Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 12, No. 1, March 2022
ISSN 2476-5880

Reshaping Content Validation: A Case Study of a Reading Achievement Test for Third-year English Majors

Huu Thanh Minh Nguyen[1]

Received: November 2021 Accepted: December 2021

Abstract
Combining Bachman's (1990) conceptualization of content validity with Messick's (1989) unifying model of construct validity, this study attempted to fill the gap in researching content validation in the context of a university reading achievement test by (a) examining traditional content validity evidence and (b) analyzing the test scores to either back-up or rebut the evidence and explore the construct of the underlying test structure. On a sample of 477 third-year English-majored test takers at a Central Vietnamese University (CVU), the study was conducted in the pre-test stage where the test content was compared with the test specification and the post-test stage where the test scores were processed via Rasch and CFA competing model analyses. Results showed that while content relevance was satisfactory and the test construct was to a large extent not threatened by construct-irrelevant variance, content coverage of the test remained problematic with instances of construct underrepresentation. Moreover, the study supported the one-factor model of general reading ability as the underlying structure of the reading achievement test over the correlated three sub-skill and higher-order factor models. The study bears implications for both test writers in designing and test takers in performing the L2 reading test.

Keywords: Content validity; Reading construct; One-factor model; Reading achievement test; English majors

1. Introduction

An achievement test is a tool in classroom assessment to determine whether learning outcomes can be reached at the end of a course (McNamara, 2000). It should adhere to qualities of validity, i.e., the extent to which a test measures what it intends to (Hughes, 2003), one of which is content validity, i.e., "representativeness or sampling adequacy of the content – the substance, the matter, the topics – of a measuring instrument" (Kerlinger, 1973, p. 458). During test design, much weight should be attached to the content because "demonstrating that a test is relevant to and covers a given area of content or ability is a necessary part of validation" (Bachman, 1990, p. 244). However, content validity is not sufficient as it does not consider the real test takers' performance (Bachman, 1990). This limitation makes it impossible to infer the

[1] Faculty of English, University of Foreign Language Studies, The University of Danang.
Email: nhtminh@ufl.udn.vn

 ISSN 2476-5880

mastery level of course objectives based on score interpretations of test takers. Therefore, the notion of construct validity is developed via Messick's (1989) unifying model – construct includes specific domains of the language ability intended to assess (Bachman & Palmer, 2000). This model embeds the evidence for content validity into the construct validity as a central concept and concerns validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). As such, this validation process based on accumulating evidence to support score-based interpretations is referred to as the "evidence-gathering approach" (Chapelle & Voss, 2013). Taken together, to address the omission of test takers' performances in content validation, this paper will reshape content validation based on combining Bachman's (1990) conceptualization of content validity and Messick's (1989) unifying model of construct validity into a concerted pursuit of examining the evidence from the test content and the test scores of a reading achievement test for third-year English majors at a Central Vietnamese University (CVU).

Validation of reading tests has concerned language testing researchers and educators (e.g., Ahmadjavaheri & Zeraatpishe, 2020; Alderson, 2000; Sheybani & Zeraatpishe, 2018; Tonekaboni et al., 2021). As stated by Messick (1989) above, the purpose of validation is to evaluate if the interpretation and use of the test scores are meaningful and are sufficiently informed by relevant empirical evidence and theoretical considerations. Therefore, in order to examine the validity of the interpretation and use of reading tests, it is essential to define the reading ability that the test is intended to measure, which is usually referred to as *construct definition* in second language assessment (Alderson, 2000). Reading construct viewed from the reader purpose perspective involves readers employing various sub-skills and cognitive processes to achieve reading purposes in their mind (Alderson, 2000; Enright et al., 2000; Rupp et al., 2006). However, the constructs of reading have been defined and operationalized from different perspectives and there are no one-size-fits-all reading purposes as different scholars adopt various approaches to conceptualizing reading purposes (Grabe, 2009; Linderholm & van den Broek, 2002; Urquhart & Weir, 1998). This is because different tests used in different contexts give rise to variations in reading purposes. Therefore, in this paper, apart from the content validation attempts, the reading construct of the reading achievement test used for third-year English majors at the CVU will also be explored based on the score interpretation.

2. Research Context

At the CVU where English majors are educated, English language skill training is skill-based, i.e., each language skill of listening, reading, writing, and speaking is instructed in a designated module. Those provisions are offered until the end of the third year of the students' program. For third-year English majors who have passed prerequisite skill-based modules at B1 and B2 levels of the Common European Framework (CEFR) in the first two years, the third-year reading module syllabus targeted at the C1 level seeks to build up a wide range of reading skills for different genres of texts ranging from personal, social to academic and professional areas. Each instructed reading skill takes up one unit in the syllabus, accompanied by corresponding

 ISSN 2476-5880

tasks to practice. Table 1 presents seven sub-skills for the construct of reading comprehension in the module and their description.

Table 1.
Reading sub-skills and description

Sub-skills	Description
Reading for main ideas	- Summarizing and understanding the main idea of a paragraph. - Summarizing and understanding the main idea of a text.
Reading for details	- Locating specific details in the text based on simple grammatical structures and elicited vocabulary in the question. - Understanding the specific details stated explicitly in the text.
Reading for reference	- Identifying pronoun antecedent. - Understanding logical relationships of ideas in the text based on cohesive devices, e.g., referent words, conjunctions, and linking words.
Reading for inference	- Inferring the situational meaning of a stance. - Identifying an implicit stance with different wording.
Understanding vocabulary in contexts	- Inferring word meaning in contexts - Inferring the meaning of idiomatic expressions in contexts.
Understanding pragmatic meaning	- Understanding the author's purpose, attitude towards an issue in a text - Understanding the author's attitude through detail in a text - Understanding the general tone of a text
Understanding the organizational pattern	- Identifying the structure of a paragraph and/or a text - Inserting a missing sentence into a paragraph in a coherent manner

(Department of English, 2017, p. 1)

Based on the module content, module instructors develop the specification of the reading achievement test to guide its design. According to the test specification, the test assesses students' reading proficiency in accordance with the seven domains or sub-skills of the syllabus and reflects each student's achievement through a single end-of-course score. The test consists of four passages, with increasing difficulty levels. There are 40 dichotomously-scored test items (hereinafter called "items") with four options, with 10 items for each passage. The genres of all passages should vary, including social, academic, professional areas; and require no prior background knowledge. Passages 1 and 2 should have a length from 300 to 400 words while passages 3 and 4 should have a length from 500 to 600 words. Regarding lexical features, passages 1 and 2 should contain more high-frequency words than passages 3 and 4 – high-frequency words are the most frequent 3000 words (Nation, 2013). Regarding syntactic features, passages 1 and 2 should accommodate a combination of simple, compound, and complex sentences while passages 3 and 4 should have most sentences as compound and complex.

To be valid in its use as an achievement test, the test content should measure what it purports to measure (Henning, 2001), i.e., the alignment of test content with test specification

 ISSN 2476-5880

informed by the syllabus. Moreover, validity, in general, refers to "inferences or actions made on the basis of the test scores" (Sireci, 2015, p. 4), considering whether the interpretation of test scores can be supported by the theory for the proposed test use. In other words, the analysis and interpretations of test scores can pinpoint whether the test scores may confirm the construct as related to and representative of what is tested. Because of great concern about reading test validity (Alderson, 2000; Bachman & Palmer, 1996; Tengberg, 2018), in this paper, content validity of the reading achievement test for third-year English majors at the CVU will be evaluated in relation to the test content and interpretation of test scores.

3. Review of Literature

3.1. Content validity and its operationalization

Content validity refers to whether the test content is "sufficiently representative and comprehensive for the test to be a valid measure of what it is supposed to measure" (Henning, 2001, p. 91). Bachman (1990) views content validity from the perspectives of content coverage and content relevance. Content relevance is concerned with "the specification of the behavioral domain in question and the attendant specification of the task or test domain" (Bachman, 1990, p. 244). That is associated with whether the items reflect the construct and the targeted domains to be measured as well as the nature of responses that test takers are expected to make. Content coverage is "the extent to which the tasks required in the test *adequately represent* the behavioral domain in question" (Bachman, 1990, p. 245, emphasis added). It is underscored at the proportionate emphasis on the domains to ensure their balanced occurrence in the test. In other words, the amount of coverage allocated to each domain should ensure the domain's significance.

Those pure attributes of content validity, however, do not consider test takers' performance (Bachman, 1990); therefore, they cannot make inferences about whether the construct tested can measure how much students master the stated course objectives based on their score interpretation. To fill this gap, Messick (1989) develops a unifying framework for construct validity that also encompasses the evidence for content validity – construct validity refers to "the extent to which performance on tests is consistent with predictions we make on the basis of a theory of abilities or constructs" (Bachman, 1990, p. 254). This framework offers a more comprehensive view of content validation by not only examining it from the perspective of traditional content validity evidence but also incorporating the score-based interpretation of the construct tested.

Construct validity, central to the trustworthiness of score meaning and its interpretation, is threatened by two aspects, i.e., construct-irrelevant variance and construct under-representation (Messick, 1989). First, construct-irrelevant variance refers to some sub-dimensions of the test that contaminate the focal construct, which can be linked to the failure to achieve content relevance as one attribute of Bachman's (1990) content validity. Construct irrelevant variance can be characterized into two strands: construct-irrelevant difficulty and construct-irrelevant easiness (Messick, 1989). Construct-irrelevant difficulty results in invalidly low scores due to including tasks that are far beyond test takers' competences. Meanwhile, construct-irrelevant easiness garners test takers an advantage in answering the items that do not necessarily require their abilities to perform, which translates into invalidly high scores. The factors that may affect construct relevance in multiple-choice reading tests include but are not limited to length (Carr, 2011); familiarity with topics of passages (Leeser

2007; Shohamy, 1984); lexical variation, lexical sophistication, and lexical density (Hu & Nation, 2000; Read, 2000); syntactic complexity (Carr, 2006); genre (Basturkmen & Elder, 2004); language of stems (Brantmeier, 2006; Haladyna, Downing, & Rodriguez, 2002; Thorndike & Hagen, 1969; Wolf, 1993); and quality of distractors (Carr, 2011; Parkes & Zimmaro, 2016). Second, construct under-representation narrows "the breadth of content specifications for a test should reflect the breadth of the construct invoked in score interpretation" (Messick, 1989, p. 35). The test measurement cannot sufficiently cover the focal construct as intended, which may be linked to the failure to achieve content coverage as a Bachman's (1990) perspective of content validity evidence.

There has been a movement to focus on content validity, suggesting it should be examined through "the qualities of test content, the interpretation and uses of test scores, the consequences of proposed score interpretation and uses, and theory refinement" (Chalhoub-Deville, 2009, p. 242). In this vein, examining the qualities of test content refers to the traditional forms of content validity evidence (Sireci, 2009), i.e., content relevance and coverage. As a result, comparing test content with the test specification is done to determine the level of alignment in terms of relevance and coverage; in other words, whether the test content is threatened by construct irrelevant-variance and construct under-representation. This is undertaken before the test to gather priori validity evidence (Weir, 2005). After the test, scores are analyzed to collect posteriori validity evidence and see "the extent to which test results are consistent in terms of content sampling, and free from bias" (Weir, 2005, p. 23). In other words, the interpretation of scores at this stage is intended to re-evaluate priori validity evidence, i.e., to both assess whether it is threatened by construct irrelevant-variance and construct under-representation and show whether the test reflects its theoretical construct through score interpretation.

Taken together, this paper followed a process from the pre-test to the post-test stage. For the pre-test stage, the test content was compared with the test specification to judge the level of compatibility and whether it was theoretically related to and covered the targeted construct. For the post-test stage, the test scores were analyzed to confirm the evidence of content validity at the pre-test stage by seeing (a) whether the gathered data could prove the fitting of the items into the construct, (b) whether the items function well in terms of their difficulty to accurately measure the students' ability in achieving learning objectives, and (c) whether the test was reflective of the theoretical construct.

3.2. Previous studies about the construct of reading

Of equal importance as validating the reading test content is to examine the alignment between the theoretically proposed and empirically derived model of reading constructs. In this study, that refers to how reading sub-skills can explain the construct of reading to reveal the underlying structure of the test. This is important because a set of reading sub-skills is a guiding framework for material design and test construction (Alderson & Lukmani, 1989; Lumley, 1993; Song, 2008). Various taxonomies of reading sub-skills have been proposed with or without empirical justification for clear description and distinction of the construct of reading proficiency (Buck, Tatsuoka, & Kostin, 1997; Carroll, 1980; Munby, 1978), which leads to the lack of consistency or empirical evidence across researchers. However, one important concern is not about how many reading sub-skills can be incorporated into the construct itself, but rather how the reading sub-skills can be identified and classified into the construct (Alderson, 2000).

 ISSN 2476-5880

This informs no consensus in determining how reading skills can be classified or even if separable skills may exist (Alderson, 2000; Rupp, 2012; Tengberg, 2018). The construct of reading can be characterized into two main strands, either as a unitary or a divisible concept.

3.2.1. The unitary view. Schedl, Gordon, Carey, and Tang (1995) examined whether the items of the Test of English as Foreign Language (TOEFL) reading test differ in terms of the reasoning level, i.e., whether the sub-skills that the items measured went beyond linguistic and general discourse competence. They concluded that this distinction was not made because all the item types were fit into the same overall measurement as a testament to the unidimensionality of the TOEFL reading test. Sawaki, Stricker, and Oranje (2009) investigated the dimensionality of the TOEFL-iBT reading test by classifying three sub-skills or traits, i.e., basic comprehension, reading to learn and inferencing as well as employing three methods for three items sets of three reading passages in a Multitrait-Multimethod analysis of four models: A correlated trait correlated method model, a correlated trait uncorrelated method model; a correlated trait model; and a correlated trait correlated uniqueness model. The single trait model was found as the final solution to describe the construct of the TOEFL-iBT reading test, which echoes Koizumi and Nakamura's (2016) study. However, the recent investigation of Min, Kyoungwon, and Howard (2021) into the English for Academic Purposes reading assessment across 1-12th graders showed that while the unidimensional model satisfactorily captured the construct of reading assessment for grade 1, the higher-order multidimensional model fitted the data better, which was consistent with the previous second-order factor applications in language assessment (e.g., Cai & Kunnan, 2018; Lee, Kim, Choi, & Kang, 2019). This is because of the large sample size that favours the more complex high-order model (DeMars, 2013).

3.2.2. The multi-trait view. The multi-trait view of reading comprehension was explored in numerous studies of L2 reading. Lumley (1993) examined nine proposed sub-skills of an EAR reading test as suggested by five experienced teachers who also judged the difficulty level of those sub-skills and their respective items. Results were that the teachers not only reached a high level of agreement on the classification of the nine sub-skills but a significant correlation was also found between the teachers' perceptions of the difficulty of the sub-skills and the item difficulty measured by Rasch analysis. As regards the reading paper of Web-based English as a Second Language Placement Exam, Song (2008) proposed and empirically tested the two sub-skill model (understanding explicitly and implicitly stated information), and the three sub-skill model (understanding main ideas, understanding supporting details, and making inference), by using structural equation modeling. It was found that the two sub-skill model had better fit indices than the three sub-skill model. Kim (2009) hypothesized and empirically tested three models of L2 reading sub-skills, using scores of 298 ESL learners on the Community English Program reading test at Teacher College-Columbia University. The models included (a) a unitary model, (b) a two sub-skill model (reading for literal meaning and reading for implied meaning), and (c) a three sub-skill model (reading for literal meaning, reading for implied meaning with endophoric reference, and reading for implied meaning with exophoric reference). While Exploratory Factor Analysis suggested a good fit with the unitary model of reading, Confirmatory Factor Analysis (CFA) showed that all three models fitted data

well. The three sub-skill model was, however, selected as being reflective of the reading construct without any cogent justifications, which made the results open to interpretations.

The above findings implied that the dimensionality of L2 reading tests has remained controversial and required more empirical evidence from different English language learning contexts to address the question of whether L2 reading is a unitary or divisible concept as implied by test scores. That motivates the current study to investigate the underlying pattern of the reading achievement test at the CVU to evaluate the level of its alignment with the theoretical construct informed by the syllabus-derived test specification.

Taken together, based on the re-shaped content validation on the basis of combining Bachman's (1990) and Messick's (1989) models, as well as the gaps identified above in relation to evaluating the construct of L2 reading tests, the study is guided by the two research questions (RQ):

(1) To what extent did the content of the reading achievement test for third-year English majors reflect the test specification and content validity through score interpretation?
(2) To what extent were the sub-skills of the reading achievement test for third-year English majors compatible with the proposed theoretical construct of L2 reading proficiency?

4. Method

4.1. Participants

Participants in this study were 477 third-year English majors of the cohort 2019-2023 at the CVU. They were selected against cluster sampling (Salkind, 2006). Twelve classes from the list of 25 classes of English majors were selected in a systematic manner by which the first two classes in the list being opted at the exclusion of the two subsequent ones until eight classes were chosen. The third-year English-majored participants, after having passed prerequisite skill-based modules at B1 and B2 levels in their first- and second- years respectively, took this reading achievement test to determine whether they could achieve reading proficiency at C1 level as targeted by the reading module in their third year.

4.2. Data collection instrument

The investigated achievement test was administered to the participants. It consisted of 40 dichotomously-scored test items with four options. There were four reading passages, each of which contained 10 items. The test was selected from the test bank developed by the CVU, which sought to ensure its objectivity. The actual reading test was not provided in this paper because of an adherence to confidentiality, but can be provided at readers' request.

4.3. Data analysis procedure

Data analysis was undertaken in the pre-test stage and the post-test stage as follows.

4.3.1. Pre-test stage. The test paper was analyzed to see how much similar the characteristics of the test were to the test specification. In other words, characteristics of reading passages and items were evaluated against the test specification. Characteristics of reading passages that were analyzed included the stated components in the test specification, i.e., (a) length, (b) lexical features whose analysis were mediated by VocabProfile (Cobb, 2009) on Lextutor (www.lextutor.ca), (c) syntactic features, (d) text genres. Characteristics of items refer

 ISSN 2476-5880

to how the items (a) ensure relevance to and coverage of the construct and targeted domains in the syllabus, (b) reflect the response types students have to make.

4.3.2. Post-test stage. First, the test scores were processed with the Rasch model (Bond & Fox, 2015), following the Item Response Theory analyzing items based on test scores (Yang & Kao, 2014). The Rasch model, using the free Jamovi software (https://www.jamovi.org/download.html), generates (a) item fit statistics showing whether the items are related to the construct and tested domains, (b) item difficulty, and a Wright map that maps item difficulty onto students' ability to determine how valid the test content is as a measure of the students' performances. Item fit statistics are expressed by logit values of infit and outfit mean square (MNSQ). Infit MNSQ index exhibits the patterns of responses to the items that are close to item difficulty and targeted on persons while outfit MNSQ is more sensitive to the outliers with difficulty far from a person (Bond & Fox, 2015; Green, 2013). According to Bond and Fox (2015), the items whose infit and out MNSQ are outside the range from .6 to 1.4 logits are misfitting.

As the dichotomously-scored test, distractor analyses of the items were also warranted to provide further critique on their content. One rule of thumb for plausible distractors is that they should be attractive to at least 10% of the examinees (Bachman, 2004). Another rule refers to the point-biseral correlation – $r_{p\text{-}bis}$, i.e., the correlation between students' scores of an item when treating each option as the correct answer and their total scores (Carr, 2011). The answer should have positive $r_{p\text{-}bis}$ and any distractor should have negative $r_{p\text{-}bis}$ so that people choosing a wrong answer tend to have lower test scores (Carr, 2011).

Second, to explore the reading construct, CFA was adopted in this study as it allows for the examination of the interrelationship among variables (or sub-skills in the current study) to see whether there is a consistency between the hypothesized model and the collected data (Byrne, 2010; Hair, Black, Babin, & Anderson, 2014; Wang & Wang, 2012). This study followed the model competing approach, i.e., evaluating comparable models to decide which one better represents the underlying test structure. The construction of models commenced with the baseline one-factor model informed by the test specification and the course syllabus. The other two competing models were then constructed based on current literature.

The baseline one-factor reading proficiency model was constructed based on the seven sub-skills identified in the course syllabus and the accordingly informed test specification. Figure 1 shows the reading construct illustrated by the circle while each latent construct is a reading sub-skill represented by the rectangles and summated by the score of test takers on the relevant items. The small circles with arrow headed towards each latent construct represent the measurement errors, i.e., the amount of variance of each sub-skill not explained by the reading construct.

ISSN 2476-5880

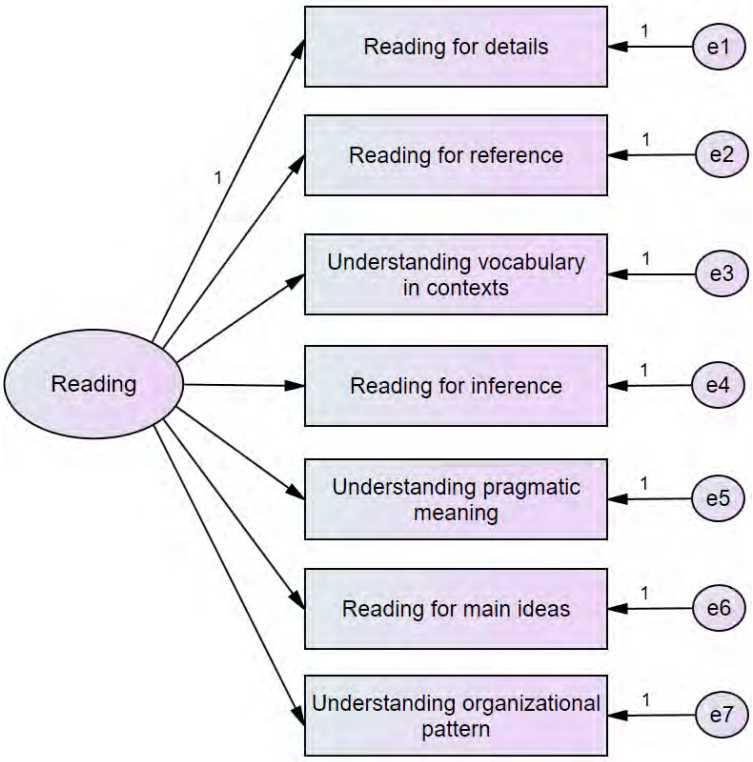

Figure 1. The baseline one-factor reading proficiency model

Figure 2 demonstrates a competing model comprised of three latent constructs having a correlation with one another, i.e., basic comprehension, inferencing, and reading to learn. This model specification was derived from the task specification developed for the L2 reading construct in the new TOEFL reading test reported by Cohen and Upton (2006). Basic comprehension is concerned with the ability to make sense of explicit information stated by a factual sentence, and reference (Cohen & Upton, 2006). Reading for details and Reading for reference are hypothesized to be indicators of this latent construct. Inferencing refers to the ability to make inference about the rhetorical purposes based on the context and textual information (Cohen & Upton, 2006). This latent construct is hypothesized to include three sub-skills, namely Understanding vocabulary in contexts, Reading for inference, and Understanding pragmatic meaning. Reading to learn is characterized by the ability to synthesize information at the discourse level and categorize and organize the information based on the relationship of ideas in a text (Cohen & Upton, 2006). The indicators of this construct are hypothesized to involve Reading for main ideas and Understanding the organizational pattern.

ISSN 2476-5880

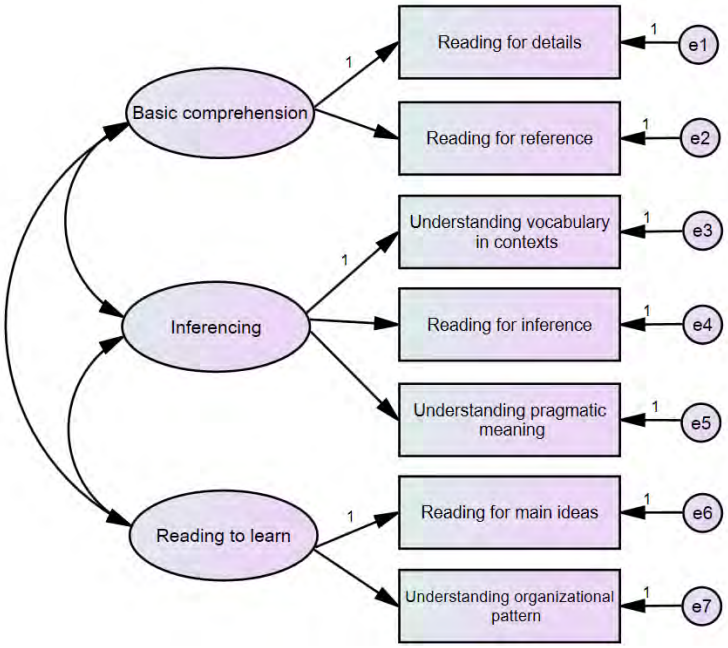

Figure 2. The correlated three sub-skill model

The other competing model in Figure 3 somewhat bears a resemblance to the model in Figure 2 in terms of the latent constructs. However, in this model, "Reading" is put as a higher-order factor, which means that all the latent constructs are now treated as distinct sub-skills and hypothetically explained by a unitary view of the reading construct rather than being correlated with one another. This model specification is reported in Sawaki et al. (2009)'s study of the TOEFL-iBT reading test, finding the unidimensionality of the reading construct and suggesting that the unitary model best represented the underlying structure of the TOEFL-iBT reading test.

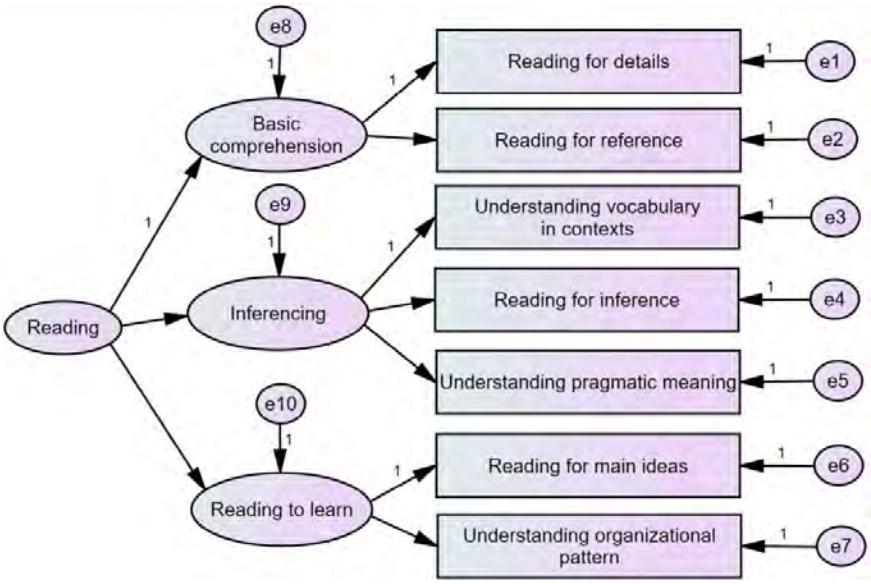

Figure 3. The higher-order factor model

Model evaluation is based on global model fit indices and the appropriateness of individual parameters. As there has been no consensus on the guiding framework for the interpretation and determination of an acceptable model fit, following Hu and Bentler (1998; 1999) and the suggestions of other researchers (e.g., Brown, 2006; Mueller & Hancock, 2010) that are of the most relevance to second language testing contexts, this study examined the model fit based on the five fit indices recommended by In'nami and Koizumi (2011), namely the chi-square statistics, SRMR, RMSEA and its confidence interval, CFI, TLI. Generally, a small χ^2 value with the non-significant *p*-value is an indication of a reasonable model fit. The cutoff values of SRMR and RMSEA not greater than .10 are suggestive of acceptable model fit (In'nami & Koizumi, 2011), with the 90% RMSEA confidence interval (CI) in the range from .00 to .09. Generally, the lower the SRMR value is, the better the fit between the data and the model is. The smaller the RMSEA is, the more parsimonious the model is. In terms of the CFI and TLI values, they should be greater than .90 to achieve a good model fit (In'nami & Koizumi, 2011).

To assess the model fit, such individual parameters as variance, covariance, and factor loadings will also be evaluated against the following two features. First, the critical ratio (C.R.) of each parameter, which is calculated by dividing the unstandardized estimate by its standard error (S.E.), should be greater than 1.96 and statistically significant at $p < .05$. Second, the sign and value of the parameters should be meaningful and adhere to the model's expectation. In other words, there should be no Heywood cases (Hair et al., 2014; Kline, 2016) where the correlation is greater than 1.00, or negative error variance is found.

5. Results

5.1. Pre-test stage

Table 2 compares the reading passages with the requirements in the test specification. It is shown that the characteristics of the examined reading passages generally met the description.

Table 2.
Characteristics of reading passages compared with test specification

	Characteristics of reading passages in the test specification	Characteristics of the reading passages in the test paper
Length	Passages 1 and 2: 300-400 words Passages 3 and 4: 500-600 words	Passage 1: 314 words Passage 2: 375 words Passage 3: 531 words Passage 4: 594 words
Language features	*Lexical features* Passages 1 and 2: more high-frequency words Passages 3 and 4: fewer high-frequency words	Passage 1: 96.8% of high-frequency words Passage 2: 96.3% of high-frequency words Passage 3: 93.2% of high-frequency words Passage 4: 87.2% of high-frequency words
	Syntactic features	Passage 1: • Simple sentences: 69% • Compound and complex sentences: 31%

 ISSN 2476-5880

	Passages 1, 2: a combination of simple, compound, and complex sentences Passages 3 and 4: compound and complex sentences as majority	Passage 2: • Simple sentences: 59% • Compound and complex sentences: 41% Passage 3: • Simple sentences: 32% • Compound and complex sentences: 68% Passage 4: • Simple sentences: 14% • Compound and complex sentences: 86%
Text genre	Social, academic, and professional areas	Passages 1, 2, and 3: social areas Passage 4: academic area

Table 3 presents the characteristics of the items and the extent to which they were in line with the targeted domains in the reading construct. It is shown that the items were related to and reflected their targeted domains.

Table 3.
Characteristics of test items

Skill domains	Items	Example questions
Reading for main ideas	1 (passage 1) 11 (passage 2) 32 (passage 4)	The main idea of this passage is____. What is the main idea of paragraph___?
Reading for details	13, 15 (passage 2) 22, 24, 26 (passage 3) 31, 37, 39 (passage 4)	According to paragraph 3, it took Geoffrey Ballard a long time to produce the first hydrogen-powered bus because_____ (item 26). The author mentions all of the following points about ____ in paragraph ____ EXCEPT___.
Reading for reference	4, 7 (passage 1) 14, 17 (passage 2) 25 (passage 3) 34 (passage 4)	The word _____ in the passage refers to_____.
Reading for inference	8, 9 (passage 1) 19 (passage 2) 27, 30 (passage 3) 36 (passage 4)	It can be inferred from the passage that_____. Which of the sentences below best expresses this sentence "_____".
Understanding vocabulary in contexts	3, 5 (passage 1) 12, 18 (passage 2) 21, 28 (passage 3) 35, 38 (passage 4)	The word "_____" is closet in meaning to_____.

ISSN 2476-5880

Understanding pragmatic meaning	6 (passage 1) 16, 20 (passage 2) 23 (passage 3) 33 (passage 4)	The author discusses "_____" in order to _____. The author's attitude about _____ in the passage is _____.
Understanding organizational pattern	2 (passage 1) 10 (passage 2) 29 (passage 3) 40 (passage 4)	The previous passage probably discussed _____. Look at the four squares [ ] that indicate where the following sentence could be added to the passage. Where would the sentence _____ best fit?

5.2. Post-test stage

5.2.1. Rash statistics. Firstly, to provide further statistical analyses of the items, the Rasch statistics illustrating item fit to the test specification were generated (Table 4). It is shown that all the items displayed their infit and outfit MNSQ in the acceptable ranges, i.e., .82 to 1.13 as well as .68 to 1.34 respectively. This means that all the items fit into the test and functioned well for measuring the reading construct. In other words, the reading construct was not threatened by any items of construct-irrelevant variance.

Table 4.
Item statistics

Item	Correct answer (%)	Measure	Infit MNSQ	Outfit MNSQ
1	54.8	-.21	.97	.96
2	78.0	-1.47	1.01	.90
3	85.9	-2.08	.98	.89
4	88.7	-2.37	.85	.79
5	88.1	-2.31	.89	.76
6	80.8	-1.67	1.04	1.05
7	87.0	-2.19	.98	.81
8	80.2	-1.62	.94	.86
9	80.2	-1.62	.87	.88
10	89.8	-2.50	.94	.92
11	75.1	-1.28	.97	.94
12	88.1	-2.31	.96	.90
13	79.7	-1.58	.92	.83
14	89.8	-2.50	1.01	1.13
15	82.5	-1.79	.82	.68
16	84.7	-1.98	.84	.69
17	81.4	-1.71	1.06	1.26
18	76.3	-1.36	.94	.82
19	69.5	-.95	.88	.82
20	89.3	-2.43	1.09	1.14

ISSN 2476-5880

21	65.0	-.71	1.04	1.10
22	35.0	.74	1.08	1.34
23	66.7	-.80	1.10	1.19
24	39.5	.51	1.11	1.20
25	55.9	-.27	1.02	1.01
26	78.5	-1.50	1.09	1.30
27	57.1	-.32	1.06	1.10
28	58.8	-.40	1.22	1.34
29	34.5	.77	1.05	1.11
30	31.6	.91	1.13	1.25
31	79.7	-1.58	1.10	.98
32	83.1	-1.84	.89	.82
33	61.0	-.51	1.09	1.11
34	90.4	-2.57	.81	.50
35	77.4	-1.43	.98	1.00
36	50.3	.00	.99	.98
37	63.8	-.66	.96	.93
38	72.9	-1.15	.93	.96
39	57.6	-.35	.99	1.01
40	48.0	.11	.96	.99

To provide further support for the claim on content relevance in the pre-test stage that the items reflected their targeted domains, each group of items was evaluated in light of the Rasch model. According to Table 5, the items in each domain had acceptable infit and outfit MNSQ values from 0.6 to 1.4 logits. The principal component analysis of Rasch residuals for each group indicated the eigenvalues of the first contrasts lower than 2 logits, all ranging from 1.34 to 1.65. This means that the data were normally distributed without any patterns of residuals situated beyond the Rasch model. Therefore, the items in each group were found to be unidimensional, i.e., only measuring their underlying domain.

Table 5.
Dimensionality of the groups of items

Skill domains	Items	Infit MNSQ	Outfit MNSQ	Eigenvalue 1st contrast
Reading for main	1	.98	.97	1.39
ideas	11	1.06	1.12	
	32	.97	.89	
Reading for details	13	.99	1.00	1.53
	15	.96	.91	
	22	1.02	1.03	
	24	1.03	1.05	
	26	1.08	1.15	
	31	.98	.93	
	37	.94	.91	

 ISSN 2476-5880

	39	1.02	1.03	
Reading for reference	4	.97	1.00	1.65
	7	1.03	1.12	
	14	.95	.85	
	17	1.02	1.13	
	25	1.07	1.09	
	34	.83	.63	
Reading for inference	8	.98	.97	1.48
	9	.96	0.94	
	19	.94	0.93	
	27	1.04	1.05	
	30	1.05	1.09	
	36	1.01	1.00	
Understanding vocabulary in contexts	3	1.05	1.04	1.61
	5	.89	.73	
	12	.96	.95	
	18	1.05	1.11	
	21	.98	.98	
	28	1.13	1.18	
	35	.90	.87	
	38	.97	1.03	
Understanding pragmatic meaning	6	1.01	1.02	1.38
	16	.97	.90	
	20	1.01	1.02	
	23	1.01	1.02	
	33	1.00	1.01	
Understanding organizational pattern and genre	2	.99	.96	1.34
	10	1.01	.97	
	29	1.01	1.03	
	40	1.00	1.00	

Secondly, in the Rasch model, the item measure (i.e., item difficulty) was also estimated and plotted onto the students' performance. Table 4 shows the item measure, with those items having higher logit values as being more difficult. Figure 4 illustrates the Wright map, arranging the test takers according to their performance and along with the item measure. As for the left scale of respondent latent trait, those in the upper part had their trait as high scorers while low scorers were towards the lower end. As for the right scale of item difficulty, more difficult items were situated towards the upper end, as opposed to less difficult items in the lower end.

ISSN 2476-5880

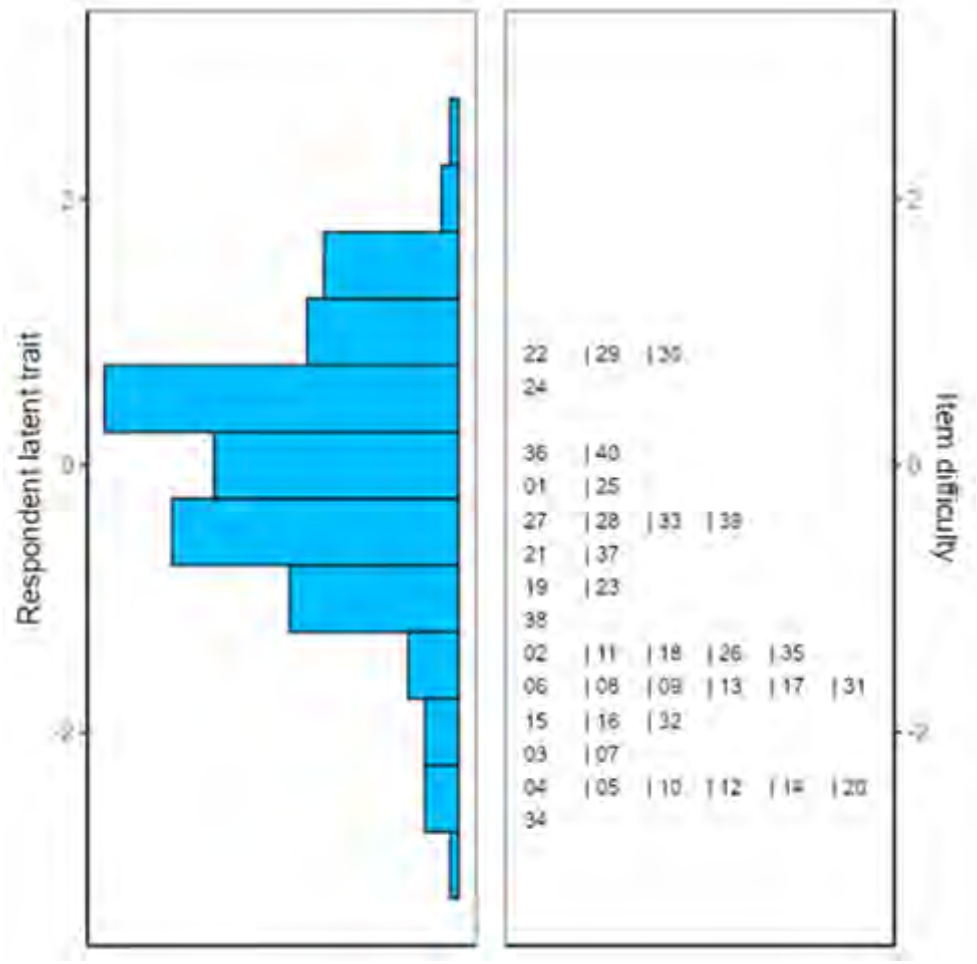

Figure 4. The Wight map

Table 4 and Figure 4 reveal that items 24, 22, 29, 30 were the most difficult, and items 3, 7, 4, 5, 10, 12, 14, 20, 34 were the easiest with their logit values below -2.00. As an overview, while most of the items in passages 3 and 4 were above the counterparts in passages 1 and 2, some others did not follow the prescribed difficulty order. For example, items 32 and 34 which should have been designed at a higher level were actually below the items in passages 1 and 2. Likewise, item 1 which should have been the easiest was actually above the items from passages 3 and 4.

Figure 4 also shows that the spread of item difficulty, to a certain extent, covered the majority of students' abilities. No items were both beyond and under the students' abilities, which means that the items' content was somehow suitable in their difficulty to elicit how well students achieved the learning objectives, and that the reading construct was not threatened by construct-irrelevant difficulty and easiness. However, there were cases in which many students at the top did not have items equivalent to their level. This could be attributed to the problem in designing distractors of items. Table 6 demonstrates distractor analysis for the items. Looking into the investigated items, while the answer to all the items had positive $r_{p\text{-}bis}$, except items 22, 24, 27, 28, 29, 30, 40 having plausible distractors attracting more than 10% of the students, the distractors in the other 33 items were problematic.

ISSN 2476-5880

Table 6.
Distractor analysis

Item	Options	Response %	$r_{p\text{-}bis}$
1	A	55	0.46
	B	31	-0.20
	C	7	-0.21
	D	7	-0.33
2	A	3	-0.25*
	B	78	0.39
	C	5	-0.21
	D	14	-0.22
3	A	10	-0.23
	B	1	-0.32*
	C	3	-0.20*
	D	86	0.39
4	A	2	-0.09*
	B	8	-0.40
	C	89	0.49
	D	1	-0.30*
5	A	3	-0.29*
	B	88	0.46
	C	4	-0.23
	D	5	-0.25
6	A	81	0.32
	B	2	-0.25*
	C	6	-0.07
	D	11	-0.23
7	A	88	0.39
	B	7	-0.28
	C	2	-0.11*
	D	3	-0.23*
8	A	3	-0.04*
	B	7	-0.38
	C	80	0.46
	D	10	-0.27
9	A	81	0.52
	B	3	-0.28*
	C	12	-0.48
	D	4	0.01
10	A	5	-0.30
	B	90	0.37
	C	1	-0.16*
	D	4	-0.16
11	A	19	-0.28
	B	1	-0.08*
	C	5	-0.33
	D	75	0.43

 ISSN 2476-5880

12	A	88	0.36
	B	3	-0.34*
	C	7	-0.16
	D	2	-0.12*
13	A	6	-0.32
	B	79	0.48
	C	11	-0.28
	D	4	-0.17
14	A	90	0.27
	B	2	-0.16*
	C	3	-0.18*
	D	5	-0.12
15	A	83	0.59
	B	5	-0.31
	C	6	-0.31
	D	6	-0.34
16	A	8	-0.40
	B	5	-0.27
	C	85	0.55
	D	2	-0.22*
17	A	8	-0.04
	B	9	-0.28
	C	82	0.27
	D	1	-0.11*
18	A	15	-0.36
	B	77	0.47
	C	7	-0.25
	D	1	-0.08*
19	A	5	-0.31
	B	18	-0.35
	C	69	0.55
	D	8	-0.21
20	A	2	-0.16*
	B	89	0.22
	C	6	-0.17
	D	3	-0.02*
21	A	10	-0.14
	B	7	-0.35
	C	65	0.37
	D	18	-0.12
22	A	25	-0.07
	B	35	0.26
	C	20	-0.13
	D	20	-0.09
23	A	11	-0.15
	B	6	-0.07
	C	67	0.31
	D	16	-0.2

ISSN 2476-5880

24	A	40	0.25
	B	28	-0.09
	C	10	-0.30
	D	22	-0.17
25	A	0	/*
	B	6	-0.53
	C	56	0.39
	D	38	-0.13
26	A	8	-0.22
	B	79	0.26
	C	8	-0.12
	D	5	-0.06
27	A	22	-0.10
	B	10	-0.40
	C	11	-0.07
	D	57	0.37
28	A	18	-0.04
	B	14	-0.15
	C	58	0.15
	D	10	-0.13
29	A	21	-0.14
	B	35	0.32
	C	28	-0.12
	D	16	-0.11
30	A	18	-0.21
	B	25	-0.02
	C	32	0.20
	D	25	-0.05
31	A	4	-0.15
	B	80	0.29
	C	15	-0.22
	D	1	-0.05*
32	A	5	-0.22
	B	84	0.49
	C	8	-0.22
	D	3	-0.38*
33	A	28	-0.16
	B	7	-0.20
	C	5	-0.16
	D	60	0.32
34	A	5	-0.29
	B	90	0.56
	C	3	-0.42*
	D	2	-0.22*
35	A	2	-0.11*
	B	77	0.41
	C	10	-0.34
	D	11	-0.19

36	A	50	0.43
	B	31	-0.07
	C	8	-0.42
	D	11	-0.22
37	A	5	-0.14
	B	15	-0.27
	C	63	0.46
	D	17	-0.26
38	A	10	-0.09
	B	14	-0.44
	C	4	-0.20
	D	72	0.49
39	A	21	-0.21
	B	16	-0.24
	C	58	0.42
	D	5	-0.17
40	A	48	0.45
	B	25	-0.07
	C	17	-0.41
	D	10	-0.13

Note: Correct answers are underlined and flagged distractors with under 3% response rate are labeled with an asterisk.

5.2.2. CFA findings

The hypothesized CFA models were evaluated against the global fit indices and estimates of their parameters. Table 7 presents the global fit indices of the investigated models.

Table 7.
Global model fit indices

Global fit indices	Baseline one-factor reading proficiency model (model 1)	Three correlated sub-skill model (model 2)	High-order factor model (model 3)
χ^2	13.796	11.978	10.457
p	.660	.487	.564
χ^2/df	.985	.827	.843
CFI	1.00	1.00	1.00
TLI	1.00	1.00	1.00
SRMR	.001	.001	.001
RMSEA	.000	.000	.000
RMSEA confidence intervals	.00 - .072	.00 - .05	.00 - .043

The baseline one-factor reading proficiency model illustrates the general latent construct of reading explained by the seven observed indicators of seven sub-skills. The fit indices suggested good model fit: χ^2/df = .985, p = .660, CFI = 1.00, TLI = 1.00, SRMR = .001, RMSEA = .000, CI [.00 - .072]. The unstandardized parameter estimates, their standard errors,

ISSN 2476-5880

critical ratios, and p-values of the model are presented in Table 8 while the standardized factor loadings are shown in Figure 5. As regards Table 8, the critical ratios of the factor loadings and error variances were significantly different from zero, higher than 1.96 to be significant at $p < .05$. This suggests that the one-factor model achieved an exceptionally good model fit.

Table 8.
Unstandardized estimates of the one-factor model

Weight	Estimate	S.E.	C.R.	*p*	Error	Estimate	S.E.	C.R.	*p*
RFR	.881	.106	8.311	***	e1	.022	.003	7.711	***
UVIC	.928	.114	8.152	***	e2	.019	.002	7.903	***
RFI	1.200	.135	8.910	***	e3	.023	.003	8.016	***
UPM	.847	.115	7.367	***	e4	.025	.003	7.336	***
RFMI	1.267	.163	7.771	***	e5	.027	.003	8.439	***
UOP	.983	.136	7.221	***	e6	.051	.006	8.245	***
RFD	1.000				e7	.039	.005	8.499	***
					Reading	.022	.004	5.121	***

** RFR = Reading for reference, UVIC = Understanding vocabulary in contexts, RFI = Reading for inference, UPM = Understanding pragmatic meaning, RFMI = Reading for main ideas, UOP = Understanding organizational pattern, RFD = Reading for details.*

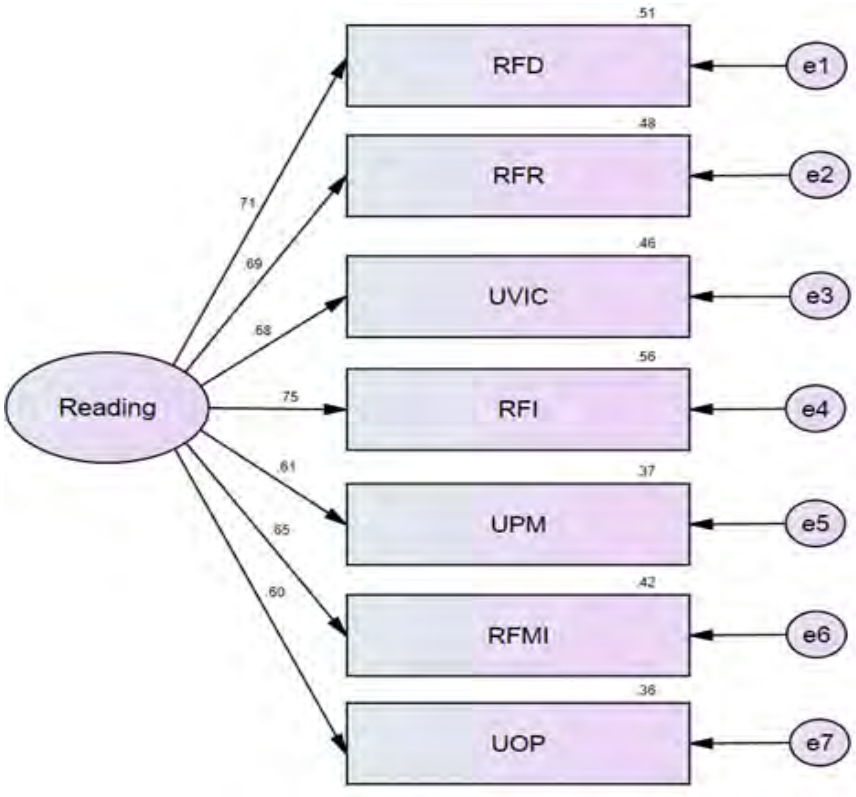

Figure 5. The baseline one-factor model with standardized estimates

The correlated three sub-skill model is explained by the distinct but correlated indicators of basic comprehension, inferencing, and reading to learn. The global fit indices were satisfactory: χ^2/df = .827, p = .487, CFI = 1.00, TLI = 1.00, SRMR = .001, RMSEA = .000, CI [.00 - .05]. The regression weights, variance, and co-variance had critical ratios greater than 1.96 at the significant $p < .001$ (Table 9). However, Figure 6 reveals a Heywood case (Hair et al., 2014; Kline, 2016) in which there was an excessively high correlation between "Inferencing" and "Reading to learn" (r = 1.04) while the other correlations were high and approximated 1.00. The Heywood case rendered this model inadmissible to be considered further.

Table 9.
Unstandardized estimates of the correlated three sub-skill model

Weight	Estimate	S.E	C.R.	*p*
UVIC ← IN	1.000			
RFI ← IN	1.291	.153	8.438	***
UPM ← IN	1.460	.205	7.110	***
RFD ← BC	1.000			
RFR ← BC	2.070	.252	8.225	***
UOP ← RTL	3.321	.493	6.743	***
RFMI ← RTL	1.000			
Variance				
BC	.020	.004	4.588	***
IN	.018	.004	4.622	***
RTL	.035	.009	3.996	***
e1	.024	.003	7.431	***
e3	.024	.003	8.400	***
e4	.028	.004	7.744	***
e5	.028	.003	8.750	***
e6	.051	.007	7.033	***
e7	.039	.005	7.724	***
e2	.020	.003	7.612	***
Co-variance				
BC <--> IN	.022	.003	6.338	***
IN <--> RTL	.027	.005	5.942	***
BC <--> RTL	.028	.005	5.870	***

** BC = Basic comprehension, IN = Inferencing, RTL = Reading to learn*

 ISSN 2476-5880

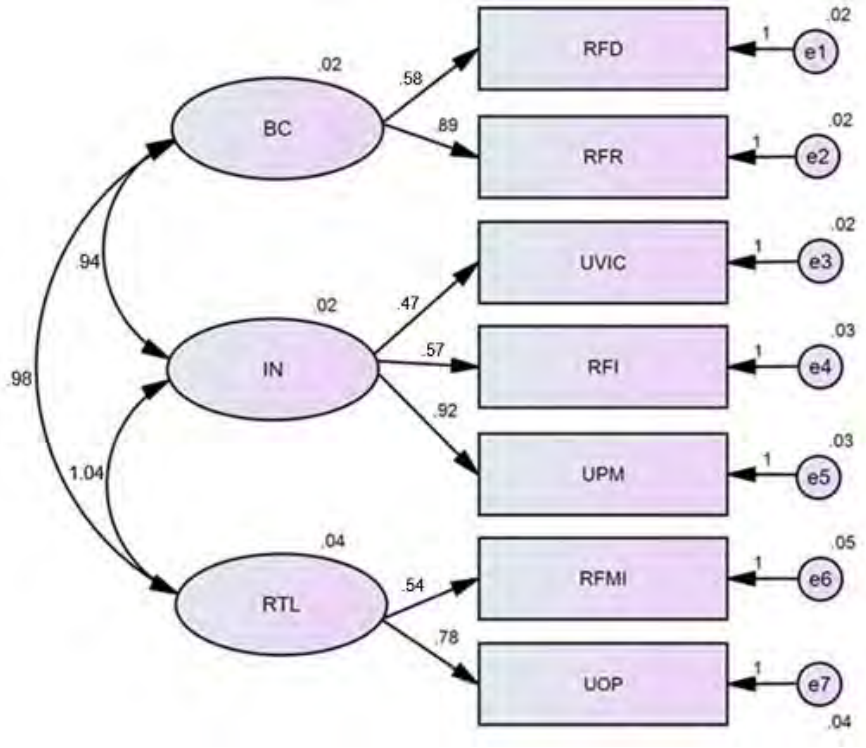

Figure 6. The correlated three sub-skill model with standardized estimates

The higher-order factor model (Figure 7) was an alternative to the correlated three sub-skill model that generated high correlations among the latent constructs. The correlations among the latent construct are now explained by the higher-order factor of reading. This model achieved acceptable fit indices: χ^2/df = .843, p = .564, CFI = 1.00, TLI = 1.00, SRMR = .001, RMSEA = .000, CI [.00 - .043]. According to Table 10, except for the error variances of the "Basic comprehension", "Inferencing", and "Reading to learn" constructs, all the other variances, and regression weights had their critical ratios different from zero at the significant $p < .001$. The error variances of basic comprehension and inferencing, other than being non-significant, were also subjected to Heywood cases (Hair et al., 2014; Kline, 2016) where their critical ratios were lower than 1.96, and the variances were negative at the same time, which made this solution inadmissible for further consideration.

Table 10.
Unstandardized estimates of the higher-factor model

Weight	Estimate	S.E.	C.R.	*p*
BC ← Reading	1.000			
IN ← Reading	1.666	.212	7.829	***
RTL ← Reading	1.251	.162	7.707	***
RFD ← BC	1.000			
RFR ← BC	2.070	.252	8.225	***
UVIC ← IN	1.000			
RFI ← IN	1.291	.153	8.438	***
UPM ← IN	1.460	.205	7.110	***

 ISSN 2476-5880

RFMI ← RTL	1.000			
UOP ← RTL	3.321	.493	6.743	***
Variance				
e8	-.002	.002	-.815	.415
e9	-.003	.002	-2.077	.038
e10	.001	.005	.214	.831
e1	.024	.003	7.431	***
e2	.020	.003	7.612	***
e3	.024	.003	8.400	***
e4	.028	.004	7.744	***
e5	.028	.003	8.750	***
e6	.051	.007	7.033	***
e7	.039	.005	7.724	***
Reading	.022	.004	5.007	***

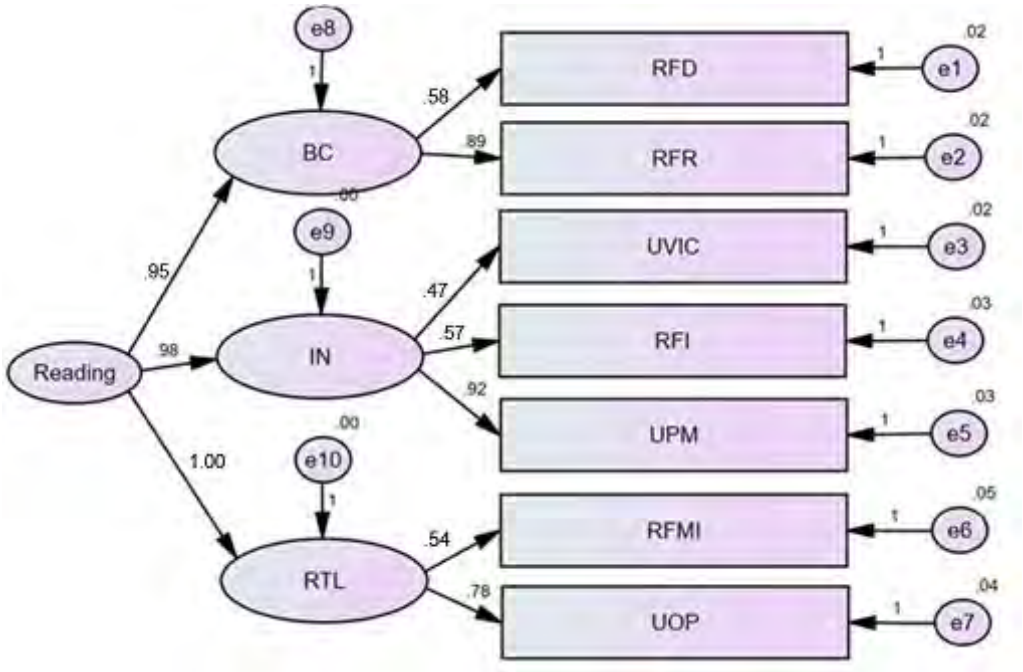

Figure 7. The higher-order factor model with standardized estimates

To conclude, this section examined the competing theoretically-proposed CFA models to determine the one that best represents the underlying structure of the reading achievement test. It is found that the one-factor reading model constructed from the test specification indicated superiority over the correlated three sub-skill and higher-order factor models.

6. Discussion

This section discusses the results in connection with the two research questions.

RQ1: To what extent did the content of the reading achievement test for third-year English majors reflect the test specification and content validity through score interpretation?

The pre-test stage showed that the reading test was compatible with the test specification as informed by the course syllabus (see Table 2). The length of the reading passages was in the required ranges. Their genres met the requirement of social and academic areas. Despite the inevitable influence of background knowledge as some students may be familiar with certain topical areas while the others are not (Shohamy, 1984), it was not the case for this test as the contexts where the input passages were situated hardly had relevance to students' life and educational background, which is aligned with the claim of the test specification.

The multiple-choice format as the response type was consistent for all items as guided by the test specification. Language of the stem, i.e., the part posing the question (Carr, 2011), was intelligible as having been encountered by the students during instructions, thereby not interfering with their comprehension to respond (Brantmeier, 2006; Wolf, 1993). For the question type "[t]he author mentions all of the following points about __ in paragraph __ EXCEPT" (see Table 3), as using the negative word "not" to ask test takers to target the correct option could have caused difficulty (Haladyna et al., 2002; Thorndike & Hagen, 1969), the stems were rephrased into statements with the capitalized "EXCEPT" to ensure the content comprehensibility. Although multiple-choice questions may contaminate content validity as students may base their answer on a loophole in options (Parkes & Zimmaro, 2016), the content of options in this test was argued not to facilitate guessing. First, the options were (a) syntactically consistent in patterns as clauses or phrases, (b) consistent with the target word in parts of speech for "understanding vocabulary in context" items to avoid revealing the answer for students' choice without having to rely on reading skills (Carr, 2011). Second, the distractors seemed to be without any two options that sounded being close to the correct answer (Carr, 2011) although the quality of distractors will be momentarily discussed.

As regards the language input of the reading passages, the lexical features revealed that the percentage of high-frequency words also satisfied the test specification, with more high-frequency words in passages 1 and 2 (i.e., 96.8% and 96.3% respectively) than passages 3 and 4 (i.e., 93.2% and 87.2% respectively). Regarding syntactic features, while simple, compound, and complex sentences were combined in the first two passages, passages 3 and 4 had the majority of the compound and complex sentences at 68% and 86% respectively. On a related note, the language input of the reading passages could offer revelation about their difficulty level in theoretical terms. It is assumed that passages 3 and 4 could be more difficult than passages 1 and 2 as described in the test specification. This is because (a) the percentage of high-frequency words in passages 3 and 4 did not cover 95% of total words as the satisfactory level of text understanding (Hu & Nation, 2000), and (b) compound and complex sentences in passages 3 and 4 outnumbered simple sentences by a wide margin. Despite such a theoretical alignment of the language input of the passages with the test specification in terms of their difficulty level, the score analysis did not support that assumption. That was evidenced by items 32 and 34 in passages 3 and 4 being actually easier than the items in passages 1 and 2, and item 1 in passage 1 being more difficult than the items from passages 3 and 4 (see Figure 4). Therefore, it is shown that those items did not perform as to the proposed difficulty level with this group of students and needed revising.

ISSN 2476-5880

The score analysis also critiqued the difficulty level of the passages in terms of their distractors. Although items 22, 24, 29, 30 met the assigned difficulty level and contained good distractors (see Figure 4 and Table 6), there were only four items of this kind which could be well-thought and handled by the top students because they found the answers to the rest discernible and easy to perform. Therefore, to ensure appropriate difficulty level in the test content and consequently give a more comprehensive snapshot of the students' ability, it is also important to improve the quality of distractors. Because the cases for highly problematic items including (a) the correct answer with negative $r_{p\text{-}bis}$, (b) distractors of over 10% response rate with positive $r_{p\text{-}bis}$ did not appear in this test, making the distractors selected by too few students more attractive will increase item difficulty enough to the targeted range and translate into a more valid test content for better reflection of the students' ability. Because of insufficient time and resources to consider revising all the 33 items, those distractors with a very low response rate at under 3% should be prioritized (Carr, 2011; see table 6 for flagged options).

In terms of content relevance, based on the example questions, the items were related to the construct and targeted domains regulated by the test specification of the syllabus (cf. Bachman, 1990; see table 3). Analyzing test scores supported the claim on content relevance that the items fit into the construct of reading comprehension and ensured unidimensionality of their intended domains (see table 5). Moreover, the score analysis also revealed that the items were not too difficult or too easy for the majority of the students (see Figure 4), which therefore essentially proposes no concern about construct-irrelevant variance in the reading construct to a large extent (cf. Messick, 1989). However, because of the problem in the quality of distractors that underestimated the difficulty level as discussed above, that is not to say that the test was completely exempted from construct-irrelevant variance, in light of occasional cases in Figure 4 where some high-achieving students did not have the items corresponding with their ability.

While content relevance was satisfactory and the test construct was to a large extent not threatened by construct-irrelevant variance, content coverage of the test remained problematic with instances of construct underrepresentation and there existed the need for revisions to guarantee this aspect. That was because, in passage 3, there was no item to test "Reading for main ideas" while three items were focused on testing "Reading for details" (see Table 3). Moreover, there was a tendency to have more concentration on certain skill domains. Fewer questions that tested "Reading for main ideas" and "Understanding organizational pattern and genre" were employed than those testing the other sub-skills (see Table 3). It is, therefore, implied that the reading construct was threatened by construct under-representation. The test writer should balance the equal emphasis on the representation of skill domains in each passage and revisions are warranted to balance and vary skills tested in each passage and the whole test (cf. Bachman, 1990; Messick, 1989).

RQ2: To what extent were the sub-skills of the reading achievement test for third-year English majors compatible with the proposed theoretical construct of L2 reading proficiency?

Results from examining the competing CFA models revealed that the one-factor model was suitable to explain the underlying structure of the reading achievement test and rejected the compatibility with the correlated three sub-skill and higher-order factor models. This result could be accounted for from the technical and language ability perspectives.

From the technical perspective, the identification of the one-factor model was aligned with the description of the test and the practice by which test scores were reported. The test

specification guided the test design based on the seven reading sub-skills which were confirmed in the pre-test and post-test stages. The superiority of the one-factor model suggested the compatibility with the test specification in that the seven sub-skills or observed indicators were only captured by one construct or factor of reading, with each indicator measured by a package of items as guided by the test specification. Furthermore, the one-factor model was in line with the score reporting practice of the test by which only one score was produced to reflect the students' achievement in their reading ability without details of the sub-skills. The finding supporting the one-factor model resonates with that of other reputed tests of English language proficiency such as the Test of English for International Communication (TOEIC) (In'nami & Koizumi, 2012), the SALT reading test (Van Steensel, Oostdam, & Van Gelderen, 2013). However, this result is not consistent with Min et al. (2021) on the EAP reading assessment for grade 1, Sawaki et al. (2009) on the TOEFL, and Song (2008) on the web-based English as a Second Language Placement Test, all of whom delineated reading as a higher-order factor explained by first-order factors of reading sub-skills. The more complex high-order model is supported over the one-factor model in those previous studies only because of the sensitivity of model estimates to the larger sample size (DeMars, 2013). This difference in the model specification is not significant given that the higher-order factor model shares a commonality with the one-factor model of reading in that only one factor of reading ability captures other constituent sub-skills. One may argue that the one-factor model is a sufficient representation of the underlying structure of the reading assessment, and the higher-order factor model adds little value for accurate calibration unless the sample size is considered. The favor of the one-factor model is, therefore, an underpinning to the higher-order factor of reading (Koizumi & Nakamura, 2016).

From the perspective of language ability, the one-factor structure of the reading construct does not fuel the clear-cut boundary in the divisibility of L2 reading proficiency. It, however, corroborates with the view on the general L2 reading ability with several lower reading sub-skills, which is reflected in previous studies (e.g., In'nami & Koizumi, 2012; Sawaki et al., 2009; Song, 2008). This finding highlights the general reading proficiency as the target of assessment, with specified lower sub-skills as the foundation upon which the item creation process is undertaken. The one-factor structure helps strengthen the argumentation of Alderson (2000) and Song (2008) that regardless of the divisibility in defining L2 reading construct, designing a reading syllabus and testing readers' proficiency should be informed by the consideration into different reading comprehension sub-skills.

7. Conclusions

This paper attempts to reshape content validation in the context of a reading achievement test for third-year English majors at the CVU by gathering evidence from its content to cross-check with the test specification and analyzing the test scores, based on the combination of Bachman's (1990) conceptualization of content validity and Messick's (1989) unifying model of construct validity. Through comparison with the test specification in the pre-test stage and Rasch analysis in the post-test stage, it is found that while content relevance is satisfactory and the test construct is to a large extent not threatened by construct-irrelevant variance, the content coverage of the test remains problematic with instances of construct underrepresentation due to an imbalance in the representation of skill domains tested. It is, therefore, implied for item

writers to reduce concentration on questions of certain sub-skills and balance the representation of sub-skills tested through reading passages. Moreover, as the test in this study is critiqued in terms of the assigned difficulty level of the items and quality of distractors, test writers should consider adherence to the prescribed difficulty level of items, if any, while looking at how distractors of items perform through a pilot at the classroom level to ensure appropriate difficult for all students and provide a more valid measure of the extent to which they fulfill the learning objectives (cf. Carr, 2011).

In light of CFA competing model analyses, the study pinpoints the one-factor model of general reading ability as the underlying structure of the reading achievement test rather than the correlated three sub-skill model and the higher-order factor model. The acceptance of the one-factor model suggests that the reading construct is a unitary concept, to the extent of this study, which involves the use of a wide range of sub-skills to reach the answer as a common practice (Rupp et al., 2006). This builds on the notion that reading is a cumulative process building upon the interaction of various lower sub-skills. This study, therefore, formulates an implication for test takers of reading from the psychometric perspective. As the items are designed to test different sub-skills, test takers necessitate maneuvering an integrative range of sub-skills from which they derive the answer rather than resorting to a single sub-skill. However, the extent to which test takers can process a range of sub-skills at once is dependent upon how easy the test is and test takers' ability to efficiently use the repertoires of their reading sub-skills (Van Steensel et al., 2013). Since many students at the top of this study do not have items equivalent to their level as revealed via the Wright map and distractor analysis, they are enabled to use integrated skills to answer the items with ease.

The limitation in this study remains as only one test from the test bank was validated for only one cohort of third-year students. Caution should be exercised when generalizing the findings of the current study to content validation studies of multiple-choice reading tests in other contexts. Moreover, the study was only restricted to quantitative methods where only Rasch analysis and CFA competing model analysis were involved. For more robustness in findings, future studies could employ expert judgments of reading sub-skills from interview accounts via verbal stimulated recall (Gass & Mackey, 2017) to triangulate with evidence from the test specification and quantitative findings (Dawadi & Shrestha, 2018; Hughes, 2003; Tengberg, 2018). Despite those limitations, this study is one of the few attempts in proposing an alternative framework to fill the gap in content validation by combining Bachman's (1990) conceptualization of content validity and Messick's (1989) unifying model of construct validity to examine the evidence from both test content and test scores as well as explore the tested construct through the score-based interpretation.

Acknowledgments

I would like to greatly thank Dr. Huahui Zhao, School of Education, University of Leeds for pointing out areas of improvement on the earlier version of this manuscript. I would also like to express my gratitude to the editor and anonymous reviewers for their insightful and critical reviews. Special thanks are extended to all the partcipants without whom the study would not have been materialized.

Declaration of Conflicting Interests

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

Ahmadjavaheri, Z., & Zeraatpishe, M. (2020). The impact of construct-irrelevant factors on the validity of reading comprehension tests. *International Journal of Language Testing*, *10*(1), 1-10.

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

Alderson, C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5*(2), 253-270.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. & Palmer, A. (2000). *Language testing in practice* (3rd ed.). Oxford: Oxford University Press.

Basturkmen, H., & Elder, C. (2004). The practice of LSP. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 672-694). Malden, MA: Blackwell Publishing.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Brantmeier, C. (2006). The effects of language assessment and L2 reading performance on advanced readers' recall. *The Reading Matrix*, *6*(1), 1-17.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*(3), 423-466. https://doi.org/10.1111/0023-8333.00016

Byrne, B. M. (2010). *Structural equation modeling with AMOS: basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.

Cai, Y., & Kunnan, A. J. (2018). Examining the inseparability of content knowledge from LSP reading ability: An approach combining bifactor-multidimensional item response theory and structural equation modeling. *Language Assessment Quarterly*, *15*(2), 109–129. http://doi.org/10.1080/15434303.2018.1451532

Carr, N. T. 2006. The factor structure of test task characteristics and examinee performance. *Language Testing*, *23*(3), 269-289. https://doi.org/10.1191/0265532206lt328oa

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Carroll, B. J. (1980). *Testing communicative competence: An interim study*. Oxford: Pergamon Press.

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 241-259). Charlotte, NC: Information Age Publishing, Inc.

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-17). Chichester, UK: Wiley Blackwell.

Cobb, T. 2009. The Compleat lexical tutor. Retrieved from http://www.lextutor.ca/.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks*. New Jersey: Wiley.

Dawadi, S., & Shrestha, P. N. (2018). Construct validity of the Nepalese school leaving English reading test. *Educational Assessment*, *23*(2), 102-120. https://doi.org/10.1080/10627197.2018.1430511

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. http://doi.org/10.1080/15305058.2013.799067

Department of English. (2017). *Reading 3 – Course materials*. University of Foreign Language Studies – The University of Danang.

Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework*. Princeton, NJ: Educational Testing Service.

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research*. London: Routledge.

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

Green, R. (2013). *Statistical analysis for language testers*. Hampshire: Palgrave Macmillan.

Haladyna, T. M, Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). London: Pearson.

Henning, G. (2001). *A guide to language testing: Development, evaluation and research*. Beijing: Foreign Language Teaching and Research Press.

Hu, L. T., & Bentler, P. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hu, M. & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403-430.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly*, *8*(3), 250-276. https://doi.org/10.1080/15434303.2011.582203

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing*, *29*(1), 131-152. https://doi.org/10.1177/0265532211413444

Kerlinger, F. N. (1973). *Foundations of behavioral research*. New York: Holt, Rinehart and Winston.

Kim, A. Y. (2009). Investigating second language reading components: Reading for different types of meaning. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *9*(2), 1-28.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. *Language Testing in Asia*, *6*(3), 1-23. https://doi.org/10.1186/s40468-016-0025-9

Lee, W., Kim, S. Y., Choi, J., & Kang, Y. (2019). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement*, *57*(2), 230–254. https://doi.org/10.1111/jedm.12248

Leeser, M. J. 2007. Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, *57*(2), 229-270. https://doi.org/10.1111/j.1467-9922.2007.00408.x

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, *94*(4), 778-784. https://doi.org/10.1037/0022-0663.94.4.778

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, *10*(3), 211-234. https://doi.org/10.1177/026553229301000302

McNamara, T. F. (2000). *Language testing*. Oxford, New York: Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY: Macmillan Publishing.

Min, S., Kyoungwon B., & Howard, G. C. (2021). Reading is a multidimensional construct at child-L2-English-literacy onset, but comprises fewer dimensions over time: Evidence from multidimensional IRT analysis. *Language Testing*. https://doi.org/10.1177/02655322211045296

Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.). *The reviewer's guide to quantitative methods in the social science* (pp. 371-383). New York: Routledge.

Munby, J. L. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Parkes, J., & Zimmaro, D. (2016). *Learning and assessing with multiple-choice questions in college classrooms*. New York, NY: Routledge.

Read, J. 2000. *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441-474. https://doi.org/10.1191/0265532206lt337oa

Rupp, A. A. (2012). Psychological versus psychometric dimensionality in reading assessment. In J. Sabatini, E. R. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 135-153). New York, NY: Rowan & Littlefield Education.

Salkind, N. J. (2006). *Exploring research* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 5-30. https://doi.org/10.1177/0265532208097335

Schedl, M., Gordon, A., Carey, P. A., & Tang, K. L. (1995). An analysis of the dimensionality of TOEFL reading comprehension items. *ETS Research Report Series*, *1995*(2), 1-26. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01662.x

Sheybani, E., & Zeraatpishe, M. (2018). On the dimensionality of reading comprehension tests composed of text comprehension items and cloze test items. *International Journal of Language Testing*, *8*(1), 12-26.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language testing*, *1*(2), 147-170. https://doi.org/10.1177/026553228400100203

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 19-39). Charlotte, NC: Information Age Publishing, Inc.

Sireci, S. G. (2015). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 1-10. https://doi.org/10.1080/0969594X.2015.1072084

Song, M. Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435-464. https://doi.org/10.1177/0265532208094272

Tengberg, M. (2018). Validation of sub-constructs in reading comprehension tests using teachers' classification of cognitive targets. *Language Assessment Quarterly*, *15*(2), 169-182. https://doi.org/10.1080/15434303.2018.1448820

Tonekaboni, R. F., Ravand H., & Rezvani, R. (2021). The construction and validation of a Q-matrix for a high-stakes reading comprehension test: A G-DINA study. *International Journal of Language Testing*, *11*(1), 58-87.

Thorndike, R. L., & Hagen, E. (1969). *Measurement and evaluation in psychology and education*. (3rd ed.). New York: John Wiley and Sons.

Urquhart, S., & Weir, C. (1998). *Reading in a second language: Process, product and practice*. London, UK: Routledge.

Van Steensel, R., Oostdam, R., & Van Gelderen, A. (2013). Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language Testing*, *30*(1), 3-21. https://doi.org/10.1177/0265532212440950

Wang, J., & Wang, X. (2012). *Structural equation modeling: applications using Mplus*. Chichester, UK: John Wiley & Son Ltd.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal*, *77*(4), 473-489. https://doi.org/10.1111/j.1540-4781.1993.tb01995.x

Yang, F., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171-177. https://doi.org/10.3969/j.issn.1002-0829.2014.03.010