課程與教學季刊　2016, 19(3),頁 205～236

On What Rational Cloze Tests Can Measure: A Revisit Applying Confirmatory Factor Analysis

Fu-Wen Lin[*]　Wen-Ying Lin[**]　Ching-Yun Yu[***]

In response to the demand for quality tests that are easy to construct and in view of the cohesive and coherent competence that is important to reading comprehension and writing development, the present study was motivated and conducted, in the hope that its results could prove informative as to whether easily constructed rational cloze tests could be customized for classroom language teachers' testing objectives with high reliability and construct validity. Employing the confirmatory factor analysis (CFA), the current study aimed to investigate whether additional evidence can be obtained to support Bachman's (1982) claim that the rational cloze test can be designed by classroom language teachers to measure an array of reading skills as intended, specifically including cohesive and coherent competence. In this study were a total of 713 participants, all college students, taking the general English courses at four universities in northern Taiwan. The participants' dichotomously-scored responses to a rational cloze test constructed by Bachman (1985) were analyzed through the CFA on the item-by-item level in the program Mplus, which provides a convenient mechanism to perform the CFA of dichotomous responses. The present study, partly following Bachman (1982), tested the general trait model, the four specific traits model, and the general plus specific traits model—the last comprising a general factor plus the four specific traits. The testing results of the present study have lent support to the hypothesized general plus specific traits model, stating that rational cloze tests can be designed to tap distinct and related language competence, such as syntactic competence, the cohesive and coherent competence that depends on contexts of across-clauses and across-sentences, and extra-textual inference, all of which could be explained by the second-order general language proficiency factor. Based on the findings of the present study, implications and recommendations were provided for future research and classroom language teachers, as well as language test constructors.

Key words: rational cloze tests, confirmatory factor analysis, second-order factor, cohesive competence, coherent competence

[*] Fu-Wen Lin, Assistant Professor, Institute of Applied English, National Taiwan Ocean University

[**] Wen-Ying Lin, Associate Professor, Department of English Instruction, University of Taipei

[***] Ching-Yun Yu, Associate Professor, Department of Psychology and Counseling, University of Taipei

Corresponding Author: Wen-Ying Lin, e-mail: wylin@utaipei.edu.tw

Motivation

To construct quality multiple-choice-item classroom English tests by following strict item-writing and item-piloting procedures is often a daunting challenge, if not a mission impossible, for English language classroom teachers who often have to try very hard to meet deadlines for various job duties or engagements. As such, there has emerged a need for the quality tests that are relatively simple to produce. In order to provide a way to cope with the tension between quality and simplicity, one recent study by Cai (2013) has demonstrated how partial dictation, one item format of language reduced redundancy (LRR) approach to language testing, could be easily constructed and administered with desirable psychometric properties. In a similar vein, the motivation of the present study was to show that another easily constructed LRR item format—rational cloze tests—can be used to assess cohesive competence and coherent competence, which have long been demonstrated and recognized to play an integral and facilitating role in reading comprehension (Chapman, 1983; Irwin, 1986; Kintsch, 1988) and writing development (Crowhurst, 1987; Kolln, 1999; Witte & Faigley, 1981). Particularly during the process of reading comprehension, according to Kintsch (1988), the reader is required to construct a coherent representation of a text in memory, which has been contended to be critical to successful reading comprehension (Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007). The reader needs to not only figure out how one piece of information is related to another, but also link everything together to develop meaning and to form a whole. The former is referred to as cohesive competence, the perception of cohesive relations in a text (Horning, 1991), and the latter as coherent competence, the understanding of how idea units in a text are interwoven together to form the web of meaning intended by the writer (Hampton, 2010). In response to the demand for the quality tests that are easy to construct and in view of the cohesive and coherent competence that is important to reading comprehension and writing development, the present study was conducted, in the hope that by transcending the confines beyond the traditional role of cloze tests in large-scale standardized general language proficiency assessment, the studying results could prove informative as to whether easily constructed rational cloze tests could be customized for classroom English teachers' testing objectives with high reliability and construct validity.

Literature Review

Cloze Tests

Cloze tests, commonly recommended for assessment in second/foreign language research, are the most important and best-known operationalization of the principle of the LRR approach to language testing (Beinborn, Zesch, & Gurevych, 2015; Klein-Braley, 1997). The crux of the LRR principle is that knowing a language involves the ability to understand an incomplete message and make educated guesses about a certain percentage of the missing linguistic information. That is, the ability to correctly restore deleted words indicates the power of comprehension (Lee, 2008). Technically speaking, the parts of a written text that are deleted in cloze tests function as noise that occurs in the surroundings of everyday language use. The performance that test-takers demonstrate under the condition is employed as an indicator of their overall language proficiency. With their proficiency in a language improving, learners can—as claimed by Spolsky, Bengt, Sako, and Aterburn (1968)—make more successful use of the redundancy inherent in the language and obtain a higher score on cloze tests. The word "cloze," according to Hinofotis (1980), comes from the concept of closure used in Gestalt Psychology and refers to the ability to fill in the gaps in an incomplete pattern. The principle of the LRR is highly involved in cloze tests in the sense that the tests reduce natural linguistic redundancies and require test-takers to utilize organizational constraints to infer meaning and fill in the blanks. As reviewed by Harsch & Hartig (2015), the assumption of the LRR theory is that the higher language proficiency level a learner has, the more gaps s/he can fill in by activating and drawing on his/her automated language skills. This alleged theoretical basis of cloze tests on the LRR principle has led to their widespread application in internationally recognized language testing as measures of general proficiency (Khodadady, 2012).

Among various forms of cloze tests, the standard or fixed-ratio cloze form and the rationally-deleted cloze form have been investigated most extensively, with a focus on their psychometric properties. The standard cloze form consists of a text, where a word is deleted after every certain number of words according to an arbitrary and fixed ratio procedure. For example, every seventh or tenth word is deleted after one or two sentences of an unbroken text. The assumption about the invariance of test results across standard cloze tests with different texts or different

deletion rates, claimed by Oller (1973), has been repeatedly investigated and questioned. To name just a few, Klein-Braley (1981) and Zarrabi (1988) demonstrated that different texts with the same deletion rate would result in different reliability estimates and different criterion-related validity coefficients with the criterion measures. This suggests that selecting different texts may produce different tests, each of which measures certain aspects more effectively than other aspects. As to the deletion rate, Alderson (1979) found that it was also an important factor affecting the results of standard cloze tests. In particular, he found that a text could produce quite different tests depending on whether, say, every fifth rather than every seventh word was deleted. Therefore, he cautioned against the use of standard cloze forms and favored the use of rationally deleted cloze forms, where the words deleted are selected rationally based on linguistic and coherence structures of texts. His suggestion has found strong echoes among numerous language testers. Farhady & Keramati (1996), for example, found that, in terms of criterion-related validity and reliability estimates, standard cloze tests generally failed to produce a better test than rationally deleted cloze tests. As such, from the perspective of reliability estimates and criterion-related validity, the bulk of existing empirical evidence generally seems to render support for the superiority of rationally deleted cloze tests over standard cloze tests.

Construct Validity of Cloze Tests

There being an extensive body of research using the correlational method that have found significant relationships between cloze tests and other measures of language abilities, another research strand has concerned a much more important issue: What specific abilities can cloze tests measure? The methodology adopted in most of the studies investigating the construct validity of cloze tests has been to vary specific aspects of cloze tests, such as the deletion ratio, the order of sentences, the difficulty of a passage, and then examine the effects of these changes on the performance of their participants. Unfortunately, the studies along this line have produced widely differing conclusions. For instance, on one side is the conclusion made by Alderson (1979) that cloze tests are capable of measuring only lower-order core proficiency skills, based on his results obtained from varying passage difficulty, scoring criteria and deletion ratios. On the other side are findings from Anderson's (1980) study and similar ones from Oller & Conrad's (1971) and Yamashita's (2003) studies: in the former, cloze tests are capable of measuring test-takers' sensitivity to

not only sentence-level grammatical structures but also cohesive relationships across sentences; in the latter, cloze tests are useful measures of higher level skills, such as the ability to negotiate language or higher order processing ability.

Up to now, very few studies have addressed the construct validity issue of cloze tests with the factor-analytic approach—an approach that comes up with specific hypotheses of what cloze tests measure, makes corresponding deletions, and then examines whether or not the response patterns agree with the hypothesis. One such study was done earlier by Weaver and Kingston (1963), followed by another similar study by Ohnmacht, Weaver, & Kohler (1970). Both studies employed the exploratory factor analysis (EFA) approach using the estimation method of principal component analysis. About a decade later came the first attempt made by Bachman (1982) to find out the extent to which rational cloze tests can be designed to measure the abilities intended by using the confirmatory factor analysis (CFA), which is powerful and stringent for testing trait components of measures. In his study, based on Halliday & Hasan's (1976) description of semantic relationships in discourse, three theoretical models were posited and tested. The first one is the general trait model, consisting of only a single general factor that accounts for most of the variance in a language test. The second one is the specific traits model, including three divisible traits (i.e. syntactic competence, which depends only on clause-level context; cohesive competence, which depends on the inter-clausal and inter-sentential cohesive context; and strategic competence, which relies on parallel patterns of coherence). The third one is the general plus specific traits model, comprising a general factor plus the three specific traits. A cloze passage with 30 rationally selected deletions was constructed and given to a total of 418 non-native English speaking students attending the University of Illinois. The results of his study showed that the third model provided the best explanation of the data, leading to the conclusion that in addition to the overall language proficiency, rationally deleted cloze tests can be designed to measure a range of abilities.

Having proved that rational cloze tests are able to assess a range of reading skills intended, Bachman (1985) conducted a subsequent study to further confirm that such a testing technique is highly valid in the sense that its test scores agree with those provided by two other language tests. As pointed out by himself, the identification of the three types of deletion (i.e. clause-level context, inter-clausal and inter-sentential cohesive context, and parallel patterns of coherence) used in his previous study (Bachman, 1982) were subjective and judgmental. He further added

that it was difficult for classroom teachers or professional test writers to apply the item classification during their construction of rational cloze tests. Therefore, in order to increase the ease of construction for his rational cloze test, Bachman proposed four selection criteria for words to be deleted: within a clause; across clauses, within a sentence; across sentences, within text; and extra-textual (Bachman, 1985). In other words, the four criteria were employed to determine the corresponding four kinds of context required for closure and thus the four types of deletions were made. The results of the study showed that the rational cloze test scores were correlated (ranging from .72 to .85) highly with various subtests' scores of the two general proficiency tests, suggesting the high criterion-related validity of the rational cloze test. However, unlike his previous study done in 1982, the scores obtained from the rational cloze test in the study were not analyzed through any sort of factor-analytic approach. As such, the question was not addressed and answered as to what extent that rational cloze tests can be designed to tap intended specific abilities, based on his four selection criteria for word deletions.

Taken together, while there has been a wealth of research on the reliability and criterion-related validity of standard or rational cloze tests, there is a scarcity of research on their construct validity using the factor-analytic approach. To make things worse, among the very few studies examining the factor structure of cloze tests, only two studies (i.e. Bachman, 1982; Saito, 2003) could be located, using the CFA. One of the CFA's major advantages over the EFA—as advocated by Bollen (1989) and Lonigan, Hooe, David, & Kistner (1999)—is its hypothesis testing capacity by empirically evaluating and statistically comparing a set of a priori specified models. In other words, the CFA has been widely acknowledged as a powerful and rigid theory-driven approach to providing empirical, or statistical, evidence confirming hypothesized factorial structures and supporting construct validity of a measure (e.g. Dimitrov, 2010; DiStefano & Hess, 2005). Besides, with the CFA allowing correlated errors of measurement, latent traits resulting from the CFA hypothesis are less confounded by measurement errors than observed variables—an advantage in which a more precise estimation can be obtained concerning the relations of the underlying traits to each other.

Unfortunately, of the two studies that have used the CFA to investigate the construct validity of rational cloze tests, the recent one by Saito (2003) was conducted in the context of large-scale standardized examination for the Certificate

of Proficiency in English, which will not be elaborated further here because of its irrelevance to the setting of the present study (i.e. classroom English tests). As to the earlier study by Bachman (1982), there was a methodological pitfall, even though he did apply the CFA to test his three models. Specifically, in order to avoid the analytic problems associated with dichotomously-scored data matrices, he formed the 30 items into 13 sets (four syntactic, seven cohesive, and two strategic) according to the item content similarity and obtained a composite score for each set by averaging the item scores in each set. A series of the CFA was then performed based on the product-moment correlations among the 13 composite scores, rather than on the tetrachoric correlations among the 30 items. Parceling items or replacing item-by-item indicators with composite scores in the CFA is likely to result in biased estimates of factor loadings, and thus the true nature of the factor structure tends to be masked, as warned by many researchers (Bandalos, 2002; Bandalos & Finney, 2001; Kim & Hagtvet, 2003; Meade & Kroustalis, 2006). Hence, given the methodological defect found in Bachman's (1982) study and the paucity of studies taking advantage of the power and strength of the CFA, Bachman's (1982) conclusion seems premature with respect to the effectiveness of rational cloze tests designed to measure an array of reading skills.

Purpose of the Study

The importance of gaining a better understanding about the construct validity of rational cloze tests warrants conducting more factor-analytic studies applying a powerful and stringent statistical technique using individual items as indicators. In response to the paucity of previous studies along this line, the present study was to substantiate the usage of rational cloze tests as classroom English tests, by seeking further evidence to confirm Bachman's (1982) claim that rational cloze tests can be designed to measure an array of crucial reading abilities as specifically intended, such as cohesive competence and coherent competence. In particular, the participants' dichotomously-scored responses were analyzed through the CFA on the item-by-item level in the program *Mplus*, which provides a convenient mechanism to perform the CFA of dichotomous responses. The program by default employs a robust weighted least squares estimator, which according to León (2011), is one of the best ways of working with categorical data modeling. In a nutshell, the present study was mainly to apply *Mplus* to delving into the reliability, validity, and factor structure of the ration cloze test that produces dichotomous responses.

Method

Participants

A total of 713 (398 female and 315 male) college students participated in this study, based on convenience sampling. Taking the general English courses at four public universities in northern Taiwan from September 2014 to June 2015, they were from a wide variety of departments, including Accounting, Applied Physics and Chemistry, Mathematics, Bioscience and Biotechnology, Business Administration, Chinese Language and Literature, Early Childhood Education, English Instruction, Finance, Food Science, Mathematics, Mechanical and Mechatronic Engineer, Medicine, Music, Social and Public Affairs, Special Education, and so on. Despite the fact that they were from four different public universities, their general English reading proficiency, as measured by a reading proficiency test (to be described later in the following section on the instruments), were not different substantially, with their means ranging from 24.46 to 25.35. The results from the one-way ANOVA (analysis of variance) with universities as the between-subject factor further confirmed that their mean differences in general English proficiency were not statistically significant, with $F(3,709) = .73$ and $p = .53$. With regard to age, the majority of the participants were between 19 and 21, with few exceptions. Like other university students in Taiwan, they all received at least around 10 years of formal English instruction prior to their college education. They were taught mainly under the grammar-translation approach, the communicative teaching approach, or a combination of both.

Instruments

The present study used two instruments, one reading comprehension test and a rationally deleted cloze test. Each of the tests is described in the following:

Reading comprehension test. Since the rational cloze test has been commonly used as a measure of general reading proficiency, the present study would like to find out the degree to which its scores agree with those of some independent measure of general reading proficiency, in addition to examining the factor structure of the rational cloze test. Therefore, to assess the participant's general English reading proficiency, the present study employed a published intermediate level sample test from the Cambridge Preliminary English Test 4 (hereafter CPET)—

which is at Level B1 of the Common European Framework of Reference for Languages (CERF), an internationally recognized benchmark of language ability. The reading section of the test included a total of 35 objectively-scored items with a reported Cronbach's alpha reliability estimate reaching as high as .88 (Cambridge English Language Assessment, 2015). In the reading section, the participants had to answer multiple choice items, select descriptions that match different text, and identify true information. The items incorporated different target language situations that gauged a range of skills involved in reading comprehension at the intermediate level (e.g. reading for gist and detailed comprehension, scanning for specific information, understanding writers' attitude opinions and purposes, and making inference). Each item was worth one point and thus the maximum possible total score was 35.

Rational cloze test. A rational cloze test, developed by Bachman (1985), was also used in the present study. In Bachman's (1985) study, a passage about automatic control mechanisms from a collection of general readings in science was selected and adopted for use as the rational cloze test in his study. The passage in the cloze test contained a total of 363 words with 30 deletions based on his four hypothesized types of context required for closure: (1) within a clause; (2) across clauses, within a sentence; (3) across sentences, within text; and (4) extra-textual. The intention of the study, as he explained, was to make the test primarily a test of cohesion and coherence, so the numbers of deletions for types 2 and 3 were maximized. Hence, out of the 30 deletions, 18 deletions were made for the two types and remaining 12 deletions were made for types 1 and 4. For a better understanding, the complete test is provided in Appendix A, together with acceptable answers and type of deletion specified for each item/blank. Although the deletions were made according to his well-hypothesized context selection criteria, Bachman did not perform the CFA to delve into the factor structure of the cloze test with the criteria and deletions he hypothesized. As such, exactly the same cloze test with the identical deletions was adopted in the present study so as to probe the unanswered question concerning its construct validity.

Administration and Scoring Procedures

All of the participants were required to take the two tests in two separate class sessions, with the reading test first and the rational cloze test following. There was around one-week interval between the two sessions for the purpose of avoiding the

participants' fatigue from the tests. For each of the tests, the time length of test administration was set around 40 minutes. The participants' answers to the cloze test were scored by exactly following Bachman's (1985) scoring approach, that is, acceptable-word scoring. In particular, the participants' answers to the deletions were compared with a prepared list of syntactically and semantically acceptable alternatives. As explained by Bachman, the list was based on the judgment of his test development team, as well as on the responses of a sample of native English speakers on whom the tests were pretested.

Analysis

Following the study by Bachman (1982), the present study also tested the three models: the general trait model (Model 1), the four specific traits model (Model 2), and the general plus the four specific traits model (Model 3). However, instead of including three divisible traits in Bachman's (1982) study, the four specific traits model posited and tested in the present study were as follows: Trait 1, syntactic competence, which depends on the within-clause context; Trait 2, cohesive and coherent competence, which depends on across-clause but within-sentence context; Trait 3, cohesive and coherent competence, which depends on across-sentence context; and Trait 4, extra-textual inference. The four traits were so named on the basis of Bachman's (1985) well-hypothesized context criteria for selecting deletions and Williams and Colomb's (2010) concept of cohesion and coherence. According to Williams and Colomb (2010), cohesion and coherence respectively refer to a sense of flow and that of the whole. The former depends on how one "sentence" ends and the next begins, and the latter how the sentences in a passage cumulatively begin (pp. 68-72). The present study was aware of the concept adopted by other scholars, like Min (n.d.) and those following Halliday & Hasan (1976) and Hasan (1984): that cohesion connects ideas at the "sentence" level, a syntactical level, and coherence at the "idea" level, a propositional or semantic level. For the discussion and comparison of theoreticians' definitions, readers are referred to Fulcher (1989) for the detail. Here, Williams and Colomb's (2010) concept was adopted in view of the definition where coherence and cohesion both connect ideas "across sentences" while differing on the ways that they link idea units, propositions, to make up the web of meaning intended by the writer. For Trait 2 hypothesized in the present study, Williams and Colomb's (2010) concept of cohesion and coherence was extended to the level of clauses because cohesion and coherence exist among not only sentences

but also clauses. A clause as well as a simple sentence is an idea unit, a proposition, semantically. As defined in Frank's (1993) discussion, a sentence and a clause both are a "full predication" that contains a subject and a predicate with a finite verb (pp. 220-229). A sentence is always independent. A clause is, by contrast, classified into two types: the independent one, which may stand alone as a sentence; and the dependent one, which must depend on, or be attached to, an independent clause. The "full predication" of a dependent clause is made to depend on an independent clause by an introductory word (say, a subordinator), or it is altered in such a way that the clause must be attached to an independent one (say, an absolute construction or a participial phrase). Given the affinity between clauses and sentences with respect to their full predication, it logically follows that cohesion and coherence also exist among clauses.

The three models—Model 1, Model 2, and Model 3—were tested through CFAs on Item-level dichotomously-scored raw data using *Mplus* version 5.1 (Muthén & Muthén, 1998-2008). For a clear picture of the models, schematic representations are shown below in Figures 1-3. The adequacy and appropriateness of the three models were compared and evaluated according to three criteria: values of selected global model fit indices, individual parameter estimates, and the principle of parsimony.

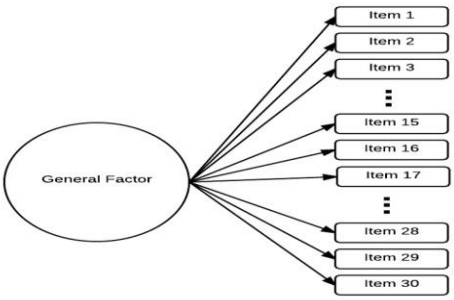

Figure 1 *Model 1: the general trait model*.

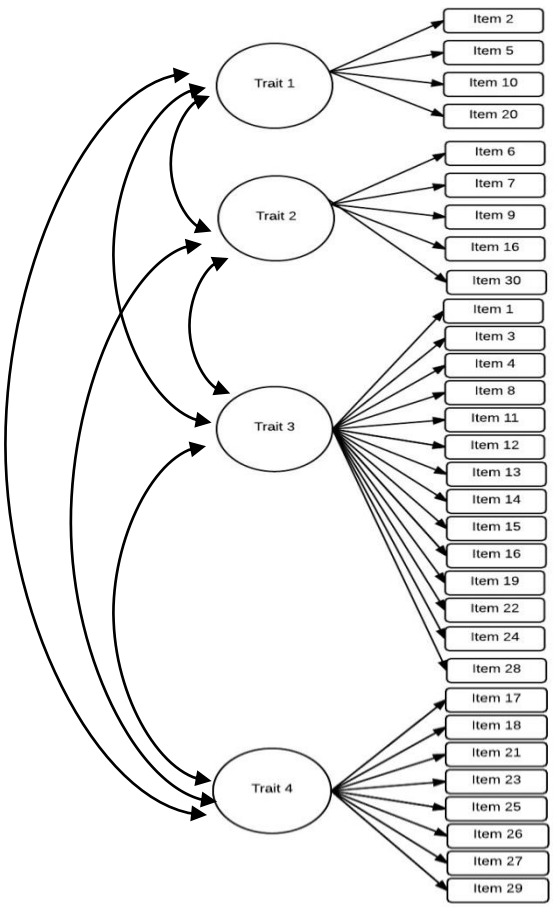

Figure2 *Model 2: the four specific traits model*.

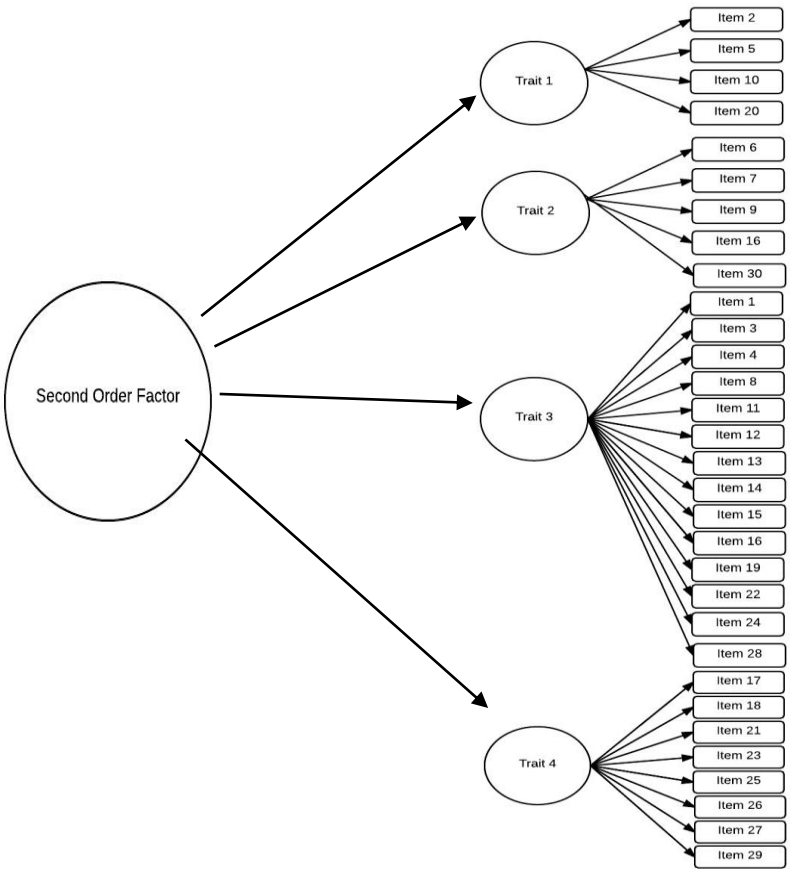

Figure3 *Model 3: the general plus the four specific traits model*.

Results and Discussion

The descriptive statistics for the two tests are given in Table 1. The mean percentage correct score of the general reading test (71.94%) was much higher than that of the rational cloze test (46.37%), suggesting that the participants on average performed better on the former test than on the latter test. The finding was not unexpected, as it was harder for the participants to supply words for the blanks in the rational cloze test, in contrast to merely choosing the correct answer from among the given options in the general reading test.

Table 1 *Summary of description statistics for the two tests (N=713).*

Test	Maximum possible score(%)	M (%) range	SD	Obtained score range (%)
Reading	35(100%)	25.18(71.94%)	5.38	7 (20%) - 35 (100%)
Cloze	25(%)	13.91(46.37%)	6.84	0 (0%) - 29 (96.67%)

In addition, the criterion-related validity of the rational cloze test was also obtained by calculating the Pearson product-moment correlation coefficient between its test scores and the scores of its criterion-measure used in the present study, the general reading test (i.e., CPET). A significantly positive correlation coefficient (r = .60, $p < .01$) was found, lending additional evidence to substantiate the claim that the two tests appeared to tap somewhat similar, yet not exactly identical aspects of the general reading proficiency construct.

The three models were tested through the CFA on item-level dichotomously-scored raw data using *Mplus* in response to the major purpose of the present study: to find out whether there is additional evidence that can be observed to substantiate Bachman's claim that the rational cloze test can be designed to measure an array of reading skills as intended. The results are presented and discussed in the following paragraphs.

Table 2 summarized various overall fit indices that are commonly used for

model evaluation and selection, including the χ^2 (chi-square) test of significance and various goodness-of-fit indices, *CFI*, *TLI*, and *RMSEA*. The χ^2 (chi-square) goodness-of-fit test is sensitive to the sample size. The larger the sample sizes, the higher probability the test will produce significant results, suggesting that the model does not fit the data (Schuster, Hammitt, & Moore, 2003). Therefore, multiple fit indices are recommended to determine the model fit. For example, values greater than .95 for the *CFI* and *TLI* are usually required to represent a good fit between the data and the hypothesized model (Hu & Bentler, 1999; Yu, 2002). As to the *RMSEA*, values less than .05 suggest a close fit, and values as high as .08 indicate an acceptable fit (Burns & Patterson, 2000; Jöreskog & Sörbom, 1993). Accordingly, as shown in Table 2, the fit of Model 1 (i.e., the one general factor model) was not satisfactory, as indicated by *CFI* = .81, *TLI*= .88, and *RMSEA* = .10. In other words, Model 1 fit the data least satisfactorily, strongly indicative of a lack of evidence in support of the hypothesis that the intercorrelations among all the items in the cloze test are attributable to the general reading proficiency factor. By contrast, both Model 2 (i.e., the four specific traits model) and Model 3 (i.e., the general factor plus the four specific traits model) revealed a better fit to the data. In particular, the *CFI*, *TLI*, and *RMSEA* values for Model 2 were .88, .95, and .07, respectively. With slightly better fit than Model2, the *CFI*, *TLI*, and *RMSEA* values for Model 3 were .90, .95, and .07, respectively. Based on the overall fit indices presented in Table 2, Model 1 was discarded and the other two models were used for further examination.

Table 2 *Fit indices for the three CFA models*.

Model	Fit indices					
	χ^2	*df*	χ^2/df	*CFI*	*TLI*	*RMSEA*
1	1007.34	120	9.39	.81	.88	.10
2	714.17	166	4.30	.88	.95	.07
2A	462.74	170	2.72	.94	.97	.05
3	616.33	149	4.14	.90	.95	.07
3A	395.61	151	2.62	.95	.98	.05

Note. Model 2A referred to Model 2 with the restrictions on zero residual correlations among a few indicators being removed. Likewise, the same removal was also applied to Model 3A.

For further fit improvement of the two models, modification indices were examined to see whether the restrictions on the corresponding zero residual correlation parameters could be removed. However, as warned by MacCallum, Roznowski, & Necowitz (1992), all modifications made to an original model have to be substantively meaningful and justifiable" (p. 491). Therefore, it turned out substantively sound and interpretable restrictions were removed for only seven zero residual correlation parameters. An example of the restriction removal was for the residual correlation parameter between item 15 and item 19. As shown in Appendix 1, the phrase "automatic control," which was the topic of the whole text, appeared in the first sentence of the first paragraph and continued appearing three more times before the two items. The answer to item 15 was "automatic," followed by an immediate clue "control"; similarly, the answer to item 19 was "control," preceded by another immediate clue "automatic." Perhaps the high co-occurrence of "automatic" and "control" in the text was the possible reason for the high residual correlation between the two items. Another reason for the correlated residuals might be due to co-text (Cai, 2013). According to Selivan (2013), co-text refers to the surrounding words in which a word is used, and the most apparent manifestation of co-text is collocations. Specifically, it was possible that the participants could simply guess and predict the word "control" from its co-text, "automatic," since one common noun collocate of "automatic" is "control." Similarly, they could also supply the word "automatic" before the word "control" as "automatic" is one common adjective collocate of "control." Collectively, the two possible reasons also held up for why most of the remaining six residual correlation parameters were freed to be estimated. As such, allowing the seven residual correlation parameters to be correlated led to two modified models, Model 2A and Model 3A. As shown in Table 2, the overall fit indices for the two modified models were more satisfactory than those for their corresponding unmodified models. Take the obtained *CFI*s for example. A *CFI* value of .95 was reported for Model 3A; it was higher than the one for Model 3, .90. Similarly, a *CFI* value of .94 was obtained for Model 2A; it improved noticeably on Model 2, whose *CFI* value only reached .88. Additionally, the fits of Model 2A and Model 3A being examined and compared, Model 3A seemed to provide better fit to the current data. Specifically, the *CFI*, *TLI*, and *RMSEA* values for Model 2A were .94, .97, and .05, respectively. Model 3A showed small degree of improvement on the fit with an increasing value of .01 in both the *CFI* and *TLI*, compared to the fit of Model 2A. Moreover, according to Chen, Sousa,

& West (2005), second-order factor models can explain the data in a more parsimonious way with fewer parameters, in comparison to first-order models with correlated factors. Hence, Model 3A, which was an example of second-order factor models, was chosen as the best fitting model, even though the overall fit advantage of Model 3A over Model 2A was found to be only slight.

In addition to the overall fit indices, the composite reliability coefficient (*CRC*), the average variance extracted (*AVE*), and the standardized factor loading (*SFL*) were also computed for both of the two modified models. However, given that Model 3A was shown above to represent the slightly better fit than Model 2A and that parameter estimates of Model 2A closely resembled those of Model 3A, Table 3 only presented the *CRC*, *AVE*, and *SFL* results from Model 3A. The *CRC* serves as an overall measure of each latent trait's reliability. As shown in Table 3, the *CRC* for each of the four traits and the general factor, ranging from .80 to .87, met the minimum acceptable criteria of .6 (Bagozzi & Yi, 1988). The *AVE* of a latent trait, as defined by Fornell & Larcker (1981), reflects the amount of variance that is captured by its indicators relative to the amount due to measurement error. With a range from .43 to .63, the *AVE*s for the traits and the second-order general factor, were either greater or close to the minimum Benchmark of .4 (Diamantopoulos & Siguaw, 2000). Accordingly, the values of *CRC* and *AVE* lent some support to the reliability and validity of the rational cloze test. Furthermore, all of the factor loadings for Model 3A were significantly different from zero, with $p < .01$. According to Fornell & Larcker (1981), a value greater than .50 is typically desired for the *SFL* because it suggests that the majority of the variance in the indicator/item can be accounted for by the latent trait. The close scrutiny of the *SFL*s for each of the four specific traits indicated that except for items 2 and 22, the values of the *SFL* for all the remaining 28 items were greater .50. Interestingly, the *SFL*s on Trait 4 were by far the highest observed, with values ranging from .61 to .89. The values of the *SFL* for the majority of the items were substantial, indicating that most of the items were all reasonably good indicators of their corresponding hypothesized trait. Likewise, all of the four 1st order traits loaded heavily on the 2nd order trait, with the values of the factor loading ranging from .75 to .83, suggesting that the general language proficiency factor (i.e. the 2nd order factor) could be well explained by the four specific traits (i.e., the 1st order factor).

Table 3 *CRC, AVE, and SFL of the traits for Model 3A.*

1st/2nd - order trait	*CRC*	*AVE*	Item/1st - order trait	*SFL*
Trait 1	.80	.50	2	.48
			5	.76
			10	.80
			20	.75
Trait 2	.80	.45	6	.74
			7	.70
			9	.61
			16	.73
			30	.57
Trait 3	.86[a]	.43	1	.58
			3	.72
			4	.64
			8	.78
			11	.68
			12	.59
			13	.52
			14	.68
			15	.59
			19	.58
			22	.48
			24	.90
			28	.70
Trait 4	.87[a]	.56	17	.77
			18	.82
			21	.88
			23	.61
			25	.72
			26	.69
			27	.81
			29	.66
General	.87	.63	Trait 1	.79
			Trait 2	.80
			Trait 3	.83
			Trait 4	.75

Note. [a] Because there were correlated errors in these latent traits, an adjusted formula (Bollen, 1980; Kano & Azuma, 2003) was applied to calculate the *CRC*

values for these traits.

Table 4 displayed the estimated correlation coefficients derived from Model 3A among the four specific traits. The correlation coefficients among the four specific traits were medium, ranging from .59 to .67, indicative of moderate strength of relationship among them. The strongest relationship (.67) was found between Trait 2 and Trait 3. This finding was not unexpected because the items for both of the two traits were intended to tap the participants' cohesive and coherent competence—with across-clause, within-sentence closure context requirement for Trait 2 and across-sentence closure context requirement for Trait 3. Another key observation to be made in Table 4 was that Trait 4 (i.e. extra-textual inference) displayed a slightly lower relationship to each of the other three traits. Again, this finding was not surprising as the ability to make extra-textual inference correctly usually would involve more diverse aspects of ability, such as test-takers' background knowledge, vocabulary size, or amount of reading outside language classes, etc.; in contrast, the other three traits involved only their grammatical knowledge and understanding of the textual relationship across clauses and sentences.

Table 4 *Correlations among the four specific traits for Model 3A*.

$1^{st}/2^{nd}$ - order trait	Trait 1	Trait 2	Trait 3
Trait 2	.63		
Trait 3	.65	.67	
Trait 4	.59	.60	.63

To sum up, the results of the present study were consistent with those of Bachman's (1982), in favor of the third model, the second-order model. In other words, the results corroborated Bachman's (1982) claim that rational cloze tests can be designed to measure distinct and related language abilities, all of which could be accounted for by the higher-order factor. Furthermore, the findings of the current study refuted the claim made by Alderson (1979) that cloze tests are capable of measuring only lower-order or basic proficiency skills, but aligned nicely with the findings from Anderson's (1980) study that cloze tests are capable of measuring test-takers' sensitivity to not only sentence-level grammatical structures but also cohesive relationships across sentences.

Conclusions

The results of the present study have lent support to the hypothesized general factor plus specific traits model, stating that rational cloze tests can be designed to tap separate and specific language competence, such as syntactic competence, cohesive and coherent competence beyond clause or sentence boundaries, and extra-textual inference, all of which could be explained by the general language proficiency factor. That is, the second-order factor of the general language proficiency underlay and accounted for the commonality or the pattern of relations among the four first-order specific language components. To put it in another way, the responses to the measurement of the general language proficiency construct could be explained by the four seemingly distinct but related first-order specific traits. This conclusion was supported by the good overall fit of the second-order model, the reasonably high factor loadings of the items on the first-order traits, the substantial high factor loadings of first-order traits on the general language factor, and the principle of parsimony.

The current study may shed some light on the longstanding theoretical discussion and numerous empirical investigations regarding the nature of language competence. The earliest theoretical framework dated back to the structuralist school of linguistics, which posited the existence of divisible components of language proficiency and took the view that learning a language involves mastering its separate elements or components (Fries, 1945; Lado, 1961, 1964). In addition to the theoretical descriptions of the components of language competence, an extensive body of research (e.g. Carroll, 1975; Gardner & Lambert, 1965; Hosley & Meredity, 1979; Lofgren, 1969; Pimsleur, Stockwell, & Comrey, 1962) has been carried out on the dimensionality of language proficiency. In the 1970s emerged a new view arguing that language proficiency consists of one general factor, based on the obtained empirical findings of the high intercorrelations among different types of language tests (e.g. Oller, 1983; Scholz, Henricks, Spurling, Johnson, & Vandenburg, 1980). However, criticism against this view was severely leveled because the interpretation of the high intercorrelations as evidence for a general factor was made based on the inappropriate use of the principle components analysis procedure, where the extracted factors included unique and common factor variances (Carroll, 1983; Farhady, 1983; Fouly, Bachman, & Cziko, 1990). Out of the dissatisfaction with this analysis technique, numerous studies (e.g. Carroll, 1983; Bachman& Palmer, 1981, 1982; Bachman, 1982; Farhady, 1983) were performed using the CFA

and the principal axis factoring method, both of which depended on only common variances in the extraction of factors. These studies not only turned down the general factor model and the previous divisible traits model, but also brought about two other hypotheses concerning the nature of language competence: the correlated-traits hypothesis, and the higher-order factor hypothesis. The former stated that the separate but correlated traits underlie the performance on language tests, while the latter stated that the separate but correlated traits are influenced by a single higher-order factor. The findings of the present study rendered additional support to the higher-order hypothesis. That is, in addition to the differentiated but related language skills, there exists a higher-order general language factor, although the nature of this general factor remains open to speculation and awaits further research (Fouly, Bachman, & Cziko, 1990; Oller, 1983). Indeed, the additional evidence from the present study may signal strong calls for more thorough and systematic in-depth investigation adopting qualitative inquiry approaches to extend the understanding about the nature of the general second-order language factor, so as to better conceptualize what the construct precisely is.

Besides, the findings of the present study also had some practical implications specifically for language teachers or language test constructors. First and foremost, rational cloze tests, based on the aforementioned satisfactorily high composite reliability estimates and reasonably high factor loadings, can be one of the good candidates for the consideration in choosing among numerous item formats to assess various distinct but related reading traits intended: syntactic competence, cohesive competence, and coherent competence. In other words, given that the latter two types of competence are perceived as higher-order comprehension skills, the findings of the present study, coupled with the ease of construction and administration, suggest that rational cloze tests can also be a useful option for classroom teachers who need to make reliable and valid tests to assess a wide range of purported reading skills, including higher-order, coherence-oriented comprehension skills. Another practical implication from the results of the present study concerns the residual correlations identified for the seven pairs of items, whose correct answers were the exact word used repeatedly for the topic (i.e. automatic control) of the whole passage. In light of this undesirable finding, classroom language teachers or test constructors are advised to avoid selecting words surrounding the co-text or words of high co-occurrence in the text for deletions, in order to guard against unwanted measurement of this sort. As suggested

by Cai (2013), for the purpose of ensuring good construct validity, blanks have to be deliberately designed to prevent test-takers from easily predicting and guessing from the "co-text" (p.195).

On a final note, this study was somewhat limited in the generalizations that could be made. This limitation, mainly resulting from the research design involved in the present study, deserves some discussion because it may point to a potential avenue for future research. Specifically, only one text of expository nature was adopted in the present study. As stated earlier, previous research (Klein-Braley, 1981; Zarrabi, 1988) has demonstrated that cloze tests with different texts would produce different tests, each of which measures different aspects of language abilities. Moreover, it is commonly believed that the way cohesive and coherence are achieved varies across texts with different genres, which may influence readers' success in comprehending the texts while they read (Boshrabadi, Biria, & Hodaeian, 2014; Hoey, 1991; Taboada, 2004). Hence, given that the present study employed only a single text, a need is definitely in order for future research to incorporate multiple texts with different genres, so as to find out whether the fit of the second-order model can be generalized to various texts of different genres.

References

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

Anderson, D. C. (1980). *Cohesion and the cloze test* (Unpublished master's thesis). University of Illinois.

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

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*, 535-555.

Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning*, *31*, 67-86.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449-465.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*(1), 74-94.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*, 78-102.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides (Ed.), *New developments and techniques in structural equation modeling* (pp. 269-296). Mahwah, NJ: Lawrence Erlbaum.

Beinborn, L., Zesch, T., & Gurevych, I. (2015). Candidate evaluation strategies for improved difficulty prediction of language tests. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1-11). Stroudsburg, PA: Association for Computational Linguistics.

Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review, 45*, 370-390.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Boshrabadi, A. M., Biria, R., & Hodaeian, M. (2014). A contrastive analysis of the links of textuality in abstracts written by Persian and English writers in clinical psychology journals. *International Journal of Applied Linguistics and English Literature*, *3*(4), 136-142.

Burns, G. L., & Patterson, D. R. (2000). Factor structure of the Eyberg Child Behavior Inventory: A parent rating scale of oppositional defiant behavior toward adults, inattentive behavior, and conduct problem behavior. *Journal of Clinical Child Psychology*, *29*(4), 569-577.

Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing*, *30*(2), 177-199.

Cambridge English Language Assessment. (2015). *Cambridge English research and*

validation. Retrieved from http://www.cambridgeenglish.org/principles/

Carroll, J. B. (1975). *The teaching of French as a foreign language in eight countries*. New York, NY: Halsted.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, MA: Newbury House Publishers.

Chapman, I. J. (1983). *Reading development and cohesion*. London: Hememann.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*(3), 471-492.

Crowhurst, M. (1987). Cohesion in argument and narration at three grade levels. *Research in the Teaching of English, 21*, 185-201.

Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL: A guide for the uninitiated*. Thousand Oaks, CA: Sage.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. J*ournal of Psychoeducational Assessment, 23*, 225-241.

Farhady, H. (1983). On the plausibility of the unitary language factor. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 11-28). Rowley, MA: Newbury House Publishers.

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

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Market Research*, *18*, 39-50.

Fouly, K. A., Bachman, L. F., & Cziko, G. A. (1990). The divisibility of language competence: A confirmatory approach. *Language Learning, 40*(1), 1-21.

Frank, M. (1993). *Modern English: A practical reference guide* (2nd ed.). Englewood Cliffs, NJ: Regen/Pearson Hall.

Fries, C. C. (1945). *Teaching and learning English as a foreign language*. Ann Arbor, MI: University of Michigan Press.

Fulcher, G. (1989). Cohesion and coherence in theory and reading research. *Journal of Research in Reading*, *12*(2), 146-163.

Gardner, R., & Lambert, W. E. (1965). Language aptitude, intelligence, and second language achievement. *Journal of Educational Psychology*, *65*, 191-227.

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

Hampton, S. (2010). *The importance of writing structures, coherence, and cohesion to writing and reading*. Retrieved from http://www.reading.org/libraries/book-supplements/bk767supp-hampton.pdf

Harsch, C., & Hartig, J. (2015). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing, 32*(1), 1-21.

Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding reading comprehension* (pp. 181-219). Newark, DE: International Reading Association.

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

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Horning, A. (1991). Readable writing: The role of cohesion and redundancy. *Journal of Advanced Composition, 11*(1), 135-145.

Hosley, D., & Meredith, K. (1979). Inter- and intra-test correlates of the TOEFL. *TESOL Quarterly*, *13*, 209-217.

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

Irwin, J. (1986). Cohesion and comprehension: A research review. In J. Irwin (Ed.), *Understanding and teaching cohesion comprehension* (pp. 31-43). Newark, DE: International Reading Association.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale: NJ: Lawrence Erlbaum Associates, Inc.

Kano, Y., & Azuma, Y. (2003). Use of SEM programs to precisely measure scale reliability. In H. Yanai, A. Okada, Y. Shigemasu, J. J. Kano, & K. Meulman (Eds.), *New developments in psychometrics* (pp. 141-148). Tokyo, Japan: Springer Verlag.

Khodadady, E. (2012). Validity and tests developed on reduced redundancy, language components and schema theory. *Theory and Practice in Language Studies*, *2*(3), 585-595.

Kim, S., & Hagtvet, K. A. (2003). The impact of misspecified item parceling on representing latent variables in covariance structure modeling: A simulation study. *Structural Equation Modeling*, *10*, 101-127.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163-182.

Klein-Braley, C. (1981). *Empirical investigations of cloze tests* (Unpublished doctoral dissertation). University of Duisburg, Duisburg, Germany.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*, 47-84.

Kolln, M. (1999). Cohesion and coherence. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: The role of teacher's knowledge about text, learning, and culture.* Urbana, IL: National Council of Teachers of English.

Lado, R. (1961). *Language testing*. New York, NY: McGraw-Hill.

Lado, R. (1964). *Language teaching: A scientific approach*. New York, NY: McGraw- Hill.

Lee, S. H. (2008). Beyond reading and proficiency assessment: The rational cloze procedure as stimulus for integrated reading, writing, and vocabulary instruction and teacher-student interaction in ESL. *System, 36*, 642-660.

León, D. A. D. (2011). *Anχálise factorial confirmatória através dos softwares Re Mplus* (unpublished manuscript). Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

Lofgren, H. (1969). *Measuring proficiency in the German language: A study of pupils in grade 7* (Didakometry No. 25). Malmo, Sweden: School of Education.

Lonigan, C. J., Hooe, E. S., David, C. F., & Kistner, J. A. (1999). Positive and negative affectivity in children: Confirmatory factor analysis of a two-factor model and its relation to symptoms of anxiety and depression. *Journal of Consulting and Clinical Psychology, 67*, 374-386.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9*(3), 369-403.

Min, Y.-K. (n.d.). *ESL: Coherence and cohesion*. Retrieved from http://www.bothell.washington.edu/wacc/for-students/eslhandbook/coherence

Muthén, L., & Muthén, B. (1998-2008). *Mplus user's guide*. Los Angeles, CA: Author.

Ohnmacht, F. W., Weaver, W. W., & Kohler, E. T. (1970). Cloze and closure: A factorial study. *The Journal of Psychology, 74*, 205-217.

Oller, J. W. Jr. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*, 105-118.

Oller, J. W. Jr. (1983). Evidence for a general language proficiency factor and expectancy grammar. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 3-28). Rowley, MA: Newbury House Publishers.

Oller, J. W., & Conrad, C. A. (1971). The cloze techniques and ESL proficiency. *Language Learning*, *21*, 183-195.

Pimsleur, P., Stockwell, R., & Comrey, A. (1962). Foreign language learning ability. *Journal of Educational Psychology*, *53*, 15-26.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading*, *11*(4), 289-312.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. In J. S. Johnson (Ed.), *Spaan fellow working papers in second or foreign language assessment* (Vol. 1, pp. 39-82). Ann Arbor, MI: The University of Michigan.

Scholz, G., Henricks, D., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Is language ability divisible or unitary? In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 24-33). Rowley, MA: Newbury House Publishers.

Schuster, R. M., Hammitt, W. E., & Moore, D. (2003). A theoretical model to measure the appraisal and coping response to hassles in outdoor recreation setting. *Leisure Sciences*, *25*, 277-299.

Selivan, L. (2013, May 5). In context or with co-text? [Web log comments]. Retrieved from http://leoxicon.blogspot.tw/2013/05/context-or-cotext.html

Spolsky, B., Bengt, S. M., Sako, E. W., & Aterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. In J. A. Upshur & J. Fata (Eds.), *Problems in foreign language testing, language learning special issue* (pp. 79-103). Ann Arbor, MI: Language Learning Research Club, University of Michigan.

Taboada, M. T. (2004). *Building coherence and cohesion*. Amsterdam: Benjamins.

Weaver, W. W., & Kingston, A. J. (1963). A factor analysis of the cloze procedure and other measures of reading and language ability. *The Journal of Communication, 13*(4), 252-261.

Williams, J. M., & Colomb, G. G. (2010). *Style: Lessons in clarity and grace* (10th ed.). Boston, MA: Pearson Education, Inc.

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication, 32*, 198-204.

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267-293.

Yu, C-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles, CA.

Zarrabi, A. (1988). *Deletion rate and test difficulty in cloze* (Unpublished master's thesis). Allameh Tabatabai University, Tehran, Tehran, Iran.

Appendix A

Rational cloze test form

The science of automatic control depends on certain common principles by which an organism, machine, or system regulates itself. Many historical developments up to the present day have helped to identify these principles.

For hundreds of years there are many (1) of automatic control systems, but no connections were recognized among them. A very early example was a device on windmills designed (2) keep their sails facing into the wind. (3) consisted simply of a miniature windmill which rotated the whole mill to face in any direction. (4) small mill was at right angles to the main (5), and whenever the latter faced in the (6) direction, the wind caught to the small mill's sails and rotated the (7) mill to the correct position. (8) automatic control mechanisms were invented with the development of steam power: first, the engine governor, (9) then the steering engine controller, (10) operated a ship's rudder in correspondence with the helm. These (11) and a few others constituted the achievement of the (12) of automatic control, up to about 50 years ago. In the past (13) decades, however, rapid technological development has created numerous urgent and complex (14). The solutions to these problems have given birth to new families of (15) control devices. For example, chemical plants needed (16) for both temperature and flow; homes needed controls for complex (17) and cooling systems; radios required control circuits which would (18) the accuracy of signals.

Historically, then, the modern science of automatic (19) has been aided by related advances in many fields. (20) now seems surprising to recall that the relationships among these developments were not originally (21). Yet we know that (22) control and regulating systems depend on common (23) which are found in both nature and human affairs.

Indeed, (24) of modern and old automatic control systems give us new insight into a wide (25) of natural and human phenomena. The results of these studies, have been very (26) in understanding how (27) is able to walk upright, how the (28) heart beats, why our economic (29) suffers from slumps and booms, and (30) the rabbit population in parts of Canada regularly fluctuates between scarcity and abundance.

Note. The rational close test was directly taken from Bachman, 1985.

Acceptable answers and type of traits for the rational cloze test

Item	Exact word	Acceptable Answer	Trait
1.	examples	kinds, types	3
2.	to		1
3.	It	this	3
4.	The	this	3
5.	one	mill, windmill	1
6.	wrong	incorrect	2
7.	main	whole, large, other	2
8.	Other	many, two, then, new	3
9.	and		2
10.	which	that	1
11.	mechanisms	inventions, devices, systems, examples, three, developments	3
12.	science	technology, field	3
13.	five	few, several	3
14.	problems		3
15	automatic		3
16.	controls	control, regulation	2
17.	heating	warming	4
18.	guarantee	control, regulate, maintain, monitor, assure	4
19.	control		3
20.	It		1
21.	recognized	known, discovered, found, noticed, understood, evident	4
22.	automatic	these, all	3
23.	principles	properties, laws	4
24.	studies		3
25.	variety	range	4
26.	helpful	useful, important	4
27.	person	man, human, child	4
28.	human		3
29.	system	development, condition	4
30.	why	how	2

Note. Trait 1 refers to syntactic competence that depends on within-clause context; Trait 2 refers to cohesive and coherent competence that depends on across-clauses but within-sentence context; Trait 3 refers to cohesive and coherent competence that depends on across-sentences context; Trait 4 refers to extra-textual inference.

專論

理性刪除克漏字測驗效度的重新審視: 驗證性因素分析方法的應用

林甫雯* 林文鶯** 游錦雲***

此研究旨在證實容易編寫的理性刪除克漏字測驗（rational cloze test）可以很有信度及效度地用來測試課堂上學生的各種不同閱讀技能。此研究藉由驗證性因數分析方法（confirmatory factor analysis），重新探討 Bachman（1982）所提出的主張：理性刪除克漏字測驗是可以用來測試課堂上學生的各種不同閱讀技能，尤其是銜接理解能力及連貫理解能力。713 位來自臺灣北部四所大學的同學，在研修一般性英文課程的課堂上，填答了 Bachman（1985）所編寫一份的理性刪除克漏字測驗。其結果顯示本研究的數據與整體特質加四個細項特質的二階模式比較吻合；也就是說，理性刪除克漏字測驗是可以用來測試課堂教師想要考的各種不同閱讀技能，如文法能力、跨子句的銜接及連貫理解能力、跨句子的銜接及連貫理解能力、及篇章文字線索外的推論能力，而且這四個細項能力可以用二階整體特質解釋。

關鍵詞：理性刪除克漏字測驗、驗證性因數分析方法、二階因素、銜接理解能力、貫理解能力

* 作者現職：國立海洋大學應用英語研究所助理教授
** 作者現職：臺北市立大學英語教學系副教授
*** 作者現職：臺北市立大學心理與諮商系副教授

通訊作者: 林文鶯，e-mail: wylin@utaipei.edu.tw