A Statistical Analysis of Reading and Writing Ability using SPSS (The CEP Placement Test at Columbia University)

Moananu, Charlton Bill*

Abstract

Acquiring effective and efficient reading skills is critical in second language education in Japan. Students take a variety of standardized examinations that require a high level of reading ability. One crucial element often overlooked when teaching reading skills is the relationship reading has with writing. In a previous paper (Moananu, 2009), ample evidence to support the hypothesis that two skills were highly correlated was provided. In addition, a theoretical model for reading and writing as well as a model depicting their correlation was put forward. However, that study was based only on theoretical facts with little practical evidence. To this end, the present paper aims to further prove this relationship through statistical analysis using SPSS. It is hoped that the results will support the hypothesis, including the theoretical models put forward in the previous paper, as well as the notion that reading and writing skills should be taught in conjunction with one another.

1. INTRODUCTION

Many theorists regard reading and writing as two interrelated processes and this notion has been supported by empirical studies that have confirmed that students who read more also write better (Janopolis, 1986). Other researchers such as Bachman and Palmer (1996) also suggested the relationship between reading and writing in that, grammatical knowledge (vocabulary, morphology and syntax); textual knowledge (rhetorical, cohesion); functional knowledge (ideational, heuristic); and sociolinguistic knowledge (register, dialect) are all important aspects that overlap in both processes (p. 256).

However, drawing hypotheses about reading and writing skills and their relationship, based only on a literature review of reading and writing seem unconvincing since only facts and theories are being put forward. To add credence to the hypothesis, an analysis of the reading and writing sections of the CEP placement test administered by Columbia University's ESL program was tested to determine if the two sections were correlated. If the analysis, through statistical evidence, showed that the two sections were correlated, and if the test was proven to be reliable and internally consistent, then one can postulate, with a high level of certainty, that both skills are correlated. The evidence would add practical evidence to theory and thus strengthen our case.

In order to achieve this, the reading and writing sections of the CEP placement test were analyzed. First, study participants, the target language use (TLU) domain, the overall specifications of the test and what

*一般教養科(准教授)
e-mail: moananu@nc-toyama.ac.jp

（平成22年４月28日受付）

Following this, a description of the writing tasks including the essential elements in evaluating the writing is followed by an explanation of the scoring procedure. Finally, various statistical test analyses on the reading and writing sections of the test are performed. The following components were tested; descriptive statistics, internal consistency reliability, standard error of measurement, item analysis (reading section), inter-rater reliability (writing section) and evidence of construct validity. Finally, the conclusions and discussions section details the findings and discussion about the results

2. TEST DEVELOPMENT and CONSTRUCTION

Test development is the entire process of creating and using a test, beginning with its initial conceptualization and design, and culminating in one or more archived test and the results of their use (Bachman & Palmer, 1996). The amount of time and effort put into making a test will depend on the situation. At one end with low stakes tests, the process may be rather informal, for example a weekly word quiz that is assigned a grade. On the other end with high stakes tests the process may be very complicated and time-consuming requiring extensive trialing and revision as well as a number of people. Such tests are characteristic of many formal examinations like; TOEFL, IELTS and the CEP placement test (which I analyze) that affect a large number of people.

Bachman & Palmer (1996) stated that regardless of the situation, all tests should adhere to careful planning of the test development process for three reasons. First, careful planning provides the best means for assuring that the test will be useful for its intended purpose. Second, careful planning tends to increase accountability: the ability to say what was done and why it is important because as teachers we must expect that test users (students, parents, and administration) will be interested in the quality of our tests. Third, careful planning increases the amount of satisfaction we experience; the effort put into the task promote a sense of accomplishment after the task is completed.

Bachman & Palmer (1996) conceptually organize test development into three stages: design, operationalization, and administration. The word “conceptually” is used because the test development process is not sequential in its implementation (p. 86). In the design stage, the test components are designed to insure that performance on the test tasks will correspond as closely as possible to the language use, and that test scores will be optimally useful for their intended purpose.

Test design includes the following components: (a) a description of the purpose(s) of the test; which makes explicit the specific uses for which the test is intended, (b) a description of the TLU domain and task types; which makes explicit the task in the TLU domain to which we want our inferences about language ability to generalize, and describes TLU task types in terms of distinctive characteristics, (c) a description of the test takers for whom the test is intended; which makes explicit the nature of the population of potential test takers, (d) a definition of the construct(s) to be measured; which makes explicit the nature of the population of potential test takers, (e) a plan for evaluating the qualities of usefulness; which includes activities that are part of every stage of the test development process, (e) an inventory of required and available resources and a plan for their allocation and management which makes explicit the resources (human, material, time) that will be required and that will be available for various activities during test development, and provides a plan for how to allocate and manage them throughout the development process (Bachman & Palmer, 1996).

Operationalization involves developing test tasks specifications for the types of test tasks to be included on the test and a blueprint that describes how test tasks will be organized to form actual tests. Also included in this stage is, specifying the scoring method which includes: defining the criteria by which the quality of the test taker's responses will be evaluated and determining the procedures that will be followed to arrive at a score.

Finally, Test administration involves giving the test to a group of individuals, collecting information, and analyzing this information for two purposes: (1) to assess the usefulness of the test, (2) to make inferences or decisions for which the test is intended (Bachman & Palmer, 1996).

A. Validation in Testing.

Validation is central to testing concerns, and the identification of a suitable construct or constructs is central to this validation. Thus reading assessments should be based on the best constructs available. However, as Alderson (2000) states, "there is no agreement on what such a construct might be" (p. 111). This is due to the major disagreements concerning higher level processing, about the nature and contribution of inferencing, the role of other cognitive processes and abilities in reading (Alderson, 2000). It seems that under the current circumstances regarding the assessment of reading, the most appropriate course of action for assessing reading according to Alderson (2000) is to apply the traditional criteria of assessing the test for reliability and validity. To the extent that such criteria apply will depend upon the purpose of the test, and whether it is high stakes or low stakes.

3. METHOD

A. Study Participants

In the statistical analysis, 156 study participants took the CEP placement test. The test takers consisted of 112 females, 42 males and 2 failed to indicate their gender. Based on a quick look at the names in the data set and also the very cosmopolitan, melting pot-like environment in New York, I would assume that the test takers come from a variety of different ethnic backgrounds.

B. Measuring Instrument

1. Target Language Use Domain of the Entire Test

The CEP program attracts a variety of individuals (nationality, native language, proficiency level, professional background, etc.) who have different purposes for entering the program and so to define the target language use domain (TLU domain) with one specific aspect is rather difficult. In addition, it is possible to list numerous target language use settings depending on the situations which students may encounter in their everyday lives. Thus, the TLU domain of the CEP test is language-instructional where academic English is emphasized. In general, students enter the program to develop academic English in writing, reading, speaking, listening and grammatical abilities in order to communicate and function in the North American context (including university context). Therefore, the test domain would include these five areas: reading, writing, listening, speaking and grammar. The TLU tasks as identified in "the purpose" of the CEP test for the reading and writing sections are: reading a passage and responding to comprehension questions (reading) and writing essays based on prompts (writing).

The reading part of the test consisted of four reading tasks that covered a wide range of themes including ants, sensory functions, and Neanderthals. The readings and the questions measured the reader's ability to understand reading for gist, reading to find detailed information, ability to make inferences about the writer's intentions and the ability to deduce vocabulary in context. The two writing tasks required students to write an organized composition based on the following two prompts; writing a post card to a friend planning to visit New York, and writing about the transportation system in New York.

2. Overall Specifications of the Entire Test

As stated earlier, the CEP placement test consists of five sections: writing, reading, speaking, listening and grammar (Purpura, 2004). The reading section consisted of four tasks with 30 selected response items (multiple-choice). The allotted time was 45 minutes. This section was scored dichotomously using an answer key. The writing section consisted of two extended production tasks. The time allotted for task one was 15 minutes and task two was allotted 30 minutes. The two tasks were scored blindly by two raters. The following chart (blueprint) provides details about the task component, task type, the time allotted for the task, the length or number of questions and the scoring criteria for all sections of the test.

C. Scoring procedure

The multiple choice questions in the reading section were scored dichotomously, using the answer key that specified a single correct answer. The scoring in this section was objective, one point was given for a correct answer and zero points for incorrect answers. The total possible score was 30 points. Tester's writing abilities were evaluated on the following elements: task fulfillment, content control, organizational control, and language control. The writing tasks were scored using a 0 - 5 **point analytic rubric rating scale**(Bachman & Palmer, 1996) (with five being the highest) where each of the four elements received a score thus the maximum score in the writing section is 20 points per task for a total of 40 points on both tasks(see Appendix 3 for the scoring rubric). The average score of all four elements for rater 1 was averaged with the average score of all four elements of rater 2 to get a final average. All essays were scored 'blind' by two trained raters, with a third adjudicating discrepancies greater than two points (Purpura, 2004, p. 204).

4. TEST ANALYSES AND RESULTS

4.1 Results for the Multiple Choice (MC) Task: Reading Ability

A. Descriptive Statistics

The number of participants was 156 (N=156). There were 30 multiple-choice items (k=30) with the maximum possible score of 30. The lowest score was 4 and the highest score was 30. The mean was 19.4 and the median was 20. The standard deviation (SD) was 6.052, the kurtosis was -0.765 and the skewness was -0.309. The range was 26. A summary of these results is shown in Table 1.

Table 1: Descriptive Statistics for the Reading Task

Statistics	
Number of participants (N)	156
Number of items (k)	30
Maximum possible score	30.00
Mean	19.39
Median	20.00
Standard deviation	6.05
Kurtosis	-.765
Skewness	-.309
Range	26.00
Minimum	4.00
Maximum	30.00

The negative skewness value of -0.309 indicated that there was a higher frequency of higher scores and a lower frequency of lower scores. This can be interpreted to mean that the test-takers may have found the reading section to be somewhat easy. The negative kurtosis value of -0.765 indicated that the distribution was somewhat platykurtic or flat. This may imply that there was slight variability or heterogeneity among all the test-takers on the reading section of the CEP test. This may also explain the relatively high standard deviation. Students performed well on the test as indicated by the central tendency, the mean of 19.397 and the median of 20.000 which is roughly equal to a grade of 66% on the reading which is fairly high. In terms of dispersion, the range was 26 from the maximum score of 30 to the minimum score of 4. This indicates that the lowest score, 4 along with 3 scores of 7, were more than 2 standard deviations below the mean which clearly indicates that the proficiency of some participants was not at the same level as that of the others. However, given the wide range of background and English ability of the participants, the result is expected. A stem-and-leaf plot is provided in Figure 4 to show the distribution of the scores including the negative skewness, negative kurtosis and the range of distribution.

Figure 1: Stem-and-Leaf Plot for the Reading Task

```
Frequency     Stem &   Leaf
.00           0.
1.00          0.   4
3.00          0.   777
8.00          0.   88889999
4.00          1.   0111
12.00         1.   222222223333
16.00         1.   4444444444445555
14.00         1.   66666666777777
14.00         1.   88888888899999
18.00         2.   000000000001111111
19.00         2.   2222222222333333333
21.00        2.   444444444555555555555
14.00         2.   66666677777777
10.00         2.   8888899999
2.00          3.   00
```

B. Internal Consistency Reliability and Standard Error of Measurement

The internal consistency reliability shows the extent to which the test results are consistent and stable. The Cronbach's Alpha (widely used with dichotomously scored items) was used to calculate the internal consistency reliability. The reliability for the 30 items in the reading test section was 0.864, which means that the scores are about 86% consistent or stable.

Table 2: Reliability Coefficient for the Reading Task

Cronbach’s Alpha	Number of Items (k)
0.864	30

Reliability values range from 0 to 1, 0 representing no reliability and 1 as perfect reliability. Thus 0.864 indicates that the items on the reading test were highly homogeneous and the reading section of the test was very consistent and reliable. I also calculated the standard error of measurement (SEM). The SEM is used to determine the probable range in which a participant’s score would fall if the test were administered to the same participant repeatedly. The formula to calculate the SEM is $SEM=S\sqrt{1-r_{xx}}$’ where S refers to standard deviation and r_{xx} refers to the reliability coefficient. Based on a standard deviation of 6.052 and a Cronbach’s Alpha of 0.864 the SEM for the reading task was calculated as 2.23. By calculating a 95% confidence interval (±2 SEM), we can estimate the range within which the participant’s score would fall with 95% confidence. For example, if a participant scored 12, we are 95% certain that his or her true score would fall somewhere between 7.54 and 16.46. This means that in this reading section, the tester’s score would fall between 8 and 16 even if the tester took the same test repeatedly.

C. Item Analyses

To assess the quality of the MC items, the difficulty index and the discrimination index were calculated for each item in the following table.

Table 3: Item Analyses for the MC Task

Item	Difficulty	Discrimination	Alpha if delete	Decision
44	0.9	0.388	0.86	Modify
45	0.87	0.227	0.863	Modify
46	0.89	0.286	0.862	Modify
47	0.83	0.314	0.861	Keep
48	0.81	0.461	0.858	Keep
49	0.87	0.328	0.861	Keep
50	0.73	0.414	0.859	Keep
51	0.76	0.5	0.857	Keep
52	0.65	0.575	0.854	Keep
53	0.81	0.37	0.86	Keep
54	0.72	0.231	0.864	Keep
55	0.54	0.443	0.858	Keep
56	0.63	0.459	0.857	Keep
57	0.58	0.413	0.859	Keep
58	0.79	0.313	0.861	Keep
59	0.31	0.198	0.865	Modify
60	0.72	0.24	0.863	Keep
61	0.5	0.417	0.859	Keep
62	0.47	0.253	0.863	Keep
63	0.69	0.559	0.855	Keep
64	0.65	0.422	0.859	Keep
65	0.63	0.406	0.859	Keep
66	0.31	0.42	0.859	Keep
67	0.51	0.321	0.862	Keep
68	0.44	0.386	0.86	Keep
69	0.37	0.523	0.856	Keep
70	0.71	0.526	0.856	Keep
71	0.44	0.418	0.859	Keep
72	0.58	0.368	0.86	Keep
73	0.67	0.447	0.858	Keep

The mean of each item indicates the difficulty index, which is the proportion of the participants who got the item correct. This value ranges from 0 to 1. 0 indicates that no one got the item correct while 1 indicates all testers got the item correct. The discrimination index examines how each item functions to discriminate between high ability testers and low ability testers. This value ranges from -1 to 1. A negative value would mean that more low ability testers got the item correct than high ability testers, which is not desired. The "Alpha if deleted" column in Table 6 shows the change in overall internal consistency reliability from the deletion of an item. Therefore, the decision on whether to keep, revise, or delete an item has to be made based on the values of the difficulty, the discrimination and the Alpha if deleted.

The difficulty index ranges from 0.31 to 0.9 and the discrimination index ranged from 0.198 to 0.575. The difficulty indices of items 44, 45 and 46 showed that about 90% of the participants answered these items correctly. Considering that the mean score was 19.4, which is 64.5%, these items seemed to have been easy for the participants. However, the difficulty index and discrimination index for item 59 was 0.31 and 0.198 and item 68 was 0.44 and 0.386 respectively, which suggests that both were relatively difficult items and they did not discriminate high performers from the low performers. A close examination of items 59 and 68 showed that both were measuring reader's ability to read for details. When examining the distracters for the two items, at least two out of the four distracters had vocabulary that was in the text and this may have confused both some high and low achievers. Also, if we look at the Alpha if deleted variable, only minimal variance is detected. As a result I suggest items 44, 45, 46 and 59 be looked at for possible modification and not deletion.

D. Distracter Analysis.

For the distracter analysis, I chose one good item, item 44 and one poor item, item 59. For both items, I looked at the difficulty, the discrimination and the reliability, "Cronbach's Alpha if deleted", in order to analyze them. The mean for item 44 was .90 or 90% of the test takers got it right, indicating that this item was relatively easy. Table 4 shows the breakdown of the item.

Table 4: Frequency Analysis of Item 44

Key	Frequency	Percent
A	2	1.28
B	5	3.20
C	141	90.38
D	8	5.12
Total	156	100.00

The discrimination index was .388 indicating that it moderately discriminated between high achievers and low achievers. The key was (C) and 141 out of 156 or 90.38% of the test takers got it correct. The distracter (D) was the most attractive distracter attracting eight or 5.12% of the participants. Distracter (B) attracted five or 3.20% of the testers and distracter (A) attracted 2 or 1.28% of the participants. If the Cronbach's Alpha if deleted is applied, the internal consistency reliability would increase to 0.86 which is very similar to the current value of .864. After further examination of the item, I decided to modify not delete the item because the items is asking for the meaning of an important vocabulary item within the context (the meaning of "fuzzy" is crucial in understanding the e-mail).

For item 59, the difficulty was 0.31 meaning only 31% of the participants got the answer correct. The discrimination index was 0.198 which is low and also an indication that this item was not discriminating low achievers from high achievers. A close examination of this item showed that the low discrimination index was due to the performance of the distracters. Table 5 provides the data for item 59.

Table 5: Frequency Analysis of Item 59

Key	Frequency	Percent
A	49	30.81
B	9	5.80
C	59	38.06
D	38	25.16
Total	155	100.00

The key was (A) with 49 or 31.6% of all participants chose (A), while 59 or 38% chose (C), 38 or 24.5% chose (D), and 9 or 5.8% chose (B). Distracter (D) attracted more participants than the key. Upon further analysis of item 59, I noticed that the distracters were functioning "too well". The question asked student to determine the best heading for paragraph three. Thus students were looking for "details" within the paragraph to determine the best heading. The answer for distracter (C) was "combination of different odors". In the first sentence of paragraph 3, there is mention of the numerous odors the sensory organ is able to detect. This may have confused some participants who may have been reading for gist or topic sentence instead of details. The answer was more distinguishable toward the end of the paragraph and some students may have just skimmed through it. The answer for distracter (D) was "the relation between smelling and reading". In the context of this paragraph, letters were used as analogies to explain how nasal sensors decode smells. Again, participants may have been confused by what appeared to be a logical answer. However, I have decided to modify not delete the items because I feel this question is important in measuring students' ability to read for details. It should be noted that one participant did not answer this question and thus there were only 155 responses.

E. Evidence of Construct Validity within the MC Task

The evidence of construct validity examines whether each of the four variables; gist, detail, inference, and vocabulary in context, in the reading test are actually measuring reading ability. In order to examine this I used the **Pearson product-moment procedure** because I used the total score for each variable thus they are interval scales. The value of the correlation coefficient can range from negative one to positive one. According to theoretical models of writing we expect to see positive correlations between all the components because they all belong to the same construct, reading. A negative correlation would be undesirable because this would indicate that the components within the construct may not be measuring the construct. For example a student may get perfect scores for the questions measuring vocabulary in context but get zeros for questions measuring detail. Table 6 provides the data.

Table 6-Correlation Matrix between Observable Variables: Reading (k=30, N=156)

Scale	Gist	Detail	Inference	Vocabulary
Gist	1.000			
Detail	.616**	1.000		
Inference	.465**	.601**	1.000	
Vocabulary	.629**	.751**	.597**	1.000

** Correlation is significant at the 0.01 level (2-tailed)

These correlations can be high (r = .75 or above), moderate (r = .5 to .74), low (r = .2 to .49), uncorrelated (r < .25), not correlated (r = 0) or negatively correlated. The correlation coefficient of .751 between detail and vocabulary shows the highest correlation which would indicate these two variables are indeed measuring reading ability. The correlation coefficients of .629 between gist and vocabulary, .616 between gist and detail, .601 between detail and inference .597 between inference and vocabulary suggests that there was a moderate correlation between these variables. Thus we can say these 3 sets of paired variables moderately measure reading ability. Finally, the correlation coefficient of .465 between gist and inference suggests that there was a low correlation between these variables. Thus paired variables of gist and inference do not measure reading as well as the previous three pairs. All the correlations are statistically significant at the 0.01 level (p < .01). Thus there is a 99% chance that the observed correlations between gist, detail, inference and vocabulary in context are not due to pure chance.

Therefore we can say that there is sufficient evidence to suggest that these variables served as successful tools for measuring reading ability. The high internal consistency reliability of .864 most likely contributed to these positive correlations.

4.2 Results for the Extended Production (EP) Task: Writing Ability

A. Descriptive Statistics

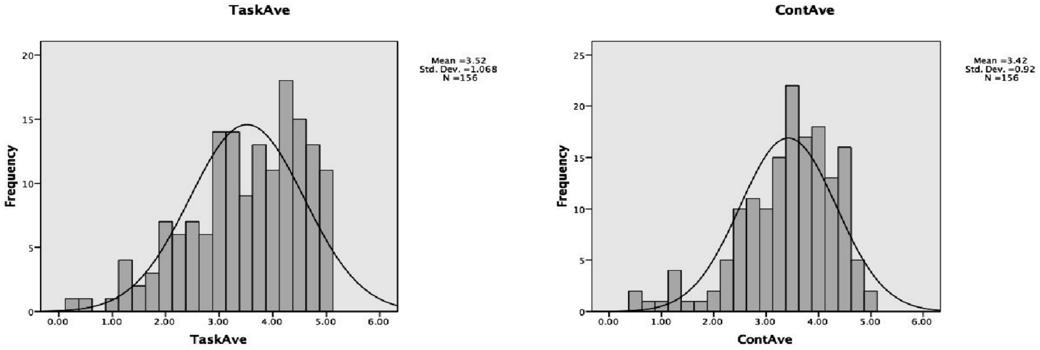
There were two raters who rated the extended production on a scale of zero to five on four observable variables; task fulfillment, content control, organizational control, and language control.　These four scores were averaged for each rater, and the average of the raters' total scores was again averaged, making 5 the maximum possible score.　The scores for the task fulfillment showed the highest mean and median (3.519 and 3.75 respectively) followed by the content control (3.4231 and 3.5), organization control (3.141 and 3.25) and language control (2.9872 and 3).　The standard deviation for each variable followed the same pattern as the mean and the median; the task fulfillment being the highest (1.06806) followed by the content (0.92047), organization (0.89089) and language control (0.78944).　The negative skewness indicates that most test takers performed well on the writing tasks. Table 7 shows the summary for each variable and the total writing score.

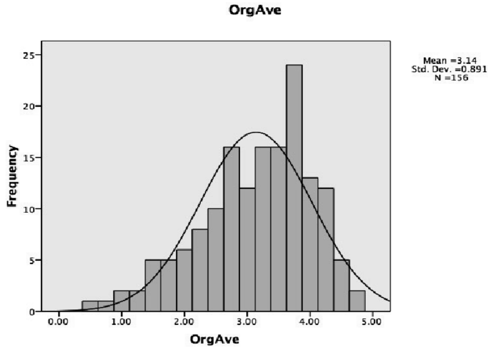
Table 7: Descriptive Statistics for the Writing Task

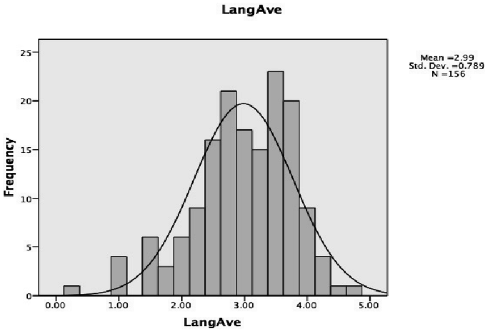
	TaskAve	ContAve	OrgAve	LangAve	WriteAve
Mean	3.5192	3.4231	3.141	2.9872	3.2676
Median	3.75	3.5	3.25	3	3.4375
Std. Dev.	1.06806	0.92047	0.89089	0.78944	0.86753
Skewness	-0.67	-0.931	-0.622	-0.667	-0.809
Kurtosis	-0.11	0.934	-0.073	0.518	0.401
Range	4.75	4.5	4.25	4.5	4.38

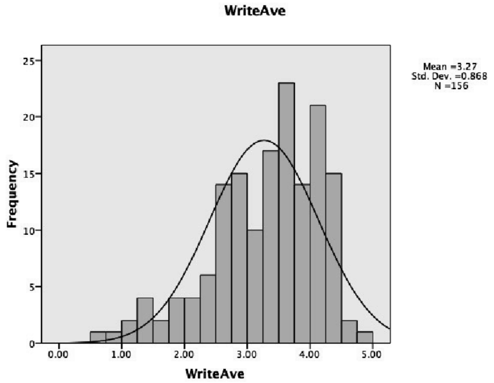
The differences in mean and median seem to demonstrate the raters' tendency in rating participants' performance; the more language specific the variable was, the stricter they became in rating. The negative skewness with all the variables indicates that there was a higher frequency of high scores in all these variables. The positive kurtosis for content (0.934) shows that it was leptokurtic or peaked and for language (0.518) suggests that it was somewhat leptokurtic or peaked. However, the value is within the range of -1 to 1 which is consistent with a more normal distribution. The negative kurtosis for task and organization indicates that the score distribution was somewhat platykurtic or flat (by definition). However, as stated earlier, because the figures for task and organization are within the -1 to 1 range, the distribution appears normal. Figure 2 depicts histograms for the descriptive statistics for the writing task

Figure 2: Histogram for Writing Average









B. Internal Consistency Reliability and Standard Error of Measurement

The internal consistency reliability of the extended production refers to the internal consistency of a test or how well the score for items on a test correlate with each other. The Cronbach's Alpha was used because it can be applied to ordinal scales as well as dichotomous items. The internal consistency reliability of the writing test was calculated by using the averaged scores of two raters for the task fulfillment, content, organization and language control variables. The alpha was 0.956, which means that the scores were about 96% consistent or stable (see Table 11). This indicates that the variables assessed in the writing test (TaskAve, ContAve, OrgAve and LangAve) were highly homogeneous and the test was trustworthy.

Table 8: Reliability Coefficient for theReading Task.

Cronbach's Alpha	Number of Items (k)
0.956	**4**

The SEM for the writing task was calculated using the same formula as before. The estimated SEM was 0.181 using the standard deviation of 0.867 and Cronbach's Alpha of 0.956. This would indicate that if the tester were to take the same test a number of times, the score would fall within the same range. The SEM is quite small indicating that the error deviation would also be small. By calculating a 95% confidence interval (+/-2 SEM), we can estimate the range within which the participant's score would fall with 95% confidence. For example, if we decided to set the passing mark of the writing test at 3, we are 95% certain that a true passing mark could fall anywhere between 2.638 and 3.362

C. Inter-Rater Reliability

Inter-rater reliability was calculated in order to examine the degree to which the first rater's scores on one variable correlated with a second rater's scores on the same variable.　To calculate the inter-rater reliability for the writing test, we used the averaged scores of rater 1 and rater 2 from the two writing tasks for four different variables (task fulfillment, content, organization and language control). Pearson Product-Moment Correlation was used for this as we were dealing with averaged scores, which are continuous. Table 12 shows that the inter-rater reliability was relatively high.

Table 9: Inter-Rater Reliability Correlation Matrix for Writing Test

NOTE: all stars follow the last digit. Thus Task R2 should read .757 and so on.**

	Task R1	Cont R1	Org R1	Lang R1	Wrt R1
Task R2	.757 **	.690 **	.673 **	.601 **	.734 **
Cont R2	.721 **	.735 **	.707 **	.666 **	.759 **
Org R2	.669 **	.668 **	.660 **	.646 **	.708 **
LangR2	.621 **	.662 **	.670 **	.668 **	.700 **
Wrt R2	.768 **	.760 **	.747 **	.709 **	.801 **

**=Correlation is significant at the .01 level (p<.01)

The correlation between TaskR1 and R2 was 0.757, ContR1 and R2 was 0.735, OrgR1 and R2 was 0.660 and LangR1 and R2 was 0.668.　The correlation between WrtR1 and R2 was much higher at 0.801. Each of these values was statistically significant at the 0.01 level (p < .01), which means that the probability that these correlations are due to chance is less than 1%.　The high correlation between the two raters can be interpreted to mean that their understanding and interpretation of the scoring rubric were relatively homogeneous. We can also assume that the raters interpreted the scoring rubric in a similar fashion and shared the understanding of what elements should be included in task fulfillment, organizational control, content control and language control. Overall, the inter-rater reliability coefficients provide further evidence that the writing test is reliable.

Looking at the internal-consistency reliability (0.956) and the inter-rater reliability coefficient (0.801) of the writing test, it appears that the latter provided a more conservative estimate of reliability. This seems to make sense because the internal-consistency measure estimates the relationships between the items or observable variables tested by the writing task while the inter-rater reliability estimates the relationships between the decisions made by the raters.　It can be said that the internal-consistency reliability represents internal reliability because it is based on the internal relationships of the subtest while the inter-rater reliability represents external reliability because even when the same test is given, the raters may be different, which may change the inter-rater reliability.

D. Evidence of Construct Validity within the Extended Production Task

In order to provide evidence of construct validity within the writing section (2 extended production tasks) of the test, a correlation analysis was performed on participants' average scores on each of the four variables as well as the overall writing average scores. As shown in Table 11, correlations were high. Strong correlations were found between task and organization (0.830), task and content (0.871), and content and organization (0.888). The weaker correlation was found between task and language (0.741). These correlations are significant at the 0.01 level (p < .01), meaning the probability is less than 1% that the correlations are due to chance.

Considering the task, which was to write a postcard to a friend who is planning to visit New York and to discuss the advantages and disadvantages of public transportation in New York, such high positive correlations are expected. Participants are encouraged to demonstrate their language, organization, and content control by following the task instructions, and the relationships among these variables should be assessed as a representation of their writing ability.

Table 10: Correlation Matrix between Variables: Writing Test (N=156)

	TaskT2	ContT2	OrgT2	LangT2
TaskT2	1	.871**	.830**	.741**
ContT2	.871**	1	.888**	.823**
OrgT2	.830**	.888**	1	.863**
LangT2	.741**	.823**	.863**	1

**=Correlation is significant at the .01 level (p<.01)

4.3 Other Evidence of Validity

A. Relationship between the Two Parts of the Test.

One of the key points in this research paper was to find a correlation between the reading and writing sections of the CEP placement test in order to confirm my hypothesis based on my research, that reading and writing skills are, to some degree, correlated. In order to address this hypothesis, I calculated the correlation between the participants' performance on the reading test and the writing test.

The result is shown in Table 14.

Table 11: Correlation Matrix between Reading and Writing Abilities (N=156)

	Reading Total	Writing Average
Reading Total	1	.727**
Writing Average	.727**	1

**= Correlation is significant at the .01 level (p< .01)

The correlation coefficient of 0.727 suggests that there is a strong correlation between the two variables. The correlation is statistically significant at the 0.01 level (p < .01). Based on this result, we can say that reading and writing, within the confines of the CEP placement test were in fact correlated. This result was both desired and expected.

5. DISCUSSION AND CONCLUSIONS

In a previous paper, Moananu (2009) hypothesized that based on empirical evidence in reading and writing skills, the two abilities were correlated. However, using only research material to show this correlation lacked practical evidence. As a result, the current paper analyzed the reading and writing sections of the CEP placement test, used by Columbia University. Using SPSS, the reading and writing sections were analyzed to determine if they were correlated. Although numerous tests were performed, it was far more important to actuate whether the CEP test was, among other things, a reliable, stable, and valid measuring instrument. Thus for each section, descriptive statistics were provided. The internal consistency reliability and the standard error of measurement were also analyzed, and evidence of construct validity was provided. In addition to this, for the writing section, inter-rater reliability consistency between the two raters (external consistency) was analyzed. Finally, in order to assess the relationship between the writing and reading sections, statistical analyses were conducted on both sections.

The results of the analysis indicate that there was a high correlation between the reading and writing sections, which was suggested by the high correlation coefficient of 0.727 with the significance at 0.001 level (p<.01). Based on the results of the analysis, both the reading and writing sections functioned well. Descriptive statistics provided means, modes and medians that were relatively high and the distribution of scores for both sections was negatively skewed. This may suggest that the participants included a relatively high number of people who were preparing to enroll in Columbia University as well as people who had been living in the community for a while, generating a high frequency of higher scores.

In the reading section, there was a high degree of internal consistency for the thirty items. In the writing sections, there were also high correlations between the four variables: task fulfillment, content control, organizational control, and language control. The high reliability estimate for both the writing and reading sections indicate that the test is consistent and stable. The construct validity in the writing section did slightly better than that in the reading section perhaps due to the fact that the writing section measured fewer variables (2 tasks with 4 scores per task for a total of 8 measured variables) than the reading section (thirty multiple choice items). It should be noted that there were four multiple-choice items in the reading section that needed to be “modified” because the distracters attracted too much attention of the test takers. However, the items were not deleted, because the change in the “alpha if deleted” was not significant.

The purpose of the CEP test, used by Columbia University, is to measure the students’ ability in five areas of which I analyzed two; the reading and the writing section. The slightly lower scores for reading would indicate that teachers in the CEP program should be cautious in making broad generalizations about the reading capabilities of the test takers.

Finally, although further research in this area is needed to provide more valuable insights about the correlation between these two skills, the finding from this paper is significant. The results of this paper are confined to the results of the CEP placement test. However, I believe the results provide teachers and researchers in the field of ESL with information that can be useful in designing curriculums or devising teaching material on reading and writing.

7. References

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

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

Bachman, L. F., (2004). Statistical Analysis for Language Testing. Cambridge: Cambridge University Press.

Janopoulos, M. (1986). The relationship of pleasure reading and second language writing proficiency. *TESOL Quaterly* 20 (4), 763-768.

Moananu, C. B. (2009) Examining the dynamics of Reading and Writing Ability including Assessment and an Analysis of their Correlation to Determine how Best to Approach these Skills in the ESL context. Research studies of Toyama National College of Maritime Technology, No. 42, 173-182.

Purpura, J. (2004). *Assessing Grammar.* Cambridge: Cambridge University Press.