Investigating the Construct Validity of the Cloze Section in the Examination for the Certificate of Proficiency in English

Yoko Saito
Teachers College, Columbia University

This study addresses the question of the construct validity of the cloze test in the Examination for the Certificate of Proficiency in English (ECPE), which is developed by the English Language Institute, The University of Michigan. Through a rigorous investigation using structural equation modeling (SEM) as a primary statistical procedure, a model composed of two factors with lexico-grammatical ability and reading ability was confirmed and accepted as the best representation of the data. Furthermore, this study empirically demonstrates that the cloze section of the ECPE measures the form and meaning of grammar. In other words, cloze items appear to measure grammatical knowledge on the sentential and suprasentential levels rather than overall language proficiency. This research also demonstrates the usefulness of the structural equation modeling for the examination of the construct validation.

Since the introduction of the cloze method in 1953 by Taylor, there have been numerous studies on cloze from both theoretical and methodological perspectives in the field of language testing. In spite of the extensive research carried out to determine the validity and reliability of cloze tests, inconsistent research findings leave the question, "what does cloze measure?" unanswered. For instance, studies on cloze indicate a wide range of reliability estimates from a low of 0.31 to a high of 0.96 (Alderson, 1979; Bachman, 1985; Brown, 1984, 1988; Hinofotis, 1980; Irvine, Atai, and Oller, 1974; Mullen, 1979; Oller, 1972; Oller and Inal, 1971; Stubbs and Tucker, 1974). Correspondingly, various results have been obtained for the validity of cloze tests (Alderson, 1979, 1980; Bachman, 1985; Brown, 1984, 1988; Hanania and Shikhani, 1986; Hinofotis, 1980; Irvine et al., 1974; Mullen, 1979; Oller, 1972; Oller and Conrad, 1971; Oller and Inal, 1971; Stubbs and Tucker, 1974). Many of the cloze validity studies show moderate to high correlations between cloze tests and standardized tests, and with their sub-tests, such as listening comprehension, reading comprehension, writing, and the Foreign Service Institute (FSI) oral interview (e.g., Hinofotis, 1980; Oller, 1973). Based on the high correlations with other criterion measures, past researchers recommended the cloze test as an integrative test of overall proficiency in English as a second language (ESL).

Similarly, Oller (1979) argues that cloze tests assess the "pragmatic expectancy grammar" that underlies language performance. The theory asserts that test-takers would use the same expectancy grammar for completing a cloze test as they would in any other language context. In other words, the reduced redundancy that results from the cloze procedure forces the test-takers to rely upon their knowledge of underlying linguistic rules and also to retain the coherence of the passage to fill in the blanks. Further, constraints imposed by the internalized rule system allow the test-takers to make predictions about the content of the passage (Laesch and van Kleeck, 1987). Based on this theory, Oller (1983) claims that cloze scores could be

interpreted as indicators of "a general language proficiency factor" (p. 3), which could be applied for various testing purposes.

Contrary to the idea that cloze tests are sensitive to constraints beyond clause boundaries, and therefore measure higher order processing abilities, some researchers argue that cloze tests provide a measure of lower order proficiency such as grammar and vocabulary (Alderson, 1979; Markham, 1985; Purpura, 1999; Shanahan, Kamil, and Tobin, 1982). Alderson (1979) examined the effect of certain methodological variables: passage difficulty, deletion ratio, and scoring criteria. The results showed that a change in each methodological variable had a significant impact on the validity of the cloze test. He concluded that if the cloze is sensitive to the changes in deletion rate, then the claim, "cloze measures higher-order skills" becomes questionable. Accounting for its sensitivity to the deletion ratio, the cloze procedure was found to be essentially "sentence-bound" (1979, p. 225), measuring lower-order skills.

A recent study conducted by Purpura (1999) draws a similar conclusion to Alderson (1979), and Shanahan et al. (1982). Purpura investigated the internal structure of the cloze used to assess English language proficiency. Although the cloze produced a high degree of internal consistency reliability, the majority of items appeared to measure lexical meaning, and to a lesser extent, morphosyntactic form. He also found that grammatical meaning and grammatical form were measured by content words and function words, respectively. Furthermore, the modals and logical connectors appeared to measure both grammatical form and meaning. Based on these results, he concluded that "the cloze task was not a single, global measure of language ability; rather, it measured two highly-related, but separate components of grammatical knowledge – grammatical form and meaning – on the sentential and suprasentential levels." (Purpura, forthcoming, p. 31).

Despite the vast amount of research, there still remain unanswered questions because the results of cloze test research presented in numerous studies differ extensively from study to study. Bachman (1985) suggests that the inconsistency among the results of these studies may be partially due to the methods used in constructing cloze tests. The cloze performance can be affected by contextual features of the test method, such as scoring systems (Alderson, 1979, 1980; Brown, Yamashiro, and Ogane, 2001), deletion ratio (Abraham and Chapelle, 1992; Alderson, 1979, 1980; Bachman, 1982, 1985; Black, 1993; Farhady and Keramati, 1996, Shanahan et al. 1982), passage difficulty (Alderson, 1979; Brown, 1984; Klein-Braley, 1983; Sasaki, 2000), number of items (Sciarone, and Schoorl, 1989), test topic (Alderson and Urquhart, 1985), and method of student response (Abraham and Chapelle, 1992; Bensoussan and Ramraz, 1984; Black, 1993; Storey, 1997). Moreover, the studies vary extensively in the quality of the theories on which they are based, in the clarity and consistency with which they apply theory, in the way sources of variance are managed, and in the way the data obtained are interpreted (Jonz and Oller, 1994). In addition to faulting the differences in test methods, Jonz and Oller argue that some of the earlier studies had serious limitations in research design, such as small sample sizes and uncontrolled for proficiency levels of the test-takers.

In addition to these variables, the choice of statistical procedure for the experimental design may also have affected the results of the research. Because most of the studies were designed in the 1970's and 80's, a majority used correlation analyses to determine the validity of the cloze procedure (except Bachman, 1982; Purpura, 1999; Turner, 1989). Now, however, more sophisticated statistical procedures, such as structural equation modeling (SEM), have increasingly gained attention among language testers (Kunnan, 1995; Purpura, 1999; Sasaki, 1993). SEM provides a means of generating models based on substantive theory, in which hypothetical relationships between latent and observed variables can be tested, evaluated, and

modified. If the SEM procedure had been used for the data in the earlier studies, rather than correlational analyses, more precise information on the underlying construct of cloze might have been revealed (Kunnan, 1998). Given the number of difficulties discussed regarding the experimental designs and the appropriate interpretation of results, the ways in which we resolve these issues will certainly provide valuable insights into the investigation on the validity of cloze tests.

Examination for the Certificate of Proficiency in English

This study examines the underlying construct of the cloze test in the Examination for the Certificate of Proficiency in English (ECPE), which is administered by the English Language Institute at The University of Michigan. The ECPE is an advanced-level ESL examination, which is designed to measure the following language abilities: speaking, listening, writing, reading, and lexical grammar. The exam consists of five components: speaking, listening, writing, grammar/vocabulary/reading and cloze. It is assumed that each section of the test measures a separate ability, and that together they determine the English proficiency of the test-takers. A certificate of proficiency is awarded only to those who obtain passing scores on all five sections of the ECPE. Unfortunately, there have not been any recognized empirical studies on the construct validation of each section to show that all five sections measure distinct abilities; thus, it may be possible that two sections of the test measure the same ability.

According to the *Michigan Certificate Examinations General Information Bulletin 2001-2002*, the ECPE cloze section is intended to assess "… an understanding of the organizational features of written text as well as grammatical knowledge and pragmatic knowledge of English, particularly knowledge about expected vocabulary in certain contexts" (English Language Institute, The University of Michigan, 2002, p. 8). This description seems analogous to the test specification for the grammar/vocabulary/reading (GVR) section. If the cloze section and the GVR section are measuring the same language ability, the test-takers should only have to pass either the cloze or the grammar section, not both. Another solution would be to combine the two sections and reduce the number of subtests in the ECPE battery to four. Either way, an investigation of the construct validity of the cloze section is needed to strengthen the validity of the ECPE.

The purpose of this study is to investigate whether the cloze section of the ECPE merits being a separate section of the ECPE battery. To provide a rationale for the cloze to be (or not to be) an independent section, the underlying trait structure of the cloze section must be compared with the underlying construct of the GVR section.

The current study addresses the following research questions:

1. To what extent do the items in each component (grammar, vocabulary, and reading) perform as a homogeneous group?
2. What is the underlying trait structure of foreign language test performance of English, as measured by the ECPE GVR section?
3. To what extent do the items in the cloze section perform as a homogeneous group?
4. What is the underlying trait structure of foreign language test performance of English, as measured by the ECPE cloze section?
5. What is the relationship between the cloze and the GVR sections?
6. Does the cloze section merit being a separate section of the ECPE battery?

Method

Participants

The data were collected from 79 different test centers all over the world in 1997. This involved 12,468 students of English as a foreign language. The majority of the participants in the study, 75 percent, spoke Greek (N = 9,237). Sixteen percent spoke Portuguese (N = 2,012); and 8 percent spoke Spanish (N = 984). The breakdown of participants by their native language is shown below (Table 1).

With regard to gender, the majority of participants in the study were female (N = 8,175), representing 65.6 percent of the population, while 30.4 percent were male (N = 3,785), and 4.0 percent failed to report the information (N = 507). Also the majority of participants (72.4%) were 22 years of age or younger. The median and the mean ages were 19 and 21, respectively, with the youngest participants being 7 and the oldest being 87. The data revealed a wide range of ages; however, there is a possibility that some of the age information may not be accurate. Although the participants were asked to darken the circles of the last two digits of the year they were born on the answer sheet, some people may have mistakenly marked the wrong circles.

Table 1. Native Language of Participants

	Number	Percent
Arabic	84	0.7
Greek	9237	74.6
Portuguese	2012	16.2
Spanish	984	7.9
Others	151	0.6
Total	12468	100.0

The ECPE Test

The test was developed by the English Language Institute of The University of Michigan (ELI-UM) for advanced-level students. The ECPE Test consists of five sections with 180 selected-response items, one writing task, and one speaking task. The ECPE is designed to measure the test-takers' English language performance levels in the different areas of language ability. The participants were given 155 minutes to complete all the sections in the exam.

Although all the sections of the ECPE need in-depth investigations of their underlying constructs, this paper examines only the GVR and cloze sections. The GVR section consists of 100 multiple-choice items measuring three types of language abilities: Grammar (40 items), Vocabulary (40 items), and Reading (20 items). The cloze section consists of a total of 40 multiple-choice items in two passages, with 20 items each. Table 2 describes the sections of the ECPE.

Procedures

The ECPE is administered annually at over 130 testing centers in about 25 countries. Writing, listening, cloze, and GVR sections are given during a single administration period, followed by the Interactive Oral Communication Section (IOCS) on a different date. In order to ensure test security and to avoid an unfair advantage to any test-taker, each participant's identity is checked, and all the test booklets are collected after the exam.

Table 2. Description of the ECPE

Tasks	Time (minutes)	Number of Items
Writing	30	(1)
Listening	25	40
Multiple-Choice Cloze	25	40
Multiple-Choice Grammar/Vocabulary/Reading	60	
Grammar (GRAM)		40
Vocabulary (VOC)		40
Reading (READ)		20
Interactive Oral Communication	15	(1)

The ECPE answer sheets are distributed first, followed by the test packet. The general instructions are read aloud in English by the test administrator. No questions regarding the test items are answered. For the GVR and cloze sections, the students read the directions and fill in their responses to each item on the provided answer sheet.

Analyses

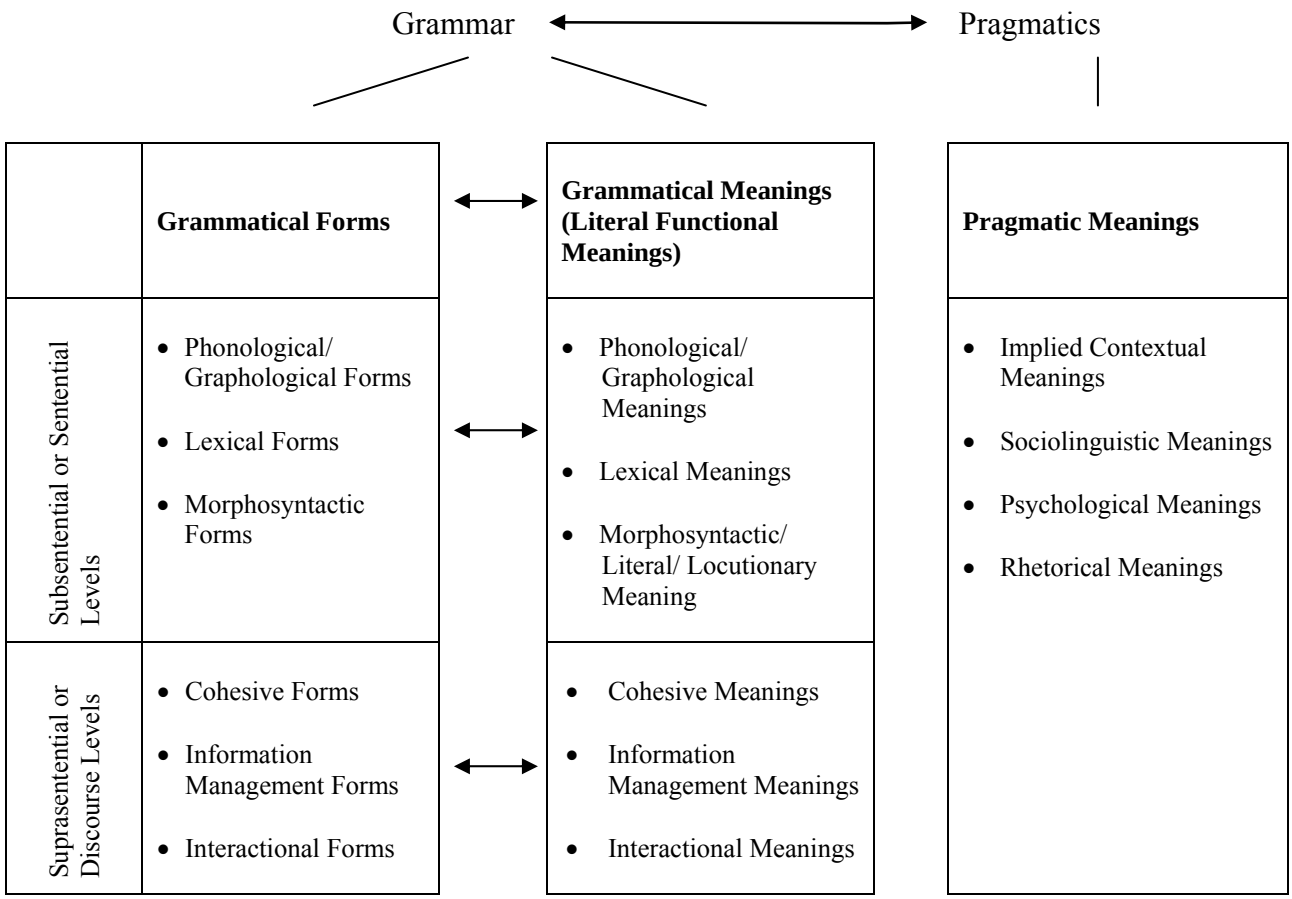
Prior to the statistical analyses, I labeled the test items according to the section of the test. For the cloze section, "C" was marked in front of the number of the item to indicate "cloze." Because there are two passages in the cloze, I labeled the first passage A and the latter B. For example, test item 41 was identified as CA41 (cloze section, passage A) and item 80 as CB80 (cloze section, passage B). Similarly, GVR items were marked with G (i.e., G100), V (i.e., V120), and R (i.e., R160), respectively.

Coding of the GV Section

Before statistical analyses were performed, all the GVR and cloze items were coded to determine what these items were measuring. For the grammar and vocabulary items in the GVR section, the coding was based on the model of grammatical ability proposed by Purpura (forthcoming), which provides a theoretical definition of grammar (See Figure 1). According to his model, language ability is primarily composed of two parts: grammatical knowledge and pragmatic knowledge. Grammatical knowledge is divided into two closely related components: grammatical form and grammatical meaning. Each knowledge component is then defined in terms of six subcomponents, including, at the sentential level, phonological or graphological forms/meaning, lexical forms/meaning, and morphosyntactic forms/meaning, and at the suprasentential level, cohesive forms/meaning, information management forms/meaning, and interaction forms/meaning. Using this model, I attempted to categorize the items according to what domain of grammatical knowledge each item was measuring. Although it is necessary to introduce the model in order to classify the items appropriately, describing each component of the model is beyond the purview of this paper. Therefore, I will only focus on the components which appear to be in the GV section of the ECPE: lexical form (LFORM), lexical meaning (LMEAN), morphosyntactic form (MFORM), cohesive form (CFORM), and cohesive meaning (CMEAN).

According to Purpura (forthcoming), "knowledge of lexical form enables us to understand and produce those features of words that encode grammar rather than those that reveal meaning" (p. 25). These include orthography, part of speech (e.g., happy, happiness), morphological irregularity, word formation (e.g., nightstand; kickoff), countability (e.g.,

Figure 1. A Theoretical Definition of Grammar



(Purpura, forthcoming, p. 84).

children; people) / gender (e.g., actress) restrictions, co-occurrence restrictions, and formulaic expressions. A co-occurrence restriction occurs when a verb or a transitive adjective is followed by a particular preposition (e.g., depend *on* X; yield *to* X) or a given noun phrase is preceded by a particular preposition (e.g., *in* my opinion) (Celce-Murcia and Larsen-Freeman, 1983). An example of lexical form (LFORM) follows:

Mary gets along _____ her roommates well.

a. with *
b. of
c. for
d. to

In this example, the word *along* is followed by the preposition *with*. This is considered the grammatical dimension of lexis, representing a co-occurrence restriction with prepositions (Purpura, forthcoming).

Another component in the model which is closely associated with lexical form is lexical meaning (LMEAN). A difference between the two is that LFORM focuses on the grammatical structure of a word, whereas the LMEAN emphasizes the literal meaning of a word. Consider the following example:

There's a serious ____ between the two university football teams.

a. competition *
b. bile
c. temper
d. exasperation

All four choices for the blank are nouns, thus this item is not measuring the form of the word. Instead, it is examining whether the test-takers understand the meaning of the word. *Competition* carries the meaning of rivalry and is the most appropriate choice in this example.

A third component which is often tested in the GV section is morphosyntactic form (MFORM). As the name of the component suggests, it focuses on a morphological and/or syntactic form of the language. The features of morphosyntactic form are the articles, prepositions, pronouns, inflectional affixes (e.g., -ed), derivational affixes (e.g., un-), simple, compound and complex sentences, mood, and voice. Consider the following example:

I had a hard time _____ for the exam this weekend.

a. studying *
b. to study
c. with study
d. study

In this example, a gerund should be included in the blank. By looking at the choices, the test-takers must recognize that *studying* is a gerund form (–ing form) that functions as a noun. This type of item on the test provides the same word with different alternatives of form in order to measure the test-takers' ability to understand the appropriate morphosyntactic form of the language.

The last components measured in the exam are cohesive form (CFORM) and cohesive meaning (CMEAN). According to Purpura (forthcoming), "knowledge of cohesive form enables us to use the phonological, lexical and morphosyntactic features of the language in order to interpret and express cohesion on both the sentential and discourse levels" (p. 27). This includes cohesive devices such as logical connectors (e.g., therefore; however), pronoun referents, and ellipses (e.g., so do I; I do too). Purpura further states that CFORM is closely associated with CMEAN through cohesive devices that make connections between cohesive forms and their referential meanings within the linguistic environment. Following is a good example of measuring CFORM and CMEAN:

"I didn't go to Jane's party last night."
"_____."

a. Neither did I *
b. I don't either
c. So do I
d. So did I

All four choices are grammatically correct; however, the inverted negative expression, *Neither did I*, is most appropriate in this context. Test-takers who choose *I don't either* as the answer understand the cohesive meaning yet fail to acknowledge the meaning (past tense) difference. On the other hand, selecting *So did I* as the answer shows awareness of the tense, yet a failure to comprehend the cohesive meaning of the sentence. In summary, this item examines whether the test-taker understands the use of cohesive form and meaning at the discourse level.

In order to determine what each item in the GV section is measuring, a total of eleven doctoral students in language testing at Teachers College, Columbia University, were asked to code the items according to Purpura's framework. They were given the descriptions of the coding scheme listed above and were asked to classify each item. When they were unsure of the appropriate classification, they were asked to use their best judgment. The students applied the same coding to the cloze sections to investigate whether the GV items and the cloze items measure common traits. Although most of the item coding was consistent among students, there were some items on which the students did not agree. Those items were carefully examined and coded by a professor who specializes in language testing at Teachers College.

As a result, the GV items were classified using three of the five components. Twenty-five items were coded lexical form (LFORM), 28 items, lexical mean (LMEAN) and 27 items, morphosyntactic form (MFORM), (see Table 3).

Table 3. Initial Taxonomy of the GV Items in the GVR Section (80 Items)

Components	Number of Items	Items
Lexical Form (LFORM)	25	81, 82, 88, 91, 102, 103, 105, 110, 115, 116, 125, 126, 134, 138, 140, 142, 145, 146, 149, 151, 154, 155, 156, 159, 160
Lexical Mean (LMEAN)	28	95, 96, 108, 121, 122, 123, 124, 127, 128, 129, 130, 131, 132, 133, 135, 136, 137, 139, 141, 143, 144, 147, 148, 150, 152, 153, 157, 158
Morphosyntactic Form (MFORM)	27	83, 84, 85, 86, 87, 89, 90, 92, 93, 94, 97, 98, 99, 100, 101, 104, 106, 107, 109, 111, 112, 113, 114, 117, 118, 119, 120

Coding of the R Section

The reading section items were divided into two question types: reading for explicit information and reading for inferential information (Purpura, 1999). For the explicit information (EXP) questions, the participants were asked about specific information in the text. The inferential information (INF) items required the participants to derive meaning not explicitly stated in the text. The coding procedure used for the grammar and vocabulary items was also used for the reading items, resulting in 12 EXP items and 8 INF items, as seen in Table 4.

Table 4. Initial Taxonomy of the Reading Items in the GVR Section (20 Items)

Components	Number of Items	Items
Reading for Explicit Information (EXP)	12	161, 162, 164, 165, 167, 168, 172, 173, 175, 178, 179, 180
Reading for Inferential Information (INF)	8	163, 166, 169, 170, 171, 174, 176, 177

Coding of the Cloze Section

Two coding systems were applied in order to investigate whether cloze tests are sensitive to constraints beyond clause boundaries (Chavez-Oller, Chihara, Weaver, and Oller, 1985; Chihara, Oller, Weaver, and Chavez-Oller, 1977; Fotos, 1991; Jonz, 1990; Oller, 1973; Oller and Conrad, 1971) or only measure sentence level processing abilities such as grammar and vocabulary (Alderson, 1979, 1983; Markham, 1985; Porter, 1983; Purpura, 1999; Shanahan et al., 1982).

I first employed the coding used by Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988) for their TOEFL cloze study. This coding assumes that cloze items measure higher order processing abilities. The second coding was based on the model of grammatical ability proposed by Purpura (forthcoming); this is the same coding used for the GV section in this study. Contrary to the Hale et al.'s (1988) coding, Purpura's categorization assumes that cloze items measure grammatical knowledge rather than global language proficiency. The two coding schemes for the cloze items were examined through the following statistical analyses in order to determine which coding more properly measures the underlying construct of the cloze test.

The Hale et al.'s (1988) classification was based on the assumption that a cloze test includes skills such as grammar, vocabulary, and reading comprehension. According to the TOEFL study, not only are these skills interrelated in certain respects, the classification scheme assumes that the reading comprehension is involved to some degree in all items (Hale et al., 1988). They developed a four-category scheme as follows:

1. Reading Comprehension/Grammar (RG)
2. Reading Comprehension/Vocabulary (RV)
3. Grammar/Reading Comprehension (GR)
4. Vocabulary/Reading Comprehension (VR)

The following description of each category with examples is taken from their study. (The following is a partial quotation.)

> <u>Reading Comprehension/Grammar (RG)</u>
>
> In this category, the task is one of understanding propositional information at an interclausal level, but answering the question also emphasizes knowledge of syntax (i.e., sequential arrangement and markers of such arrangements) rather than of lexicon.
>
> Example: A ballad is a folk song; however, a folk song is not a ballad _____ it tells a story.
>
> a. because
> b. if
> c. whether
> d. unless
>
> <u>Reading Comprehension/Vocabulary (RV)</u>
>
> In this category, the problem is one of long-range constraints, but a lexical choice is required to solve it. The reader's task is basically one of understanding the text and getting the propositional information out of elements that may be some distance apart (usually across clause boundaries) yet a lexical choice is also required.
>
> Example: … known as the Lost Sea. It is listed (in) the <u>Guinness Book of World Records</u> as the world's largest underground _____.

a. water
b. body
c. lake
d. cave

(Parentheses denote another place where a word has been deleted, the correct response here being the word "in.")

Grammar/Reading Comprehension (GR)

Here the source of item difficulty involves relatively short-range grammatical constraints -- usually a few words on either side of the blank, or within a single grammatical phrase or clause. The item primarily taps knowledge of surface syntax, and reading comprehension is involved primarily because the reader must understand within clause propositional information.

Example: It is generally understood that a ballad is a song that tells a story, (but) a folk song is not so _____ defined.

a. easy
b. easily
c. ease
d. easier

Vocabulary/Reading Comprehension (VR)

The primary aspect of this category is vocabulary (including idioms and collocations) although it also invokes reading comprehension to the extent of understanding the information presented within clause boundaries. The main source of difficulty, from the examinee's standpoint, is vocabulary -- not grammar and not the understanding of long-range textual constraints.

Example: In fact, there are folk songs for many occupations -- railroading, _____ cattle, and so on.

a. following
b. mustering
c. concentrating
d. herding

(Hale et al., 1988, pp. 11-12)

A total of four people (three testing experts at the English Language Institute at The University of Michigan and I) participated in coding the items. We used the coding scheme described above to classify each item. Because some items were difficult to classify, we indicated our degree of certainty on a four-point scale for each item (very certain (4), somewhat certain (3), somewhat uncertain (2), and very uncertain (1)). When we did not rate the items as very certain (4), we indicated the other classification(s) the item could receive.

All four judges placed 15 out of the 40 cloze items in the same group, and three judges placed 20 other items in the same group. The fourth judge gave 5 of these 20 items a secondary rating that agreed with the other judges' primary rating. (For example, three judges rated an item grammar/reading (GR). The fourth judge rated it reading/ grammar (RG); however, s/he marked GR as the second choice.) The remaining 5 items were controversial among judges; two judges picked one category while the other two judges chose different categories. (For

example, two judges rated an item GR, but the other two judges marked it RG and RV. In short, the judges appeared to have some difficulty coding the items.)

Each of the 40 items was placed into a category based on the judges' ratings. Where three or four judges indicated the same category (35 of the 40 items), the item was assigned to that category. Where two judges marked one category yet the other two judges were split between two other categories, the item was assigned to the first category. Table 5 summarizes the coding of the cloze section.

Table 5. Initial Taxonomy of the Cloze Section Based on the TOEFL Study Coding

Components	Number of Items	Items
Reading-Grammar (RG)	10	42, 43, 45, 48, 51, 59, 60, 70, 71, 74
Reading-Vocabulary (RV)	12	44, 47, 49, 54, 57, 61, 63, 64, 69, 76, 77, 80
Grammar-Reading (GR)	9	41, 50, 56, 58, 62, 68, 72, 73, 75
Vocabulary-Reading (VR)	9	46, 52, 53, 55, 65, 66, 67, 78, 79

The second coding scheme using Purpura's (forthcoming) model of grammatical ability produced the following taxonomy of the items in the cloze section (see Table 6). While RG/RV/GR/VR coding uses four components, this coding uses five components: lexical form (8 items), morphosyntactic form (2 items), cohesive form (6 items), lexical meaning (23 items), and cohesive meaning (1 item).

Table 6. Second Taxonomy of the Cloze Section Based on Purpura's Model

Components	Number of Items	Items
Lexical Form (LFORM)	8	49, 52, 53, 54, 60, 66, 68, 73
Morphosyntactic Form (MSFORM)	2	72, 78
Cohesive Form (CFORM)	6	41, 51, 58, 62, 70, 75
Lexical Mean (LMEAN)	23	43, 44, 45, 46, 47, 48, 50, 55, 56, 57, 59, 61, 63, 64, 65, 67, 69, 71, 74, 76, 77, 79, 80
Cohesive Mean (CMEAN)	1	42

Descriptive Statistics

To examine the central tendency and dispersion, I calculated descriptive statistics using SPSS Version 10 for the PC. The standard deviations were checked to identify items with no variability. Subsequently, to examine the item distribution, I calculated the kurtosis and skewness of each variable. This allowed for an examination of any potential violations to the assumption of normality. The kurtosis and skewness were expected to range from –3 to +3. If any of the items lay outside the acceptable limits, the items were flagged and further examined to determine whether they could be deleted from the test analyses.

Reliability Analysis

To examine consistency of measurement, the internal consistency reliability estimates were calculated to examine the homogeneity of the test items in each component of the coding schemes for the GVR and cloze sections. I performed reliability analyses on the data using SPSS Version 10.1 for the PC to examine (a) how each item correlated with the other items in the component and (b) how the items in each scale performed as a group. The item-total correlations for each item as well as the overall estimate of the scale reliability were investigated. For the GVR and the cloze sections, I used Cronbach's alpha and the adjusted alpha for the scale if the item was to be deleted. As for the cloze, the reliability estimates may be overestimated because the cloze items may violate the assumption of independence (Bachman, 1990).

Exploratory Factor Analysis

Subsequent to the coding of the items, I performed a series of exploratory factor analyses (EFAs), using Mplus Version 2 for the PC to examine the patterns of correlations among the items within and across each component of the coding schemes for the cloze and GVR sections. In other words, I used EFA to determine whether the items in each component were measuring the same underlying construct and if each component represented an independent construct.

Following Kim and Muller's (1978) and Purpura's (1999) procedures for EFA, I followed three steps in performing the EFAs: (1) preparation of the matrix to be analyzed, (2) extraction of the initial factors, and (3) rotation and interpretation. First, for both cloze and GVR, because of the dichotomous nature of the variables I produced a matrix of tetrachoric correlations among the various items. The data were analyzed and evaluated for factor analytic appropriateness. I based all appropriateness decisions on the determinant of the correlation matrix.

With regard to the extraction, I examined the eigenvalues obtained from the initial extraction, which provided a preliminary indication of the number of factors represented by the data. Consequently, these initial extractions together with the theoretical design of the cloze and the GVR section were used to determine the ultimate number of underlying factors to be extracted.

Following the determination of the minimum and maximum number of factors to extract, the extractions were rotated to an orthogonal solution using a varimax rotation and to an oblique solution using a promax rotation (Purpura, 1999). To determine whether to interpret the orthogonal or the oblique solution, I examined the interfactor correlation matrices. In sum, I used simple structure and meaningful interpretation as final criteria for deciding the best number of factors to extract.

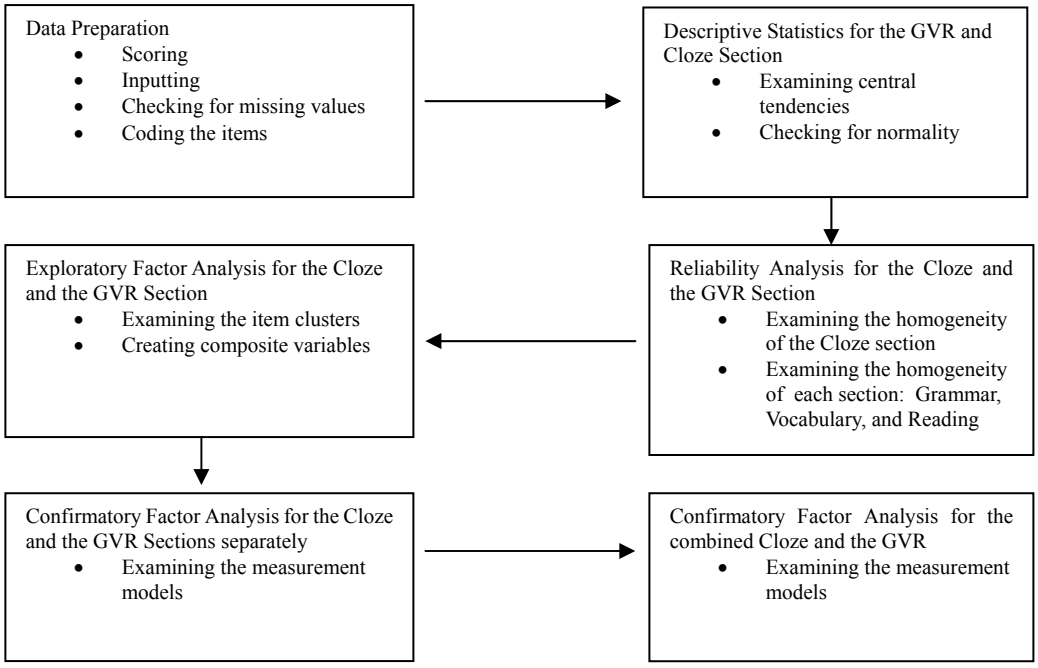
Finally, I performed the reliability analyses with the revised taxonomy of the ECPE based on the results of the EFA. Once again, the reliability analyses examined (a) the homogeneity of the items in the cloze and the GVR section and (b) the degree of consistency of each section.

Structural Equation Modeling

The primary statistical procedure used in this study was structural equation modeling (SEM). SEM is a means of representing interrelationships between observed and content variables and among latent variables based on substantive theory (Purpura, 1999). Each relationship in the model is defined by a set of mathematical equations, and the entire model is empirically tested for overall model-data fit. SEM involves two steps in the analyses:

validating the measurement model, and fitting the structural model (Purpura, 1999). The former is called Confirmatory Factor Analysis (CFA), which examines the hypothesis of linkages between observed variables and their latent variables in the individual measurement model. The latter refers to the procedures for testing the hypotheses of linkages among latent variables. In the current study, I will examine the CFA in order to answer the question of how the items in the GVR and the cloze sections compose the underlying structure of the exam. A flow chart of these procedures is seen in Figure 2.

Figure 2. A Flow Chart of Statistical Procedures Used in this Study



(Adapted from Purpura, 1999)

Results

The results are discussed in three sections, with each section containing four subsections (descriptive statistics, reliability analysis, exploratory factor analysis, and structural equation modeling). First, the results of the GVR section are presented, followed by the results of the cloze section. Finally, the results of the combined GVR and cloze sections are discussed.

The GVR Section

Descriptive Statistics

First, I analyzed the item-level data from the GVR section based on all 12,468 test-takers (see Appendix A). The means for the grammar section ranged from 0.30 to 0.99, suggesting a wide range of item-difficulty levels. The standard deviation ranged from 0.11 to 0.50. Nine items (G89, G90, G96, G97, G98, G100, G102, G112, G117) had means above 0.91, and the values for skewness and kurtosis of those nine items were beyond +/- 3. However, it is perfectly normal to expect high kurtosis values for these items because the mean values were also extremely high.

The means for the vocabulary section ranged from 0.20 to 0.85, again suggesting a wide range of item-difficulty levels, and the standard deviations ranged from 0.36 to 0.50. All values for skewness and kurtosis were within the accepted limits, indicating that all the items appeared to be normally distributed. Compared to the grammar section, this section contains well-balanced item difficulty levels with no item means over 0.85.

The means for the reading section ranged from 0.50 to 0.98, suggesting a moderate range of item-difficulty levels, and the standard deviations ranged from 0.15 to 0.50. All values for skewness and kurtosis were within the accepted limits except for two items (R161 and R165), indicating univariate normality.

There were many items that did not fall within the accepted limits of the descriptive statistics described in the Method section. For instance, items with a mean value higher than 0.90 had kurtosis and skewness values over the acceptable limit. A range of difficulty levels was necessary in the test, and the items that fell outside the accepted descriptive statistic limits did not appear to cause any substantial threats to normality when the reliability analyses were performed. Hence, these items were kept for the subsequent analyses.

Internal Consistency Reliability for the GVR Section

Reliability analyses were performed to examine the extent to which the items in each component of the coding schemes (grammar, vocabulary, and reading) performed as a homogeneous group and the extent to which the items related to other items in the GVR section of the ECPE.

In answer to the first research question, the results showed that all original sections yielded alphas of 0.50 or more: grammar ($\alpha = 0.80$), vocabulary ($\alpha = 0.74$), and reading ($\alpha = 0.72$). The standard error of measurement for each section was then examined to estimate an average of the distribution of error deviations across each section. The standard error of measurement for grammar, vocabulary, and reading were 2.37, 2.81, and 1.71, respectively.

On the sub-section level, the grammar section had a high internal consistency reliability (over 0.80); however, the alpha of the vocabulary section and the reading section were somewhat moderate. In other words, the items in the grammar section appeared to measure the same construct within the section more than the items in the vocabulary and reading sections did. When the reading items were fixed to 40 items, using the Spearman-Brown Prophecy formula, reliability increased to 0.84. This indicates that when the number of items is held constant to 40 items, READ has the highest internal consistency reliability among the three types of items.

In order to examine the internal consistency reliability for the GVR section of the ECPE, all 100 items were included in the analyses. The results yielded an alpha of 0.87 with a standard error of measurement of 4.10. The reasonably high alpha of 0.87 suggests that the items in the exam appear to be measuring the same construct: English as a foreign language (EFL) grammar, vocabulary, and reading comprehension test performance.

A summary of the Cronbach's alpha reliability estimates for internal consistency of the GVR section is presented in Table 7. Although the present reliability analyses provide invaluable information on the homogeneity of the items as well as the degree of consistency of each section, the information on the underlying trait structure of the GVR section remains unknown. In order to answer the second research question (see p. 3), exploratory factor analysis was performed. It is discussed in the following section.

Table 7. Reliability Estimates for the GVR Section

GVR Section	Number of Items	Reliability Estimates	Corrected Reliability Estimates	Standard Error of Measurement
Grammar	40	0.80		2.37
Vocabulary	40	0.74		2.81
Reading	20	0.72	0.84	1.71
Total	100	0.87		

Exploratory Factor Analysis

To investigate the trait structure of the GVR section of the ECPE, a matrix of tetrachoric correlations using all 100 items was generated in Mplus Version 2. Then a series of EFAs was performed on the GVR section. The following presents a summary of the findings.

I first performed EFAs on all 100 items in the GVR section. These analyses produced a three factor promax rotation that seemed to maximize parsimony and interpretability. Although each component of the GVR was designed to measure three factors, each representing one of the three components identified in the test content specifications (grammar, vocabulary, and reading), many of the vocabulary and grammar items loaded on the same factors. This indicates that some of the vocabulary items were measuring the same trait as many of the grammar items, and vice versa. Table 8 presents the initial 3-factor loadings of the GVR section.

While the grammar and vocabulary items combined in loading on two factors, all 20 reading items loaded on one factor, indicating that they appeared to be measuring one underlying construct. This implies that reading items measure a distinct ability from grammar and vocabulary. Based on the results of the EFAs, the GVR section of the ECPE was considered to measure two underlying construct abilities: reading ability and lexico-grammatical ability (L-G).

The next step was to analyze the GRAM and VOC (GV) items together to examine how the lexico-grammatical ability was measured. The EFAs on the GV items produced a three factor promax rotation that maximized parsimony and interpretability. In the course of these analyses, 37 items (G82, G83, G84, G85, G88, G92, G95, G96, G97, G98, G99, G101, G102, G104, G106, G107, G108, G109, G110, G113, G114, G120, V121, V124, V129, V130, V131, V132, V134, V139, V141, V144, V145, V147, V149, V157, and V159) produced extremely low factor loadings (lower than 0.3) and double loadings. This may be due to the moderate values of reliability for grammar and vocabulary sections ($\alpha = 0.80$ and $\alpha = 0.74$, respectively). If the items in each section had been performing more homogeneously, the EFA may not have produced this many items with low factor loadings and double loadings. Because these items distract from investigating the construct validity of the ECPE, 37 items were dropped from further analyses. Some items with a factor loading of a little less than 0.30 were kept because they clearly loaded on a factor. For example, Item V158, with a loading value of 0.276, was kept in the analysis even though it did not make the cut-off line of 0.30. A reason for keeping the item was that the loadings for morphosyntactic form and lexical form were extremely low: 0.095 and –0.074, respectively. In other words, when the factor loadings were compared, Item V158 was clearly measuring lexical meaning. Ultimately, 43 items were retained to measure the three underlying factors. Through a series of EFAs, a promax solution again produced three factors, as seen in Table 9. The factor correlation matrix is shown in Table 10. Based on the Purpura's grammatical ability model, these three factors are the following.

Table 8. The Initial EFA Results of GVR Section: Promax Rotation

Item	F1	F2	F3	Item	F1	F2	F3	Item	F1	F2	F3
G81	0.545	-0.187	0.091	G114	-0.338	0.550	-0.021	V147	0.285	0.197	-0.056
G82	0.248	0.186	0.018	G115	0.731	-0.281	0.054	V148	-0.693	0.819	-0.049
G83	0.468	0.074	-0.089	G116	0.671	-0.117	0.053	V149	0.140	0.086	-0.032
G84	0.625	-0.164	0.016	G117	0.143	0.074	0.188	V150	-0.117	0.558	-0.084
G85	0.373	0.023	0.027	G118	0.025	0.419	0.072	V151	0.239	0.042	-0.073
G86	0.064	0.153	0.044	G119	0.182	0.124	0.168	V152	-0.107	0.401	-0.043
G87	0.497	-0.096	0.068	G120	0.548	-0.028	-0.064	V153	-0.162	0.518	0.193
G88	0.198	0.216	0.180	V121	0.252	0.125	0.037	V154	0.499	-0.109	0.072
G89	-0.025	0.361	0.113	V122	-0.055	0.466	-0.009	V155	0.343	-0.061	-0.129
G90	0.192	0.048	0.219	V123	-0.088	0.539	0.001	V156	0.301	-0.053	0.192
G91	0.539	-0.150	0.075	V124	-0.111	0.366	-0.094	V157	0.371	-0.033	0.063
G92	0.358	0.083	0.109	V125	0.255	0.055	-0.048	V158	-0.061	0.298	0.033
G93	0.162	0.168	0.109	V126	0.314	0.022	0.073	V159	0.471	-0.017	0.134
G94	0.313	0.277	0.007	V127	0.048	0.444	-0.037	V160	0.602	-0.330	0.128
G95	0.285	0.109	-0.066	V128	0.077	0.432	-0.049	R161	0.093	-0.087	0.370
G96	0.304	-0.043	0.183	V129	0.331	0.041	0.012	R162	-0.011	-0.015	0.363
G97	0.698	-0.142	0.117	V130	0.281	0.055	-0.068	R163	0.144	-0.063	0.459
G98	0.252	0.013	0.167	V131	0.162	0.276	-0.062	R164	-0.027	0.042	0.387
G99	0.035	0.367	0.031	V132	0.209	0.176	0.002	R165	0.150	-0.090	0.571
G100	0.295	0.096	0.098	V133	0.017	0.390	0.023	R166	-0.007	0.072	0.449
G101	0.239	0.051	0.178	V134	-0.388	0.535	-0.017	R167	-0.058	0.211	0.367
G102	0.462	-0.135	0.142	V135	-0.062	0.319	0.282	R168	-0.003	0.040	0.400
G103	0.801	-0.384	0.114	V136	-0.258	0.603	-0.022	R169	-0.124	0.119	0.390
G104	0.600	-0.043	0.076	V137	0.223	0.291	0.102	R170	-0.155	0.173	0.534
G105	0.448	0.066	-0.053	V138	0.643	-0.144	-0.058	R171	0.028	-0.064	0.556
G106	0.576	-0.140	0.037	V139	0.345	0.145	-0.027	R172	-0.008	0.015	0.557
G107	0.390	0.228	0.007	V140	0.529	-0.053	-0.107	R173	0.031	0.023	0.409
G108	0.507	-0.061	0.085	V141	0.340	0.215	-0.037	R174	0.095	-0.128	0.418
G109	0.487	-0.113	-0.015	V142	0.442	0.049	-0.023	R175	0.165	-0.131	0.543
G110	0.148	0.193	0.065	V143	-0.114	0.360	0.115	R176	0.094	0.051	0.371
G111	0.283	0.280	-0.037	V144	0.344	0.215	-0.073	R177	0.041	0.070	0.452
G112	0.208	0.201	0.146	V145	0.260	0.105	-0.103	R178	0.003	0.062	0.502
G113	0.420	0.308	-0.050	V146	0.577	-0.061	-0.014	R179	-0.045	0.104	0.461
								R180	0.135	-0.037	0.371

Table 9. EFA Results of GV Section: Promax Rotation

Item	Code	F1: MF	F2: LF	F3: LM
G118	MF	0.525	-0.163	0.177
G93	MF	0.505	-0.015	-0.023
G112	MF	0.497	0.062	0.033
G90	MF	0.446	0.099	-0.049
G100	MF	0.445	0.124	-0.038
G89	MF	0.440	-0.165	0.181
G87	MF	0.430	0.298	-0.200
G117	MF	0.425	0.040	-0.055
G94	MF	0.385	0.149	0.116
G111	MF	0.338	0.104	0.151
G86	MF	0.313	-0.058	0.010
G119	MF	0.297	0.113	0.065
G103	LF	0.096	0.776	-0.234
G115	LF	0.147	0.657	-0.193
V160	LF	-0.029	0.654	-0.136
V138	LF	-0.085	0.640	0.017
G116	LF	0.173	0.596	-0.052
V146	LF	0.052	0.542	0.030
V140	LF	-0.092	0.509	0.068
V154	LF	0.077	0.487	-0.027
G81	LF	0.186	0.476	-0.157
G91	LF	0.206	0.462	-0.130
V142	LF	0.036	0.412	0.122
V155	LF	-0.231	0.383	0.100
V156	LF	-0.231	0.342	0.019
V126	LF	0.031	0.339	0.097
G105	LF	0.123	0.339	0.077
V151	LF	-0.130	0.276	0.134
V125	LF	-0.081	0.265	0.142
V148	LM	0.103	-0.707	0.616
V150	LM	-0.141	-0.050	0.610
V123	LM	-0.071	-0.018	0.589
V136	LM	0.012	-0.243	0.549
V153	LM	0.138	-0.090	0.507
V133	LM	-0.119	0.105	0.495
V127	LM	-0.008	0.071	0.465
V143	LM	-0.007	-0.034	0.393
V128	LM	0.100	0.027	0.380
V122	LM	0.148	-0.099	0.378
V152	LM	0.026	-0.118	0.364
V135	LM	0.135	0.055	0.357
V137	LM	0.166	0.222	0.307
V158	LM	0.095	-0.074	0.276

LF = Lexical Form, MF = Morpho-Syntactic Form, LM = Lexical Mean

Table 10. GV Section Factor Correlation Matrix

	MFORM	LFORM	LMEAN
MFORM	1.00		
LFORM	0.31	1.00	
LMEAN	0.36	0.23	1.00

1. A form factor (MFORM), which includes items dealing with morphosyntactic form.
2. Another form factor (LFORM), which consists of lexical forms with word formation, co-occurrence restrictions, and formulaic expressions.
3. A meaning factor (LMEAN), which includes denotational (literal) meanings, meanings of formulaic expressions, semantic fields, etc.

An inspection of the interfactor correlation matrix indicates that there is a low correlation among three factors. Based on these analyses, 14 items were used to form the LMEAN composite variables for subsequent analyses. For LFORM and MFORM, 17 and 12 items were used, respectively, in order to form composite variables.

After performing the EFAs on the GV section, I performed a separate series of EFAs on the reading section of the test. The reading section is composed of four discrete passages with five corresponding items per passage. The EFAs produced a 4-factor promax solution that seemed to maximize parsimony and interpretability in spite of the fact that two factors, each representing (1) reading for explicit information and (2) reading for inferential information, were expected (Purpura, 1999). The 4-factor solution produced an interesting result: the items loaded according to the passage, which clearly indicates that the items are text dependent (see Table 11). For example, items 161-165 loaded on the first passage in the reading section, items 166-170 loaded on the second passage, and so forth. An inspection of the interfactor correlation matrix (Table 12) shows that all four factors are moderately correlated, with correlation coefficients over 0.50.

Post-EFA Reliability Analyses

Following the exploratory factor analysis, I performed reliability analyses with the revised taxonomy, which contains two sections with four variables in the GVR section of the ECPE (see Table 13). The sections are grammar/vocabulary (GV) and reading (READ). The GV section contains three components, which are morphosyntactic form (MFORM), lexical form (LFORM), and lexical meaning (LMEAN). The reading section contains only one component, READ. In examining the new GVR section for its overall reliability and the degree to which the items in the scale related with the others, I found that all four components were moderate: MFORM ($\alpha = 0.52$), LFORM ($\alpha = 0.72$), LMEAN ($\alpha = 0.64$), and READ ($\alpha = 0.72$). The standard errors of measurement for each component were examined to estimate an average of the distribution of error deviation across each component. The standard error of measurement for MFORM, LFORM, LMEAN, and READ were 1.21, 1.71, 1.55, and 1.71, respectively.

In addition to the reporting of reliability estimates for all 63 items, corrected reliability estimates using the Spearman-Brown Prophecy formula for MFORM, LFORM, and LMEAN items are provided (see Table 13). The purpose of using corrected reliability estimates is to compare reliability estimates for tests with differing lengths, as the formulas are dependent on test length (Hatch and Lazaraton, 1991; Henning, 1987). For instance, an alpha of 0.8 on a 100-item test is not comparable to an alpha of 0.8 on a 10-item test. Corrected reliability

Table 11. EFA Results of READ Section: Promax Rotation

Item	Code	F1 Passage 1	F2 Passage 2	F3 Passage 3	F4 Passage 4
165	EXP	0.782	0.118	-0.005	-0.097
164	EXP	0.445	0.048	0.042	-0.022
161	EXP	0.424	-0.101	-0.004	0.171
162	EXP	0.379	0.063	-0.005	0.009
163	INF	0.333	0.074	0.154	0.063
169	INF	0.027	0.529	-0.075	-0.007
170	INF	0.007	0.523	0.083	0.031
166	INF	0.052	0.412	0.079	0.051
168	EXP	0.016	0.369	0.146	-0.015
167	EXP	0.093	0.365	-0.014	0.086
175	EXP	0.060	-0.018	0.570	0.080
172	EXP	0.044	0.028	0.558	0.024
171	INF	0.011	0.113	0.552	-0.032
173	EXP	-0.078	0.065	0.533	-0.023
174	INF	-0.014	0.006	0.412	0.085
177	INF	-0.017	0.019	-0.019	0.662
176	INF	-0.088	0.039	-0.020	0.625
179	EXP	0.002	0.159	-0.030	0.465
178	EXP	0.000	0.049	0.154	0.464
180	EXP	-0.015	-0.089	0.195	0.426

EXP = Reading for Explicit Information, INF = Reading for Implicit Information

Table 12. Read Section Factor Correlation Matrix

		Passage 1	Passage 2	Passage 3	Passage 4
F1	Passage 1	1.00			
F2	Passage 2	0.55	1.00		
F3	Passage 3	0.54	0.60	1.00	
F4	Passage 4	0.51	0.53	0.66	1.00

Table 13. Reliability Estimates for the GVR Section (63 items)

Section		Number of Items	Reliability Estimates	Corrected Reliability Estimates	Standard Error of Measurement
GV	MFORM	12	0.52	0.86	1.21
	LFORM	17	0.72	0.78	1.71
	MEAN	14	0.64	0.84	1.55
READ	READ	20	0.72	0.72	1.71
	Total	63	0.81	0.87	

estimates for MFORM, LFORM, LMEAN components were fixed to 40 items since that was the original number of items for both the grammar and vocabulary sections. According to the corrected reliability, MFORM, LFORM, and LMEAN produced fairly high reliability estimates, indicating that the items in each section are highly homogeneous. Corrected reliability

estimates for the reading section are not required because no items were deleted from the section.

The internal consistency reliability for the revised GVR section yielded an alpha of 0.81 with a SEM of 3.20. The corrected reliability estimate for the overall GVR was fixed to 100 items, which produced an alpha of 0.87. Despite the extensive decrease in the number of items, the alpha value did not change much from the original value. This suggests that the items remaining in the GVR section appear to measure the construct consistently.

Structural Equation Modeling

According to the EFA results, the GVR section is hypothesized to be composed of two components measuring lexico-grammatical ability and reading ability. Table 14 presents a summary of descriptive statistics for the seven factors in the GVR section. Since lexico-grammatical ability appears to measure only one type of MEAN items (LMEAN), the lexical mean items are specified as MEAN in the further analyses.

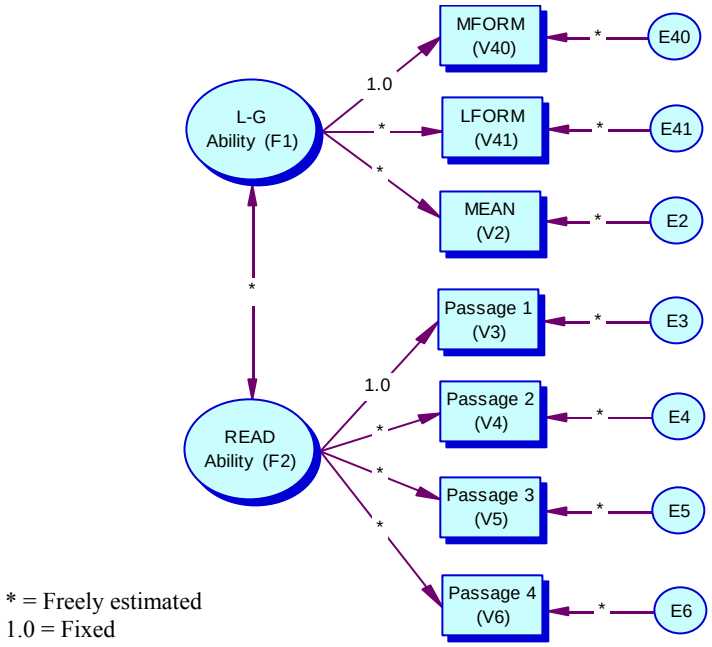
Table 14. Distributions of the GVR Section (63 items)

Variable	Mean	Std. Dev.	Kurtosis	Skewness	Min	Max	# Poss.
Lexico-Grammatical Ability							
MFORM	9.37	1.75	0.49	-0.69	0	12	12
LFORM	10.61	3.24	-0.19	-0.55	0	17	17
MEAN	8.51	2.58	-0.45	-0.08	0	14	14
Reading Ability							
READ 1	4.48	0.77	2.89	-1.65	0	5	5
READ 2	3.79	1.19	0.17	-0.88	0	5	5
READ 3	4.00	1.15	0.82	-1.14	0	5	5
READ 4	3.04	1.45	-0.88	-0.27	0	5	5

Based on the results of the EFAs, the abbreviated GVR section of the ECPE was represented as a two-factor model of foreign language test performance of English and contains two intercorrelated factors, lexico-grammatical ability (L-G) and reading ability. Between the two factors, there are seven observed variables (Morphosyntactic Form, Lexical Form, Lexical Mean, Passage 1, Passage 2, Passage 3, and Passage 4), with each variable hypothesized to load on only one factor (see Figure 3). This is a first-order confirmatory factor analysis designed to test the multidimensionality of foreign language test performance of English as measured by the abridged GVR section. Before exploring the trait structure of the GVR section, I first investigated the statistical assumptions underlying the estimation procedure used in these analyses and then proceeded to assessing model-data fit. The statistical analysis showed that the variables in this model were univariately normally distributed; thus, further statistical analyses were produced. Model 1.1 in Figure 3 addresses the following research question: (2) what is the underlying trait structure of foreign language test performance of English as measured by the ECPE GVR section?

Model 1.1 is similar to the model presented in Purpura's study (1999). This is a first-order confirmatory factor analysis designed to investigate the multidimensionality of the foreign language test performance of English measured by the GVR section. Prior to exploring the trait

Figure 3. Initially Hypothesized 2-Factor Model of the GVR Section: Model 1.1



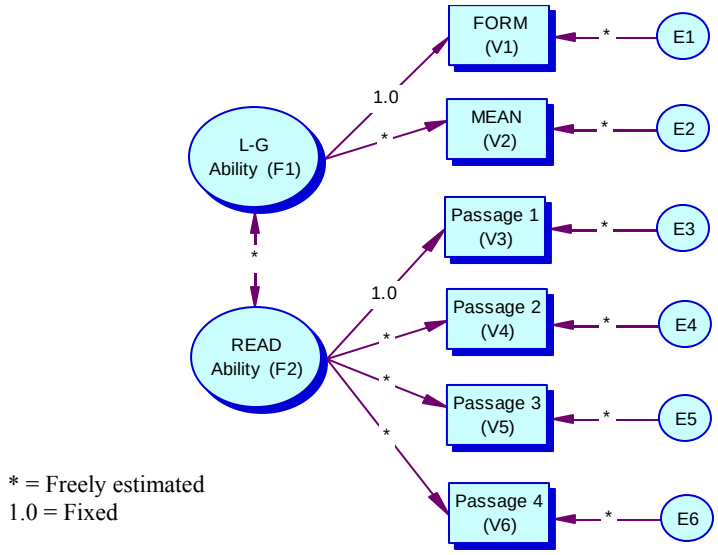
structure of this model, I examined the univariate and multivariate sample statistics for sample normality. As shown in Table 14, the skewness and kurtosis values were within the acceptable limits, indicating that these variables are normally distributed. Consequently, by using Mplus Version 2, I assessed the hypothesized model to determine to what extent the model fit the sample data. The model-data fit statistics for Model 1.1 produced a chi-square of 873.562 with 13 degrees of freedom ($p < 0.0000$) and a CFI of 0.92, as seen in Table 15. A root mean square error of approximation (RMSEA) of 0.07 indicates a degree of global misfit. Although this model fit is not completely unsatisfactory, it does not provide substantial evidence for acceptance of Model 1.1. As Model 1.1 was not the best model, I made no interpretation of the individual parameter estimates. Instead, I further investigated a model which better represents the sample data.

Table 15. Results for Initially Hypothesized 2-Factor Model of the GVR Section: Model 1.1

Goodness of fit summary:	
Comparative fit index (CFI)	0.92
The Tucker-Lewis index (TLI)	0.92
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.04
Root Mean Square Error of Approximation (RMSEA)	0.07
Chi-square test of model fit:	
Value	873.562
Degrees of Freedom	13
P-Value	0.0000

Based on the results of Model 1.1, I reconceptualized the model using a series of fitting procedures. Numerous models were examined; however, most were misfitting or substantively irrelevant. Through rigorous investigation, it was revealed that one model appeared to represent the sample data well from both a substantive and statistical point of view. The two observed variables under lexico-grammatical ability are both assumed to measure form (lexical form and morphosyntactic form); therefore, these variables were combined to construct one variable; FORM. In the revised model, the GVR section is hypothesized to be composed of two factors with six observed variables (Model 1.2, Figure 4).

Figure 4. The Revised 2-Factor Model of the GVR Section: Model 1.2



Prior to assessing the model-data fit, the descriptive statistics for FORM were reexamined, as seen in Table 16. All values for skewness and kurtosis were within the acceptable limits, indicating the variable is unvariately normally distributed; therefore, the SEM was performed on the Model 1.2.

Table 16. Distributions of the FORM Variable in the GVR Section (63 items)

Variable	Mean	Std. Dev.	Kurtosis	Skewness	Min	Max	# Poss.
FORM	19.98	4.00	0.13	-0.56	0	29	29

Using Mplus Version 2, I evaluated the model for overall model data fit, as seen in Table 17. The goodness-of-fit index for the revised two-factor model of foreign language test performance of English produced a chi-square value of 408.302 with 8 degrees of freedom, representing a substantial drop in overall chi-square ($\Delta\chi^2(5) = 465.26$) from the initially hypothesized model. This reduction in χ^2 demonstrated a substantial improvement in goodness of fit. Along with the chi-square, the CFI (0.96) also reflected an improvement in model-data fit ($\Delta = 0.04$). Although a smaller RMSEA may be preferred, Model 1.2 is a better representation of the data compared to the previous model.

Table 17. Results for the Revised 2-Factor Model of the GVR Section: Model 1.2

Goodness of fit summary:	
Comparative fit index (CFI)	0.96
The Tucker-Lewis index (TLI)	0.97
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.03
Root Mean Square Error of Approximation (RMSEA)	0.06
Chi-square test of model fit:	
Value	408.302
Degrees of Freedom	8
P-Value	0.0000

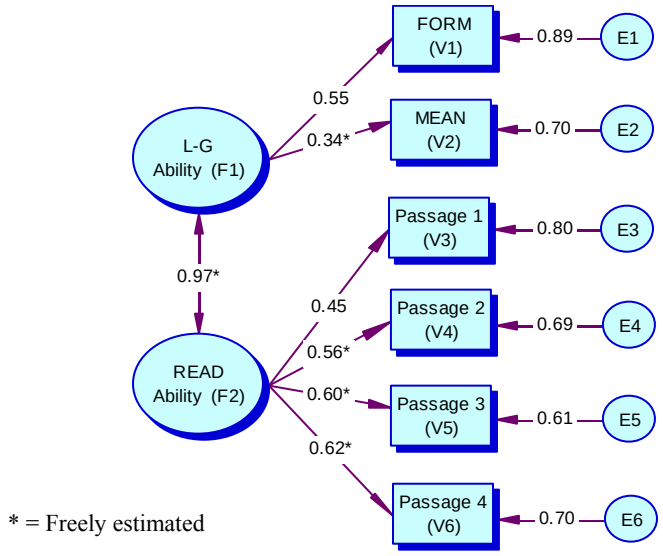
The standardized solution shown in Table 18 presents Model 1.2 in the form of substantive relationships represented by mathematical equations. For instance, the equation for FORM shows that the FORM (V1) items depend on one latent variable, lexico-grammatical ability, and one error term (E1), which accounts for any measurement error in this variable as well as any specific systematic component of the variable not captured in the latent variables. I evaluated the feasibility of the individual parameter estimates and discovered all to be reasonable and statistically significant at the 0.05 level. This indicates that the underlying factors are well measured by the observed variables, and that these variables are measuring lexico-grammatical and reading ability. Moreover, the variances of the error terms as well as all the parameter estimates were statistically significant. The loadings in the standardized solution were somewhat moderate, ranging from a low 0.34 to a moderate 0.62. Model 1.2, along with the standardized parameter estimates, is presented in Figure 5.

In sum, Model 1.2 provides strong evidence for acceptance of the two-factor solution of foreign language test performance measured by the abbreviated GVR section as a reasonable explanation of the correlations among the observed variables. This solution asserts the notion that the shortened GVR section of the ECPE consists of two underlying factors: lexico-grammatical ability and reading ability. According to this model, reading ability results vary because the reading passages are different, not because the item types measure different skills (i.e., reading for explicit information and reading for inferential information). On the other hand, lexico-grammatical ability results vary because the item types measure different skills. Furthermore, this solution produced a high interfactor correlation (r = 0.97) between lexico-grammatical ability and reading ability, suggesting that these abilities are not purely independent. Instead, these two abilities are inextricably related.

Table 18. Parameter Estimates for Model 1.2

Standardized Solution:						
FORM	=	V1	=	0.55 F1	+	0.89 E1
MEAN	=	V2	=	0.34 F1	+	0.70 E2
Passage 1	=	V3	=	0.45 F2	+	0.80 E3
Passage 2	=	V4	=	0.56 F2	+	0.69 E4
Passage 3	=	V5	=	0.60 F2	+	0.61 E5
Passage 4	=	V6	=	0.62 F2	+	0.70 E6

Figure 5. The Revised 2-Factor Model of the GVR Section with Standardized Parameter Estimates: Model 1.2



Cloze Section

Descriptive Statistics

First, I analyzed the item-level data from the cloze section based on all 12,468 test-takers (see Appendix B). The means for the cloze section ranged from 0.31 to 0.97, suggesting a wide range of item-difficulty levels. The standard deviation ranged from 0.16 to 0.50. There are four items (CB62, CB72, CB75, CB78) that had means above 0.94, and the values for skewness and kurtosis for those four items were beyond +/- 3. However, it is entirely logical to expect high kurtosis values for these items because the mean values were so high.

Although some items did not fall within the accepted limits of the descriptive statistics described in the Method section, they did not appear to cause any substantial threats to normality when performing the reliability analyses, thus were kept for the subsequent analyses.

Internal Consistency Reliability

Reliability analysis was performed in order to examine the internal consistency reliability for the cloze section, as seen in Table 19. The results show that the first passage and the second passage yielded alphas of 0.60 and 0.50, respectively. The overall cloze section yielded an alpha of 0.70. Such a moderate value is surprising because higher reliability is normally expected for an exam like the ECPE with over 12,000 participants and 40 items. Compared to the reliability estimate in the GVR section ($\alpha = 0.87$), 0.70 seems fairly low, and suggests that the items in the cloze section do not appear to measure strongly a homogeneous construct. In order to answer the fourth research question (what is the underlying trait structure of foreign language test performance of English, as measured by the ECPE cloze section?), I performed exploratory factor analysis, which is discussed in the following section.

Table 19. Reliability Estimates for the Cloze Section

Section	Number of Items	Reliability Estimates
First Passage	20	0.60
Second Passage	20	0.50
Total	40	0.70

Exploratory Factor Analysis

To investigate the factorial structure of the cloze section, a matrix of tetrachoric correlations using all 40 items was generated in Mplus Version 2. Then, a series of EFAs was performed on the cloze section, which produced a two factor promax rotation that seemed to maximize parsimony and interpretability. Although it was initially hypothesized that the cloze items were measuring four factors, each representing one of the four components identified in the coding (RG, RV, GR, VR), there seemed to be no substantial EFA results to support the idea of four factors. In the course of these analyses, 15 items (CA41, CA47, CA52, CA53, CA55, CA57, CA58, CB60, CB65, CB67, CB68, CB70, CB71, CB74, and CB76) were dropped due to extremely low (lower than 0.3) factor loadings and double loadings (see Table 20). The same

Table 20. The Initial EFA Results of Cloze Section: Promax Rotation

Item	Code	F1	F2	Item	Code	F1	F2
C41	CF	0.241	0.118	C61	LM	0.117	0.255
C42	CM	-0.187	0.409	C62	CF	0.306	0.117
C43	LM	-0.076	0.411	C63	LM	0.024	0.348
C44	LM	0.087	0.248	C64	LM	-0.080	0.337
C45	LM	0.122	0.319	C65	LM	0.079	0.197
C46	LM	0.154	0.231	C66	LF	0.411	-0.254
C47	LM	0.162	0.105	C67	LM	0.135	0.148
C48	LM	-0.009	0.301	C68	MF	0.214	0.161
C49	LF	0.418	0.087	C69	LM	-0.093	0.441
C50	LM	0.109	0.306	C70	CF	0.159	0.061
C51	CF	0.343	0.190	C71	LM	0.204	0.128
C52	LM	0.209	0.121	C72	MF	0.541	-0.153
C53	CF	0.159	0.139	C73	LF	0.233	0.078
C54	CF	0.357	0.180	C74	LM	0.158	0.183
C55	LM	0.181	0.193	C75	CF	0.583	-0.084
C56	LM	0.012	0.476	C76	LM	0.018	0.035
C57	LM	0.062	0.130	C77	LM	-0.021	0.353
C58	MF	0.145	0.139	C78	MF	0.686	-0.213
C59	LM	0.184	0.336	C79	LM	-0.132	0.680
C60	MF	0.263	0.196	C80	LM	-0.065	0.396

LF = Lexical Form, MF = Morpho-Syntactic Form, CF = Cohesive Form,
LM = Lexical Mean, CM = Cohesive Mean

procedure as for the GVR analysis was used for dropping items that produced a factor loading lower than 0.30. The deletion of 15 items was due to the moderate value of reliability for the cloze section. If the items in the cloze section had been performing more homogeneously, the EFA would not have produced this many items with low factor loadings and double loadings.

Although deleting 15 items from the subsequent analyses seems excessive, it was necessary to remove the low and/or double loading items in order to avoid distracting from investigating the construct validity of the ECPE. Ultimately, 25 cloze items were used to measure the two underlying factors (Table 21). The items loading on the first factor appeared to be measuring grammatical forms, and those on the second factor appeared to be measuring grammatical meaning based on Purpura's theoretical model.

Table 21. EFA Results of the Shortened Cloze Section: Promax Rotation

Item	Code	F1 FORM	F2 MEAN
CB78	MF	0.662	-0.149
CB75	CF	0.569	-0.036
CB72	MF	0.521	-0.099
CA49	LF	0.387	0.131
CB66	LF	0.365	-0.205
CA54	CF	0.326	0.219
CA51	CF	0.309	0.229
CB62	CF	0.294	0.149
CB73	LF	0.243	0.098
CB79	LM	-0.104	0.664
CA56	LM	0.012	0.472
CB69	LM	-0.062	0.428
CA43	LM	-0.069	0.400
CB80	LM	-0.041	0.390
CA42	CM	-0.168	0.388
CA59	LM	0.143	0.353
CB63	LM	0.021	0.345
CB77	LM	-0.023	0.343
CA45	LM	0.114	0.333
CB64	LM	-0.073	0.323
CA50	LM	0.098	0.316
CA48	LM	-0.003	0.298
CB61	LM	0.109	0.265
CA44	LM	0.088	0.259
CA46	LM	0.137	0.246

LF = Lexical Form, MF = Morpho-Syntactic Form, CF = Cohesive Form,
LM = Lexical Mean, CM = Cohesive Mean

An inspection of the interfactor correlation matrix indicated that there is a moderate correlation between two factors: 0.55. This moderate correlation indicates that the FORM items and the MEAN items are not measuring the same construct, yet they are somewhat interdependent. Based on these analyses, 16 items were used to form the MEAN composite variables, whereas 9 items were used to form the FORM composite variables for the analyses.

Post EFA Reliability Analyses

Following the exploratory factor analysis, I performed a reliability analysis with the revised cloze section taxonomy containing two factors (FORM and MEAN). The analysis

yielded the alphas of 0.40 and 0.59 for FORM and MEAN items, respectively, (see Table 22). The Spearman-Brown formula was used in order to examine the corrected reliability (the test length fixed to 40 items), yielding values within acceptable limits of 0.50 or more (0.75 and 0.78 for FORM and MEAN, respectively). The corrected reliability estimate for the entire shortened cloze section increased to 0.74 from the original cloze section reliability estimate of 0.70.

Table 22. Reliability Estimates for the Revised Cloze Section: FORM and MEAN

	Number of Items	Items Kept	Reliability Estimates	Corrected Reliability Estimates
FORM	9	49, 51, 54, 62, 66, 72, 73, 75, 78	0.29	0.64
MEAN	16	42, 43, 44, 45, 46, 48, 50, 56, 59, 61, 63, 64, 69, 77, 79, 80	0.58	0.77
Total	25		0.58	0.70

Structural Equation Modeling (SEM)

Based on the results of the EFAs, the shortened foreign language cloze test performance of English appeared to be measured by two variables: FORM and MEAN. I attempted to perform SEM on these factors by treating FORM and MEAN as observed variables. However, the model is under-identified because there are only two observed variables in the one-factor model. An under-identified model has one or more parameters which may not be uniquely determined due to insufficient information in the matrix (Schumaker and Lomax, 1996). Therefore, such a model produces unreliable parameter estimates. In order to compensate for this limitation, item level SEM was performed for the EFA-generated factors. In other words, FORM and MEAN were considered as underlying factors while the items were treated as observed variables in this model. Model 2 addresses the following research question: what is the underlying trait structure of foreign language test performance of English measured by cloze?

Using Mplus Version 2, I evaluated the hypothesized model to determine to what extent the model fit the sample data for the item level analysis. Table 23 presents the summary of the model fit. The data produced the standardized root mean square residual of 0.041. The goodness of fit index for this model produced a chi-square value of 1862.276 with 255 degrees of freedom and a CFI of 0.90. Although a CFI of 0.95 and above is preferred, the RMSEA is acceptable (RMSEA = 0.022), indicating that this model is a good representation of the data.

Table 23. Results for EFA-Generated Cloze Section 2-Factor Model: Model 2

Goodness of fit summary:	
Comparative fit index (CFI)	0.90
The Tucker-Lewis index (TLI)	0.91
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.04
Root Mean Square Error of Approximation (RMSEA)	0.02
Chi-square test of model fit:	
Value	1862.276
Degrees of Freedom	255
P-Value	0.0000

Considering that there are 25 observed variables in the model, CFI = 0.90 seems to be an acceptable model fit for this data.

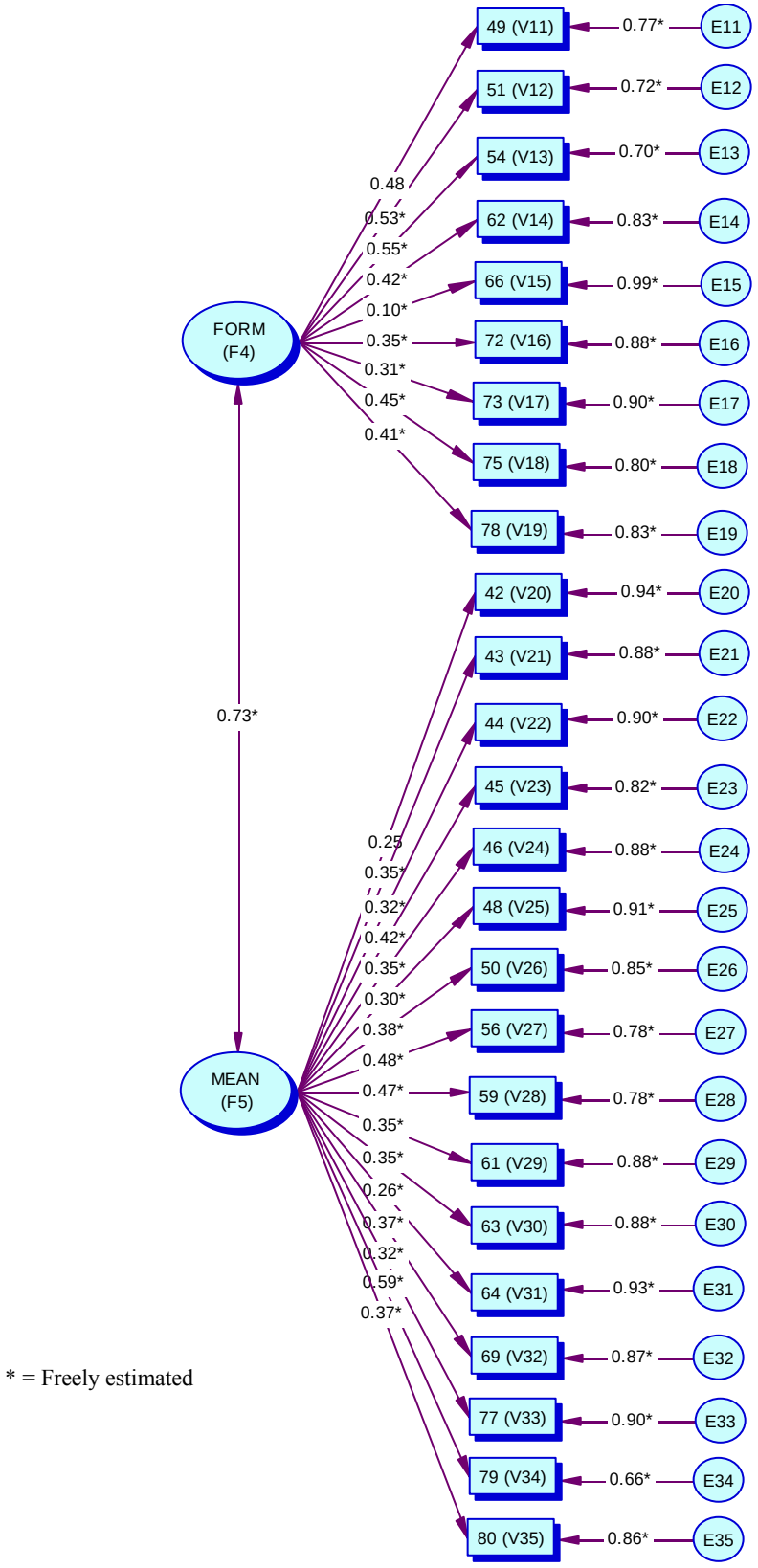
I then evaluated the feasibility of the individual parameter estimates and discovered all to be substantively reasonable and statistically significant at the 0.05 level. This indicates that the underlying factors are reasonably measured by the observed variables. In other words, the items in the shortened cloze section appear to measure two underlying factors: FORM and MEAN. The standardized solution in Table 24 shows that the factor loadings for Model 2 were somewhat moderate, ranging from a low of 0.10 to a high of 0.59. Model 2, along with the standardized parameter estimates, are presented in Figure 6.

Table 24. Parameter Estimates for Model 2

Standardized Solution:						
49	=	V11	=	0.48 F1	+	0.77 E11
51	=	V12	=	0.53 F1	+	0.72 E12
54	=	V13	=	0.55 F1	+	0.70 E13
62	=	V14	=	0.42 F1	+	0.17 E14
66	=	V15	=	0.10 F1	+	0.99 E15
72	=	V16	=	0.35 F1	+	0.88 E16
73	=	V17	=	0.31 F1	+	0.90 E17
75	=	V18	=	0.45 F1	+	0.80 E18
78	=	V19	=	0.41 F1	+	0.83 E19
42	=	V20	=	0.25 F2	+	0.94 E20
43	=	V21	=	0.35 F2	+	0.88 E21
44	=	V22	=	0.32 F2	+	0.90 E22
45	=	V23	=	0.42 F2	+	0.82 E23
46	=	V24	=	0.35 F2	+	0.88 E24
48	=	V25	=	0.30 F2	+	0.91 E25
50	=	V26	=	0.38 F2	+	0.85 E26
56	=	V27	=	0.48 F2	+	0.78 E27
59	=	V28	=	0.47 F2	+	0.78 E28
61	=	V29	=	0.35 F2	+	0.88 E29
63	=	V30	=	0.35 F2	+	0.88 E30
64	=	V31	=	0.26 F2	+	0.93 E31
69	=	V32	=	0.37 F2	+	0.87 E32
77	=	V33	=	0.32 F2	+	0.90 E33
79	=	V34	=	0.59 F2	+	0.66 E34
80	=	V35	=	0.37 F2	+	0.86 E35

To summarize, Model 2 provides sufficient evidence for the acceptance of the two-factor solution of foreign language cloze test performance of English as a reasonable explanation of the correlations among the observed variables. This model suggests that the selected items in the cloze section are measuring two underlying constructs: grammatical forms and grammatical meanings.

Figure 6. EFA-Generated 2-Factors Model of the Cloze Section: Model 2



* = Freely estimated

The GVR and the Cloze Sections

The primary purpose of this study was to determine how the cloze items relate to the various parts of the GVR section and to investigate whether the cloze section merits being a separate section of the ECPE battery. The first step in answering this research question was to examine the correlations among the cloze scores and the three parts in the GVR section. Table 25 presents the Pearson product-moment correlations of the total cloze score, the total GVR score, and the scores of the three parts (grammar, vocabulary, and reading) in the GVR section.

Table 25. Correlations of Cloze and GVR Scores

	GVR			
	Grammar	Vocabulary	Reading	Total GVR
Total Cloze	0.59*	0.49*	0.55*	0.66*

* significant at the 0.01 level (2-tailed)

The correlations show that the cloze items appear to measure more of the grammatical aspects of the language than the vocabulary and reading comprehension aspects; however, the differences in correlations for all three parts in the GVR were not substantial (0.49 to 0.59). The correlation between the total cloze score and the total GVR score was 0.66, indicating that there is a moderate degree of overlap in the processes measured by the cloze and the GVR. In other words, these two sections appear to measure a homogeneous construct to some extent but not so robustly.

To further investigate the relationship between the cloze and the GVR sections, the correlation between these two sections based on the SEM analyses was examined. According to the SEM results, both the abbreviated cloze and the grammar and vocabulary parts of the GVR measure FORM and MEAN. Thus, the extent to which the cloze form/mean items correlate with the GVR form/mean items was investigated. Table 26 presents the results of the correlational analysis.

Table 26. Correlations of the Cloze Form/Mean Items and the GVR Form/Mean Items

	GVR FORM	GVR MEAN
Cloze FORM	0.45*	0.07*
Cloze MEAN	0.34*	0.43*

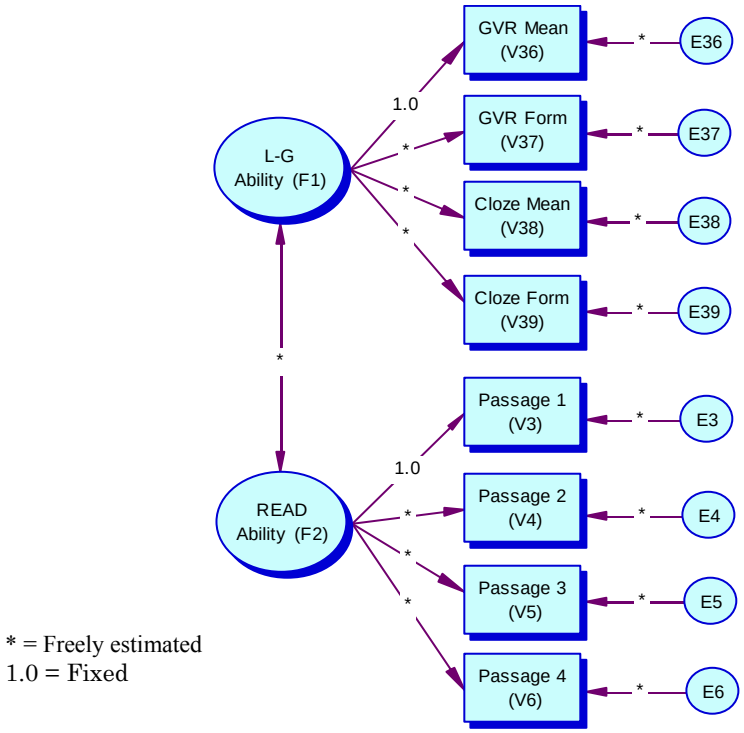
* significant at the 0.01 level (2-tailed)

The cloze FORM and the GVR FORM correlated the highest with a value of 0.45. The second highest correlation was between the cloze MEAN and the GVR MEAN, with the value of 0.43. Although the correlations are not particularly high, this analysis shows that the FORM items in the cloze and GVR sections correlate, and the MEAN items in the cloze and GVR sections correlate. The cloze MEAN and the GVR FORM showed a low correlation of 0.34, suggesting a weak relationship between the two item types. Both correlations between different item types across the cloze and GVR sections produced low correlations (0.07 and 0.34), indicating a weak relationship between the FORM and the MEAN items.

Following the correlation analyses, the overall SEM model for the cloze and GVR sections was examined in order to investigate the underlying construct of the combined sections. Based on the results of the SEMs, the following model (Figure 7) was initially hypothesized as the

overall model. It contains two intercorrelated factors (L-G = lexico-grammatical ability and READ = reading ability) with eight observed variables (GVR Mean, GVR Form, Cloze Form, Cloze Mean, Reading passage 1, passage 2, passage 3, and passage 4), and each observed variable is hypothesized to load on only one factor. Errors associated with each observed variable (E36 through E39 and E3 through E6) are assumed to be uncorrelated.

Figure 7. Initially Hypothesized Model of the Overall Cloze and the GVR Section: Model 3.1



Model 3.1 is a first-order confirmatory factor analysis designed to examine the multidimensionality of the foreign language test performance of English measured by the abbreviated cloze and GVR sections. Due to the exploratory nature of this study, I examined the relationships among the variables with the objective of generating the best fitting and most substantively meaningful model, rather than simply confirming or rejecting this particular model.

Prior to exploring the trait structure of this model, the univariate and multivariate statistical assumptions underlying the maximum likelihood estimation procedure were examined. The univariate values for skewness and kurtosis were satisfactorily normally distributed. Then, the trait structure of the hypothesized model was examined to investigate the extent to which the model fit the sample data. With regard to model adequacy as a total, the data produced a root mean square error of approximation (RMSEA) of 0.10, indicating a degree of global misfit (see Table 27). Furthermore, the goodness of fit index for the initially hypothesized two-factor model produced a chi-square value of 2188.00 with 19 degrees of freedom ($p<0.0000$). This again suggests a weakly fitting model. Along with the RMSEA and the chi-square value, the comparative fit index (CFI) of 0.88 confirms that this model does not provide compelling

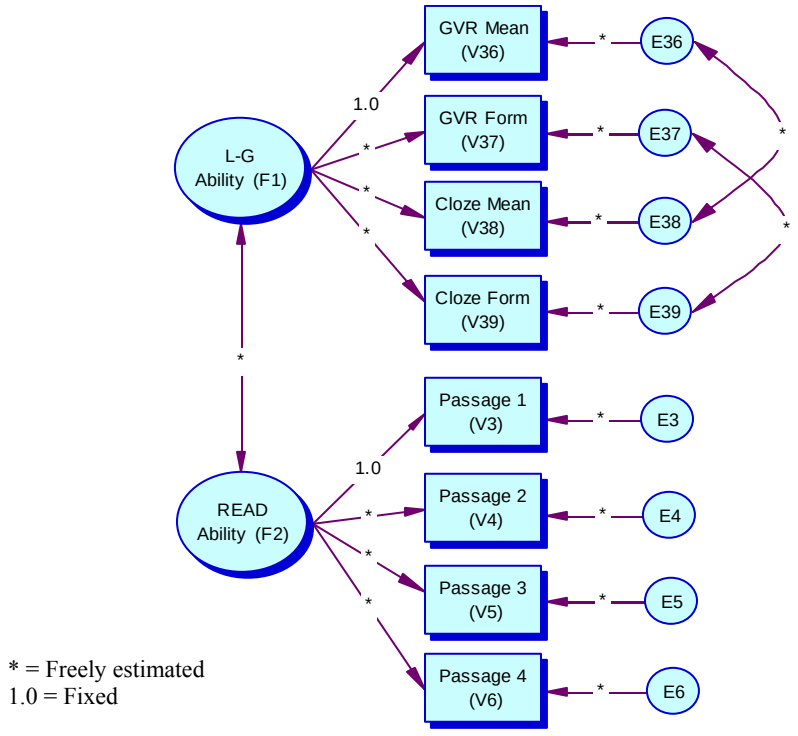
evidence for acceptance. Although the individual parameters of this model were evaluated, I did not interpret these parameters due to inadequate fitting of the overall model.

Table 27. Results for Initially Hypothesized Overall Cloze/GVR Section Model: Model 3.1

Goodness of fit summary:	
Comparative fit index (CFI)	0.88
The Tucker-Lewis index (TLI)	0.89
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.05
Root Mean Square Error of Approximation (RMSEA)	0.10
Chi-square test of model fit:	
Value	2188.00
Degrees of Freedom	19
P-Value	0.0000

Based on the results of Mode 3.1, I performed a series of post hoc fitting procedures in order to discover a better fitting model. My primary concern was, if the FORM items in the cloze and GVR sections are measuring the same trait, to what extent are they correlated? The same inquiry was raised for the MEAN items. Therefore, in the revised model, the error terms associated with Cloze FORM-GVR FORM and Cloze MEAN-GVR MEAN are hypothesized to be correlated. Model 3.2, presented in Figure 8, was built based on both a substantive and statistical point of view.

Figure 8. The Revised Model of the Overall Cloze and the GVR Section: Model 3.2



All eight variables produced satisfactory skewness and kurtosis values based on the sample statistics; thus, further analysis proceeded. With respect to goodness of fit, Model 3.2 produced a root mean square error of approximation (RMSEA) of 0.05, indicating an insignificant degree of misfit (see Table 28). It also produced a chi-square statistic of 751.721 with 17 degrees of freedom, representing a drastic decrease in overall chi-square ($\Delta\chi^2(2) = 1436.279$) from the initially hypothesized model. This reduction in χ^2 exhibited a substantial improvement in goodness of fit. Along with the chi-square, the CFI (0.96) also reflected an extensive improvement in model-data fit ($\Delta = 0.08$).

Table 28. Results for the Overall Cloze and the GVR Section: Model 3.2

Goodness of fit summary:	
Comparative fit index (CFI)	0.96
The Tucker-Lewis index (TLI)	0.97
Standardized residual matrix:	
Standardized Root Mean Square Residual (SRMR)	0.03
Root Mean Square Error of Approximation (RMSEA)	0.05
Chi-square test of model fit:	
Value	751.721
Degrees of Freedom	17
P-Value	0.0000

These statistics provide strong evidence for acceptance of Model 3.2. As seen in Table 29, the loadings in the standardized solution ranged from a low 0.31 for GVR Mean to a moderately high 0.62 for Passage 4 in the reading section. Nonetheless, all factor loadings were found to be statistically significant.

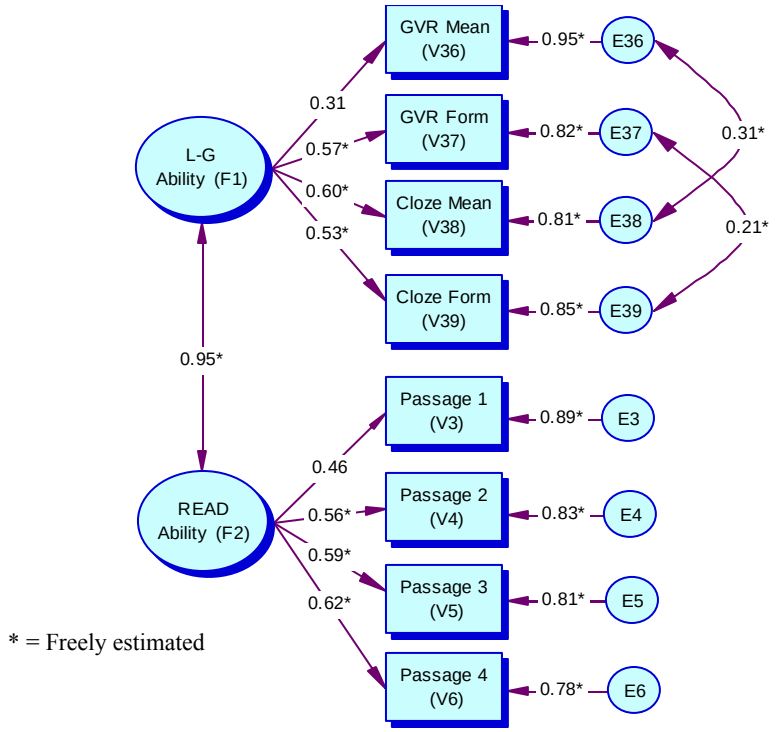
Table 29. Parameter Estimates for Model 3.2

Standardized Solution:						
GVR Mean	=	V36	=	0.31 F1	+	0.95 E36
GVR Form	=	V37	=	0.57 F1	+	0.82 E37
Cloze Mean	=	V38	=	0.60 F1	+	0.81 E38
Cloze Form	=	V39	=	0.53 F1	+	0.85 E39
Passage 1	=	V3	=	0.46 F1	+	0.89 E3
Passage 2	=	V4	=	0.56 F1	+	0.83 E4
Passage 3	=	V5	=	0.59 F1	+	0.81 E5
Passage 4	=	V6	=	0.62 F1	+	0.78 E6

Figure 9 provides a diagrammatic representation of Model 3.2, in which the standardized parameter estimates are indicated. An inspection of Model 3.2 illustrates that the ECPE foreign language test performance of English for the selected items is represented by two highly related underlying factors measured by eight observed variables. The high (0.95) interfactor correlation suggests that lexico-grammatical ability and reading ability are closely related. The most significant finding in this model is that the two error terms are significantly

related to each other. Although the correlations of 0.21 and 0.31 are not high, this indicates that there is some redundant content being measured across the cloze and GVR sections.

Figure 9. Results for the Overall Cloze and the GVR Section with Standardized Parameter Estimates: Model 3.2



In summary, Model 3.2 provides a reasonable explanation of the underlying construct of the shortened ECPE cloze and GVR sections. With its two intercorrelated factors, eight measured variables and two correlated errors, this model generally supports the hypothesis that cloze is measuring form and meaning, thereby supporting the notion that cloze does not measure processing abilities beyond the clause level (Alderson, 1979; Shanahan et al., 1982; Markham, 1985). Factor 1, lexico-grammatical ability, is represented by items assessing forms and meanings in both the cloze and the GVR sections. Factor 2, reading ability, is represented by items assessing reading in four different passages in the GVR section.

Discussion

The present study investigated six research questions concerning the underlying trait structure of the ECPE cloze and GVR sections. The first research question investigated the extent to which the items in each component (grammar, vocabulary, and reading) in the GVR section performed as a homogeneous group. The reliability analysis indicated that the grammar items were the most homogeneous among the three components, with a Cronbach's alpha of 0.80. However, when the number of items for all three components was held constant to 40 items, the reading items had the highest internal consistency, with an alpha of 0.84. The internal consistency reliability range of 0.72 to 0.84 suggests that the items reasonably measure the

same construct within each component. The internal consistency reliability of the overall GVR section produced a high alpha of 0.87, indicating that the items in the GVR section appear to measure reliably second language GVR test performance of English. The information provided by the reliability analysis proved valuable in determining whether or not to proceed with the exploratory and the confirmatory factor analyses, which were designed to analyze the hypothesized underlying structures of the GVR and cloze sections, as well as to analyze whether or not the composite variables were measuring the language ability they were designed to measure.

The second research question examined the underlying trait structure of foreign language test performance of English measured by the GVR section. In order to answer the question, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used. The EFA indicated that the selected grammar items measure morphosyntactic form and lexical form, while the selected vocabulary items measure lexical form and lexical meaning. In other words, lexical form is measured in both grammar and vocabulary sections, whereas lexical meaning is only measured in the shortened vocabulary section, and morphosyntactic form is measured only in the shortened grammar section. This is not a surprising result, since lexical form and lexical meaning are closely associated by definition (Purpura, forthcoming).

The CFA proved valuable in confirming the hypothesized two-factor underlying trait structure, indicating moderate loadings between observed variables and their hypothesized factors. This confirmed that test takers' performance on the abbreviated GVR section was explained by two hypothesized factors (lexico-grammatical ability and reading ability), and that lexico-grammatical ability is measured by two observed variables: form and meaning.

Although the reading items were expected to load on the explicit and implicit reading factors, they loaded according to passages. This may suggest that reading items in this test are text dependent rather than item type dependent. There is a need for further investigation of how the choice of reading passages affects the measure of reading ability.

The interfactor correlation between lexico-grammatical ability and reading ability was extremely high ($r = 0.97$). This suggests that these two factors are not purely independent. Rather, lexico-grammatical ability and the variables that measure the lexico-grammatical ability appear to be closely related to reading ability and vice versa.

The third and fourth research questions addressed the underlying trait structure of the cloze section. With regard to the third question, the internal consistency reliability of the cloze items was investigated, resulting in moderate estimates of 0.57 and 0.45, which indicates that items in each passage did not perform in a homogeneous way. The internal consistency reliability estimate of the overall cloze section was 0.65, suggesting a weak homogeneity of the cloze items. Compared to the reliability for the grammar and the vocabulary parts in the GVR section, the cloze reliability seems rather low. This may suggest that cloze items are not measuring a single underlying construct. As previously discussed, past studies have indicated a wide range of reliability estimates (0.31 to 0.96). Considering this wide range, the reliability of 0.65 seems reasonable. However, considering the ECPE is a high-stakes exam with over 12,000 subjects and a reasonable number of items, 0.65 seems rather low. In order to understand the reason for the low reliability, the cloze items and their distractors should be studied further.

The fourth research question examined the underlying trait structure of foreign language test performance of English measured by the cloze section, using EFA followed by CFA. According to the *Michigan Certificate Examinations General Information Bulletin*, "the cloze section is intended to assess the understanding of the organizational features of written text as well as grammatical knowledge and pragmatic knowledge of English, particularly knowledge

about expected vocabulary in certain contexts" (English Language Institute, The University of Michigan, 2002, p. 8). In other words, the cloze section is intended to assess higher-order processing abilities. However, based on the rigorous investigation using SEM in the present study, the cloze test items selected appeared to be accounted for by two factors: form and meaning. The findings indicate that the abbreviated cloze section appears to measure only lower processing skills and not the skill of comprehending organizational features of written text.

The fifth research question, examining the relationship between the cloze and GVR sections, was addressed by composing a model that would fit both statistically and substantively. As a result, the final model identified two underlying factors, each with four observed variables and two correlated errors. The reading ability factor is represented by four reading passage variables in the GVR section. The lexico-grammatical ability factor on the other hand is represented by two FORM variables and two MEAN variables measured in the cloze and GVR sections. The model indicates moderate relationships between variables and their respective hypothesized factors, as well as a high interfactor correlation, indicating that these abilities are inextricably related. The correlated errors provide evidence of some redundancy in the content being measured across the cloze and GVR sections. However, low error values (0.21 and 0.31) suggest that form and meaning measured in the cloze and GVR sections are different. This finding leads to the last research question, which asks whether the cloze section merits being a separate section of the ECPE battery.

In light of the above observations, the answer to the sixth research question is that the GVR and the cloze items may be integrated into one section. Although the correlation of the FORM/MEAN items was low between the cloze and the GVR sections, many of the cloze items and the items in the grammar and the vocabulary sections were essentially measuring form and meaning. Given this observation, it may be unnecessary to expect the test-takers to pass both the cloze and GVR sections along with the remaining three sections of the ECPE to obtain the certificate of proficiency.

The pre- and post-EFA reliability analyses showed that reliability stayed constant after 37 items in the GVR were deleted (observing the corrected reliability based on the Spearman-Brown formula). The deletion of 15 cloze section items also resulted in an unchanged reliability estimate (0.70). Therefore, if the cloze and the GVR sections were to be combined, test developers could reduce the number of items without substantially decreasing the reliability of the section.

Conclusion

This study investigated the underlying construct of the 1997 cloze section in the ECPE, which was developed by the English Language Institute, The University of Michigan. The question to be answered was, if the cloze section measures the same construct as the GVR section, then why have the cloze section be a distinct section of the ECPE battery? In order to answer this question, I attempted to identify the underlying construct of the cloze section and compare it with the trait structure of the GVR section.

Through a rigorous investigation using structural equation modeling, I determined that the cloze section appears to measure grammar forms and meaning rather than overall language proficiency. When the cloze and the GVR items were included in the same model, the cloze again measured forms and meaning, along with the GV items.

Although this study provides beneficial information regarding the underlying trait structure of the ECPE cloze and GVR sections, it is imperative to recognize the study's limitations. Though the underlying structure using Purpura's model may appear to be a plausible representative of second language test performance of English measured by the cloze and GVR sections, the results should not be generalized. There may be other models that would be a better representative of the ECPE cloze and GVR sections. In order to create an accurate and full representation of the cloze and GVR sections, this study should be replicated and the results should be confirmed by other studies with ECPE tests from different years. Furthermore, different models need to be rigorously tested for a better model fit.

In addition, it is important to recognize that distractor efficiency was not evaluated prior to performing any statistical procedures. This may have affected the credibility of the subsequent statistical analyses. When writing distractors for multiple-choice items, one should be careful to be accurately assessing the intended test-taker abilities. One of the guidelines to follow in writing grammar distractors is that all choices should belong to the same grammatical category. For instance, the distractors should not include prepositions when the correct answer is a conjunction. There were some questions that violated this guideline. These questions may have caused a decrease in reliability of the cloze section. Because the initial cloze test reliability produced a moderate value of 0.65, there is a possibility that the cloze test was not strongly measuring a homogeneous construct from the beginning. This may have affected the results of EFA and SEM analyses. Revisions in distractors may increase the reliability and produce a more accurate representation of the underlying construct of the cloze section.

It is also imperative to acknowledge that there are various other research questions to be asked and answered, especially concerning the ethnicity, age, and gender of the subjects. Further research could analyze these variables to develop more detailed descriptions or explanations of second language test performance for these populations. A reliability analysis with a consequent confirmatory factor analysis could again be used to determine the underlying construct and strength of relationships among the language ability variables for these groups. This analysis may present different factor loadings or even a disparate underlying trait structure.

From a methodological point of view, this study has demonstrated the significance of using various statistical procedures, especially structural equation modeling. SEM has presented evidence that it can be a powerful research tool for investigating the underlying construct of latent factors and for providing insights into the interrelationships among the latent factors as well as the observed variables.

Despite the limitations, the findings of this study have contributed to a deeper understanding of the construct validity of cloze items, which has been debated for many decades. According to Oller and Jonz (1994), the cloze procedure contributes to the understanding of the "basic theoretical questions about human mental abilities as well as urgent practical questions about designing curricula. We are convinced that answers to such fundamental questions about meaningfulness will have countless invaluable applications" (p. 12). It is hoped that this study will encourage ESL test administrators to apply these findings to improve the validity of cloze tests.

References

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, *76*(4), 468-479.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, *13*(2), 219-227.

Alderson, J. C. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning*, *30*(1), 59-76.

Alderson, J. C. (1983). The cloze procedure and proficiency in English as a foreign language. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 219-228). Rowley, MA: Newbury House.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, *2*(2), 192-204.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*(1), 61-70.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535-556.

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

Bensoussan, M., & Ramraz, R. (1984). Testing EFL reading comprehension using a multiple-choice rational cloze. *Modern Language Journal*, *68*(3), 230-239.

Black, J. H. (1993). Learning and reception strategy use and the cloze procedure. *The Canadian Modern Language Review*, *49*(3), 418-445.

Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. A. Orem, & B. P. Taylor (Eds.), *On TESOL '83. The question of control* (pp. 109-119). Washington, DC: Teachers of English to Speakers of Other Languages.

Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing*, *5*(1), 19-31.

Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The emperor's new cloze: Strategies for revising cloze tests. In T. Hudson & J. D. Brown (Eds.), *A focus on language testing development*. (Technical report No. 21). Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

Celce-Murcia, M., & Larsen-Freemen, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). Boston, MA: Heinle & Heinle.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J.W. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning*, *35*(2), 181-206.

Chihara, T., Oller, J. W., Weaver, K. A., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning*, *27*(1), 63-73.

English Language Institute, The University of Michigan (2002). *Michigan Certificate Examinations general information bulletin 2001-2002*. Ann Arbor, MI: English Language Institute, The University of Michigan.

Farhady, H., & Keramati, M. N. (1996). A text-driven method for the deletion procedure in cloze passages. *Language Testing*, *13*(2), 191-207.

Fotos, S. S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning*, *41*(3), 313-336.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1988). *Multiple-choice cloze items and the test of English as a foreign language* (Research Reports No. 26). Princeton, NJ: Educational Testing Service.

Hanania, E., & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: Standardized ESL, cloze, and writing. *TESOL Quarterly*, *20*(1), 97-109.
Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle & Heinle.
Henning, G. (1987). *A guide to language testing*. Boston, MA: Heinle & Heinle.
Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 121-128). Rowley, MA: Newbury House.
Irvine, P, Atai, P., & Oller, J. W. (1974). Cloze, dictation, and the test of English as a foreign language. *Language Learning*, *24* (2), 245-252.
Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly*, *24* (1), 61-83.
Jonz J., & Oller, J. W. (1994). A critical appraisal of related cloze research. In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 371-407). London: Bucknell University Press.
Kim, J. O., & Muller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it.* Newbury Park, CA: Sage University Press.
Klein-Braley, C. (1983). A cloze is a cloze is a question. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 218-228). Rowley, MA: Newbury House.
Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural equation modeling approach*. Cambridge, UK: Cambridge University Press.
Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing*, *15*(3), 295-332.
Laesch, K. B., & van Kleeck, A. (1987). The cloze test as an alternative measure of language proficiency of children considered for exit from bilingual education programs. *Language Learning*, *37*(2), 171-189.
Markham, P. L. (1985). The rational deletion cloze and global comprehension in German. *Language Learning*, *35*(3), 423-430.
Mullen, K. (1979). More on cloze tests as of proficiency in English as a second language. In E. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 21-32). Washington DC: Teachers of English to Speakers of Other Languages.
Oller, J. W. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal*, *56*(3), 151-158.
Oller, J. W. (1973). Cloze tests of language proficiency and what they measure. *Language Learning*, *23*(1), 105-118.
Oller, J. W. (1979). *Language tests at school*. London: Longman Group Limited.
Oller, J. W. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 3-10). Rowley, MA: Newbury House.
Oller, J. W., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning*, *21*(2), 183-194.
Oller, J. W., & Inal, N. (1971). A cloze test of English prepositions. *TESOL Quarterly*, *5*(4), 315-326.
Oller, J. W., & Jonz J. (1994). Why cloze procedure? In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 1-20). London: Bucknell University Press.
Porter, D. (1983). The effects of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63-74). London: Academic Press.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach.* Cambridge, UK: Cambridge University Press.
Purpura, J. E. (Forthcoming). *Assessing grammar*. Cambridge, UK: Cambridge University Press.
Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude and intelligence: A structural equation modeling approach. *Language Learning*, *43*(3), 313-344.
Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing*, *17*(1), 85-114.
Shumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.
Sciarone, A. G., & Schoorl, J. J. (1989). The cloze test: Or why small isn't always beautiful. *Language Learning*, *39*(3), 415-438.
Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*(2), 229-255.
Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing*, *14*(2), 214-231.
Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal*, *58*(5-6), 239-241.
Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 414-438.
Turner, C. E. (1989). The underlying factor structure of L2 cloze test performance in Francophone, university-level students: Casual modeling as an approach to construct validation. *Language Testing*, *6*(2), 172-197.

Appendix A

GVR Section - Descriptive Statistics

Item	Mean	Std. Dev	Kurtosis	Skewness
G81	0.84	0.36	1.55	-1.89
G82	0.84	0.37	1.37	-1.83
G83	0.52	0.50	-1.99	-0.09
G84	0.86	0.35	2.33	-2.08
G85	0.70	0.46	-1.21	-0.89
G86	0.76	0.43	-0.54	-1.21
G87	0.77	0.42	-0.29	-1.31
G88	0.36	0.48	-1.65	0.59
G89	0.93	0.26	8.95	-3.31
G90	0.93	0.26	8.86	-3.30
G91	0.84	0.36	1.63	-1.91
G92	0.81	0.39	0.63	-1.62
G93	0.79	0.41	0.04	-1.43
G94	0.30	0.46	-1.19	0.90
G95	0.36	0.48	-1.66	0.58
G96	0.92	0.27	7.85	-3.14
G97	0.91	0.29	6.05	-2.84
G98	0.91	0.29	5.66	-2.77
G99	0.79	0.41	-0.01	-1.41
G100	0.96	0.20	19.74	-4.66
G101	0.76	0.43	-0.58	-1.20
G102	0.95	0.21	16.93	-4.35
G103	0.86	0.35	2.31	-2.08
G104	0.83	0.38	1.07	-1.75
G105	0.62	0.49	-1.78	-0.47
G106	0.86	0.35	2.34	-2.08
G107	0.36	0.48	-1.64	0.60
G108	0.72	0.45	-1.08	-0.96
G109	0.87	0.33	3.04	-2.25
G110	0.80	0.40	0.31	-1.52
G111	0.80	0.40	0.21	-1.49
G112	0.93	0.26	8.65	-3.26
G113	0.31	0.46	-1.35	0.81
G114	0.41	0.49	-1.88	0.35
G115	0.87	0.34	2.73	-2.17
G116	0.69	0.46	-1.35	-0.81
G117	0.99	0.11	75.99	-8.83
G118	0.51	0.50	-2.00	-0.04
G119	0.72	0.45	-1.09	-0.96
G120	0.60	0.49	-1.85	-0.39

GVR Section - Descriptive Statistics cont.

Item	Mean	Std. Dev	Kurtosis	Skewness
V121	0.77	0.42	-0.43	-1.25
V122	0.50	0.50	-2.00	-0.01
V123	0.85	0.36	1.93	-1.98
V124	0.26	0.44	-0.84	1.08
V125	0.53	0.50	-1.98	-0.14
V126	0.49	0.50	-2.00	0.03
V127	0.48	0.50	-1.99	0.09
V128	0.38	0.48	-1.75	0.50
V129	0.77	0.42	-0.31	-1.30
V130	0.68	0.47	-1.43	-0.76
V131	0.31	0.46	-1.35	0.80
V132	0.35	0.48	-1.62	0.61
V133	0.70	0.46	-1.27	-0.85
V134	0.41	0.49	-1.88	0.35
V135	0.80	0.40	0.18	-1.48
V136	0.20	0.40	0.28	1.51
V137	0.75	0.43	-0.61	-1.18
V138	0.73	0.45	-0.98	-1.01
V139	0.38	0.49	-1.76	0.49
V140	0.39	0.49	-1.81	0.44
V141	0.53	0.50	-1.98	-0.13
V142	0.61	0.49	-1.81	-0.43
V143	0.84	0.37	1.28	-1.81
V144	0.27	0.44	-0.94	1.03
V145	0.42	0.50	-1.89	0.33
V146	0.65	0.48	-1.62	-0.62
V147	0.49	0.50	-2.00	0.06
V148	0.38	0.48	-1.74	0.51
V149	0.42	0.49	-1.89	0.33
V150	0.59	0.49	-1.86	-0.38
V151	0.26	0.44	-0.75	1.12
V152	0.41	0.49	-1.88	0.35
V153	0.79	0.40	0.13	-1.46
V154	0.63	0.48	-1.70	-0.55
V155	0.27	0.44	-0.87	1.06
V156	0.75	0.43	-0.63	-1.17
V157	0.55	0.50	-1.95	-0.21
V158	0.84	0.37	1.41	-1.85
V159	0.81	0.40	0.37	-1.54
V160	0.59	0.50	-1.87	-0.36

GVR Section - Descriptive Statistics cont.

Item	Mean	Std. Dev	Kurtosis	Skewness
R161	0.93	0.26	9.31	-3.36
R162	0.86	0.34	2.54	-2.13
R163	0.81	0.39	0.54	-1.59
R164	0.90	0.31	4.70	-2.59
R165	0.98	0.15	39.40	-6.43
R166	0.76	0.43	-0.47	-1.24
R167	0.75	0.43	-0.69	-1.15
R168	0.83	0.38	0.95	-1.72
R169	0.67	0.47	-1.48	-0.72
R170	0.78	0.41	-0.10	-1.38
R171	0.84	0.36	1.57	-1.89
R172	0.82	0.39	0.67	-1.64
R173	0.68	0.47	-1.42	-0.76
R174	0.77	0.42	-0.36	-1.28
R175	0.86	0.35	2.33	-2.08
R176	0.50	0.50	-2.00	0.00
R177	0.57	0.50	-1.93	-0.27
R178	0.76	0.43	-0.51	-1.22
R179	0.63	0.48	-1.72	-0.53
R180	0.58	0.49	-1.89	-0.33

Appendix B

Cloze Section - Descriptive Statistics

Item	Mean	Std. Dev	Kurtosis	Skewness
CA41	0.87	0.34	2.64	-2.15
CA42	0.51	0.50	-2.00	-0.03
CA43	0.68	0.47	-1.41	-0.77
CA44	0.48	0.50	-1.99	0.09
CA45	0.40	0.49	-1.85	0.39
CA46	0.76	0.43	-0.57	-1.20
CA47	0.69	0.46	-1.32	-0.82
CA48	0.82	0.39	0.69	-1.64
CA49	0.87	0.34	2.64	-2.15
CA50	0.51	0.50	-2.00	-0.03
CA51	0.74	0.44	-0.85	-1.07
CA52	0.80	0.40	0.28	-1.51
CA53	0.58	0.49	-1.90	-0.33
CA54	0.62	0.48	-1.74	-0.51
CA55	0.74	0.44	-0.78	-1.11
CA56	0.70	0.46	-1.21	-0.89
CA57	0.79	0.40	0.11	-1.46
CA58	0.86	0.35	2.21	-2.05
CA59	0.57	0.50	-1.93	-0.26
CA60	0.70	0.46	-1.21	-0.89
CB61	0.83	0.38	0.93	-1.71
CB62	0.90	0.30	4.95	-2.64
CB63	0.46	0.50	-1.98	0.16
CB64	0.35	0.48	-1.59	0.64
CB65	0.68	0.47	-1.43	-0.76
CB66	0.56	0.50	-1.95	-0.23
CB67	0.31	0.46	-1.36	0.80
CB68	0.61	0.49	-1.81	-0.43
CB69	0.81	0.40	0.39	-1.55
CB70	0.64	0.48	-1.64	-0.60
CB71	0.52	0.50	-1.99	-0.10
CB72	0.95	0.21	16.55	-4.31
CB73	0.77	0.42	-0.37	-1.28
CB74	0.74	0.44	-0.81	-1.09
CB75	0.97	0.16	31.28	-5.77
CB76	0.63	0.48	-1.73	-0.52
CB77	0.35	0.48	-1.60	0.64
CB78	0.94	0.25	10.47	-3.53
CB79	0.44	0.50	-1.94	0.24
CB80	0.51	0.50	-2.00	-0.06