
where SW: Writing Score, WT: Writing Text message, WE: Writing Emails, WF: Writing Forms, WC: Writing via Online Chatting, WL: Writing Shopping Lists, WN: Writing Notes, GEND: Gender, AGE: Age, EDUC: Education, TIME: Time to learn English before joining the Columbia ESL program

The estimation results are presented in Table 20:

TABLE 20
Regression Analysis of Writing Performance

Variable	Unstandardized Coefficient (B)	Std. Error	T-statistic	Significance of T-statistic
Constant	7.84	4.57	1.72	0.10
WT	-1.00	1.23	-0.81	0.43
WE	0.44	0.70	0.63	0.54
WF	-0.20	0.74	-0.27	0.79
WC	-0.20	1.61	-0.12	0.91
WL	-0.29	1.07	-0.27	0.79
WN	0.72	0.48	1.50	0.15
EDUC	-0.32	0.54	-0.60	0.56
AGE	-0.96	0.70	-1.37	0.19
GEND	1.04	1.46	0.72	0.48
TIME	0.46	0.20	2.33*	0.03
R^2: 0.63				

*p<.05. **p<.01.

Table 20 confirms that prior study time was an important factor in explaining writing ability. It is significant at a 95% confidence level and the sign of the coefficient is positive. R^2 of the estimation was relatively high (0.63), meaning that 63% of variance in writing scores was explained by the ten independent variables. These results imply we may have better explanations for writing performance than for reading performance given the model specified in this study. Improved regression results could come about through the revision of the survey questions, an increase in the number of participants, and development of a better theoretical construct of reading and writing abilities. We will leave this task for future study.

DISCUSSION AND CONCLUSION

To summarize, the current study reviewed previous research on L2 reading and writing ability and provided a conceptual framework based on these two defined constructs. Placement procedures including test development, a pilot study, and test analyses were explored in the context of the Community English Program. Four components were defined for reading ability: Gist, Detail, Inference, and Vocabulary in Context. Each component contained three items. A total of 12 reading multiple-choice items were included. Writing ability was measured across three domains: Form, Organization, and Content. An extended production task was given to the 29 examinees.

In order to investigate the internal consistency of L2 reading MC items, internal consistency reliability was performed and results showed that the 12 items measured the reading constructs with a moderate degree of consistency. When the three poor items (Items 3, 8, and 12) were deleted, the reliability coefficient increased. As discussed previously, the numbers of items and examinees affected the reliability estimate; thus, an increased number of items and a larger sample size are required for a higher reliability estimate. On the other hand, the three variables of the writing task showed high internal consistency reliability, indicating that the three variables measured the construct of writing ability consistently.

Inter-rater reliability analysis revealed that the two raters were consistent in scoring the three variables of the writing task; all correlation coefficients between the two raters across the three domains were statistically significant. In addition, the total writing scores from raters 1 and 2 were highly correlated and the correlation was also statistically significant.

Furthermore, construct validation for both the reading and writing tests provided some evidence that the underlying theoretical definitions for the constructs being measured were valid. In the multiple-choice task, correlations after poor items were deleted were statistically significant at the $\alpha = 0.05$ level for inference and gist ($r = 0.469$). Even though all other correlations were not statistically significant, they were relatively low ranging from 0.223 to 0.469. This, in turn, implies that the variables were indeed testing different skills. Likewise, the statistically significant correlation coefficients at the $\alpha = 0.01$ level, which were found among all writing variables, demonstrate that the components were measuring the same construct of writing ability. Finally, we also tried to identify differences among the three proficiency levels and to find determinants of reading and writing ability using ANOVA and multiple regression analysis.

We hope our work provides practical skills for ESL teachers and administrators in the very important yet often overlooked area of assessment: the design, evaluation, and analysis of an ESL placement test. The study, however, was not without limitations. First, only a limited number of CEP classes were available at each proficiency level with several researchers conducting separate investigations. In fact, one of the intermediate level 4 classes had taken another placement test one day before the administration of our test. Test-takers may have been influenced during the administration of this particular test by practice or the emotional discomfort of taking another test.

Second, the Likert scale, which is a five-point scale used to elicit opinions in surveys, may have been more appropriate for this test in measuring learner's reading and writing activities. The interval scale used in our questionnaire may have misrepresented participants' reading and writing habits. Moreover, the respondents, in their haste to complete the survey, may have answered survey items too quickly resulting in responses that did not necessarily reflect their real opinions. It would also be desirable to check on the reliability and validity of the survey instrument itself. The questionnaire may be a more reliable instrument if it is administered both before and after the test, to determine whether participants' responses are consistent across time. Validity may be established by correlating each survey item and discarding those items that show low or negative correlations. This would ensure not only that the survey items are homogeneous, but also that they measure the same underlying trait.

Next, the number of participants for this study was small. Only eight beginning level students and three advanced level students took the test. Furthermore, the sampling procedure may not have been proportional. An effort to increase the validity of this study by focusing on the highest level for each group (i.e., B-4, I-4, A-4) inevitably resulted in a small sample size.

As a result of the limitations stated above, it is difficult to make generalizations about participants' reading and writing ability. If given the opportunity to readminister this test in the future, we would address the aforementioned changes in order to develop a more comprehensive placement test.

ACKNOWLEDGEMENTS

We would like to acknowledge the invaluable comments given by Dr. James Purpura and the anonymous reviewers. We would also like to thank Yen-Fen Alick Liao for feedback on earlier versions of the paper.

REFERENCES

Aebersold, J., & Field, M. (1997). *From reader to reading teacher*. Cambridge, UK: Cambridge University Press.

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

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*, 79-113.

Anderson, N. (1999). *Exploring second language reading: issues and strategies*. Boston: Heinle & Heinle.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge, UK: Cambridge University Press.

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

Cumming, A. (1990). Expertise in evaluating second-language compositions. *Language Testing, 7*, 31-51.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. In R.M. Manchon (Ed.), *International Journal of English Studies, 1*, 1-23.

Garrett, P., Griffiths, Y., James, C., & Scholfield, P. (1995). The development of a scoring scheme for content in transactional writing: some indicators of audience awareness. *Language and Education, 9*, 179-93.

Horowitz, D. (1986). Process not product: Less than meets the eye. *TESOL Quarterly*, *20*, 141-144.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, *10*, 211-234.

Matsuda, P. K. (2003). Second language writing in the twentieth century: A situated historical perspective. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 15-34). Cambridge, UK: Cambridge University Press.

Purpura, J. E., & Pinkley, D. (2000). *On Target 1 Intermediate*. White Plains, NY: Longman.

Raimes, A. (1991). Out of the woods: Emerging traditions in the teaching of writing. *TESOL Quarterly*, *25*, 407-430.

Weir, C.J. (1997). The testing of reading in a second language. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language education: Vol. 7. Language testing and assessment* (pp. 39-49). Dordrecht, The Netherlands: Kluwer.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal, 77*, 473-489.

APPENDIX 1

Design Statement

1. Test purposes	
A. *Inferences*	
About test takers' language ability in a wide range of everyday communicative domains in which reading and writing are necessary	
B. *Decisions*	
1. Stakes Relatively high; test results used to make decisions about test takers' proficiency level 2. Individuals affected Test takers, teachers, CEP administrators 3. Specific decisions to be made a. Placement Assign an appropriate proficiency level to test takers b. Diagnosis For CEP teachers and administrators, test results will be used for diagnosing whether the placement test functions properly.	
2. Description of TLU domain(s) and task types	
A. *Identification of tasks*	
1. TLU domain: real-life communication 2. Identification and selection of TLU tasks for considering as test tasks: The TLU tasks to be analyzed were identified on the basis of consideration of the CEP population. Because of the diversity of the population, common, probable tasks were selected for both the reading and writing sections: getting information from two reading texts and writing an informal postcard.	
B. *Description of TLU task types* See Appendix 2.	
3. Description of characteristics of test takers	
A. *Personal characteristics*	
1. Age	19 and above
2. Sex	Male and female
3. Nationalities	Widely varied
4. Immigrant status	Immigrants and temporary residents
5. Native languages	Widely varied
6. Level and type of general education	Middle school, High school, College/ University, Graduate school, Post-graduate
7. Type and amount of preparation or prior experience with a given test	Widely varied
B. *Topical knowledge of test takers*	
In general, relatively widely varied topical knowledge due to cultural, educational and occupational differences	

C. *Levels and profiles of language knowledge of test takers*
1. General level of language ability Beginning to advanced (Although there is originally no targeted proficiency level in the placement test, the items were developed for the intermediate level for the current study.) 2. Specific reading and writing ability Widely varied, beginning to advanced
D. *Possible affective responses to taking the test*
1. Highly proficient test takers should be likely to feel positive about the test since they might be placed at a high level of class at the CEP. 2. Less proficient test takers should be likely to feel frustrated since the test items are too challenging for them.
4. Definition of construct(s)
A. *Language ability*
Theoretical model-based and curriculum-based construct definition. The goals of the CEP specify that the students should improve their general English communicative ability, and the CEP placement test aims at placing the students at an appropriate level in order to maximize its goals. Therefore, unlike an achievement test, the CEP placement test has quite broad descriptions of its constructs because it defines wide-ranging communicative ability. 1. <u>Reading ability</u> a. Gist b. Detail c. Inference d. Vocabulary in context 2. <u>Writing ability</u> a. Content control b. Organization control c. Language control
B. *Strategic competence*
Not included in the construct
C. *Topical knowledge*
Not included in the construct

APPENDIX 2

Student Survey

Directions: Please fill in the blanks and/or check the appropriate item(s).

1. What is your gender? Male __ Female __
2. How old are you? 20s __ 30s __ 40s __ 50s __ Other (please specify: _______)
3. What is your educational background?
   Elementary School __ Middle School __ High School __
   2-year college __ 4-year college __ Graduate degree __

Doctorate __ Other (please specify:__________________)

4. What is your nationality?
Brazilian __ Chinese __ German __
Japanese __ Korean __ Peruvian __
Polish __ Portuguese __ Russian __
Spanish __ Taiwanese __ Turkish __
Other: (_______________________)

5. What is your first or native language? _______________________

6. How many years have you been studying English? ____ years

7. Which of the following skills would you like to improve on while studying English? (Please check all items that apply.)
Reading __ Writing __ Speaking __ Listening __

8. How many hours do you engage in the following activities per week? (Please check all the boxes that are applicable.)

	Less than 1 hr.	**1-2 hrs.**	**2-3 hrs.**	**3-4 hrs.**	**More than 4 hrs.**
READING in English					
Book					
Internet Surfing					
Magazine					
Newspaper					
WRITING in English					
Cell phone texting					
Email					
Fill in forms					
Internet chat room writing/ IM (Instant Messaging)					
Shopping lists					
Taking notes					

APPENDIX 3

Reading Section of CEP Placement Test

Theme: History of New York State
Length: **20 minutes** to read 2 passages and answer 12 questions
Passage 1: Growth of New York State
Passage 2: History of West Point Military Academy

Reading Task 1

Instructions: Read the passage and answer the questions.

George Washington first proposed a military academy in 1783, but critics opposed this new idea of a special school to train army officers as too European. They believed it contradictory with democratic institutions, fearing the creation of a military aristocracy. Finally, two decades after Washington's first proposal, on 16 March 1802, Congress approved legislation establishing the United States Military Academy at West Point, one of the oldest military service academies in the world. It stood on a commanding plateau overlooking the Hudson River at West Point, New York, 50 miles north of Manhattan.

During the Revolutionary War, both the colonists and the British realized the importance of gaining possession of the Hudson River valley, and West Point became the key to its defense. It was a very important defense commanding the Hudson with artillery and a 136-ton chain strung across the river to prevent enemy ships from passing. In the Civil War, 90% of the commanders on both sides were West Point graduates.

Colonel Sylvanus Thayer, superintendent at West Point from 1817 to 1833 is credited with establishing the high standards of discipline and scholarship for which the Academy is known today. Under Thayer's term, cadets, students in the armed forces, were trained as civil engineers as well as soldiers. After graduation from West Point, the graduates put their technical skills to work for the U.S. government in construction of canals, roads, railroads, and other internal improvements needed to facilitate westward expansion.

Both Ulysses S. Grant and Robert E. Lee were educated at West Point. Other famous graduates include Union generals George H. Thomas and George A. Custer, President of the Confederate States of America Jefferson Davis, World War I hero General John J. Pershing, and Dwight David Eisenhower, Supreme Allied Commander at the time of the D-day invasion during World War II and the thirty-fourth U.S. President.

(Adapted from http://americanhistory.si.edu/westpoint &
http://memory.loc.gov/ammem/today/mar16.html)

1. In line 3, what does 'proposed' mean?
   a. advised
   b. offered
   c. intended
   d. suggested (VOCABULARY IN CONTEXT)

2. Who approved legislation establishing the West Point?
   a. Congress (DETAIL)
   b. Judiciary
   c. Assembly
   d. Parliament

3. In line 12, what does 'defense' mean?
   a. responding to an emergency after attack
   b. avoiding prosecution from a criminal act
   c. preparing for the battle against the attackers

d. preventing an army from conquering territory (VOCABULARY IN CONTEXT)

4. Colonel Sylvanus Thayer recognized the importance of ________________.
   a. civilian propriety
   b. academic standard (INFERENCE)
   c. political standpoint
   d. military punishment

5. What is the best topic sentence for the last paragraph?
   a. Many renowned people attended the academy. (GIST)
   b. Many war heroes graduated from the academy.
   c. Many former presidents studied at the academy.
   d. Many commanders were trained at the academy.

6. What is the main idea of the passage?
   a. West Point graduates led many of the expeditions westward.
   b. West Point graduates set high standards of military leadership.
   c. West Point graduates played an integral role in the U.S. history. (GIST)
   d. West Point graduates dominated the highest ranks during the war.

Reading Task 2

Instructions: Read the passage and answer the questions.

In the last years of the 1700s large tracts of land in central and western New York, a state in the Middle Atlantic region of the United States, had been opened for settlement. An area extending from just below Ithaca to Lake Ontario, called the Military Tract, was reserved for veterans of the American Revolution. Lands west of Seneca Lake that had formerly been owned by Massachusetts were turned over to New York and sold to business leaders and speculators. After the revolution, New England's farmers, discouraged by stony soil and high taxes, moved west into New York. In 1820 half the state's inhabitants were New Englanders or their descendants. Central and western New York were quickly settled, but the north country remained a sparsely populated wilderness for many years.

From the 1820s to 1860, New York state and especially New York City were transformed by a flood of immigrants and unprecedented urban growth. The United States experienced a wave of immigration from Europe, starting in 1820 and reaching a peak in 1845. For most immigrants, their point of entry was New York City, and huge numbers of them stayed there. By 1860 New York City was the nation's largest city, with a population of 1 million, and nearly half of those residents were foreign-born. New York state's population also exploded during this period, from 340,120 people in 1790 to more than 3.8 million in 1860. Nearly one-fifth of those residents were foreign-born. In contrast to colonial times, when most New Yorkers lived on farms, by 1860 about half of them lived in cities and towns.

The largest group of immigrants was the Irish, especially after a severe potato famine struck their homeland in the 1840s. Many settled in New York City, while German immigrants tended to settle upstate, especially in Buffalo and Rochester. This influx of people included skilled European craftsmen and a huge pool of low-wage laborers that enabled New York to develop diverse industries. But low wages, long hours, and harsh working conditions made life difficult for many industrial workers. To try to improve their conditions, many skilled artisans and factory workers, who were mostly women and children, joined labor unions. With such rapid population growth, New York City and other urban areas faced problems of inadequate water supplies, sanitation and housing.

By 1830 New York ranked first among the states in population, manufacturing, trade, commerce, and transportation. New York City emerged as the primary center for textile manufacturing and ready-made clothing, banking, imports, insurance, and the stock exchange. From 1825 to the late 1850s the Genesee Valley was a national center for growing wheat. Other important products included livestock, corn, barley, oats, and hops. When the Midwest became the major source of grain, New York's farmers turned to dairy products, fruits and vegetables. They supplied great quantities of milk, butter and cheese to the growing cities.

(Adapted from Encarta, www.encarta.msn.com)

7. Choose the best title for this passage.
   a. Growth of New York State (GIST)
   b. Industry of New York State
   c. Immigrants of New York City
   d. Urban Growth of New York City

8. In line 5, what does "reserved for" mean?
   a. taken by (VOCABULARY IN CONTEXT)
   b. booked by
   c. prepared for
   d. organized for

9. Which of the following is **correct** according to the passage?
   a. At the beginning of settlement, people moved to northern New York first.
   b. Half of New York City's residents was comprised of immigrants by 1860. (DETAIL)
   c. In the mid-1800s New York City became a national center for agriculture.
   d. The number of New York urban area residents decreased in the mid-1800s.

10. Which of the following factors caused Irish immigrants to move to New York state?
   a. Severe potato famine (DETAIL)
   b. Stony soil and high taxes
   c. Harsh working conditions
   d. Inadequate water supplies

11. What can be inferred from the passage about the role of labor union?
   a. It fought for the rights and welfare of employees.
   b. It helped improve employees' working conditions. (INFERENCE)
   c. It set the minimum living standard for employees.
   d. It helped change women's social status in the society.

12. In 1800s New York City can be characterized as a _______________.
   a. center for a variety of industries (INFERENCE)

b. place for immigrants' settlement
c. home for foreign skilled workers
d. place for diverse group of people

APPENDIX 4

Writing Section of CEP Placement Test

Directions: Imagine you are back in your home country after studying English at the Community English Program. Your classmates from the program love to travel and want to visit your home country. Write a postcard to your classmate describing the 3 most interesting places to visit in your home country (20 minutes).

Make sure that your descriptions are clear and organized using correct grammar

To: Jenny Yoo
509 W. 121st Street, APT #101
New York, NY 10027
U.S.A.

____

____

____

____

____

____

____

____

____

APPENDIX 5

A Sample Response in the Writing Section of the Placement Test

Dear Jenny

I am very happy that I came back to my country but I miss you so much. I would like to invite you to Poland so I would be able to show you our cities and history.

The first place I would like you to see is Krakow. This city is the south part of Poland and it was capital of Poland in past ages. The main attraction there is Wawel castle and its symbol dragon. Also Old Town is really nice. You must see St. Mary's church and Market Square-the main point of Old Town. You can't leave this city without listening to the "hejnal," which is a melody played at every full hour from the top of Mary's church.

The second place I would like you to see is Wieliczka-old salt mine. You can go underground to see all caves, pools and ways used to transport the salt. There is a hotel and tennis court there also.

The third place I would recommend for you is Zakopane. It is a city on the south end of Poland surrounded by Tatry Mountains. This place is good for hiking in the winter as well as hiking in the summer. But there are much more outdoor activities. Zakopane is known as a city of mountains for people who keep their culture, dialect, dance and songs.

You have to come and see all of these. Hope to see you soon.

Kasia

APPENDIX 6

Analytic Scoring Rubric for Writing

FORM	**Level of Control**	**Descriptors**
	5 Complete control	Language control is generally accurate; meanings not affected by minor errors; range of grammatical structure and vocabulary use is sophisticated and varied
	4 Extensive control	Language shows satisfactory but inconsistent control; meanings generally not affected by errors; range of grammatical structure and vocabulary use shows adequate variety
	3 Moderate control	Language shows inconsistent control; meanings sometimes affected by errors; simple grammatical structure; vocabulary use shows a lack of variety
	2 Limited control	Language shows inconsistencies that distract the reader; meanings often impeded by major errors; limited linguistic control. vocabulary use is restricted
	1 Little or no control	Language control frequently distracts the reader; meanings obscure due to major errors; lack of linguistic control; vocabulary use is highly restricted and/or inaccurate

ORGANIZATION	Level of Control	Descriptors
	5 Complete control	The response shows excellent rhetorical control and displays unity; ideas are balanced with support that is organized according to the content; textual elements are connected
	4 Extensive control	The response shows strong rhetorical control; ideas are generally balanced with support; strong control of organization that is appropriate to the content; textual elements are generally well connected
	3 Moderate control	The response shows acceptable rhetorical control; appropriate organization; ideas may not be balanced with support; a lack of connectors sometimes interferes with fluency
	2 Limited control	The response shows a lack of rhetorical control much or most of the time; the organization is difficult to discern; organization suggests a lack of balance of support; partial success for transitions within and across sentences
	1 Little or no control	The response shows little rhetorical control, no evidence of organization; underdeveloped or nonexistent connections and transitions

CONTENT	Level of Control	Descriptors
	5 Complete control	The response is admirably thorough and complete.
	4 Extensive control	The response is thorough and complete.
	3 Moderate control	The response lacks thoroughness but complete.
	2 Limited control	The response lacks both thoroughness and completeness.
	1 Little or no control	The response fails to develop and support an argument related to the topic.