研究論文

The Trait Structure of the Edinburgh Project on Extensive Reading Placement/Progress Test

EPERプレイスメント・プログレステストの特性構造

Kiyomi Yoshizawa　Atsuko Takase　Kyoko Otsuki
吉澤清美　髙瀬敦子　大槻きょう子

多読は近年第2言語・外国語教育の一つのアプローチとして注目されてきており、多読が言語能力の発達に寄与しているという実証研究も国内外で多く報告されている。エジンバラ大学多読プロジェクトチームによって開発されたEPERプレイスメント・プログレステストは短い物語文にほぼ任意に空白が設けられたクローズテストであり、学習者が読むのに適した本のレベルを決める、読みの発達を見極めるために多読実践者によって使われている。しかしながら、その特性構造については研究が皆無である。本研究はBachman（1982/1994）に基づき、EPERテストの項目を3つのタイプに分類し、5つの特性モデルを提唱し、データ分析を行った。Bifactorモデルがデータへの適合が最もよかった。

Keywords:
trait structure, extensive reading, L2 reading, Edinburgh Project on Extensive Reading Placement/Progress Test（EPER PPT）

キーワード
特性構造、多読、L2リーディング、多読プレイスメント・プログレステストエジンバラプロジェクト（EPER PPT）

1. Introduction

1.1 Extensive Reading and L2 Teaching

Extensive Reading (henceforth, ER) is defined as the reading of a large quantity of comprehensible materials “in order to gain a general understanding of what is read” (Richards & Schmidt, 2011, p. 212). Through ER, learners develop good reading habits, build knowledge of vocabulary and structure, and develop a liking for reading (Richards & Schmidt, 2011). Over the past 30 years, a great number of studies have reported that ER contributes to various aspects of learners' English proficiency, including reading, writing, vocabulary, reading fluency, and even listening and speaking (e.g., Beglar et al., 2011; Cho & Krashen, 1994; Elley & Mangubhai, 1981; Grabe, 2004; Hafiz & Tudor, 1989; Huffman, 2014; Irvine, 2011; Kobayashi et al., 2010; Mason & Krashen, 1997; Mermelstein, 2015; Nation & Waring, 2020; Robb & Susser, 1989; Suzuki, 1996; Takase 2004, 2008, 2009, 2010; Yamashita, 2004). Since ER has started to be acknowledged as one of the most effective methods to improve learners' reading fluency and English proficiency, a number of teachers have implemented or have been thinking of implementing ER in their English curricula in Japan.

To date, the effectiveness of ER on learners' English proficiency has been shown by practitioners at different age groups who have used various formal tests: Assessment of Communicative English (ACE) for junior and senior high school students (Furukawa, 2008); Edinburgh Project on Extensive Reading (EPER, 1992) Placement/Progress Test (EPER PPT) for university students (Akao, 2015; Takase, 2008; 2009; Takase & Otsuki, 2012; Yamashita, 2008); Global Test of English Communication (GTEC) for high school students (Sakamoto, 2014; Watanabe, 2014); Secondary Level English Proficiency (SLEP) for high school students (Takase, 2007); Test of English for International Communication (TOEIC) for university students and adults (Nishizawa et al., 2010); and Test of Practical English Proficiency (EIKEN) for junior high school students (Takase, 2010). Among these tests, the EPER PPT is the only test that indicates to learners and instructors which level of books to start reading extensively and when to raise the level of books based on the learners' progress for an effective and successful ER practice in class (Day & Bamford, 1998).

1.2 EPER Placement/Progress Tests

The EPER PPT has often been used by many ER practitioners on Japanese university students. The EPER PPT is a cloze test developed by the Edinburgh Project on Extensive Reading team at Institute for Applied Language Studies in University of Edinburgh. The purpose of this test is to place learners in appropriate reading levels to start reading English books comfortably without depending on a dictionary or translation. If you administer the EPER PPT several times in an ER program, you can also monitor learners' progress in reading as they engage themselves in ER. Originally several forms of EPER PPT were created, but at present Forms A and E are most often used. These are random-deletion cloze tests where deletions are not at evenly-spaced intervals. Cloze passages were taken from different levels of obsolete graded readers.

As such, EPER PPT is considered as most appropriate to assess learners' reading levels at the onset of an ER program and to assess their progress in the program. However, it has been pointed out that EPER PPT has some drawbacks: “The primary one is that it is a modified cloze test and as such may not indicate a learner’s fluent reading level but is open to guessing and involves writing” (Lemmer et al., 2012, p. 23). Little research exists concerning its trait structure or its construct. If the inference about test-takers’ abilities is made based on the performance of EPER PPT, its users (i.e., ER teachers and administrators) need to know what trait(s) EPER PPT measures. The present study aims to examine the trait structure of EPER PPT Form A.

In language testing, studies on the trait structure of cloze tests showed conflicting results in the past. Some see that test-takers use discourse processing ability while engaged in cloze tasks and cloze can measure overall language proficiency (Oller, 1979). Hinofotis’ (1980) study supported that cloze testing could offer a “viable alternative procedure for placement and proficiency testing” (p. 121). Chihara et al. (1977/1994) concluded that cloze tests measure the sensitivity to discourse constraints across sentences. On the other hand, Alderson (1979, 2000) concluded that cloze tests measure only lower-level skills. Reviewing the literature on the construct validity of cloze tests, Chapelle and Abraham (1990) state that the fixed-ratio (i.e., random deletion) cloze can measure written grammatical competence in some cases and textual competence in others and that the inconclusive findings are due to the fact that random deletion was used.

Several research studies took qualitatively approaches to this controversial issue concerning what cloze procedures measure. For example, Storey (1997) created a discourse cloze to examine the processes EFL learners undergo for successful completion of each deletion. He used a think-aloud approach. One of the common phenomena which emerged from the analyses of the introspective protocols was that the participants were “utilizing a number of information sources” for successful completion of deleted expressions (p. 222).

Bachman (1982/1994) analyzed the trait structure of a cloze test using confirmatory factor analysis. Rational deletions were made. There are three types of deletions: syntactic, cohesive, and strategic. To respond syntactic items correctly, test-takers were supposed to depend on clause-level context. Similarly, they were assumed to depend on “the interclausal or intersentential cohesive context” (p. 63) to respond cohesive items correctly. Finally, test-takers depended on the information on “parallel patterns of coherence” (p. 63) to respond strategic items correctly. Bachman hypothesized three models of theoretical interest which would underlie the cloze test scores. The first model named “general trait model” (p. 64) posits a single general factor, hypothesizing it would account for the most of the variances in the cloze test scores.

This model represents the “indivisibility hypothesis” (Oller, 1979, p. 425). A second model named “specific trait model” (p. 64) posits three independent factors: syntactic, cohesive, and strategic abilities. This model represents the “completely divisible competence hypothesis” (Bachman, 1982, p. 64). A third model named “general plus specific trait model” (p. 64) posits a general factor with three specific trait factors (i.e., syntactic, cohesive, and strategic). The results showed that the first two models did not fit the data; however, the third model provided the best explanation. Based on the findings, Bachman suggested that a rational-deletion cloze test could measure textual relationships both within and “beyond clause boundaries” (p. 66). He further stated that the debate over what abilities random-deletion cloze passages measure could be addressed by identifying deletion types of random-deletion cloze passages and analyzing response patterns using factor analytic procedures.

1.3 Research Question

The present study aims to examine the trait structure of EPER PPT Form A: What is the trait structure of EPER PPT Form A? EPER PPT is a random-deletion cloze test. Bachman (1982/1994) is the only study which applied a confirmatory factor analysis to examine the trait structure of a cloze test. Bachman used a rational-deletion method and created three types of deletions. We adopt his approach to analyze the trait structure of EPER PPT Form A.

2. Method

2.1 Participants

A total of 442 first- and second-year Japanese university students participated in the study. They were all learning English as a foreign language. They were in intact classes. 270 of them were in the ER classes: they involved themselves in extensive reading in and out of class for 10 months. The remaining 172 students were in the classes which focused more on reading texts with a deeper understanding of their grammar/discourse structures and interpretation of inference which the texts convey. The participants were taught in two different courses with different types of textbooks. A textbook in one course was a collection of passages about interesting places, people, events and customs in the world based on *National Geographic Magazine* articles. The textbook which the other course used was an anthology of British short stories, which included unabridged work by twentieth century’s authors such as Saki, James Joyce, and Muriel Spark. In the course, the participants were required not only to follow the storyline of each work, but also to appreciate the literary effects of wordings and characters’ feelings in particular situations.

2.2 EPER PPT: Content analysis and classification of cloze deletions

There are twelve short independent passages each of which consists of 91 words on the average. The twelve passages are arranged in the order of ascending difficulty. 99% of the sentences are in active voice and only 1% in passive voice. The average readability of the passages is 92.7 according to Flesch Reading Ease formula. Flesch-Kincaid Grade Level Index shows the readability of an English text in terms of grade levels（K-12）in the United States and the EPER PPT Form A indicates a grade level of 2.3. In terms of Lexile measures（henceforth, L）, the EPER reading passages of are 470 L. The passages in the Japanese junior high-school English textbooks range from 220 L to 480 L（Negishi, 2015）and those in the first year senior high school English textbooks range from 630 L to 932 L（Ota, 2015）. Based on these previous studies, the reading levels of the passages in the EPER PPT Form A are equivalent to those of the passages used in Japanese junior high school English textbooks.

Each passage has 10 to 15 deletions and there are 141 deletions in total. A deletion is made every 4 to 12 words and the average deletion interval is 6.3 words. Test takers are asked to fill in each blank with one English word.

Table 1 on page 77 presents the grammatical categories of deleted words. Verbs combined with verb aspects（i.e., perfect, progressive, and perfect progressive）are deleted most frequently（37 items, 26.2%）, followed by common nouns（21 items, 14.9%）. The third most frequently deleted word group includes pronouns（18 items, 12.8%）. Those are personal pronouns and only one deletion of interrogative pronouns. No relative pronoun is deleted.

Bachman（1982/1994）established a classification scheme with three categories of deletions in cloze tests: syntactic, cohesive, and strategic. Similarly, in the current study, all 141 items in EPER PPT Form A were categorized into three groups: syntactic, cohesive, and strategic. Each group is defined as follows: syntactic cloze items are items which require test-takers to use grammatical knowledge in the clause-level context. Cohesive items require test-takers to use interclausal or intersentential context. Strategic items require test-takers to have the ability to utilize macro-level information in restoring deletions and to reach the most suitable answer in the context of alternative answers, including vocabulary and grammar knowledge. The following paragraph presents examples of each item type.

Simon looks at the people in the station.（1）can see students in jeans, and men（2）suits. He can see families and children. He cannot see any spies. Simon's train goes（3）11.00, and

> it is 10.57 now. Simon (4) to the train. There is an old woman with an umbrella near Simon. She is walking very fast. Simon does not see her. He does not see her bag. (*Simon and the Spy*, Penguin Readers)

In order to answer item 1, learners need to refer to Simon in the previous sentence, which leads to the correct answer, the third person singular masculine pronoun "he." Item 1 is classified as cohesive. Concerning item 3, learners need to have grammatical knowledge about a preposition before the time expression (i.e., 11:00). Item 3 is classified as syntactic. Before item 4, the information about the setting (i.e., at the station) and the main character (i.e., Simon) is presented to the readers. Also, it is presented that there is little time before the departure of Simon's train, which leads to consider that Simon is in a rush. Item 4 is categorized as strategic.

The other theoretical rationale for the present categorization, especially concerned with the cohesive relationship in text, is provided by Halliday and Hasan (1976), whose view on discourse is that a certain linguistic item in discourse is interpreted in reference to another. Thus, based on Bachman (1982/1994) and Halliday and Hasan (1976), 141 deletions were categorized into three groups, yielding 45 syntactic deletions, 43 cohesive deletions and 53 strategic deletions. All three authors categorized the same deletions independently first and discrepancies were discussed until we reached a consensus. Further, in order to ensure that the subjective assessment of the categorization of deleted items into the three groups is in tune with one another, an interrater reliability analysis using the Kappa statistic was performed. 36 items, about 26% of the deleted items, were classified by an experienced EFL instructor who was not involved in the present research using the above-mentioned scheme: syntactic, cohesive, and strategic items. The Kappa statistic was .87 ($p < .001$). According to Landis and Koch (1977), the value of Kappa greater than .80 is considered a good level of agreement.

2.3 Procedures

The EPER PPT Form A was administered to the participants in the first week of the first semester in 2014–2015 school year. The testing time was 45 minutes. The participants were requested to answer as many items as they could during the testing time.

2.4 Scoring

Acceptable response scoring was applied and the original and acceptable words were counted as correct. In the previous studies where the EPER PPT Form A was used (Yoshizawa,

Table 1 *The grammatical categories of deleted words in the EPER PPT Form A*

Grammatical Categories	Frequency	Percentage
Verbs	32	23%
Common nouns	21	15%
Pronouns	18	13%
Prepositions	17	13%
Adverbs	15	11%
Adjectives	11	8%
Articles	11	8%
Aspect（*be*+ing, *have*+pp）	5	4%
Conjunctions	5	4%
Demonstratives	3	2%
Auxiliary verbs	2	1%
Interrogative pronouns（Wh-Questions）	1	1%
Voice（*be*+pp）	0	0%
Total	141	

Takase, & Otsuki, 2012 & 2013), a scoring rubric was created. At the beginning, all three authors marked the same answer sheets independently and listed correct and incorrect answers for all the blanks. Discrepancies were discussed until we reached a consensus. After the initial scoring rubric was created, we marked different answer sheets independently. The scoring rubric was revised each time we agreed upon new alternative answers or incorrect answers. The same scoring rubric was used in the current study.

2.5 Data analyses

Initially, descriptive statistics for the items were examined. Furthermore, since the average interval of word deletions is 6.3 words, ranging from 4–12 words, redundant items were examined for possible local item dependencies. There were nine pairs of dependent items. Nine items（i.e., one item from each pair）were deleted from the subsequent analyses.

Bachman（1982/1994）used composite scores in order to deal with the problems associated with analyzing binary data. The items were grouped based on the similarity of their content and composite scores were calculated by averaging the item scores in each group. The use of aggregated score or its average is called “parceling” and its use in factor analyses and structural equation modeling has been a topic of controversy. One of the reasons for the preference of parcel-level data over item-level data is related to sample sizes. When items are used to define a construct and many constructs are involved, factor analyses and structural equation modeling require a large sample. By contrast, parcel-level data can be used when sample sizes are relatively small (Little et al., 2002; Shimizu & Yamamoto, 2007). In addition, according to Little et al. (2002), models with parcel-level data are more parsimonious, have fewer chances for correlated residuals and dual loadings, and have reduced sources of sampling error compared with item-level data (p. 155).

In the current study, we used parceling. The reasons for our decision were two-fold. First, the sample size of the current study is not large enough to conduct factor analyses with item-level data. The ratio of participants to items is 3.13 (i.e., 442 participants divided by 141 items). Nunnally (1978) recommends ten cases for one item; Tabachnick and Fidell (2013) recommend five cases for one item. The ratio of participants to items of the current study did not fulfill either recommendation. Second, we preferred to replicate the methods used in Bachman's study as much as possible so that we could indicate possible differences between Bachman (1982/1994) and the current study would be due to those in deletion methods (i.e., random vs. rational deletions). Since there has been no established procedure about how to parcel items (Little et al., 2002; Shimizu & Yamamoto, 2007), we used the results of the content analyses of each item.

In this study, the same type of deletions (i.e., syntactic, cohesive, or strategic items) were grouped in each passage. Composite scores were calculated by summing the item scores in each group: each passage had one syntactic, cohesive, and strategic parcels, except for passage 8, which had only syntactic and strategic parcels. A total of 35 parcels were used for confirmatory factor analyses.

Prior to specifying specific models, we based our understanding of reading comprehension abilities on Grabe (2009) and Grabe and Stoller (2002, 2020), who outlined the reading comprehension processes activated when skilled readers read a longer text for general comprehension. According to Grabe and Grabe and Stoller, reading comprehension processes are divided into two processes: lower-level processes and higher-level processes. The former refer to lexical access, syntactic parsing, semantic proposition formation and working memory activation. The latter refer to text model of comprehension, situation model of reader interpretation, background knowledge use and inferencing, and executive control processes. As a reader continues to read and understand the text, he/she develops a set of main ideas of the text. This is defined as the text-model of comprehension. At the same time that the reader builds the text-model of comprehension, he/she starts to interpret the information from the emerging text model. Grabe and Grabe and Stoller emphasize that the classification of processes do not indicate one level of processes are more difficult than the other. Also, they mention that "reading comprehension processes work in parallel when some skills are relatively automatic" (Grabe & Stoller, 2002, p.

29).

Five models were hypothesized in the current study. These models were based on Bachman's (1982/1994) and the current reading theories (Grabe, 2009; Grabe & Stoller, 2002, 2020). What follows is a brief description of each model:

> *Model A: A single first-order factor model.* This model hypothesizes one single first-order factor underlies the performance of EPER PPT Form A scores and all measured variables load freely on the single first-order factor.
>
> *Model B: A second-order factor model.* This model hypothesizes that three separate first-order factors underlie the performance of EPER PPT Form A scores and those three factors are influenced by one second-order factor.
>
> *Model C: A three independent-factor model.* In this model, three first-order factors are hypothesized and they were uncorrelated with each other.
>
> *Model D: A three correlated-factor model.* The model hypothesizes that three separate traits underlie the performance of EPER PPT and these traits are correlated with each other.
>
> *Model E: A bifactor model.* This model hypothesizes a single first-order factor with three specific factors underlies the performance of EPER PPT Form A scores. A single first-order factor directly load onto all of the observed variables. Further, three specific factors load onto subgroups of the same set of the observed variables. Those specific factors are correlated with each other (Dunn & McCray, 2020).

Figures 1–5 present schematic representations of the five hypothesized models. In each figure, a square denotes an observed variable and a circle denotes a latent variable. E denotes observed variable errors. D denotes latent-variable errors. A single-headed arrow denotes the direct effect from one variable to another and the double-headed arrow denotes a correlation. For example, in Figure 1, single-headed arrows are directed from a latent variable named Text Processing Ability on the left to all of the nine observed variables on the right. This hypothesizes that these observed variables define the latent variable, text processing ability. SYN, COH, and STR refer to the syntactic, cohesive, and strategic parcels, respectively. Numbers following the letters in each observed variable refer to a passage number in the EPER PPT. Figure 1 represents a single first-order factor model (Model A); Figure 2, a second-order factor model (Model B); Figure 3, a three independent-factor model (Model C); Figure 4, a three correlated-factor model (Model D); and Figure 5, a bifactor model (Model E).

Confirmatory factor analyses were conducted using EQS for Windows Version 6.2 (Bentler

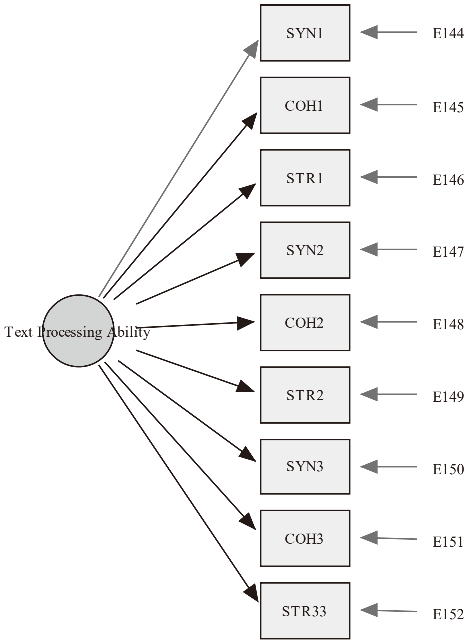

Figure 1. Model A: A single first-order factor model

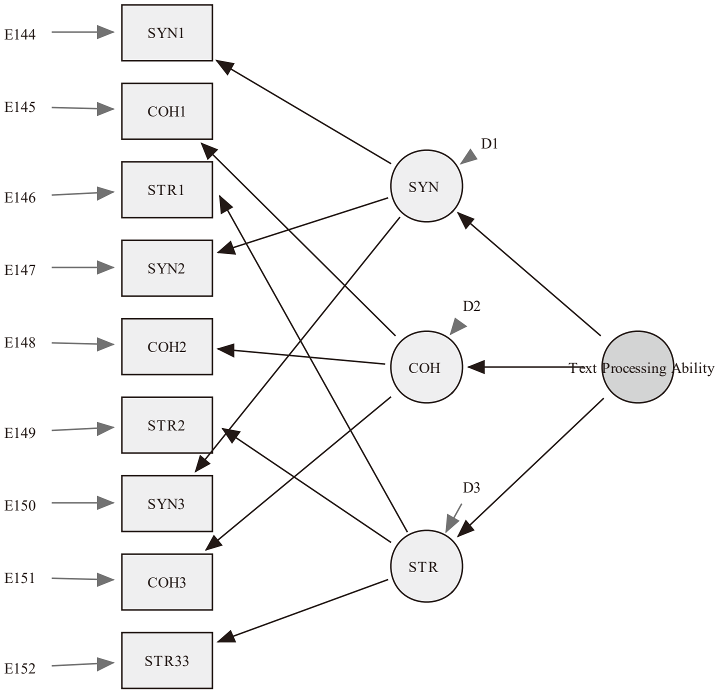

Figure 2. Model B: A second-order factor model

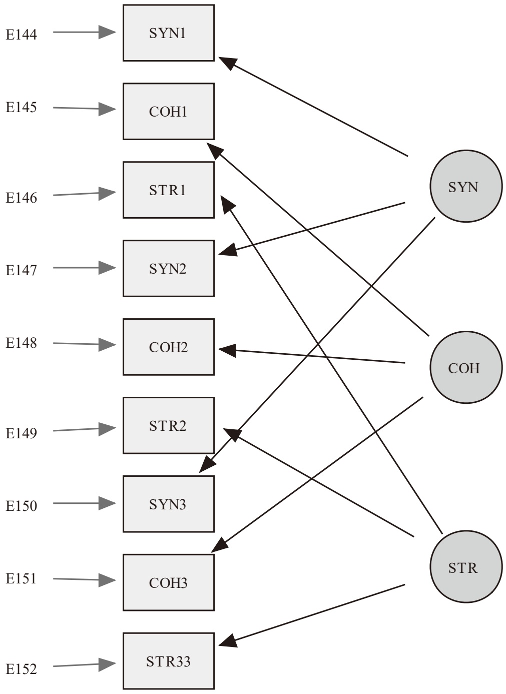

Figure 3. Model C: A three independent-factor model

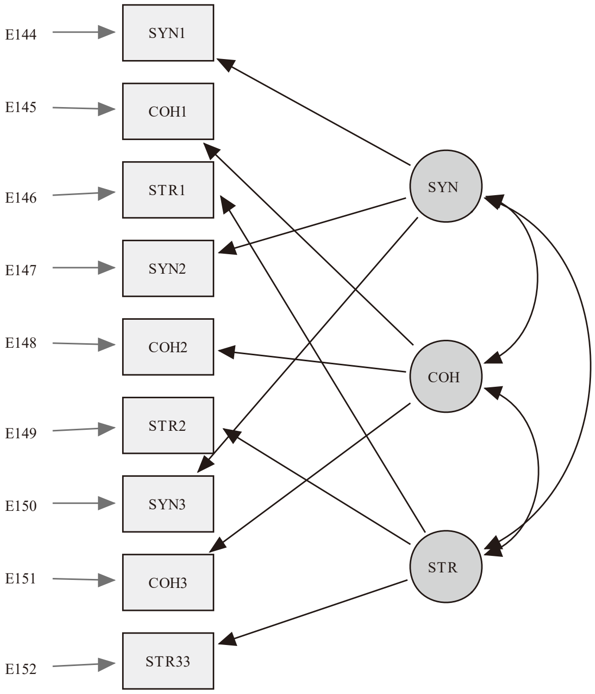

Figure 4. Model D: A three correlated-factor model

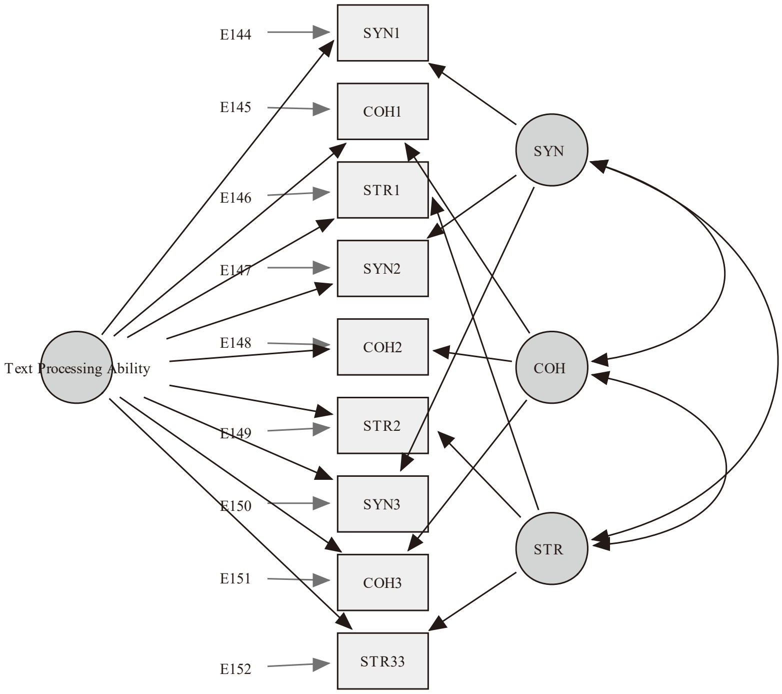

Figure 5. Model E: A bifactor model

& Wu, 2008) statistical package. Because some of the composite scores were non-normal, Satorra-Bentler Scaled chi-square statistic (Satorra, 1990; Hoyle, 1995) was used for model fit. Also, proposed models were estimated using maximum likelihood robust as well as maximum likelihood techniques.

The fit between the hypothesized models and the sample data was evaluated based on the following criteria:

a. The ratio of Satorra-Bentler model chi-square to model degrees of freedom ($\chi^2_{S\text{-}B}$/df): 3.0 or below are suggested as a good model fit (Kline, 1998).
b. Comparative Fit Index (CFI): Although a CFI of .90 or above was considered an adequate model fit, one close to .95 is advised for a well-fitting model (Byrne, 2008, p. 97).
c. Root Mean Square Error of Approximation (RMSEA): RMSEA assesses the model fit, taking into account model complexity. Values less than .05 are considered as good fit; values less than .08 are adequate fit (Kano & Miura, 2002).

Also, the statistical significance of each parameter was examined by dividing the parameter estimate by its standard error. Based on an alpha level of .05, values greater than +/− 1.96 are considered statistically significant (Byrne, 2008). Post-hoc analyses were not conducted.

3. Results

The coefficient of alpha reliability of the EPER PPT was .94. The means of the syntactic, cohesive, and strategic parcels were calculated for each of the 12 texts. The mean of the syntactic parcels was .41 (SD = .25), that of the cohesive parcels was .52 (SD = .21), and that of the strategic parcels was .19 (SD = .13). Figure 6 shows the means of the item parcels. The x-axis indicates the text numbers in the EPER PPT and the y-axis indicates the mean of each parcel. In general, the means of strategic parcels were lower than those of syntactic and cohesive parcels. The only exception was that the mean of the syntactic parcel in Text 4 was .27, that of the strategic parcel was .31, and the former was slightly lower than the latter. In addition, there was a tendency for the means in the second half of the EPER PPT, especially Texts 8–12, to be lower than those in the first half (i.e., Texts 1–7).

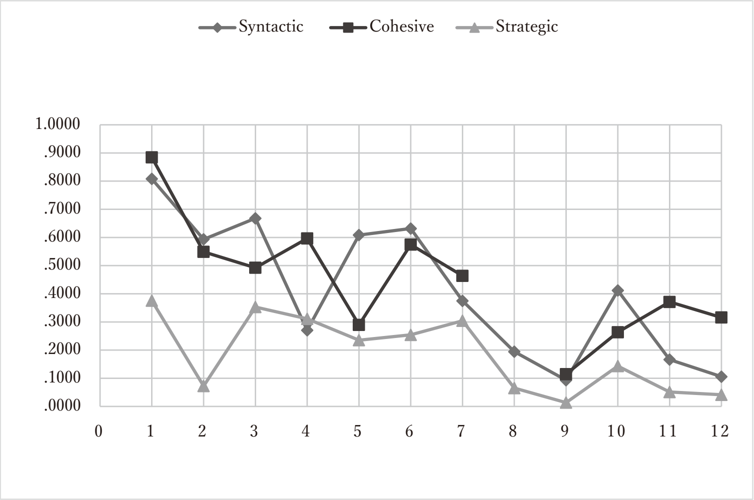

Figure 6. Means of Item Parcels

The descriptive statistics indicated eight parcels (SYN12, COH1, COH9, STR2, STR8, STR9, STR11, and STR12) had non-normal univariate distributions and they were deleted from the subsequent analyses.

Table 2 shows the fit indices for the five hypothesized models. The results indicated that Model C, the three independent-factor model, had poor fit. This model hypothesizes that three traits are independent of each other. The results of Models A, B, and D showed that the hypothesized models adequately described the sample data. The differences among those three models were considered negligible. In contrast Model E, the bifactor model, indicated better model fit than the other models in terms of the three criteria set in the current study.

Table 2 *The fit indices for the five hypothesized models*

Model	$\chi^2_{S\text{-}B}$	df	$\chi^2_{S\text{-}B}$/df	CFI	RMSEA	(90% CI)
Model A	617.648	324	1.906	0.916	0.046	.041 – .051
Model B	617.211	321	1.923	0.911	0.046	.040 – .051
Model C	1607.935	324	4.963	0.615	0.095	.090 – .099
Model D	613.941	321	1.913	0.912	0.045	.040 – .051
Model E	399.11	294	1.378	0.968	0.028	.021 – .035

Note. Model A: A single first-order factor model; Model B: A second-order factor model; Model C: A three independent-factor model; Model D: A three correlated-factor model; Model E: A bifactor model.

Following the examination of the fit indices, the fit of individual parameters in the Models A, B, D, and E were examined. Concerning Models A, B and D, reviewing the unstandardized solution, all estimates are statistically significant and all standard errors are considered to be in good order.

Table 3 shows the parameter estimates and standard errors of Model E, the bifactor model from the initial analysis: the loadings of each parcel on their respective factors (specific factors), their loadings on the general factor, and error variances. Concerning the loadings of each parcel on the specific factors, the unstandardized solution results indicate only eight out of 27 loadings were significant: SYN1, SYN10, SYN 11, COH10, COH11, COH12, STR7, and STR10. Also, those significant loadings were all from the latter half of the EPER PPT except for SYN1. On the other hand, all the loadings on the general factor were significant. Concerning the correlations between the specific factors, the syntactic and cohesive factors showed 1.00, .832 between the syntactic and strategic factors, and .604 between the cohesive and strategic factors.

Table 3 *Loadings on the specific and general factors and error variances: the bifactor model（Standardized solution）*

Variables	Specific factors	General factor	ER	Variables	Specific factors	General factor	ER
SYN1	−0.154	0.371	0.916	COH2	—	0.701	0.708
SNY2	—	0.699	0.709	COH3	—	0.621	0.772
SYN3	—	0.630	0.772	COH4	—	0.526	0.846
SNY4	—	0.589	0.808	COH5	—	0.529	0.848
SYN5	—	0.620	0.784	COH6	—	0.662	0.747
SNY6	—	0.611	0.792	COH7	—	0.476	0.871
SYN7	—	0.532	0.845	COH10	0.385	0.505	0.773
SNY8	—	0.536	0.840	COH11	0.460	0.435	0.774
SNY9	—	0.496	0.860	COH12	0.410	0.437	0.801
SYN10	0.435	0.643	0.631	STR1	—	0.216	0.974
SNY11	0.222	0.501	0.837	STR3	—	0.583	0.807
COH2	—	0.701	0.708	STR4	—	0.527	0.850
COH3	—	0.621	0.772	STR5	—	0.539	0.829
COH4	—	0.526	0.846	STR6	—	0.590	0.797
COH5	—	0.529	0.848	STR7	0.213	0.603	0.769
				STR10	0.433	0.380	0.817

Note. SYN = Syntactic parcels; COH = Cohesive parcels; STR = Strategic parcels.

4. Discussion

The present study classified the deletion patterns of EPER PPT to examine its trait structure. Our findings indicate results similar to and different from those of Bachman（1982/1994）, who analyzed a rational-deletion cloze test. In Bachman, three models were hypothesized, and they were named “general trait model,” “specific trait model,” and “general plus specific trait model”（p. 64）. In Bachman’s study, only the “general plus specific trait model” showed a significantly good fit to the data, while neither the “general trait model” nor the “specific trait model” showed an adequate fit to the data. In addition, concerning the “general plus specific trait model,” it was observed that “the composites load most heavily on the general factor, with lesser loadings on specific trait factors”（p. 65–66）in most cases. In the present study, the three independent factor models did not show adequate fit to the data. The single first-order factor model, second-order factor model, and three-correlated factor model indicated a good fit to the data. The differences among the three models were negligible. However, the bifactor model indicated a better fit to the data than these three models.

According to Rindskopf and Rose（1988）, the single first-order factor model, second-order factor model, three correlated factor model, and bifactor model are in the hierarchy of decreasing restrictions. The one-factor model is generally the most restricted, whereas the bifactor model is the least restricted. Placing restrictions on the parameters of one model changes the model to one above it. For example, setting all loadings of the general factor to zero in the bifactor model changes it into a three-correlated factor model. Rindskopf and Rose state that "any data which is consistent with one model will be consistent with a less restricted model in the hierarchy' (p. 56). The authors further state that the selection of a model is based on "theoretical plausibility, parsimony, or a statistical test of the difference' (p. 56).

In terms of "theoretical plausibility," the best fit of the bifactor model may reflect how the respondents cope with deletions in the EPER PPT. The test consists of 12 short narrative texts with an average number of 91.4 words per text. Each text starts with a lead in one or two short sentences before deletion. Although each text is a self-contained unit of discourse, the information in the lead-in is limited, and the readers have to figure out the setting, main characters, their relationships, and the event each text describes or the topic of the conversation among the characters. Thus, the length of each text, especially those with deletions, is unlikely to provide sufficient information for readers to process text information for general understanding. In particular, when cloze texts are in ascending difficulty in the EPER PPT, readers may not be able to process text information, as described in Grabe (2009) and Grabe and Stoller (2002, 2020). This may result in significant loadings of each parcel on the specific factors in the latter half of the test but not in the first half of the test.

Based on the results of the current study, we suggest that a single first-order factor and three specific correlated factors underlie the performance of the EPER PPT. The single first-order factor is called text- or discourse-processing ability. For successful completion of EPER PPT deletions, readers have to utilize not only grammatical and textual knowledge but also pragmatic and social linguistic knowledge. Text-processing ability is the ability to process text at the discourse level, not only at the clausal level. To understand the text and fill in the blanks, interpretation at the clause level (i.e., sentence meaning or literal meaning) is not sufficient; test-takers need to interpret the clause at the context level, that is, contextual meaning or speaker meaning. For instance, an anaphora requires a contextual-level interpretation to identify antecedents. Thus, this identification is straightforward. However, if there is more than one possible antecedent in the text, it is necessary to understand the passage as a whole in order to find out the accurate combination between an anaphoric expression (e.g., *this, they*) and its antecedent. In fact, common nouns and pronouns, which require test takers to figure out the right antecedent for interpretation, are the second- and third-most grammatical categories of deleted words in EPER PPT Form A. This is why EPER PPT Form A is recognized as demanding by many Japanese test takers, despite its relatively low reading levels indicated by the Flesch Reading Ease formula and Flesch-Kincaid Grade Level Index, which are based on word and sentence lengths.

The current study showed a different finding from that of Bachman (1982/1994). The different results are likely due to the differences in deletion patterns: rational deletions in Bachman (1982/1994) and random deletions in the current study. However, two factors may have contributed to the different findings. First, the deletions in Bachman's (1982/1994) study were based on a 365-word expository text from an introductory social psychology textbook, while the EPER PPT consists of 12 short texts. Another factor that might have affected the cloze test performance concerns test-taker characteristics. The participants in Bachman's (1982/1994) study were English language learners at an American university from a wide variety of L1 backgrounds and a wider age range who were learning English as a Second Language. In contrast, the participants in the current study were Japanese learners of English at the university level, who had been learning EFL. Differences in test-taker characteristics such as proficiency level, learner motivation, and attitudes toward the target language and learning might have influenced their test performance.

5. Limitations and conclusion

The current study has some limitations. The participants were from the intact classes. Thus, the generalizability of our findings is limited. It is necessary to examine whether a bifactor model applies to another group of participants in a learning context similar to that in the current study.

Through ER, learners can have abundant exposure to their target languages. In ER programs, learners are provided with books appropriate for their reading levels, and their instructors monitor them and decide whether the learners are reading the materials appropriate for their reading levels, guiding them appropriately when it is time to upgrade the level of their reading materials. In this way, learners can experience a flood of comprehensible inputs through their ER programs. This is quite important, especially for learners in an EFL context (Iwahori, 2008), since their exposure to linguistic input of the target language is limited. Learners need to be provided with books appropriate to their proficiency levels, and their progress needs to be monitored on a regular basis. The EPER PPT is the only measurement designed for these purposes (Day & Bamford, 1998).

It is our hope that the current study has shed some light on the structure of the EPER PPT. Further study is needed to examine the trait structure of the EPER PPT so that stakeholders in ER programs (i.e., learners, teachers, and administrators) will know exactly what ER programs affect and how they affect learners.

Acknowledgments

We are grateful to Prof. Lyle Bachman for his kind responses regarding his categorization of cloze items in Bachman (1982/1994). Any faults related to the interpretation of his study are solely ours. This research was conducted as a part of a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (No. 18K00766).

References

Akao, M. (2015). Tadoku to eigoryoku nobi no kanrensei—daigaku sairishu kurasu niokeru tadoku jyugyo [Relationship between extensive reading and improvement of English proficiency—ER at university repeater classes]. *JERA Bulletin 8*, 39–50.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly 13*(2), 219–223.

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

Bachman, L. F. (1994). The trait structure of cloze test scores. In J. Oller and J. Jonz (Eds.), *Cloze and coherence* (pp. 177–187). Bucknell University Press. (Reprinted from *TESOL Quarterly,* 16, 61–70, 1982)

Beglar, D., Hunt, A., & Kite, Y. (2011). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Teacher, 61*(4), 1–39.

Bentler, P. M., & Wu, E. J. C. (2008). *EQS 6 for Windows user's guide.* Encino, CA: Multivariate Software, Inc.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872–882.

Chapelle, C., & Abraham, R. (1990). Cloze method: What difference does it make? *Language Testing 7* (2), 121–146.

Chihara, T., Oller, J., Weaver, K., & Chavez-Oller, M. (1994). Are cloze items sensitive to constraints across sentences? In J. Oller and J. Jonz (Eds.), *Cloze and coherence* (pp. 135–148). Bucknell University Press. (Reprinted from *Language learning, 27*(1), 63–73, 1977).

Cho, K. S., & Krashen, S. D. (1994). Acquisition of vocabulary from the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading, 37*(8), 662–667.

Day, R., & Bamford, J. (1998). *Extensive reading in the second language classroom.* Cambridge University Press.

Dunn, K. J., & McCray, G. (2020). The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. *Frontiers in Psychology, 11*, 1–16. doi.:10.3389/fpsyg.2020.01357

Edinburgh Project on Extensive Reading. (1992). The EPER guide to organising programmes of extensive reading. University of Edinburgh, Institute for Applied Language Studies.

Elley, W. B., & Mangbhai, F. (1981). *The impact of a book flood in Fiji primary schools*. New Zealand Council for Educational Research and Institute of Education: University of South Pacific.

Furukawa, A. (2008). Extensive reading program from the first day of English learning. *Extensive Reading in Japan 1*(2), 11–15.

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

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading* (2nd ed.). Pearson Education.

Grabe, W., & Stoller, F. L. (2020). *Teaching and researching reading* (3rd ed.). Routledge.

Hafiz, F. M., & Tudor, I. (1989). Extensive reading and the development of language skills. *ELT Journal, 43*(1), 4–13.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 121–128). Newbury House Publishers.

Huffman, J. (2014). Reading rate gains during a one-semester extensive reading course. *Reading in a foreign language, 26*(2), 17–33

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Sage.

Irvine, A. (2011, September 3–6). *Extensive reading and the development of L2 writing* [Paper presentation]. The First Extensive Reading World Congress, Kyoto, Japan.

Iwahori, Y. (2008). Developing reading fluency: A study of extensive reading in EFL. *Reading in a Foreign Language 20*(1), 70–91.

Kobayashi, M., Kawachi, T., Fukaya, M., Sato, T., Tani, M, (2010). *Tadoku de hagukumu eigoryoku plus α* [Potentials of extensive reading: Promoting English learning and much more]. Seibido.

Kano, H., & Miura, A. (2002). *Gurafikaru tahenryokaise*i [Graphical multivariate analyses]. Gendai Sugakusha.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* The Guilford Press.

Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal, 73*(iv), 440–464.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lemmer, R., Brierley, M., Reynolds, B., & Waring, R. (2012). Introduction to the extensive reading foundation online self-placement test. *Extensive Reading World Congress Proceedings, 1*, 23–25.

Little, T. D., Cunningham, W. A., & Shahar, G. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151–173.

Mason, B., & Krashen, S. (1997). Extensive reading in English as a foreign language. *System, 25*(1), 99–102

Mermelstein, A. D. (2015). Improving EFL learners' writing through enhanced extensive reading. *Reading in a Foreign Language 27*(2), 182–198.

Nation, I. S. P., & Waring, R. (2020). *Teaching extensive reading in another language*. Routledge.

Negishi, M. (2015). Lexile measure-niyoru chu-kou-dai-no eigo kyokasho-no tekisuto nanido-no kenkyu [Lexile measures of textbooks used in a secondary and tertiary education in Japan]. *ARCLE Review, 9,* 6–16.

Nishizawa, H., Yoshioka, T., & Fukada, M. (2010). The impact of a 4-year extensive reading program. In A. M. Stoke (Ed.), *JALT2009 conference proceedings*, 632–640. JALT.

Ohta, E. (2015). Lexile measure-de arawasu koukou eigo kenteikyoukasho-no nanido: komyunike-shon-eigo-to eigo-no hikaku [Lexile measures of English textbooks used at senior high schools]. *Shiroyama eibei bungaku, 40*, 41–56.

Oller, J. (1979). *Language tests at school: A pragmatic approach*. Longman.

Richards, J. C., & Schmidt, R. W. (2011). *Longman dictionary of language teaching and applied linguistics,* (*4th ed.*). Pearson Education.

Rindskopf, D., & Rose, (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*(1), 51–67. https://doi.org/10.1207/s1532

Robb, T., & Susser, B. (1989). Extensive reading vs. skills building in an EFL context. *Reading in a Foreign Language, 5*(2), 239–251.

Sakamoto, A. (2014, June 28). *Tachou/tadoku de tsuchikawareru hasshinryoku* [Productive skills which come with extensive listening/reading] [Paper presentation]. Japan Extensive Reading Association Kyushu Seminar, Fukuokajyogakuin, Fukuoka, Japan.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent development. *Quality & Quantity, 24*, 367–386.

Shimizu, K., & Yamamoto, R. (2007). Shohokashita hensuniyoru personality kouseigainenkanno kankeiseino moderuka [Modeling of the relationships among the constructs of personality using parceling method: Big Five, state-trait anxiety, and mood states]. *Journal of Sociology, 3*, 61–95.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing, 14*(2), 214–231.

Suzuki, J. (1996). Dokusho no tanoshisa wo keiken saseru tamemo reading shido [Teaching reading for enjoyment]. In T. Watanabe (Ed.), *Atarashii yomi no shido* [*New approach to teaching reading*] (pp. 116–123). Sanseido.

Tabachnick, G., & Fidell, S. (2013). *Using multivariate statistics*, (6th ed.). Pearson.

Takase, A. (2004). Investigating students' reading motivation through interviews. *Forum for Foreign Language Education, 3*, 23–38.

Takase, A. (2007). Japanese high school students' motivation for extensive L2 reading. *Reading in a Foreign Language 19* (1), 1–18.

Takase, A. (2008). The two most critical tips for a successful extensive reading. *Kinki University English Journal, 1*, 119–136.

Takase, A. (2009). The effects of SSR on learners' reading attitudes, motivation, and achievement: A quantitative study. In A. Cirocki (Ed.), *Extensive Reading in English Language Teaching* (*pp.* 547–560). Lincom.

Takase, A. (2010). *Eigo tadoku tacho shido manual* [Teaching manual for extensive reading and listening]. Taishukan Shoten.

Takase, A., & Otsuki, K. (2012). New challenges to motivate remedial EFL students to read extensively. *Apples – Journal of Applied Language Studies 6(2)*, 75-94. University of Jyväskylä, Finland.

Watanabe, M. (2014). Kouritsu chutokyoiku ni okeru tadoku shidou no seika──GTEC, Shinken moshi no data kara. [The effects of extensive reading at prefectural secondary school──Results of GTEC and Shinken mock test]. *JERA Bulletin 7,* 27-28.

Yamashita, J. (2004). Reading attitudes in L1 and L2, and their influence on L2 extensive reading. *Reading in a Foreign Language, 16*(1), 1-19.

Yamashita, J. (2008). Extensive reading and development of different aspects of L2 proficiency. *System36,* 661-672.

Yoshizawa, K., Takase, A., & Otsuki, K. (2012, October 27). *Can we treat the EPER Form A and Form E as alternate forms?* [Paper presentation]. The 16th annual conference of Japan Language Testing Association, Kawasaki, Kanagawa, Japan.

Yoshizawa, K., Takase, A., & Otsuki, K. (2013, September 21). *Comparison of the EPER Form A and Form E: Do they work as alternative forms?* [Paper presentation]. The 17th annual conference of Japan Language Testing Association, Tokyo, Japan.