Investigating Second Language Reading Components: Reading for Different Types of Meaning

Ah-Young (Alicia) Kim[1]
Teachers College, Columbia University

ABSTRACT

The nature of second language (L2) reading ability is extremely complex and its components are yet to be agreed upon. The current study hypothesized that L2 reading ability consists of (1) reading for literal meaning (LIT), (2) reading for implied meaning with endophoric reference (IMP-EN), and (3) reading for implied meaning with exophoric reference (IMP-EX). In addition, different reading passages were assumed to affect test performance in varying degrees. Participants were 298 incoming students to an adult English as a second language (ESL) program. They took a reading placement test consisting of four reading passages and 30 items. Participants' reading scores were analyzed using the confirmatory factor analysis (CFA) approach in structural equation modeling (SEM). Findings suggest that L2 reading ability consists of the three constructs of LIT, IMP-EN, and IMP-EX. Moreover, the three reading constructs are on a continuum with LIT having the most direct relationship between the text and the response, while IMP-EX having the most indirect relationship. However, reading passages did not have a significant influence on test performance.

INTRODUCTION

Context of the Problem

In everyday and academic settings, reading ability is a critical tool for obtaining knowledge. Due to its importance, the nature of reading ability has been the subject of much research over the years in applied linguistics, education, and psychology (Urquhart & Weir, 1998). In the field of both first language (L1) and second language (L2) reading, scholars have asked questions such as "What is the nature of reading ability?" In fact, the componentiality of reading ability is still an on-going controversy, especially in L2 reading research. While some L1 researchers (e.g., Lunzer, Waite, & Dolan, 1979) suggest that reading ability is a unitary trait, the majority of L2 researchers (e.g., Freedle & Kostin, 1993; Grabe, 1991; Grabe & Stoller, 2002) believe that reading ability consists of multiple components. These components, or *skills*, are part of the cognitive ability used when interacting with written texts (Urquhart & Weir, 1998). If L2 reading ability is a divisible trait, the question remains as to how many components can be identified and

[1] Ah-Young Kim is a doctoral student in the Department of Applied Linguistics and TESOL at Teachers College, Columbia University in New York City. Her research interests include second reading assessment and cognitive diagnostic assessment. Correspondence should be sent to Ah-Young Kim, Teachers College, 525 W 120th Street, Box 66, New York, NY 10027. Email: ayk2101@columbia.edu

also what the relationship is among them. Examining the components of L2 reading ability can benefit not only teaching, but also assessment. Reading can be taught by focusing on appropriate exercises of identified skills (Williams, 2004). Likewise, reading tests can be designed based on those components of reading ability.

Previous theoretical (e.g., Grabe, 1991) and empirical studies (e.g., Lumley, 1993) in L2 reading have identified such major reading comprehension skills as understanding the main idea, finding the detail, making an inference, and comprehending vocabulary. However, some problems are inherent within this list of reading skills. First, there is considerable overlap among the skills. For example, certain main ideas are not explicitly stated in the text and therefore need to be inferred. Similarly, the meaning of some details is literally stated in the text while others are implied in the passage. This highlights the insufficiency of having a simple list of skills that are taught and tested. Rather, reading skills need to be categorized within a framework that distinguishes between *literal meaning* and *implied meaning* in the text (e.g., Alderson & Lukmani, 1989).

Second, the definition of an inference skill needs more clarification. There are two different types of inferencing. One is to draw implied meaning from within the text; that is, the meaning has *endophoric reference* (Halliday & Hasan, 1989). The other type of inferencing is obtained by drawing implied meaning not only from within but also from outside of the text, which is termed *exophoric reference*. To understand implied meaning with exophoric reference, readers need to use their prior knowledge to understand the pragmatic meaning encoded in the text. There are five different types of pragmatic meanings: contextual (e.g., meaning conveying the setting or topic), sociolinguistic (e.g., meaning indicating social identity), sociocultural (e.g., meaning associated with specific cultures), psychological (e.g., meaning of the tone), and rhetorical knowledge (e.g., meaning regarding the structure and purpose of the text) (Purpura, 2004). Often the two components of reading — *reading for implied meaning with endophoric reference* and *reading for implied meaning with exophoric reference* — are grouped into one concept and are measured by one type of inference item. However, as this paper will suggest, a distinction should be made between the two components of reading when designing test items.

Numerous L2 reading tests are designed specifically to evaluate whether advanced learners are competent to study in English speaking universities (e.g., TOEFL, IELTS) or work for companies (e.g., TOEIC). However, many of these tests have been constructed based on a list of skills that does not clearly distinguish among different types of inferencing skills. Inferencing is believed to be more challenging than skills that deal with the literal meaning of text (Kintsch, 1998). Therefore, incorporating various types of inference items can lead to tests that better differentiate among advanced readers.

Thus, there is a need for a new conceptual framework of L2 reading ability that can provide implications for L2 reading test design. The framework should be based on three different components of reading — (1) reading for literal meaning (LIT), (2) reading for implied meaning with endophoric reference (IMP-EN), and (3) reading for implied meaning with exophoric reference (IMP-EX). The feasibility and the usefulness of implementing the framework will be investigated by providing a validity argument of an L2 reading test developed on the basis of this framework.

Furthermore, it is necessary to understand how reading texts affect reading performance. According to Bachman and Palmer (1996), test method factors, which are methods used to elicit test performance, can influence reading scores. In reading, the text (i.e., the input) is a major method factor. Consequently, the characteristics of reading texts may influence test-takers'

reading performance. A case in point is the idea of having one task that contains a group of items, which constitutes a *testlet*. Specifically, Lee (2004) found that having different reading passages in one reading test led to the *testlet effect*. In this light, it is essential to uncover potential passage effects. The current study aims to fill the gaps in L2 reading literature by examining L2 reading components and passage effects.

Purpose of the Study and Research Questions

The purpose of this research is to examine the nature of L2 reading ability by investigating its components. L2 reading ability is hypothesized to consist of three components: LIT, IMP-EN, and IMP-EX. This framework was used to develop the reading placement test of the Community English Program (CEP) at Teachers College, Columbia University. The CEP offers English as a second language (ESL) classes to adults in the local community, and the reading placement test scores are used for grouping incoming students into appropriate levels. This paper examines the underlying trait structure of reading ability in the CEP reading test and provides a validity argument for the new L2 reading framework.

Confirmatory factor analysis (CFA) is used to investigate the underlying construct of L2 reading ability. CFA is a powerful procedure that can explore the nature of L2 reading ability based on substantive theoretical rationales. More specifically, CFA can be used to investigate the relationship among multiple underlying factors at the same time and compare various models set a priori. Moreover, it is possible to measure method effects (e.g., passage effect) on test performance using the multi-trait multi-method (MTMM) approach in CFA (Byrne, 2006). Since the text is a major method factor in reading, the characteristics of reading texts may influence test-takers' reading performance. Therefore, it is necessary to uncover the potential passage effects on reading performance.

The current study investigates the following research questions:

1. What is the nature of the L2 reading ability defined in terms of (1) reading for literal meaning (LIT), (2) reading for implied meaning with endophoric reference (IMP-EN), and (3) reading for implied meaning with exophoric reference (IMP-EX) in the context of the Community English Program (CEP) reading placement test?
2. Is there any passage effect in the CEP reading placement test?

In order to answer these questions, the literature on both first language (L1) and second language (L2) reading research will be reviewed. Studies regarding text variables that may affect test performance will also be presented. The CEP reading test will then be explained with a focus on its reading passages and items. Procedures for data collection, scoring and analyses will follow. Afterwards, results and conclusions will be discussed.

LITERATURE REVIEW

The Nature of L2 Reading Ability

To discuss the nature of reading ability, it is first necessary to define what reading is. In the narrowest sense, reading is decoding, i.e., the transformation of printed words into spoken words (Perfetti, 1985). More specifically, decoding requires one to sound out the phonemes of a word and blend them into a recognizable word. One may be able to decode without knowing the

definition of a word. Since decoding may not always involve understanding the meaning of words, it cannot be equated with reading (Urquhart & Weir, 1998). Rather, reading should be seen as a cognitive activity, where the reader interacts with the text to derive meaning. Thus, in the current paper, reading ability will be defined as reading comprehension, or "the ability to receive and interpret information encoded in language form via the medium of print" (Urquhart & Weir, 1998, p.22). Reading comprehension is a much broader term than decoding and emphasizes the act of obtaining meaning from the text.

It is first necessary to discuss L1 reading ability in order to discuss L2 reading ability, considering that L2 reading research stems from relevant L1 research (Grabe & Stoller, 2002). While some L1 reading researchers (e.g., Lunzer et al., 1979) have found that reading ability is a unitary trait, the majority (e.g., Grabe, 1991; Hoover & Gough, 1990; Lumley, 1993) believe that reading consists of multiple components. For example, proponents of the *Simple View of Reading* (e.g., Curtis, 1980; Hoover & Gough, 1990; Stanovich, Cunningham, & Feeman, 1984) suggest that reading is composed of two skills: decoding and linguistic comprehension**.** Decoding, here, refers to word recognition, and linguistic competence is "the ability to take lexical information and derive sentence and discourse interpretations" such as parsing, bridging, and discourse building (Hoover & Gough, 1990, p. 131). In an empirical study, Stanovich et al. (1984) examined the nature of reading using third- and fifth- grade data. They found that decoding respectively accounted for 19% and 38% of variance in scores for each grade and linguistic comprehension respectively accounted for 14% and 13% of variance in scores. However, the Simple View of Reading is somewhat simplistic and does not reflect the intricate nature of reading comprehension.

Indeed, numerous L1 reading studies (e.g., Drum, Calfee, & Cook, 1981; Davey, 1988; Davis, 1941; 1968; Spearitt, 1972) indicate that reading ability consists of more than just two components. Davis (1968) conducted a study to examine the components of reading ability and identified eight skills. Participants were 988 American high school students who took a reading test with 320 items, represented by eight skills: (1) recalling word meaning, (2) drawing word meaning from context, (3) finding explicit information, (4) integrating ideas of text, (5) drawing inferences from text, (6) recognizing the author's purpose and tone, (7) identifying the author's techniques, and (8) identifying the structure of text. Results from factor analyses indicated that the skills were independent from each other, suggesting that reading measured multiple traits. However, the number of skills remains controversial.

In L2 reading, Grabe (1991) classified reading into six components: (1) automatic recognition skills, (2) vocabulary and structural knowledge, (3) formal discourse structure knowledge, (4) content and world background knowledge, (5) synthesis and evaluation skills, and (6) metacognitive skills. Later, Grabe and Stoller (2002) listed the following key factors in L2 reading: vocabulary, discourse organization, main idea comprehension, motivation, and social context. However, the key point is not the number of skills that may theoretically exist, but how many can be identified and realized on tests (Alderson, 2000).

In this connection, a number of empirical studies (Alderson & Lukmani, 1989; Alderson, 1990; Freedle & Kostin, 1993; Lumley, 1993) investigated the components of reading ability. Alderson (1990) extended Alderson and Lukmani's (1989) study and examined if reading skills could be classified according to their difficulty levels. In the study, 18 experienced ESL teachers were asked to decide which of the 14 skills listed in Weir (1983, cited in Alderson, 1990) were measured by 15 items on the Test of English for Educational Purposes (TEEP). However, they

were not able reach an agreement. Moreover, they could not determine which skills were more difficult than others.

On the other hand, Lumley (1993) was able to clearly identify different L2 reading skills. Initially, 158 English as a second language (ESL) students took an English for academic purposes (EAP) test with 116 items. Using Rasch analysis, he generated difficulties for all items. On a post hoc analysis, five experienced teachers rated the difficulty of nine reading skills and matched them to 22 items from the test. The nine skills were (1) vocabulary, (2) identifying explicit information, (3) identifying implicit information, (4) explaining a fact, (5) selecting the main idea, (6) examining a causal or sequential relationship, (7) drawing a conclusion, (8) transcoding information to a diagram, and (9) understanding grammatical and semantic reference. There was not only agreement regarding the difficulty of skills among raters, who were experienced teachers, but also a significant correlation between item difficulty and raters' perception of item difficulty.

Through an in-depth examination of the above studies, several problems are identified. First, the reading tests were not developed based on a theoretical model of reading. Often, the tests measured a list of skills, which were not clearly distinguishable from each other. For example, Lumley (1993) found nine distinct reading skills, but there was much overlap among them. It is unclear how *identifying explicit information* or *identifying implicit information* differ from *explaining a fact*. Since facts are either explicitly or implicitly stated in the text, *explaining a fact* should belong to either the item type of *identifying explicit information* or *identifying implicit information*. In other words, certain types of details are literally stated in the text whereas others are implied in the text. In this case, it is questionable whether an item can be categorized as a detail item or an inference item using the traditional list of skills. Similarly, certain main ideas are explicitly stated in the text while others are implied (Lumley, 1993). Moreover, not all vocabulary items are the same: some test the ability to recall word meaning while others require the ability to infer word meaning from the text (Davis, 1968). Therefore, there is a need to classify items that require *reading for literal meaning* from those that elicit *reading for implied meaning*.

Second, the definition of an inference item needs elaboration. Inferencing is the act of using information from the text to produce new information (Kintsch, 1998). Chikalanga (1992) defines inferencing as a "cognitive process a reader goes through to obtain the implicit meaning of a written text on the basis of two sources of information: the "propositional content of the text" (i.e., the information explicitly stated) and "prior knowledge" of the reader" (p. 697). Based on this definition, he classifies inferencing into *propositional inference* and *pragmatic inference*. Propositional inference is "logically derived from the semantic content of the text" (p. 704) whereas pragmatic inference necessitates reader's prior knowledge. In other words, a propositional inference item requires one to read for implied meaning from the text and therefore has *endophoric reference* (Halliday & Hasan, 1989). On the other hand, a pragmatic inference item requires readers to use background knowledge or features of the external context and thus has *exophoric reference* (Halliday & Hasan, 1989). For example, it may ask examinees to *recognize the author's purpose or tone.*

In sum, it is necessary to establish a framework of L2 reading that clearly distinguishes between *reading for literal meaning* versus *reading for implied meaning*. Furthermore, reading for implied meaning needs to be subcategorized into *reading for implied meaning with endophoric reference* versus *exophoric reference.* Several researchers have provided a classification that differentiates between *reading for literal meaning* and two types of reading for

implied meaning. In L1 reading, Gray (1960) popularized the terms *reading the lines*, *reading between the lines*, and *reading beyond the lines*. In addition, Herber (1978) categorized L1 reading into *literal*, *interpretive* and *applied* levels. Literal level reading refers to identifying the information conveyed in the text through paraphrasing or translating of information. Readers will experience difficulty in this stage of reading if they lack vocabulary or grammatical knowledge. At the interpretive level, readers derive meaning from the text by interpreting the intrinsic relationships among statements. Readers must understand implicitly stated ideas in the text by putting together pieces of information. At the applied level, readers combine the literal and the interpretive understanding of the text with their prior knowledge and experience in order to derive a deeper understanding of the text, thereby comprehending ideas which extend beyond the immediate context of the text (Herber, 1978).

In L2 reading, Pearson and Johnson (1978) proposed three different types of responses based on the relationship between questions and responses in reading tests: *textually explicit*, *textually implicit*, and *scriptally implicit responses.* A textually explicit response refers to cases where the information for the question is provided in the text. A textually implicit response is when the reader needs to make a logical inference to answer the question. A scriptally implicit response refers to when the reader needs to use his/her prior knowledge to answer the question. However, Pearson and Johnson (1978) do not provide a clear definition of *script*. In addition, both Herber (1978) and Pearson and Johnson (1978) do not discuss in detail the different types of prior knowledge. This calls for a new framework for L2 reading ability, which better defines the three components of reading and the three different levels of meaning.

A New Framework for L2 Reading Ability

Building upon prior research, the framework of L2 reading ability described here attempts to clearly define the three components of reading and provide a list of items that might measure each component. The three components of L2 reading ability are (1) *reading for literal meaning* (LIT), (2) *reading for implied meaning with endophoric reference* (IMP-EN) and (3) *reading for implied meaning with exophoric reference* (IMP-EX) (refer to Table 1). Reading for literal meaning refers to understanding an explicitly stated concept from the text, whereas reading for implied meaning with endophoric reference requires comprehending implicitly conveyed concept from the text. In both reading for literal meaning and reading for implied meaning with endophoric reference, all the necessary information has endophoric reference — information is derived from the text (Halliday & Hasan, 1989). However, reading for implied meaning with exophoric reference requires additional knowledge outside the text. It refers to deriving an implicitly stated concept from outside the text (Halliday & Hasan, 1989). In other words, reading for implied meaning with exophoric reference involves comprehending pragmatic meaning — implied meanings encoded in text such as contextual, sociolinguistic, sociocultural, psychological, and rhetorical meanings. Contextual meaning is derived by inferring the setting, including time, place, and theme of the event. Sociolinguistic meaning can be understood through inferring the social identity of characters in the text, identified through factors such as politeness or formality of the characters' relationships. Sociocultural meaning, which somewhat overlaps with sociolinguistic meaning, can be inferred by using one's prior knowledge of cultural norms. Psychological meaning can be obtained by understanding the emotional or attitudinal stance of the text. Lastly, rhetorical meaning can be inferred based on the organizational structure of the text, such as the genre or the purpose of the passage in concern (Purpura, 2004).

Table 1 includes examples of items that could be used to elicit different components of L2 reading ability. For instance, a reading for literal meaning item measures readers' ability to find (1) the reference of a word (cohesion), (2) explicit detail from the text, (3) essential information of a lengthy sentence, and (4) the explicitly stated main idea of a paragraph or the whole text. A reading for implied meaning with endophoric reference item assesses the ability to infer (1) the meaning of a word, (2) detailed information, (3) the title/heading of a paragraph or the entire text, and (4) the main idea. It also tests test-takers' knowledge of coherence by asking them to insert a sentence into the passage. Reading for implied meaning with exophoric reference item tests the ability to infer (1) contextual meaning, (2) sociolinguistic meaning, (3) sociocultural meaning, (4) psychological meaning, or (5) rhetorical meaning in the text.

TABLE 1
Three Components of L2 Reading and Their Item Types

Endophoric Reference		**Exophoric Reference**
Literal meaning (Understanding an explicitly stated concept from the text)	**Implied meaning** (Understanding an implicitly stated concept from the text)	**Implied meaning** (Understanding an implicitly stated concept from the context)
Reference (=cohesion) (e.g., What does "X" refer to?)	Inferring vocabulary meaning from the text (e.g., What does "X" mean in this text?)	Deriving contextual meaning (e.g., Where is the event taking place?)
Detail (e.g., What is true according to the passage?)	Inferring detail (e.g., What does the sentence imply?)	Deriving sociolinguistic meaning from the context (e.g., What is the relationship between the two people?)
Sentence simplification (e.g., What is the essential information of the following sentence?)	Inserting sentence into the passage (coherence) (e.g., Where would the following sentence best fit?)	Deriving sociocultural meaning from the context (e.g., What can be inferred about the American culture?)
Main idea (e.g., What is the main idea of paragraph X?)	Inferring title/heading (e.g., What is the best heading for the paragraph?)	Deriving psychological meaning from the context (e.g., What is the author's attitude?)
	Inferring main idea (e.g., What is the main idea of the paragraph?)	Deriving rhetorical meaning from the context (e.g., Why does the author mention "X"?)

Furthermore, different components of reading ability will be required depending on whether the expected response to an item has a *direct* or an *indirect* relationship to the text. A direct relationship between the input and the response refers to cases where the response information (answer to an item) is supplied in the input (text). On the other hand, an indirect relationship refers to when the response information is not supplied in the input (Bachman & Palmer, 1996). Accordingly, the most direct relationship between the text and the response will require reading for literal meaning and a somewhat direct relationship will necessitate reading

for implied meaning with endophoric reference. The most indirect relationship will generate reading for implied meaning with exophoric reference.

Method Factors that Affect L2 Reading Performance

Bachman (1990) provides a theoretical framework of factors that affect language performance, which is easily applicable to L2 reading. He categorizes four sets of factors that affect language performance: (1) communicative language ability, (2) test method facets, (3) personal attributes, and (4) random factors. He states that while some factors are within the test-developer's control, others are not. Factors such as test-taker's personal attributes (e.g., gender, age and cultural background) and random factors (e.g., equipment failure) are uncontrollable. However, test-taker's language ability and test method facets are factors that can and must be controlled in designing language tests. Of the two controllable factors, test method facets are directly under the control of test developers (Bachman & Palmer, 1996).

Test method factors, also known as test-task characteristics, refer to methods used to elicit test performance. Building on Bachman (1990), Bachman and Palmer (1996) provide a framework of five major task characteristics: (1) setting (physical circumstances), (2) test rubric (structure for tasks and procedures to accomplish them), (3) input (materials contained in a task), (4) expected response (language use elicited from a task), and (5) relationship between input and response. Among the task characteristics, input is the material which readers must process and respond to (Alderson, 2000). On a reading test, the major source of input used for tapping into the learner's reading ability is the text. Since reading is primarily an interaction between the reader and the text, it is vital to understand the characteristics of text and how it affects reading.

There are numerous text variables that affect L2 reading performance. Even the usage of different reading passages creates a unique task effect called the *testlet effect*. A testlet is defined as a "group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (Wainer & Kiely, 1987, p.190). Generally, items under a passage are assumed to maintain *local independence* from each other. Local independence assumes that an examinee's response to one item does not influence the response to other items on the test (Hambleton & Swaminathan, 1985). However, when a cluster of items are more related to each other due to some common features shared by items, such as a shared stimulus or text, local independence is violated (Stienberg & Thissen, 1996).

A few studies have investigated the testlet effect in reading assessment. Using item response theory (IRT), Lee (2004) found some testlet effect in an English as a foreign language (EFL) reading comprehension test. Participants were 1,857 Korean high school students, who completed a reading comprehension test with 40 items. The data were analyzed using MULTILOG and IRTNEW. The results indicated that positive local item dependence (LID) was found among items sharing the same text, but LID was not seen among the same item types.

Although the testlet effect describes the degree of item dependence among items within a testlet, it does not provide much information regarding how much each item interacts with the text. It is difficult to assume that all items in a testlet will be influenced by the input in the same degree. It is more plausible that items will be affected differently. For example, a reading for implied meaning item may depend more on the text compared to a reading for literal meaning item because reading for implied meaning items generally require more processing of the text.

Thus, there is a need to examine if certain item types are easily affected by the text compared to others.

This section provided a summary of the literature leading up to the new framework of L2 reading ability and method factors that may affect reading performance. The new framework of L2 reading ability suggests that L2 reading consists of three components of LIT, IMP-EN, and IMP-EX. The following section will discuss the methods used for examining the proposed nature of L2 reading ability.

METHOD

The purpose of the current study is to investigate the nature of L2 reading ability by examining the three components of *reading for literal meaning* (LIT), *reading for implied meaning with endophoric reference* (IMP-EN), and *reading for implied meaning with exophoric reference* (IMP-EX). An *ex post factor* research design was used to address the research questions above.

Context of the Study

The Community English Program (CEP) at Teachers College offers English as a Second Language (ESL) classes to adults with diverse cultural and linguistic backgrounds. The CEP mainly targets the immigrant population residing in the New York City area, but a number of learners in the program are international students or their spouses, and international corporate executives. Hence, students demonstrate a wide range of reading proficiency. The role of the CEP placement test is to place the new incoming students into one of 13 different levels. The CEP placement exam consists of five parts: grammar, listening, speaking, reading, and writing.

Participants

The participants were 298 new incoming students to the CEP in Summer 2007. Of these, 215 (71.2%) students were females and 83 (27.8%) were males. As seen in Table 2, the two largest first languages spoken were Japanese (32.8%), and Spanish (22.5%), followed by Korean (17.1%), Portuguese (5.3%), Chinese (5%), and Polish (3.6%). Other languages (14%) were Bengali, Czech, Farsi, German, Romanian, Russian, and Ukrainian. Table 3 shows that the participants' length of stay differed considerably as well. At the time of the study, 179 (60%) students had been in the U.S. for less than a year; 86 (29%) were in the U.S. for one to five years; 15 (5%) students were in the U.S. for five to ten years; and 18 (6%) were in the U.S. for over ten years.

TABLE 2
Native Language of Participants

First Language	Frequency	Percentage
Japanese	98	32.8%
Spanish	67	22.5%
Korean	51	17.1%
Portuguese	16	5.3%
Chinese	15	5.0%

TABLE 3
Length of Stay in the U.S.

Polish	11	3.6%
Others	40	13.7%
Total	298	100%

Length of Stay	Frequency	Percentage
Less than 1 year	179	60.1%
1 to 5 years	86	28.9%
5 to 10 years	15	5.0%
Over 10 years	18	6.0%
Total	298	100%

Instrument

There were four reading passages in the CEP reading test. As seen in Table 4, the passages differed in a number of ways: topic (competition and cooperation, complaints, mysteries of smell, demise of Neanderthals), text structure (compare/contrast, problem/solution, description, causation), and length (192 to 528 words).

TABLE 4
Analysis of the Reading Tasks

Task	Topic	Text Structure (content structure)	Length	Number of items
Task 1	Competition and Cooperation	Compare/contrast	313 words	5
Task 2	Complaint Letters	Problem/solution	192 words	5
Task 3	Mysteries of Smell	Description	412 words	10
Task 4	Demise of Neanderthals	Causation; Compare/contrast	528 words	10

L2 reading ability was operationalized in terms of (1) reading for literal meaning with literal meaning (LIT), (2) reading for implied meaning with endophoric reference (IMP-EN), and (3) reading for implied meaning with exophoric reference (IMP-EX). The three components were measured using 30 multiple-choice questions, as seen in Table 5.

TABLE 5
Original Taxonomy of the CEP Reading Test

Construct	Number of Items	Item Numbers on the Test
LIT	11	54, 55, 56, 60, 64, 65, 69 71, 74, 78, 79
IMP-EN	16	53, 57, 58, 59, 62, 63, 66, 68, 70, 72, 73, 75, 76, 77, 81, 82
IMP-EX	3	61, 67, 80

The 30 reading items were coded by three coders, who were graduate students in Applied Linguistics. They received a 30-minute coding training session and coded the items into LIT, IMP-EN, and IMP-EX. The inter-coder reliability was calculated using Fleiss' Kappa. The Fleiss' Kappa for the reading items was high at .771, indicating a relatively high degree of agreement in categorizing items into the three components.

Data Collection & Scoring

The reading test was administered as part of the CEP placement test in Summer 2007. The participants took the reading test after completing the listening and the grammar sections of the placement test. The writing and the speaking sections followed the reading test. Examinees were given 45 minutes to complete the reading section.

Each item was scored dichotomously: one point for a correct response and zero point for an incorrect response. The total possible score for the reading section was 30 points, which comprised 15% of the total placement test score.

Data Analysis

The CEP reading data were analyzed in the following order: (1) descriptive statistics, (2) reliability analysis, (3) exploratory factor analysis, and (4) confirmatory factor analysis.

Descriptive Statistics and Reliability Analysis

Descriptive statistics (e.g., mean, standard deviation, skewness, and kurtosis) were computed using SPSS version 12. Results were used to inspect the univariate and multivariate normality of the data.

The internal consistency reliability was examined using Chronbach's alpha. The reliability estimates provided an estimate of the extent to which the items on the test related to the other items on the test.

Exploratory Factor Analysis

Exploratory factor analysis (EFA) was conducted to examine the number of factors by identifying how observed variables clustered. EFA is a statistical procedure that examines how observed variables can be represented by a smaller number of underlying variables or factors (Kim & Mueller, 1978). As a preliminary step for confirmatory factor analysis, EFA, was performed to examine the clustering of observed variables using Principal-axes factoring. In the beginning, appropriateness for factor analysis was determined by examining three indices: (1) Bartlett's test of sphericity, (2) Kaiswer-Meyer-Olkin (KMO) measure, and (3) the determinant of the correlation matrix. Subsequently, the eigenvalues and the scree plot obtained from the initial extractions were examined as an initial indication of the number of factors represented by the data. Principal axes were extracted with the number of factors equal to, one above, and one below the number of factors indicated by the initial extraction. These extractions were rotated using both the Promax rotation (for correlated factors) and the Varimax rotation (for uncorrelated factors) procedures. The best EFA model was chosen based on the simplicity and the meaningfulness of the structure.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA), a type of SEM, was used to investigate the underlying construct of L2 reading ability. CFA is a set of statistical procedures that tests hypothesized models set a priori. It was hypothesized that reading consists of the three

components of LIT, IMP-EN, and IMP-EX. Based on this substantive evidence and the EFA results, several trait-only models were first analyzed using EQS version 6.1. Kline (2004) noted that factorial structures identified by EFA may not have the same fit in CFA. That is, CFA may not always confirm the findings of EFA. Therefore, models created based on substantive theory of reading ability were initially analyzed. Afterwards, alternative models, including the one suggested by EFA, were compared against each other.

First, multivariate normality of the current data was checked using both Mardia's coefficient and normality index. The robust method was used if the data did not meet the normality assumption. The quality of the model was determined using the chi-square (χ^2) values, the ratio of chi-square over degrees of freedom (χ^2/df), and fit indices — comparative fit index (CFI) and the root mean square error of approximation (RMSEA) index. Although chi-square values are commonly used for model comparison, they are sensitive to sample size and have led to problems in data interpretation (Byrne, 2006). Thus, the CFI and RMSEA values were used to determine the goodness-of-fit of the models. The CFI measures the proportionate improvement in fit by comparing the hypothesized model and an independent model. Its values range from zero to 1 (Byrne, 2006). A CFI index close to .95 is advised to be the cutoff point for determining a good model fit (Hu & Bentler, 1999). The RMSEA is a type of "absolute misfit index," which decreases as the goodness-of-fit improves (Browne, Hudson, Norris, & Bonk, 2002). It is an error of approximation in the population and estimates how "the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available" (Browne & Cudeck, 1993, pp.137-138). Because the RMSEA is measured per degree of freedom, it is sensitive to the number of estimated parameters in the model (Byrne, 2006). A value of less than .05 represents good fit and an index between .05 and .08 is considered to be reasonable (Browne & Cudeck, 1993).

After assessing the model as a whole, individual parameter estimates were inspected for (1) the appropriateness of the estimates and (2) statistical significance. Examples of unreasonable estimates are correlations above 1.0, negative variances, and covariance or correlation matrices that are not positive definites (Byrne, 2006). In addition, the parameter estimates need to be statistically significant with their Z-statistics above the absolute value of 1.96 (alpha level of .05). Statistically insignificant estimates may indicate a small sample size. With the exception of error variances, statistically insignificant estimates are meaningless to the model and should be dropped (Byrne, 2006).

The best-fitting model was chosen based on fit statistics: χ^2/df, CFI, and RMSEA index. By extending the best-fitting trait model, a few multi-trait multi-method (MTMM) models were further analyzed to investigate method effects in the CEP reading test. For example, if the parameter estimates of the method factors are higher than those of the trait factors, the tasks are affecting test-takers' performance more than the traits. However, if the parameter estimates of the method factors are statistically insignificant, they are not influencing the test scores.

The methods section explained how the CEP reading placement test scores from 300 participants was collected and analyzed, using EFA and CFA. The following section reports the results from data analysis.

RESULTS

Descriptive Statistics and Reliability Analysis

The total possible score obtainable on the CEP reading placement test was 30. As seen in Table 6, the mean score was 19.3154 (64.38%) and the standard deviation was quite large at 5.44713. The data had a skewness of -.363 and a kurtosis of -.537. These results indicate that on average more students did well than poorly on the test. The minimum score obtained was 4 and the maximum score was 30 points, producing a range of 26 points, which is typical for a placement test.

TABLE 6
Descriptive Statistics of the Reading Test Scores (N=298)

Date	Mean	St. Dev.	Skewness	Kurtosis	Min	Max
Total (K=30)	19.3154	5.44713	-.363	-.537	4.00	30.00
Item 54 (LIT)	.7919	.40660	-1.446	.091	.00	1.00
Item 55 (LIT)	.8087	.39397	-1.578	.493	.00	1.00
Item 56 (LIT)	.8624	.34504	-2.115	2.489	.00	1.00
Item 60 (LIT)	.8624	.34504	-2.115	2.489	.00	1.00
Item 64 (LIT)	.6544	.47638	-.652	-1.585	.00	1.00
Item 65 (LIT))	.7617	.42673	-1.235	-.478	.00	1.00
Item 69 (LIT)	.4094	.49255	.370	-1.875	.00	1.00
Item 71 (LIT))	.4899	.50074	.040	-2.012	.00	1.00
Item 74 (LIT)	.7651	.42465	-1.257	-.423	.00	1.00
Item 78 (LIT)	.4094	.49255	.370	-1.875	.00	1.00
Item 79 (LIT))	.6879	.46412	-.815	-1.344	.00	1.00
Item 53 (IMP-EN)	.7685	.42253	-1.279	-.366	.00	1.00
Item 57 (IMP-EN)	.8557	.35198	-2.035	2.155	.00	1.00
Item 58 (IMP-EN)	.8893	.31434	-2.493	4.246	.00	1.00
Item 59 (IMP-EN)	.9329	.25064	-3.478	10.162	.00	1.00
Item 62 (IMP-EN)	.8322	.37430	-1.787	1.202	.00	1.00
Item 63 (IMP-EN)	.5403	.49921	-.162	-1.987	.00	1.00
Item 66 (IMP-EN)	.2852	.45229	.956	-1.093	.00	1.00
Item 68 (IMP-EN)	.5302	.49993	-.122	-1.999	.00	1.00
Item 70 (IMP-EN)	.4597	.49921	.162	-1.987	.00	1.00
Item 72 (IMP-EN)	.7651	.42465	-1.257	-.423	.00	1.00
Item 73 (IMP-EN)	.4832	.50056	.067	-2.009	.00	1.00
Item 75 (IMP-EN)	.3725	.48428	.530	-1.731	.00	1.00
Item 76 (IMP-EN)	.5537	.49795	-.217	-1.966	.00	1.00
Item 77 (IMP-EN)	.4329	.49631	.272	-1.939	.00	1.00
Item 81 (IMP-EN)	.4799	.50043	.081	-2.007	.00	1.00
Item 82 (IMP-EN)	.6477	.47851	-.621	-1.625	.00	1.00
Item 61 (IMP-EX)	.8691	.33783	-2.200	2.859	.00	1.00
Item 67 (IMP-EX)	.7047	.45695	-.902	-1.194	.00	1.00
Item 80 (IMP-EX)	.4094	.49255	.370	-1.875	.00	1.00

Six items (#56-61) were found to have skewness and kurtosis values larger than ±2. This was attributed to high mean scores above .85. In particular, item 59 had a very high mean score (.9329), skewness (-3.478), and kurtosis (10.162) value, indicating that the item was too easy and would not help in discriminating the proficiency among the test-takers. Except for item 56, the highly skewed items were from Task 2, suggesting that the content of the task may have been too easy for the group. However, it is acceptable to have a set of easy items since the CEP reading test is a placement test.

Cronbach's alpha for the reading section was very high at .830. In other words, the test items were highly related to each other and also somewhat homogeneous, making the test internally consistent.

Exploratory Factor Analysis

As a preliminary step to confirmatory factor analysis (CFA), exploratory factor analysis (EFA) was performed to examine how the observed variables grouped among themselves under different constructs. The three indices of (1) Bartlett's test of sphericity, (2) Kaiser-Meyer-Olkin (KMO) measure, and (3) the determinant of the correlation matrix all indicated that the data were appropriate for factor analysis: Bartlett's test of sphericity was significant; KMO was above .75 at .821; and the determinant was above zero at .007.

Principal-axis factoring (PAF) was performed on the 30 reading test items. The initial factor extraction yielded ten factors with eigenvalues greater than 1.0, which accounted for 55% of the total variance. However, the scree plot suggested a 2-factor solution, which was more economical and closer to the hypothesized 3-factor model. It is customary to extract the number of factors equal to, one above, and one below the number of factors indicated by the scree plot. Therefore, based on the scree plot results and substantive evidence, 1-factor, 2-factor, and 3-factor extractions were obtained. Moreover, the initial factors obtained by the 1-, 2- and 3-factor analyses were rotated using both Promax rotation (for correlated factors) and Varimax rotation (for uncorrelated factors).

The results indicated that the 1-factor analysis maximized parsimony and interpretability by producing a clear loading on the unitary reading comprehension factor. In the 2- and 3-factor extractions, a number of cross loadings were identified. Thus, the 1-factor extraction was determined to be the best-fitting factor solution for the CEP reading test. The highlighted section in Table 7 indicates significant eigenvalues over .3 that loaded on the 1-factor model. Except for four items — item 71, 70, 56, and 66 — all items significantly loaded on the reading comprehension factor. A closer analysis of the four items indicated that most of the items were either vocabulary or reference items: item 71 (vocabulary), item 70 (vocabulary), item 56 (reference). Their low loading indices suggest two possible explanations. First, vocabulary or reference items may not be the best way to measure reading comprehension ability. Secondly, vocabulary or reference items may measure another dimension of reading.

The EFA results suggest that reading in this test is better represented by a unitary factor rather than the hypothesized 3-factor model. This implies that reading cannot be divided into multiple components. However, it is possible for the factorial structures identified by EFA and CFA to differ (Kline, 2004). Since a CFA model cannot always be specified by the EFA results, the CFA model in the current paper was constructed based on the new framework of L2 reading ability.

TABLE 7
EFA Results for the Reading Test: 1-factor Model

Item Number	L2 Reading Comprehension
Item 53 (IMP-EN)	.542
Item 82 (IMP-EN)	.528
Item 78 (LIT)	.523
Item 61 (IMP-EX)	.492
Item 62 (IMP-EN)	.491
Item 54 (LIT)	.439
Item 68 (IMP-EN)	.434
Item 58 (IMP-EN)	.433
Item 60 (LIT)	.430
Item 59 (IMP-EN)	.401
Item 73 (IMP-EN)	.395
Item 76 (IMP-EN)	.387
Item 79 (LIT)	.387
Item 72 (IMP-EN)	.386
Item 64 (LIT)	.383
Item 80 (IMP-EX)	.378
Item 74 (LIT)	.375
Item 81 (IMP-EN)	.350
Item 75 (IMP-EN)	.344
Item 67 (IMP-EX)	.337
Item 57 (IMP-EN)	.337
Item 65 (LIT)	.335
Item 77 (IMP-EN)	.330
Item 69 (LIT)	.315
Item 63 (IMP-EN)	.313
Item 55 (LIT)	.308
Item 71 (LIT)	.294
Item 70 (IMP-EN)	.268
Item 56 (LIT)	.258
Item 66 (IMP-EN)	.211

Confirmatory Factor Analysis (CFA)

The main purpose of this research was to investigate the nature of L2 reading ability by examining the underlying traits of the CEP reading test constructed based on the new L2 reading framework. First, a number of substantively plausible reading ability models were tested using trait-only CFA models, in which the reading components were represented as trait factors. After selecting the best-fitting trait-only CFA model, an MTMM model was examined to explore if texts affected test-takers' scores in the CEP reading test. In the MTMM model, the reading components were treated as trait factors, whereas the four passages were defined as method factors.

The reading data were first analyzed based on the initially hypothesized 3-factor model, followed by two alternative models: a 2-trait and a 1-trait model. After finding the best fitting trait model, an MTMM model was further examined to inspect possible method effects on the CEP reading test.

Trait Only Models

Model 1: 3-trait model.

To address the first research question, regarding the nature of L2 reading ability, the CEP reading data were first analyzed using a 3-trait model of reading ability operationalized in terms of LIT, IMP-EN, and IMP-EX (refer to Figure 1). The three factors (indicated by the three circles) were believed to correlate with each other (illustrated by the two headed arrows). Thirty observed variables (shown as rectangles) were hypothesized to load on one of the three factors. The errors associated with each observed variable were postulated to be uncorrelated.

First, univariate and multivariate sample statistics were examined. The multivariate values were high—Mardia's coefficient was at 61.5620 and the normalized estimate was higher than 5 at 12.1266, indicating that the data are non-normally distributed. Thus, the robust maximum likelihood (ML) method was used to further analyze the data. All other statistical assumptions of the estimation procedure, such as number of iterations for conversion, did not indicate significant violations in the data.

To determine how well the model fit the sample data, standardized residuals and fit statistics were examined. The data produced a low average off-diagonal absolute standardized residual value of .0435, indicating a good fit. Standardized residual values bigger than 2.58 are considered large (Joreskog & Sorbom, 1988). In addition, the Satorra-Bentler scaled chi-square value was 524.5379 with 402 degrees of freedom ($p<.00004$), making the chi-square over degrees of freedom (χ^2/df) ratio 1.3, or lower than the suggested ≤ 2.0. The root mean-square error of approximation (RMSEA) value was .032, lower than the recommended value of $\leq .5$. However, the Satorra-Bentler comparative fit index (CFI) was .868, smaller than the recommendation of $\geq .95$ (Hu & Bentler, 1999). These results are presented in Table 8.

TABLE 8
Results for the 3-factor Model of Reading Ability

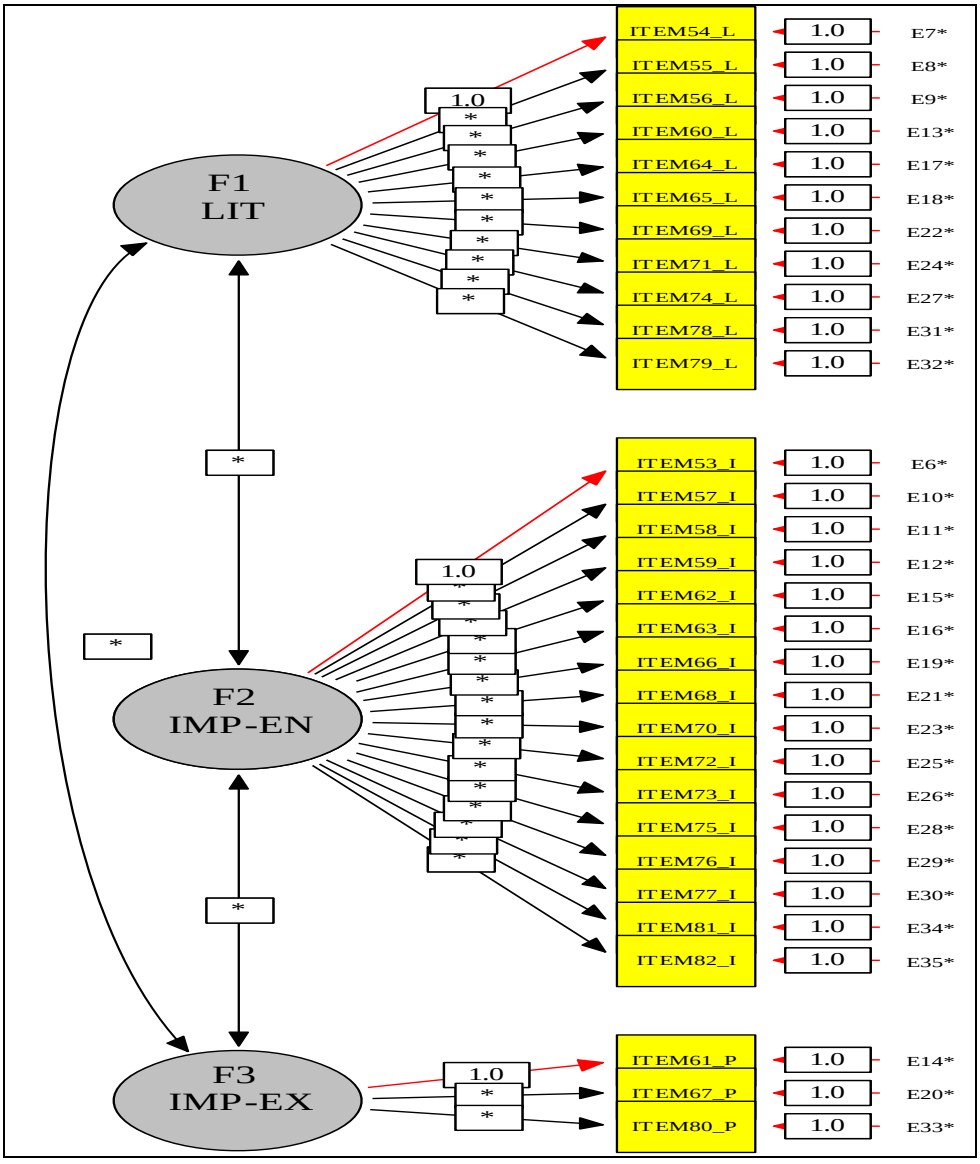
Goodness of fit summary	
Satorra-Bentler Scaled Chi-square (on 402 degrees of freedom; p= .00004)	524.5379
Comparative fit index (CFI)	.868
RMSEA	.032

Because the CFI was lower than .95, each item was reanalyzed to check if the items were properly categorized. Three items—items 59, 69, and 71—were determined to be somewhat problematic. Item 69, which was designed to be a LIT item, was found to load higher on the IMP-EN factor. Therefore, the item was analyzed in detail. The stem read

69. In line 24, what does "one" refer to?

Line 24: "Proust referred to taste and smell as one entity and, indeed, **one** would not be much without the other."

FIGURE 1
Initially Hypothesized 3-trait Model of Reading Ability



Note.
Variable names:
1. L: reading for literal meaning
2. I: reading for implied-endophoric meaning
3. P: reading for implied-exophoric meaning

The item was designed as a *reference* item, asking the test-taker what "one" refers to in line 24. The item may have loaded higher on the IMP-EN factor rather than the LIT factor due to the multiple meanings of "one." "One" can have the general meaning of the number, but it can also refer to "taste" within the sentence. Therefore, test-takers may need to infer the meaning of "one" from the context. Moreover, the item requires some grammatical knowledge of what "one" and "the other" generally refer to: "one" refers to the first noun and "the other" refers to the second noun that was previously mentioned. Overall, it was unclear what item 69 was measuring. Thus, it was excluded from further analysis.

In addition, item 71, a vocabulary item, was determined to be a poorly designed item. Each vocabulary item was designed to measure the ability to infer the correct meaning in the text. Therefore, only words with multiple meanings were tested. However, item 71 tested the meaning of "waft," which had only one possible definition. Thus, it was necessary to exclude the item from further analysis. Item 59 was also omitted from the analysis because it had an unusually high skewness and kurtosis value, as was mentioned in the descriptive analysis, and was suspected of having outliers.

Moreover, a Lagrange Multiplier (LM) test was conducted to inspect if one or more restrictions better represented the data. The LM test revealed a number of potential restrictions that could be added to the model. As seen in Figure 2, five correlations were added between the following five error sets: errors 11 (item 58) and 7 (item 54); errors 17 (item 64) and 11 (item 58); errors 35 (item 82) and 34 (item 81); errors 23 (item 70) and 15 (item 62); and errors 15 (item 62) and 13 (item 60). The first three pairs of errors were part of item pairs that measured the main idea or the title of the text or a paragraph. For example, the stems for both item 54 and 58 measure the main concept of the text:

> 54. The main idea in the second paragraph (lines 6-13) is ____.
> 58. What is the best subject for the first e-mail?

Conversely, items 62 and 70 were detail-oriented:

> 62. What is implied in the second e-mail?
> 70. In line 26, what does "meager" mean?

To elaborate, in order to answer items 62 and 70 correctly, it is necessary to find detailed information from the text. On another note, item 60 and 62 had very similar item stems. As seen in the following, both items had the wording of "What is ___ in the ___ e-mail?":

> 60. What is included in the first e-mail?
> 62. What is implied in the second e-mail?

Indeed, dropping the three items and adding five error correlations into the 3-trait model improved the χ^2/df ratio to 1.128, CFI value to .951, and the RMSEA to .021 (refer to Figure 2).

The correlations among the three- trait factors, measured by 27 observed variables, ranged from .87 to .98. The correlation between the LIT and the IMP-EN factors was .94, indicating that the two factors were highly correlated and were quite similar in nature. Likewise, the correlation between the IMP-EN and the IMP-EX factors was high at .98, suggesting that the two factors may be measuring very similar traits. The correlation between the LIT and the IMP-EX factors was .87, indicating that the two factors were highly correlated, yet separate. These

findings suggest that the three components of reading are on a continuum from LIT to IMP-EX, with LIT having the most direct relationship between the text and the response, and the IMP-EX having the most indirect relationship.

Once the model was assessed as a whole, individual parameters were evaluated. Most individual parameter estimates were generally within a reasonable range (beyond .3) and all estimates were statistically significant at the 0.05 level. The covariance between the three factors and the variances of error terms were also statistically significant. This implied that the underlying factors were well measured by the observed variables and that these variables were measuring reading ability. Three items — item 56, 66, and 70 — had somewhat low factor loadings ranging between .2 and .3. Interestingly, these were the items that had low item loadings in EFA, confirming that they were either (1) not measuring reading ability as well as the other items or (2) functioning somewhat differently from the rest of the items. The standardized solution in Table 9 presents the model in mathematical equations. The table shows that most factor loadings ranged from .30 and .56 and were statistically significant at the .05 level. For example, the equation for item 53 (V6), an IMP-EN item, shows that the item had a moderately high loading of .559 on the IMP-EN factor (F2). Also, an error term (E6) of .829 accounts for any measurement error or specificity. Similarly, the equation for item 54 (V7), a LIT item, indicates that the item had a moderately high loading of .436 on LIT (F1).

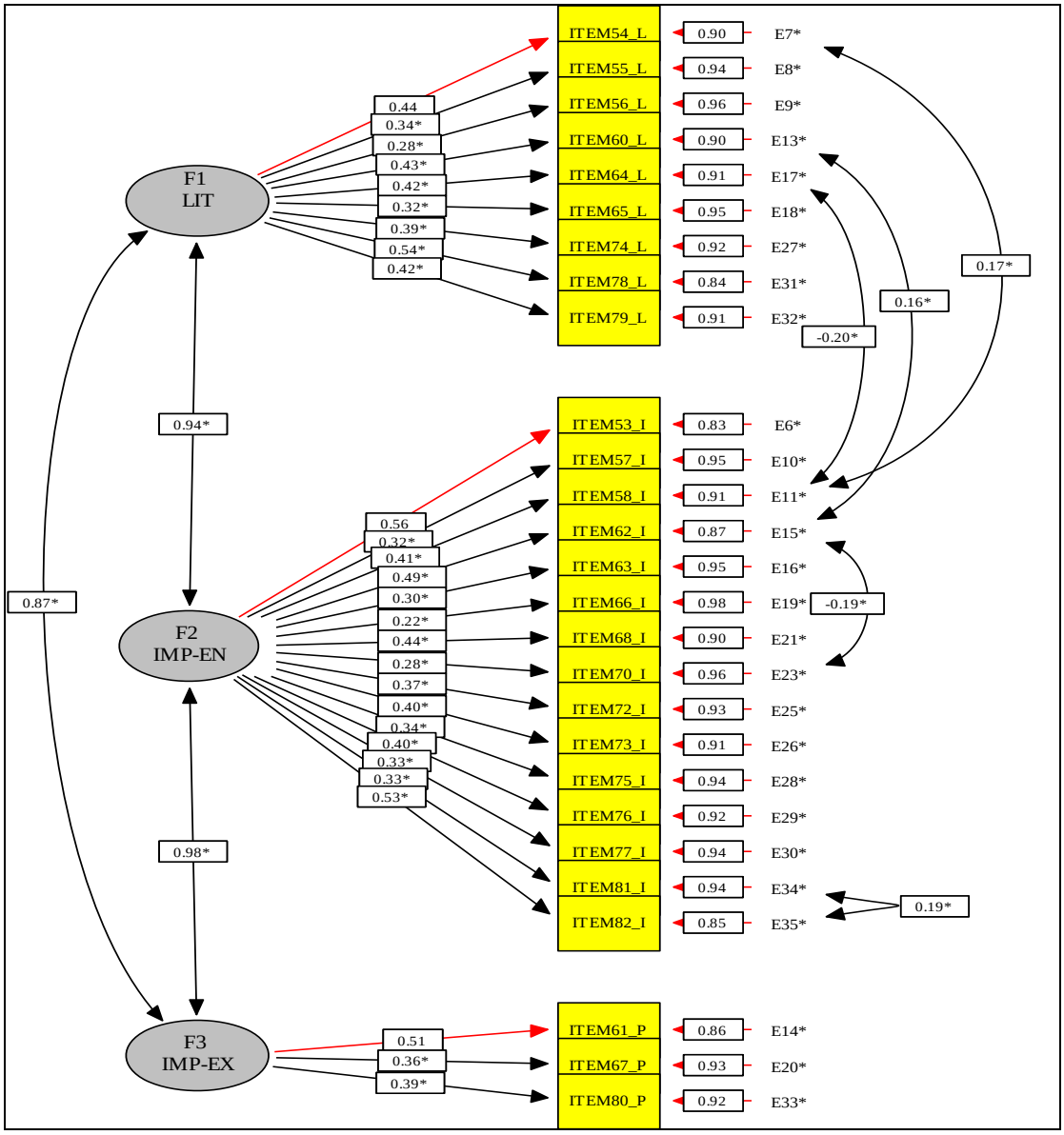
A second-order 3-trait model was then modeled and analyzed to examine if the three traits could be represented by an all encompassing higher-order factor. The results showed that the factor loading on the second-order from the IMP-EN factor was 1.0, an unacceptable parameter estimate value. Therefore, the second-order 3-trait model was rejected as a plausible representation of the data.

Alternative models: 2-trait model and 1-trait model.

The originally hypothesized 3-trait model fit the current data well. However, two alternative models — a 2-factor model and a 1-factor model — were analyzed as well. In the 3-factor model, the correlation between the IMP-EN and the IMP-EX factors was very high at .98, suggesting that the two factors may be measuring very similar constructs. In fact, both factors measure the ability of *reading for implied meaning*. Therefore, the 2-factor model with LIT and IMP factor was further analyzed. The two correlated traits of LIT and IMP were assessed by 27 items and five correlated errors (Figure 3). Furthermore, reading ability may be a unitary trait as was suggested by the EFA results. Thus, a 1-factor model, a unitary reading comprehension factor, was measured by 27 items and five correlated errors (Figure 4).

The two alternative models of 1- and 2-trait models respectively produced a proper solution. That is, the models were all empirically identified and all parameter values were within the acceptable range between zero and one. Table 10 provides a summary of the fit statistics of the modified 3-factor model and the two alternative models. The three models all had good fit statistics with similar CFI indices. The χ^2/df ratios for the models were all under 2.0, the CFIs were higher than .95, and the RMSEA values were all below .05.

FIGURE 2
Modified 3-factor Model (χ^2/df ratio:1.128; CFI:.951; RMSEA:.021)



Note.
Variable names:
1. L: reading for literal meaning
2. I: reading for implied-endophoric meaning
3. P: reading for implied-exophoric meaning

TABLE 9
Parameter Estimates for the 3-trait Modified Model of Reading Ability

Standardized Solution
ITEM53_I=V6 = .559 F2 + .829 E6
ITEM54_L=V7 = .436 F1 + .900 E7
ITEM55_L=V8 = .343*F1 + .939 E8
ITEM56_L=V9 = .280*F1 + .960 E9
ITEM57_I=V10 = .323*F2 + .946 E10
ITEM58_I=V11 = .406*F2 + .914 E11
ITEM60_L=V13 = .426*F1 + .905 E13
ITEM61_P=V14 = .514 F3 + .858 E14
ITEM62_I=V15 = .486*F2 + .874 E15
ITEM63_I=V16 = .300*F2 + .954 E16
ITEM64_L=V17 = .421*F1 + .907 E17
ITEM65_L=V18 = .324*F1 + .946 E18
ITEM66_I=V19 = .222*F2 + .975 E19
ITEM67_P=V20 = .365*F3 + .931 E20
ITEM68_I=V21 = .443*F2 + .896 E21
ITEM70_I=V23 = .277*F2 + .961 E23
ITEM72_I=V25 = .369*F2 + .929 E25
ITEM73_I=V26 = .405*F2 + .914 E26
ITEM74_L=V27 = .385*F1 + .923 E27
ITEM75_I=V28 = .345*F2 + .939 E28
ITEM76_I=V29 = .395*F2 + .919 E29
ITEM77_I=V30 = .333*F2 + .943 E30
ITEM78_L=V31 = .545*F1 + .839 E31
ITEM79_L=V32 = .417*F1 + .909 E32
ITEM80_P=V33 = .386*F3 + .923 E33
ITEM81_I=V34 = .333*F2 + .943 E34
ITEM82_I=V35 = .527*F2 + .850 E35

Note.
Variable names:
1. L: Reading for literal meaning
2. I: reading for implied-endophoric meaning
3. P: reading for implied-exophoric meaning

In addition, since the models were non-nested, the Akaike's Informational Criteria (AIC) values were compared (Kline, 2004). AIC is not used for comparing statistically significant difference between the models. Rather, it is used as a criterion for decision making (ZenCaroline, 2007). The differences were minimum among the three models. Among the three, Model 2 had the lowest AIC index, indicating that it was the best model, while Model 1 had the highest AIC value.

TABLE 10
Fit Statistics for Trait-only Models

Model	χ^2	Df	X^2/df	CFI	RMSEA	AIC
Model 1 (3-trait)	356.4366	316	1.128	.951	.021	-275.563
Model 2 (2-trait)	357.0152	318	1.123	.953	.020	-278.985
Model 3 (1-trait)	360.0055	319	1.129	.950	.021	-277.995

Although the current data provided empirical evidence for the acceptance of all three models, the 3-trait model was chosen for substantive reasons. It illustrated the directness of the relationship between the text and the response on a continuum among the three components of reading ability; LIT had the most direct relationship between the text and the response, IMP-EN had an indirect relationship, and IMP-EX had the most indirect relationship between the text and the response.

In sum, the current data provided empirical and substantive evidence for acceptance of the 3-factor solution with LIT, IMP-EN, and IMP-EX traits. However, the 3-factor model only focuses on how traits affect the test and not on how method effects might be accounted for. Therefore, an MTMM approach was used to examine the interactions among traits and method factors.

Multi-trait Multi-method (MTMM) Models

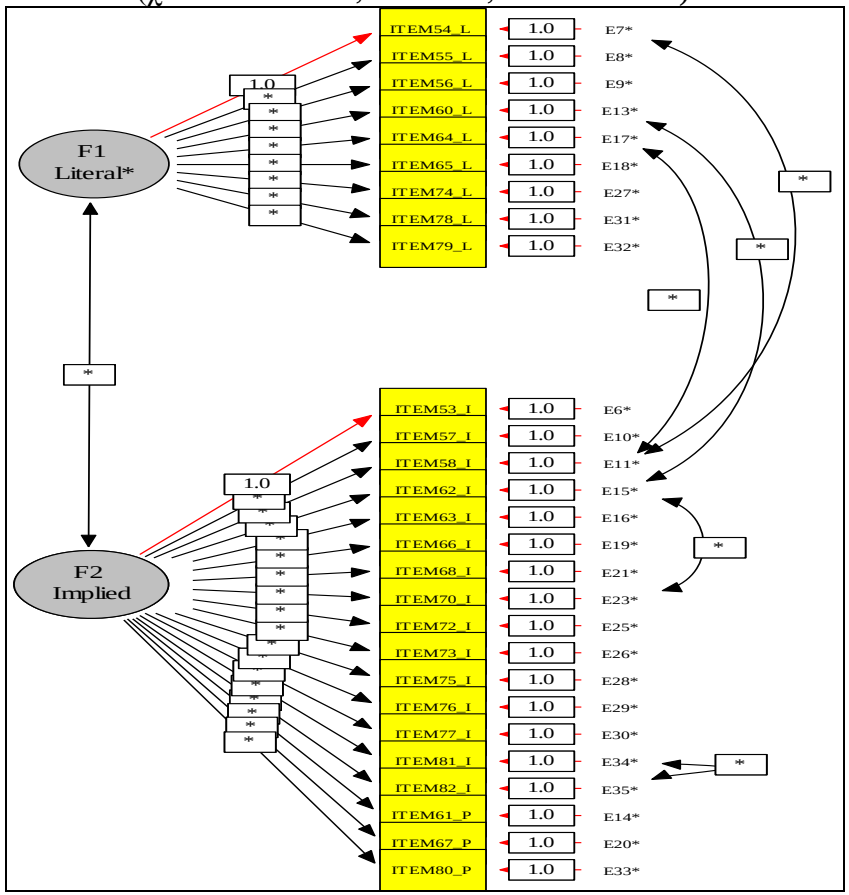
To investigate the second research question regarding the passage effects on the CEP reading placement test, a multi-trait multi-method (MTMM) model was analyzed. MTMM examined the degree to which both traits and methods, i.e., different reading passages, affected test scores. Certain texts may affect test scores more than the others due to their particular text structure or text topic. In other words, items for certain texts may have higher loading than for others. In addition, certain item types may be affected by the texts more than others. For example, an IMP-EX item may require more processing of the content than an LIT item, and thus have a higher loading on the text.

As seen in Figure 5, the MTMM model was constructed by adding method factors—Text 1, 2, 3, and 4—onto the 3-trait model—with LIT, IMP-EN and IMP-EX factors. Each of the 27 observed variables loaded on one trait factor and one method factor. In addition, the three trait factors were postulated to correlate with each other, but the four method factors were not. Five error sets (errors 11 and 7; errors 17 and 11; errors 35 and 34; errors 23 and 15; and errors 15 and 13) were designed to be correlated.

The MTMM model produced good fit statistics: the chi-square/df ratio was below 2 at .98; the CFI was beyond .95 at 1.0; and the RMSEA was below .05 at .0. In addition, the parameter estimates for the trait factors were generally above .03 and the parameter estimates for the method factors were below .3, indicating a small degree of text effect. However, the factor loadings on the text factors were statistically non-significant at the .05 level. This may have been due to a relatively small sample size. Therefore, the parameter estimates were considered unimportant to the model and were deleted. In other words, method factors did not affect the test scores. Deletion of method factors left the reading ability model with the three traits seen in Figure 2, making it the best-fitting model for the CEP reading placement test.

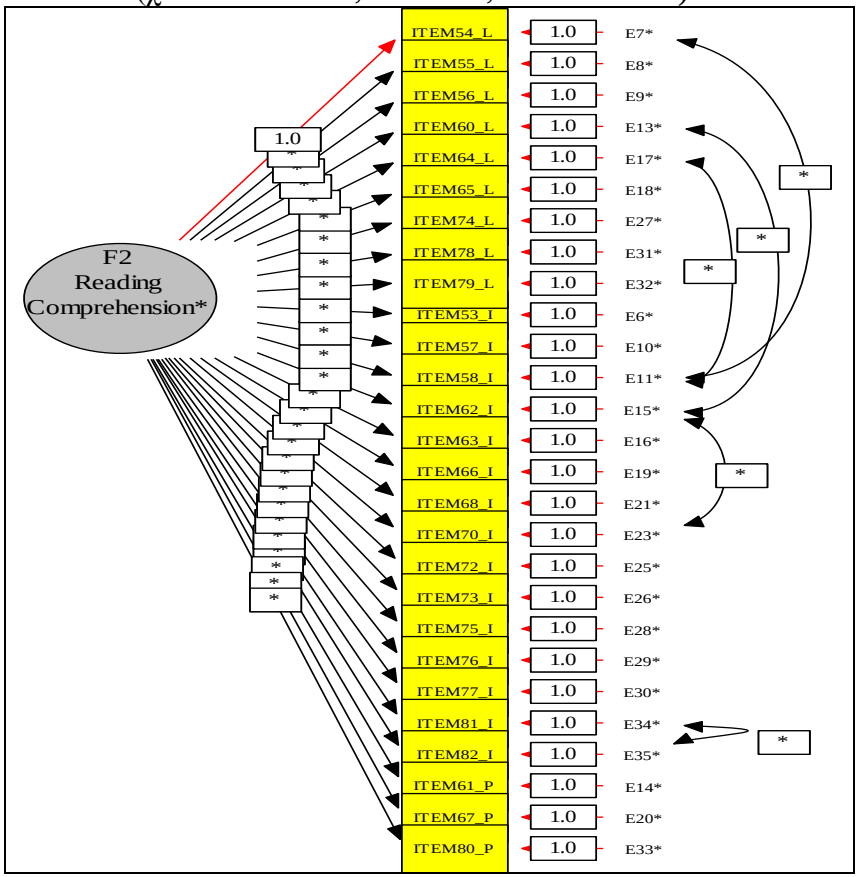
In sum, the present section provided the statistical findings from the study. While EFA results suggested the 1-factor model for L2 reading ability, CFA findings supported the 3-factor model. That is, L2 reading ability consists of multiple components of three different types of reading: LIT, IMP-EN, and IMP-EX. However, contrary to the expectations, passages were not found to have an effect on L2 reading performance. The next section discusses the findings and concludes the paper by describing the limitations of the study and suggestions for future research.

FIGURE 3
2-factor Model of Reading Ability
(χ^2/df ratio:1.123; CFI:.953; RMSEA:.020)



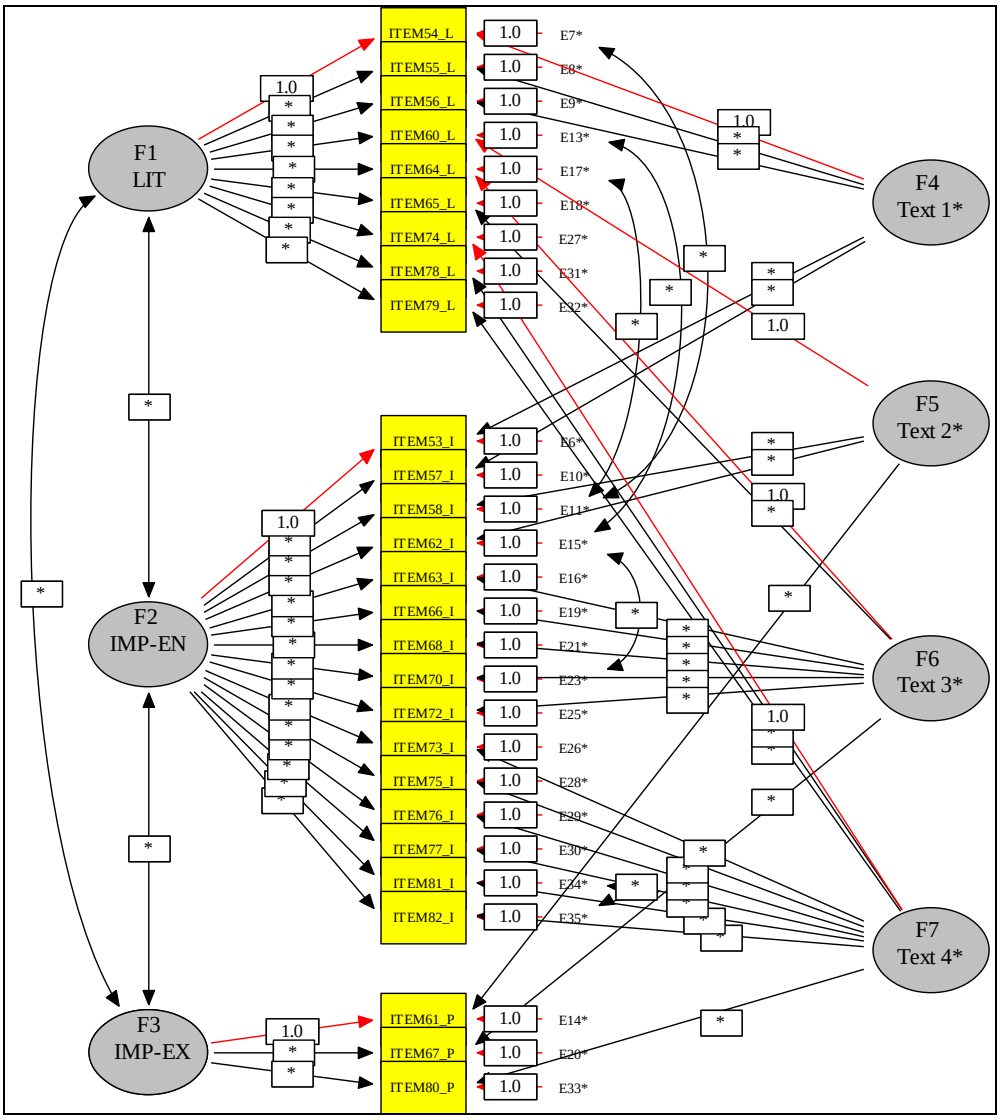
Note.
Variable names:
1. L: reading for literal meaning
2. I: reading for implied-endophoric meaning
3. P: reading for implied-exophoric meaning

FIGURE 4
1-factor Model of Reading Ability
(χ^2/df ratio:1.129; CFI:.950; RMSEA:.021)



Note.
Variable names:
1. L: reading for literal meaning
2. I: reading for implied-endophoric meaning
3. P: reading for implied-exophoric meaning

FIGURE 5
Hypothesized MTMM Model of Reading Ability with 3-traits & 4-method Factors



Note.
Variable names:
1. L: reading for literal meaning
2. I: reading for implied-endophoric meaning
3. P: reading for implied-exophoric meaning

DISCUSSION AND CONCLUSION

The current paper examined the nature of L2 reading ability by investigating the construct of the Community English Program (CEP) reading placement test, developed based on a new L2 reading framework with three components of reading—reading for literal meaning (LIT), reading for implied meaning with endophoric reference (IMP-EN), and reading for implied meaning with exophoric reference (IMP-EX). In addition, reading passage effects were

examined. This section will summarize the results from the study as they relate to the research questions. Implications of the study and suggestions for future research will be discussed as well.

The first research question addressed the nature of L2 reading ability defined in terms of LIT, IMP-EN, and IMP-EX. Confirmatory Factor Analysis (CFA) results confirmed that reading ability could be divided into the three components. In addition, the components showed high correlations among themselves, indicating that they are all measures of L2 reading ability. These findings suggest that L2 reading ability consists of multiple elements as suggested by numerous researchers (e.g., Grabe, 1991; Lumley, 1993). In addition, the results indicate that different types of meanings can be tested via multiple-choice reading items.

Moreover, research findings indicated that LIT, IMP-EN, and IMP-EX were on a continuum in terms of the relationship between the text and the response: LIT had the most direct relationship between the text and the response, whereas IMP-EX had the most indirect relationship. In order to respond to IMP-EX items, readers must not only fully understand the literal meaning of the text and infer from it, but also use their pragmatic knowledge to infer from the context. Therefore, IMP-EX necessitates a deeper understanding of the text compared to LIT or IMP-EN.

The second research question explored whether there was any passage effect in the CEP reading test, using the Multi-trait Multi-method (MTMM) approach in CFA. The MTMM model with three traits (LIT, IMP-EN, and IMP-EX factors) and four method factors (Text 1, 2, 3, and 4 factors) produced a good model fit. However, most of the loadings on the method factors were not statistically significant, suggesting that the texts did not affect reading performance. These findings contradict previous studies (e.g., Lee, 2004), which suggest large influence of texts on reading performance. It could be that L2 readers' ability to comprehend different levels of meaning may not be significantly affected by the characteristics of text.

In sum, L2 reading ability can be seen to consist of multiple components, represented by reading for three different levels of meaning—LIT, IMP-EN, and IMP-EX. Interestingly, the three elements are not distinctly separate, but are correlated. They are on a continuum with LIT having the most direct relationship between the text and the response, and IMP-EX having the most indirect relationship. As the relationship between the test and the response become more indirect, the reader must not only depend on the information provided on the text, but also on his/her background knowledge (e.g., sociolinguistic, sociocultural, psychological, and rhetorical knowledge) to obtain an in-depth understanding of the text. Overall, findings from this research provide evidence for the validity argument of the CEP reading placement test and the new framework of L2 reading ability.

Implications of the Study

The study findings provide some implications for testing and teaching of L2 reading. In terms of testing, the results suggest that L2 reading tests should be designed to measure all three types of reading: LIT, IMP-EN, and IMP-EX. Large-scale multiple-choice tests tend to measure only LIT and IMP-EN. However, reading tests should incorporate more IMP-EX items since IMP-EX is an important construct of L2 reading ability. Furthermore, inclusion of IMP-EX item types, which are generally more difficult than others, could better differentiate among advanced-level readers. However, IMP-EX items should be developed with caution by carefully considering the purpose of the test. For example, if the purpose of a reading test is to screen test-

takers into a British University, it may not be appropriate to measure test-takers' sociocultural knowledge of the American culture.

Likewise, the three different types of reading should be explicitly taught in the classroom. IMP-EX skill is not often explained to students, partly due to the overall difficulty of teaching pragmatics. In terms of sociocultural norms, perception of what is appropriate or polite may vary among individuals or geographic locales. In addition, there are regional differences in viewing politeness. Nevertheless, IMP-EX is an important component of reading that cannot be ignored. Thus, teachers should teach this particular reading skill (the most difficult of the three) to improve students' reading ability. Considering that instruction is often driven by assessment, the inclusion of all three types of reading on reading tests will perhaps encourage teachers to incorporate them in their reading instruction.

Limitations and Suggestions for Future Research

More research is necessary in the future to support the validity argument of the new L2 reading framework. One of the limitations of this study is the rather small number of IMP-EX items involved in the CEP reading placement test. While 11 and 16 items respectively measured LIT and IMP-EN skills, only three were designed to assess IMP-EX. Three items may not be sufficient to provide strong evidence for the current L2 reading framework. Theories (e.g., Purpura, 2004) indicate that there are five types of meanings involved in IMP-EX: contextual, sociolinguistic, sociocultural, psychological, and rhetorical meanings. Thus, having more IMP-EX items that cover all five types of pragmatic meaning will not only provide a better understanding of IMP-EX items in general, but also clarify the differences between the five types.

Also, the non-significant parameter estimates in the MTMM model may have been a result of having a relatively small sample size. A database consisting of 300 test scores may have not been enough to find stable results due to the complexity of MTMM modeling. A larger number of data (e.g., 1000) may provide slightly different results. In sum, a study with both more IMP-EX items and more data should be performed in the future.

AKNOWLEDGEMENTS

I would like to thank my advisor, Professor James Purpura, for his continuous guidance throughout this paper. He introduced me to confirmatory factor analysis (CFA), which inspired me to conduct a study using it. Also, I would like to thank the two anonymous reviewers for their insightful comments on this article. Any remaining mistakes are my own.

REFERENCES

Alderson, J. C. (1990). Testing reading comprehension skills (Part One). *Reading in a Foreign Language, 6*, 425-438.

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

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*, 253-270.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Browne, J. D., Hudson, T. D., Norris, J. M., and Bonk, W. (2002). *An investigation of second language task-based performance assessments.* Honolulu, HI: University of Hawaii Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 445-455). Newbury Park, CA: Sage.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications and programming.* Mahwah, NY: Lawrence Erlbaum Associates.

Chikalanga, I. W. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in a Foreign Language, 8*, 697-710.

Curtis, M. (1980). Development of components of reading skill. *Journal of Educational Psychology, 72*, 656-669.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Journal of Experimental Education, 56*, 67-76.

Davis, F. B. (1941). Fundamental factors of comprehension in reading. *Psychometrika, 9*, 67-76.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3*, 499-545.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16*, 486-514.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading items difficulty: Implications for construct validity. *Language Testing, 10*(2), 133-170.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*, 375-397.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching: Reading*. Harlow, UK: Longman.

Gray, W. S. (1960). The major aspects of reading. In H. Robinson (Ed.), *Sequential development of reading abilities, Supplementary Educational Monographs No. 90,* (pp. 8-24). Chicago: University of Chicago Press.

Halliday, M. A. K., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a socio-semiotic perspective.* Oxford: Oxford University Press.

Hambleton, R. K., & Swaminatahn, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Herber, H. L. (1978). *Teaching reading in content areas*. Englewood Cliffs, NJ: Prentice Hall.

Hoover, W., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*, 127-160.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.

Joreskog, K. G., & Sorbom, D. (1988). *LISREL 7: A guide to the program and applications*. Chicago, IL: SPSS Inc.

Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it.* Newbury Park, CA: Sage University Press.

Kintsch, W. (1998). *Comprehension*. Cambridge: Cambridge University Press.

Kline, R. B. (2004). *Structural equation modeling*. New York: Guildford Press.

Lee, Y-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*, 74-100.

Lumley, T. J. N. (1993). Reading comprehension sub-skills: teachers' perceptions of content in an EAP test. *Melbourne Papers in Applied Linguistics*, *2*, 25-55.

Lunzer, E., Waite, M., and Dolan, T. (1979). Comprehension and comprehension tests. In E. Lunzer and K. Garner (Eds.), *The effective use of reading* (pp.37-71). London: Heinemann Educational Books.

Pearson, P. D., & Johnson, D. D. (1978). Teaching reading comprehension. New York: Holt, Rinehart, and Winston.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

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

Spearitt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly*, *8*, 92-111.

Stanovich, K. E., Cunningham, A., & Feeman, D. (1984). Intelligence, cognitive skills and early reading progress. *Reading Research Quarterly, 19*, 278-303.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81-97.

Urquhart, S., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice.* New York: Longman.

Wainer, H., & Keily, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Williams, E. (2004). Literacy studies. In D. Davies and E. Catherine (Eds.), *The handbook of Applied Linguistics* (pp. 576-603). Malden, MA: Blackwell Publishing.

ZenCaroline. (2007, August 27). Non-nested SEM model comparison. Message posted to http://zencaroline.blogspot.com/2007/08/non-nested-sem-model.html