Lim *Language Testing in Asia* (2019) 9:6
https://doi.org/10.1186/s40468-019-0082-y

Language Testing in Asia

RESEARCH Open Access

Test format effects: a componential approach to second language reading

Hyojung Lim

Correspondence: lim@kw.ac.kr
Department of English Language and Industry, Kwangwoon University, Kwangwoon-ro 20, Nowon-gu, Seoul 01897, South Korea

Abstract

Background: This study aims to empirically answer the question of whether the role of sub-reading skills changes depending on the test format (e.g., multiple-choice vs. open-ended reading questions). The test format effect also addresses the issue of test validity—whether the reading test properly elicits construct-relevant reading skills or ability. The research questions guiding the study are as follows: (1) Do test scores differ systematically depending on the test format? (2) Do the predictors of test scores differ systematically depending on the test format?

Methods: Ninety Chinese ESL students participated in the study at the post-secondary level and took two TOEFL practice testlets, one with multiple-choice (MC) questions and the other with stem-equivalent open-ended (OE) questions. In addition to the reading comprehension test, the participants completed a vocabulary test, grammar test, word recognition task, sentence processing task, working memory test, and strategy questionnaires (reading and test-taking strategies).

Results: The participants performed better on the MC questions than the corresponding OE questions, regardless of the text effect. More importantly, an L2 reading test in a different format involved different sub-reading components; vocabulary knowledge was the only significant predictor of MC test scores, whereas for the OE reading test, grammar knowledge, word recognition skills, and possibly inferencing strategies were found to be significant predictors.

Conclusion: Despite a number of limitations, the value of this study lies in the effort to empirically test format effects by taking a componential approach to reading. The findings suggest the possibility that differently formatted reading questions may tap into different sub-reading component skills. To accurately reveal the underlying structure of the reading construct being tested in MC and OE tests, however, we call for a larger scale data collection with mixed research methods employed.

Keywords: Test format effect, Componential analysis, Second language reading

Background

Numerous componential analyses have been performed to explore the variables that contribute to language learners' second language (L2) reading comprehension (Jeon & Yamashita, 2014; Kim & Cho, 2015; Kremmel, Brunfaut, & Alderson, 2015; Shiotsu & Weir, 2007; van Gelderen et al., 2004). Because reading involves various sub-reading

skills, knowledge, and processes intertwined in a complex way, this componential approach has gained significant popularity over the years among language testers as well as researchers seeking to argue for or against the construct validity of a reading test and/or examining the developmental aspect of L2 reading ability. Jeon and Yamashita (2014) summarized the componential studies of L2 reading in recent decades; the major contributors discovered thus far, namely, the high-evidence correlates, are vocabulary and grammar knowledge, as well as other language-related variables, including decoding, phonological awareness, and orthographic and morphological knowledge. These elements play more significant roles than language-independent variables, such as metacognition and working memory (WM), labeled "low-evidence correlates." Since this meta-analysis, L2 researchers have continued to explore additional predictive variables (e.g., Kremmel et al., 2015) or aimed to identify confounding factors (e.g., Kim & Cho, 2015) that may affect the relationships between sub-reading components and reading constructs. Along these lines, this study revisits the popular issue of the sub-reading component that contributes to L2 reading comprehension while highlighting the effects of the test format. Prior studies have primarily focused on pinpointing independent variables that predict an individual's reading ability. Of special interest to this study, however, is the manner in which reading comprehension is measured—that is, whether a different test format is connected to a different dimension of the reading construct. That is, the main objective of this paper is to empirically answer the question of whether the role of sub-reading skills differs systematically depending on the test format.

Component skills approach

The component skills approach assumes that reading ability is a collection of several distinctive, empirically separable sub-reading skills that are orchestrated to yield reading comprehension as a product. The componential analysis is twofold (Koda, 2005): one aspect involves identifying the components of the reading construct, and the other involves comparing each component's relative contribution to reading comprehension. The component skills approach is particularly valuable in L2 reading research because it reveals the developmental aspect of reading ability and helps identify the cause of individual differences among L2 readers.

The results of componential analyses have varied depending on the target population, experiment settings, and predictor variables involved. While considering moderator effects, Jeon and Yamashita (2014) conclude that language-dependent variables likely play more important roles than language-independent variables in L2 reading; thus, L2 reading is a language problem. Among language variables, however, whether vocabulary or grammar knowledge is more influential remains uncertain, although these two are usually considered the most important determinants of L2 reading ability. Kim and Cho (2015) noted that the relative importance of vocabulary and grammar knowledge fluctuates as readers' language proficiency advances: for advanced EFL readers, grammar knowledge has emerged as more important, whereas for intermediate readers, vocabulary knowledge has taken precedence over grammar knowledge. Shiotsu and Weir (2007) alluded to the effect of learners' first language (L1) in observing that Japanese undergraduate students' English syntactic knowledge had a greater influence on their

English reading comprehension than did their English vocabulary knowledge. Apart from the either-or debate, Barnett (1986) underscored the interaction between the two variables: vocabulary knowledge significantly contributed to recall scores if and only if readers attained a high level of syntactic knowledge, whereas learners who had low or medium syntactic knowledge benefitted less from vocabulary knowledge.

Given the role of processing skills in relation to L2 reading comprehension, many reading researchers have called for examination at both the lexical and sentential levels (Grabe, 2010). The componential approach and many other reading models value the role of processing skills as critical for accurate and fluent reading in L2. For instance, the simple view of reading defines reading comprehension as a product of decoding skills and listening ability. In the cognitive processing approach (Khalifa & Weir, 2009), the core reading processing begins with bottom-up processes that involve word decoding, lexical access, and syntactic parsing, and it gradually integrates higher-level cognitive processes (e.g., inference). For the successful operation of such linguistic processing, strong vocabulary and grammar knowledge should be instantly available to the reader. Although the important role of learners' processing skills in L2 reading is commonly acknowledged, experimental data supporting this view seem to be scant. Gui (2013) reported in his factor analysis that word recognition skills had a small effect but still contributed to L2 reading comprehension. In van Gelderen et al. (2004), both word recognition and sentence processing skills failed to predict L2 reading comprehension when knowledge components were controlled.

The roles of language-independent variables, such as WM and metacognition, in L2 reading have been controversial. One major function of WM is to store and compute incoming information simultaneously (Daneman & Carpenter, 1980). Thus, common sense tells us that readers with high WM capacity should be better at processing, understanding, and synthesizing textual information. L1 reading researchers have provided empirical evidence supporting the strong link between WM and reading performance (Calvo, 2001; Carretti, Cornoldi, De Beni, & Romanò, 2005). However, the role of WM in L2 reading does not appear to be straightforward. Leeser (2007) reports that the effect of WM on L2 reading performance is mediated by topic familiarity among high-beginner L2 Spanish learners; only those who were familiar with the given topic took advantage of high WM. In contrast, Alptekin and Erçetin (2011) did not find such an interaction between WM and content familiarity among advanced Turkish undergraduate students of English. Another important issue concerns methods of operationalizing the construct of WM. Most reading studies measured learners' WM using a reading span task (Daneman & Carpenter, 1980) where learners are asked to judge the semantic plausibility of an English sentence and memorize the last word of each sentence simultaneously. Numerous scholars have noted overlaps between the reading span task and the reading comprehension test, both of which require participants to read for meaning (Kintsch, 1998).

The role of metacognition, which involves learners' metacognitive knowledge of people, task and strategies, and cognitive/affective experiences (Flavell, 1979), has also attracted the attention of many reading researchers (e.g., Phakiti, 2003). In Khalifa and Weir's (2009) cognitive processing model, metacognitive activities such as goal setting and goal monitoring play a significant role in determining the types and levels of reading and thus the relative importance of associated mental processes. In van Gelderen et

al. (2004), metacognitive knowledge was the strongest explanatory factor for learners' reading comprehension among five sub-reading components (vocabulary and grammar knowledge, lexical and sentential processing skills, and metacognitive knowledge). Conversely, Jeon and Yamashita (2014) classified metacognition as a low-evidence variable alongside WM. It is worth noting that under testing conditions, research on learners' metacognition has revolved around the knowledge and use of reading strategies, including planning, inferencing, and synthesizing (Schoonen, Hulstijn, & Bossers, 1998).

In the discussion of the componential skills approach, however, an essential question appears to be missing: how is one's reading ability assessed, or how are reading components operationalized? Given that reading ability is a latent variable and is thus estimated only by means of indirect testing methods, different test formats are likely to tap into different aspects of a reading construct or even involve construct-irrelevant elements. The componential studies of L2 reading have used various reading test formats, such as MC test questions (Alptekin & Erçetin, 2010; Jiang, Sawaki, & Sabatini, 2012; Nassaji, 2003; van Gelderen et al., 2004; Zhang & Koda, 2012), recall tests (Leeser, 2007), and cloze tests (Mancilla-Martinez & Lesaux, 2010). A few have adopted a combination of several different test formats (e.g., Kremmel et al., 2015; Shiotsu & Weir, 2007). To my knowledge, however, none of those studies discusses the effects of test format on the componential analyses. Meanwhile, language testers have paid much attention to test format effects in the context of test validation while insisting that learners' reading ability should be assessed in a variety of test formats (e.g., Schmitt, Jiang, & Grabe, 2011). The following section will review format effects in detail and introduce the subsequent problem—the effect of learners' test-taking strategies.

Test format effects

L2 reading has been assessed by examining how learners process objectively scored items (multiple-choice (MC) questions, true/false questions, or matching exercises) and/or free-response items (such as open-ended (OE) questions, summary statements, and recollection of idea units). Conceivably, to provide correct answers, especially for objectively scored items, learners must employ additional mental processes, also known as item-response processes. In this regard, the test format effect is often associated with the question of test validity—whether the test measures what it is supposed to measure. The assumption is that if a test is valid, these additional mental processes that are involved in reading test items but are not necessarily relevant to the construct to be tested should not significantly affect test scores. The format effect also alludes to the effect of test-taking strategies because learners' use of test-taking strategies is often induced by the test format (Sarnaki, 1979). Some test-wise learners can take advantages of MC items by using clues embedded in options and other question items (Allan, 1992).

The comparison of MC and OE items has a long history, frequently questioning the validity of MC tests, the format widely used in language testing. Martinez (1999) stated that MC test items, compared to OE test items, are likely to provoke a limited range of cognition in readers. Rauch and Hartig (2010) corroborated that test formats differ in their cognitive demands; OE items may be more appropriate for assessing higher

reading processes, whereas MC items may be limited to basic reading processes. A similar idea is echoed in Ozuru, Briner, Kurby, and McNamara (2013); while OE reading test items activated a wide range of processes, MC items were primarily associated with readers' topical knowledge. In terms of test authenticity, Prince (2014) posited that tests with OE items are better than MC tests because the former better replicate real-world tasks. Similarly, Field (2011) claimed that test-taking strategies specific to the MC format could undermine a test's cognitive validity.

In'nami and Koizumi (2009) observed in their meta-analysis study that MC formats were easier than OE formats in the L1 reading context, but this was not necessarily the case for L2 reading. However, a test format effect in favor of MC items emerged even in L2 reading if and only if an empirical study involved between-subjects designs, random assignment, stem-equivalent items, or advanced learners.

The presence of test format effects suggests the possible involvement of test-taking strategies (Rogers & Harley, 1999). Test-taking strategies are the consciously selected processes upon which test takers draw to manage both language and test-response demands (Cohen & Upton, 2007). Learners' use of test-taking strategies has been shown to distort otherwise natural reading processes (Paulson & Henry, 2002). For instance, in Rupp, Ferne, and Choi (2006), when given a text with MC questions, test takers first read to obtain an overall idea of the given text and/or to understand the questions by scanning or skimming. By assessing the perceived difficulty of the text and questions, test takers started to deploy particular strategies. Students primarily scanned the text for keyword matching instead of carefully reading for detailed comprehension. The authors also noted that "for questions with distractors that are very close in meaning and plausibility, comprehension of the text content or general argument structure might have been subordinate to logical reasoning" (p. 468), which may eventually alter normal reading processes.

Admittedly, the format effect is expected in any form of language assessment, and thus, caution must be exercised in the interpretation of componential analyses. Concerning the role of vocabulary knowledge in L2 reading comprehension, Jeon and Yamashita (2014) noted that the relationship between vocabulary knowledge and reading comprehension could be mediated by the vocabulary test format; the correlation between the two was much higher when vocabulary was tested in embedded items rather than discrete items. Likewise, the relationship between sub-reading components and reading comprehension can vary depending on how learners' reading ability is assessed. As suggested in prior studies, a different test format likely provokes different knowledge sources and cognitive skills. In other words, it is highly likely that different sub-reading components may come into play in response to a particular test format.

The study

This study compares learners' performance on the MC test to their performance on the stem-equivalent OE test in an attempt to examine whether the dynamics between sub-reading components change depending on the test format. It is hypothesized that test takers would deploy different knowledge sources, skills, and strategies in response to differently formatted question items, and thus,

sub-reading components would have different degrees of predictive power for ME and OE test scores. The predictable variables for analysis are chosen based on prior L2 reading studies (e.g., Jeon & Yamashita, 2014): learners' vocabulary knowledge, grammar knowledge, lexical processing skills, sentence processing skills, working memory capacity, reading strategies, and test-taking strategies. The research questions guiding the current study are listed below.

Research questions

1. Do test scores differ systematically depending on the test format?
2. Do the predictors of test scores differ systematically depending on the test format?

Methods

Participants

Ninety Chinese ESL students attending a large Midwestern university participated in the study. The participants included both undergraduate and graduate students with academic statuses varying from provisional to regular (48 undergraduate, 28 graduate, and 14 provisional students). After the missing values were deleted, 81 (for the MC test) and 82 (for the OE test) students were included in the final analysis. None of the participants had lived in the USA for more than 5 years. A summary of the demographic information is presented in Table 1.

Instruments

Reading test

Two reading passages, which various readability indices have confirmed to be comparable, were selected from the TOEFL iBT complete practice test volume 24 (http://toeflpractice.ets.org/). The results of the text analysis, conducted by the automatic text analysis tool *Coh-Metrix* and the vocabulary analysis tool *Compleat Lexical Tutor*, are summarized in Table 2. Because different text types lend themselves more readily to disparate testing skills and strategies (Weir, 2005), the text genre in this study was restricted to expository texts.

After the two MC testlets were chosen, an OE test version was created for each testlet. The question stems were kept as equal as possible except for the negative factual information questions. Transforming the negative factual information questions on the MC test into factual information questions on the OE test was inevitable because it is unreasonable to ask test takers to write what is not true about the text in the OE question form, which also causes scoring problems. Although

Table 1 Description of participants ($n = 90$)

	Mean	SD	Range
Age	20	2.81	19–31
Onset of learning (year)	9.7	2.60	3–16
Age of arrival (year)	20.44	3.23	16–30
Length of residence (year)	1.25	1.30	0.08–5
Number of test-taking experiences (TOEFL)	2.76	1.39	1–8

Table 2 Text analysis of reading passages

		Text W	Text S
Genre		Expository	Expository
Text structure		Description	Description
Title		Lake water	Breathing during sleep
Number of words		737	709
Readability	Flesch reading ease score (0–100)	51.53	45.91
	Flesch-Kincaid grade level (0–12)	12.18	12.04
	Coh-Metrix L2 readability	13.54	12.05
Vocabulary complexity	K1+K2 word percentage	84.05	80.40
	AWL percentage	7.04	8.04
	Type and token ratio	0.41	0.40
Syntactic complexity	Left embeddedness, words before main verb, mean	6.10	6.47
	Mean number of modifiers per noun-phrase	1.04	0.97
	Sentence syntax similarity, all combinations, across paragraphs, mean	0.054	0.085

the wording in the MC and OE stems differs in this pair, the OE item was considered the stem-equivalent of the MC item in that test takers are asked to focus on details and understand explicit information equally in both formats. In addition, items that were inappropriate for OE questions, such as the sentence simplification question and the "insert text" question (i.e., the cohesion question), were excluded from the OE version of the MC test. Consequently, fewer questions were included in the OE test; in total, 9 OE items were included in text Water (W), and 8 OE items were included in text Sleep (S). Although the OE test had fewer questions than the MC test, more time was allowed for the OE test (20 min and 30 min, respectively). The time allotted for the OE version was determined based on native speakers' performance during a pilot test; on average, American undergraduate students took one and a half times longer finishing the OE test than the MC test.

The four versions of the testlets (form A = text W with MC questions, form B = text S with MC questions, form C = text W with OE questions, and form D = text S with OE questions) were devised in HTML format to enable participants to take the tests on a computer screen. As in the actual TOEFL iBT test, one question appeared at a time, the text remained on the right side, and test takers could freely return to previous questions for correction. The test forms were counterbalanced across the participants; half of the participants took forms A and D, and the other half took forms B and C. The Cronbach's alpha values for the 13 MC items with text W and for the 12 MC items with text S were .678 and .685, respectively, which are acceptable given the limited number of items and the relatively homogenous group of test takers.

For the comparison of MC and OE items, however, only the MC items that could be successfully transformed into OE items were considered, which reduced the number of MC items to 9 with text W and 8 with text S, respectively, for

regression analyses. Although the number of question items differs across texts, the total scores remained the same at 10 points. The discrepancy is due to the last question of each testlet, which has a summary-type item in which test takers are asked to fill out a table by making multiple choices among given statements. The last question for text S requires more selections and thus deserves more points compared to that for text W. In case of the OE versions, higher scores were still assigned to the summary question for text S, as it requires more writing. Both the MC and OE tests, therefore, awarded 2 points to the last question for text W (forms A and D) and 3 points to that for text S (forms B and C). Partial credit was allowed in both formats. To score OE items, the researcher first created the set of possible answers for each item, primarily consulting the answer keys for the MC items. The second rater, an English teacher with an MA degree in TESOL, graded 20% of the OE responses. The level of agreement reached 99%; a few disagreements were resolved through discussion. Tables 3 and 4 summarize the results of the item analysis of the MC item comparison. The last question was excluded from the analysis, since it does not follow the binary scoring system.

Vocabulary test

The vocabulary test was adopted from Schmitt, Schmitt, and Clapham (2001). During piloting, 15 Chinese ESL students were invited to complete the original version of the test that contained five different levels—2000, 3000, 5000, and 10,000 levels and an academic vocabulary list—with 30 items per level. In the test, students had to choose one of six words to match a given definition. The students obtained an almost perfect score at the 2000 level, but they missed too many items at the 10,000 level; that is, the items at the lowest and highest levels seemingly failed to yield much variation among the Chinese ESL students. Consequently, the researcher decided to exclude those items, primarily to save time and reduce participants' fatigue. The final version of the vocabulary test therefore has 90 items in total, 30 items each from the 3000 and 5000 levels and the academic word list. The reported Cronbach alphas for each level are .93, .92, and .93, respectively (Schmitt et al., 2001).

Table 3 Item analysis of the 8 MC reading items from form A ($n = 47$)

	Item type	Mean	Std. deviation	Item discrimination	Corrected item-total correlation	Cronbach's alpha if item deleted
Item 1	Reference	.64	.486	.65	.459	.636
Item 2	Vocabulary	.32	.471	.26	.130	.686
Item 3	Inference	.64	.486	.34	.197	.677
Item 4	Factual	.49	.505	.70	.496	.629
Item 5	Factual	.49	.505	.45	.256	.669
Item 6	Vocabulary	.81	.398	.35	.393	.650
Item 7	Factual	.83	.380	.29	.196	.674
Item 8	Negative Factual	.85	.360	.35	.320	.660
Item 9	Summary					

Table 4 Item analysis of the 8 MC reading items from form B ($n = 43$)

	Item type	Mean	Std. deviation	Item discrimination	Corrected item-total correlation	Cronbach's alpha if item deleted
Item1	Factual	.58	.499	.29	.222	.681
Item2	Vocabulary	.54	.505	.64	.457	.643
Item 3	Factual	.30	.465	.21	.231	.679
Item 4	Rhetorical	.35	.482	.71	.480	.640
Item 5	Negative factual	.56	.503	.43	.257	.676
Item 6	Factual	.37	.489	.57	.200	.684
Item 7	Inference	.28	.454	.36	.246	.676
Item 8	Summary					

Grammar test

The grammar test was obtained from Shiotsu (2003). The original question items were made available by the author. The original version consisted of 35 MC questions with four options. However, it was advised that 3 items (questions 12, 18, and 21) functioned improperly (private communication with Shiotsu, September 4, 2012). The current analysis therefore excluded the problematic items. Notably, this grammar test was carefully designed to examine learners' grammar knowledge apart from their vocabulary or other linguistic knowledge (Shiotsu & Weir, 2007). The Cronbach's alpha for the 32 MC items was .83.

Lexical processing task

The lexical processing task was adopted from Lim and Godfroid's (2015) semantic classification task. The participants had to decide as quickly and accurately as possible whether the word on the computer screen referred to a living being (e.g., a boy) or a non-living artifact (e.g., a book). Segalowitz and his associates argue that compared to traditional lexical decision tasks, animacy judgment tasks better estimate learners' word recognition skills because they invoke relatively strong semantic processing and are thus authentic (see Lim & Godfroid, 2015, for more detail). Shorter reaction times are considered as faster, more automatized processing skills. The task began with 6 practice items, followed by 46 test items.

Sentence processing task

The sentence processing task was also borrowed from Lim and Godfroid's (2015) sentence construction task. The participants were asked to construct part of a sentence in their minds, similar to how they would produce a sentence in written or oral form. On the first screen, the beginning of a sentence was provided (e.g., "After some time…"), and on the next screen, two possible options appeared (e.g., A. "works", B. "she"). The participants could read the beginning part at their own pace; then, they were asked to quickly choose the option that best continued the earlier phrase. They were told that the options would not necessarily complete the sentence. The reaction times (RTs) and accuracy rates involved in decision-making were collected. The length of the sentence beginning was kept short (range, one to three words), whereas each option constituted one or two words (either a function

word or a content word). With a minimal length of stimuli (both the stems and options), the probability of a heavy semantic analysis intervention was minimized (see Lim & Godfroid, 2015 for more details). As in the lexical processing task, shorter reaction times are considered as faster, more automatized, processing skills. The task began with 6 warm-up trials, followed by 50 test items.

Working memory task

This study used the automated symmetry span task (Redick et al., 2012) to measure learners' WM. This study did not adopt the reading span task to forestall the potential effects of L2 proficiency on the reading span task or the construct overlap between the WM task and the reading comprehension test. The operation span task was not an option either, given that Chinese students' mental arithmetic skills could distort the WM test scores. In the symmetry span task, the participants made symmetry judgments of pictures while memorizing spatial locations. The reported Cronbach's alphas for the symmetry span task were .81 for partial scores and .73 for absolute scores (Engle, Tuholski, Laughlin, & Conway, 1999).

With regard to the format effect, it was assumed that WM could play a more important role in the OE test than in the MC test. Test takers who can hold more information and process it better were expected to construct their responses to OE items more efficiently, especially under the testing conditions. In cases of MC items for which cues are presented in options, however, individual differences in WM were assumed not to affect test scores to such an extent. Previous empirical studies also alluded to the test format effect relevant to the role of WM in L2 reading; Leeser (2007) reported an significant interaction between learners' WM and topic familiarity in L2 reading, which was not the case in Alptekin and Erçetin (2011). Notably, the former used a free recall task to assess reading, while the latter administered an MC test.

Strategy questionnaires

Two strategy questionnaires were administered online immediately after the MC reading test, the reading strategy and the test-taking strategy questionnaires. Six-point Likert scale questionnaires were adopted from Cohen and Upton (2007), where 1 = "seldom" and 6 = "very frequently." The original reading questionnaire contained 28 strategy statements, whereas the test-taking questionnaire included 28 test management strategies and 3 test-wise strategies. During piloting, 40 Chinese students were recruited to complete the questionnaires. To reduce the number of items, an exploratory factor analysis was conducted: First, the items associated with the factors that showed relatively small eigenvalues (i.e., close to 1) were deleted. Second, the items that loaded on more than one factor were excluded to avoid overlaps between the factors. Finally, we ensured that there were 3 items for each factor. As a result, the final version had 15 reading strategy and 26 test-taking strategy statements. For the analysis, three groups of reading strategies were considered: (1) approaches to reading the passage, (2) using discourse knowledge, and (3) making inferences about the meanings of new words. The test-taking strategies were categorized as follows: (1) test management strategy and (2) test-wise strategy. The internal consistency values were Cronbach's alpha = .64 and

Cronbach's alpha = .81 for the reading strategy and test-taking strategy questionnaires, respectively.

Procedure

To minimize the participants' fatigue, the students were asked to finish the tasks on two different days with a 1-week interval between the sessions. The data were collected on an individual basis in a language lab. On day 1, the researcher explained the entire procedure to the participant and collected a consent form. The vocabulary test and the sentence processing task were completed in that order. Then, the reading test, with either MC or OE questions, was administered on a computer. Upon finishing the MC test, the participant finished the strategy questionnaires online. The second session began with the WM test and the untimed grammar test. The second reading test followed; those who took the OE test in session 1 were given the MC test on day 2, and vice versa. The participants who took the MC test completed the strategy questionnaires and the lexical processing task. Participation was voluntary, and $20 was given as compensation. Each session took approximately an hour and a half; 20 min were allowed for the MC test, while 30 min were allotted for the OE test. The time allotted for the OE version was determined based on native speakers' performance during a pilot test; on average, American undergraduate students took one and a half times longer finishing the OE test than the MC test.

Data analysis

A dummy variable regression was performed to answer the first research question: do test scores differ systematically depending on the test format? The regression model included test scores as a dependent variable, with the test version corresponding to one of four possible cases: (A) text W with the MC format, (B) text W with the OE format, (C) text S with the MC format, and (D) text S with the OE format. The test scores were between 0 and 10 for all test versions. The main specification included a test format dummy as an independent variable with the value 0 for MC and 1 for OE. A non-zero coefficient of the dummy variable reflects the existence of a test format effect. No additional independent variables are necessary for two aspects of our research design: (1) participants were randomly assigned to different test versions, and (2) various readability indices confirmed the comparability of the two reading texts. As a robustness check, I relaxed the assumption of reading text comparability and performed a modified regression analysis. The extended specification included an additional text dummy variable with the value 0 for text W and 1 for text S. A non-zero coefficient of the text dummy variable controlled for the systematic difference between the two reading texts, if any. The robustness of the conclusion can be assessed by investigating the coherency of the coefficient estimates for the test format dummy variable.

Predictive regression analysis was performed for both MC and OE to answer the second research question—the componential analyses: do the predictors of test scores differ systematically depending on the test format? Suppose that both MC and OE involve identical sub-reading skills such that no test format exists. In such a case, one of the immediate implications is that MC and OE test scores should have an identical set of component skills as predictors. Conversely, finding different predictors for MC and OE

test scores can constitute empirical evidence against the null hypothesis of no test format effect. The predictive regression model included test scores as a dependent variable and a variety of sub-reading skill indices as candidate predictors, such as vocabulary, grammar knowledge, lexical and sentential processing skills, WM, and aggregate strategy scores. Among a number of alternative data-driven model selection methods, stepwise forward regression was adopted because of its balanced properties. The resulting predictor sets for MC and OE were compared to investigate the existence of a test format effect.

For the processing skills task, the coefficient of variation (CV) instead of RTs was considered as an index of learners' automatic word recognition skills (see Lim & Godfroid, 2015 for more details). The missing data analysis was not conducted—first, because there were few missing data (less than 5%) and, second, because previous empirical data showed little or no difference before and after data cleaning.

Results

Test format effects

A dummy variable regression was performed to investigate whether test scores differ systematically depending on test formats. Overall, participants performed better in the MC tests (Table 5); with the Water text, students scored an average of 5.76 in the MC format and of 4.55 in the OE format, while with the Sleep text, of 4.32 in the MC format and of 2.92 in the OE format. As summarized in Table 6, the results of the main specification with a test format dummy indicated that a significant regression model was estimated ($F = 19.5$, $R^2 = .1$).[1] Furthermore, a statistically significant coefficient was estimated for the test format dummy variable ($\beta = -.316$, $t = -4.413$, $p < .001$). The coefficient remained significant ($\beta = -.304$, $t = -4.562$, $p < .001$) even when I relaxed the assumption of text comparability by adding a text dummy variable. Therefore, the test score data confirmed the systematic difference depending on the test format, and the result was robust to the assumption regarding the comparability of the two reading texts used in the test.

The componential analysis with the MC-based L2 reading test

A pairwise correlation table (Table 7) indicates that a set of potential predictors for the MC reading test score exists, with the highest correlate (in absolute value) being vocabulary knowledge ($r = .526$, $p < .001$), followed by grammar knowledge ($r = .425$, $p < .001$), test-wise strategies ($r = -.305$, $p = .004$), and word recognition skills ($r = -.204$, $p = .056$) in that order. Note that word recognition skills and test-wise strategies were negatively correlated with MC test scores; participants with shorter response times obtained higher scores on the MC test, while those who adopted more test-wise strategies scored lower. Table 7 also indicates a potential risk of multicollinearity. Vocabulary knowledge, for example, shows high pairwise correlations with other potential predictors, such

Table 5 Descriptive statistics of test scores

	Form A (Water, MC)	Form B (Sleep, MC)	Form C (Water, OE)	Form D (Sleep, OE)
Mean	5.76	4.32	2.92	4.55
SD	2.07	2.21	1.54	1.78

Table 6 Summary of dummy variable regression analysis for variables predicting reading test scores ($N = 90$)

Variable	Model 1			Model 2		
	B	SE B	β	B	SE B	β
Constant	5.067	.218		5.806	.245	
Format	− 1.360	.308	− .316**	− 1.308	.287	− .304**
Text				− 1.529	.287	− .355**
R^2	.10			.225		
F	19.471**			25.468**		

*$p < .05$, **$p < .001$

as grammar knowledge ($r = .763$, $p < .001$), test-wise strategies ($r = - .327$, $p = .003$), and word recognition skills ($r = - .252$, $p = .019$). Considering the number of potential predictors and the limited information from the sample, data-driven model selection approaches can lead to a parsimonious model with enhanced prediction accuracy and interpretability. Table 8 shows that the stepwise forward regression method identified a model with vocabulary knowledge as the sole predictor ($\beta = .091$, $t = 5.462$, $p < .001$), explaining 27.7% of the total variation in the data. Other potential predictors, such as grammar knowledge, test-wise strategies, and word recognition skills, contributed less than expected from pairwise correlations, presumably due to their high correlations with vocabulary knowledge.

The componential analysis with the OE-based L2 reading test

A pairwise correlation table (Table 9) indicates the same set of potential predictors for the OE reading test score as well, with the highest correlate (in absolute value) being grammar knowledge ($r = .573$, $p < .001$), followed by vocabulary knowledge ($r = .513$, $p < .001$), word recognition skills ($r = - .257$, $p = .016$), and test-wise strategies ($r = - .255$, $p = .019$) in that order. However, the stepwise regression selected a different set of predictors than that in the MC-based reading test. The final model in Table 10 includes grammar knowledge ($\beta = .553$, $t = 6.044$, $p < .001$) and word recognition skills ($\beta = - .201$, $t = - 2.192$, $p = .031$) as the most significant predictors of OE test scores, and the parsimonious

Table 7 Correlations among sub-reading components and MC reading test scores

	VK	GK	WR	SP	WM	RS1	RS2	RS3	TTS1	TTS2	MCtest
VK	1.000	.763**	− .252*	− .131	.003	.000	.161	− .030	.008	− .327*	.526**
GK		1.000	− .216*	− .193	.036	− .027	.275*	− .033	.012	− .284*	.425**
WR			1.000	.327*	− .179	.189	− .015	.115	.154	.136	− .204
SP				1.000	− .188	.073	.009	.038	.107	.095	− .131
WM					1.000	.225*	.038	.092	.162	.025	.081
RS1						1.000	.327*	.073	.302*	.135	.005
RS2							1.000	.312*	.324*	.046	.173
RS3								1.000	.303*	.134	.098
TTS1									1.000	.333*	− .035
TTS2										1.000	− .305*
MCtest											1.000

VK vocabulary knowledge, *GK* grammar knowledge, *WR* word recognition skill, *WM* WM, *RS1* approaches to reading the passage, *RS2* discourse knowledge, *RS3* inferences about the meanings of new words, *TTS1* test management strategy, *TTS2* test-wise strategy
*$p < .05$, **$p < .001$

Table 8 Summary of the stepwise regression analysis for the MC test

Variable	Model 1		
	B	SE B	β
Constant	−1.394	1.204	
VK	.091	.017	.526**
R^2	.277		
F for change in R^2	29.832**		

VK vocabulary knowledge
*$p < .05$, **$p < .001$

model could explain 39.7% of the total variation in the data. Taken together, the predictors for MC and OE test scores were not identical even though the exact same data-driven model procedure was applied. This result is not consistent with the null hypothesis of no test format effect, under which an identical set of sub-reading skills should be the predictors for both MC and OE.

Discussion

Test format effect

The primary reason for conducting the dummy variable regression analysis was to examine whether test scores differ systematically depending on a test format. The results revealed that the test format indeed had an effect, even when the text effect was taken into account. The second regression model suggests that participants performed better in the MC format and on the "Lake Water" reading passage (text W).

It is unfortunate that the study failed to select two parallel texts for the experiment. The results of the readability analyses, with which the vocabulary levels and syntactic complexity are mainly concerned, seemingly explain that the two texts are comparable but not identical. The Flesch-Kincaid grade level and the Coh-Metrix L2 readability index even indicate the possibility that text W could be slightly more difficult than text S, which contradicts the outcome of the simple regression analysis. Possible sources of this unexpected result are, first, that the participants might have been more familiar with one topic than the other (e.g., Carrell & Eisterhold, 1983) or, second, that they might have perceived the comprehension questions associated with text S as more challenging (see Tables 3 and 4).

More important, however, is that the test format effect still emerges when the text difficulty variable is controlled; the participants scored higher on the MC test than on the OE test. This finding is reminiscent of the conclusion that In'nami and Koizumi (2009) drew from their meta-analysis: in general, the format effect did not exist in L2 reading; however, the MC test was indeed easier than the OE test if and only if the study involved a between-subjects design, random assignment, stem-equivalent items, or advanced learners. The present study satisfied all of the conditions and invariably found the test format effect. The presence of the format effect also raises the possibility that students who are well versed in test-taking strategies specific to an MC format may take advantage of the MC question items and perform differently on the OE test. Learners' use of test-taking strategies and its effects will be discussed in greater detail later in this section.

Table 9 Correlations among sub-reading components and OE reading test scores

	VK	GK	WR	SP	WM	RS1	RS2	RS3	TTS1	TTS2	OE scores
VK	1.000	.763**	−.252*	−.131	.003	.000	.161	−.030	.008	−.327*	.513**
GK		1.000	−.216*	−.193	.036	−.027	.275*	−.033	.012	−.248*	.573**
WR			1.000	.327**	−.179	.189	−.015	.115	.154	.136	−.257*
SP				1.000	−.188	.073	.009	.038	.107	.095	−.055
WM					1.000	.225*	.038	.092	.162	.025	−.064
RS1						1.000	.327*	.073	.302*	.135	−.065
RS2							1.000	.312*	.324*	.046	.198
RS3								1.000	.303*	.134	−.148
TTS1									1.000	.333*	−.085
TTS2										1.000	−.255*
OE scores											1.000

VK vocabulary knowledge, *GK* grammar knowledge, *WR* word recognition skill, *WM* WM, *RS1* approaches to reading the passage, *RS2* discourse knowledge, *RS3* inferences about the meanings of new words, *TTS1* test management strategy, *TTS2* test-wise strategy
*$p < .05$, **$p < .001$

The subsequent componential analyses—one with the MC test and the other with the OE test—provide additional evidence for the format effect. Vocabulary knowledge alone predicted the MC test scores, while grammar knowledge and word recognition skills (and possibly inferencing strategies) together predicted the OE test scores. Although the regression results do not answer the question of what sub-reading components actually determine test scores in each format, the findings could be interpreted as indicating that the sub-reading components have different levels of power in contributing to test scores depending on the test format.

With regard to the significant predictive power of vocabulary knowledge for MC reading tests, the procedure of MC item writing and test takers' strategic approach to MC items may provide a useful explanation. To create keys and distractors, item writers often manipulate wording by paraphrasing target words or replacing them with synonyms in the options. Accordingly, matching key words in the text and question items was reported as the most frequently used strategy among Chinese learners in the TOEFL iBT reading section (Cohen & Upton, 2007). Test takers' word matching or word checking behaviors, especially with vocabulary items in the iBT TOEFL reading section, were also illustrated in Lim's (2016) eye-tracking study. For the OE test, the relative superiority of grammar knowledge (to vocabulary knowledge) in predicting test scores can be interpreted as suggesting

Table 10 Summary of the stepwise regression analysis for the OE test scores

Variable	Model 1			Model 2		
	B	SE *B*	*β*	*B*	SE *B*	*β*
Constant	− 1.351	.788		.816	1.252	
GK	.220	.033	.599**	.203	.034	.553**
WR				− 6.278	2.864	− .201*
R^2	.359			.397		
F for change in R^2	43.059**			24.996**		

GK grammar knowledge, *WR* word recognition skill
*$p < .05$, **$p < .01$

either that the absence of options on the OE test may force test takers to read carefully beyond the lexical level or that test takers have to capitalize on their grammar knowledge to a great extent to construct responses. The contributing role of word recognition skills in the OE test might reflect test takers having more time pressure for reading. Compared to MC items, OE items impose two different time-consuming tasks on test takers: reading and writing. Given that ESL learners are slow writers and frequently adopt time-management strategies during a test (Abbasian & Hartoonian, 2014), some participants could have tried to read faster in the OE format to secure more time for writing. Further qualitative investigations, including focused interviews and eye-tracking, are recommended to identify the reasons.

Research has documented a dilemma regarding the relative contributions of vocabulary and grammar knowledge to L2 reading comprehension. In Shiotsu and Weir (2007), grammar knowledge better predicted Japanese learners' English reading ability than did vocabulary knowledge, regardless of the language learning setting. Zhang (2012) demonstrated, however, that vocabulary knowledge related to Chinese EFL learners' reading ability more strongly than did grammar knowledge. In Zhang's SEM study, both the depth and breadth of vocabulary knowledge were measured, while both implicit and explicit grammar knowledge were assessed. In a recent perceptron artificial neural network study, Aryadoust and Baghaei (2016) again confirmed that vocabulary knowledge was more strongly associated with L2 reading comprehension than grammar knowledge among Iranian EFL students. Despite a number of possible confounding factors (e.g., L1-L2 distance, learners' proficiency level) affecting the link between vocabulary, grammar knowledge, and reading comprehension, the current study suggests that the test format effect could be ascribed to the varying dynamics between the sub-reading components and reading ability. Shiotsu and Weir (2007) did not have MC items on the reading test, whereas Zhang (2012) and Nergis (2013) used only MC items from the retired TOEFL test and the University of Teheran English proficiency test, respectively.

Other sub-reading components

Contrary to expectations, other sub-reading components, including WM, sentence processing skills, and learners' use of strategies, failed to predict either the MC or the OE reading test scores. To measure the learners' WM, this study adopted the symmetry test, which involves neither vocabulary nor grammar knowledge, whereas prior componential analyses typically used the reading span test. This approach was an attempt to disentangle learners' L2 language proficiency from the WM construct, but it might have consequently resulted in little or no relation between learners' WM and reading test scores. This finding glosses over a number of considerations and thus calls for further investigation; is the significant correlation between learners' WM and reading ability due to the overlap between the two constructs, or does the correlation in effect imply causation? Is the nature of the WM data specific, such that reading performance engages only the WM that is oriented towards language? Other possible reasons for the lack of relationship between WM and reading test scores are that the text is not sufficiently long for the role of WM to appear (Andreassen & Bråten, 2010) or that the text

remains on the left side of the screen such that learners do not need to memorize any details in either the MC or the OE test.

Learners' sentence processing skills did not contribute to either MC or OE reading test scores. While theoretical reading models never doubt the contribution of sentence processing skills to L2 reading ability, this field of study appears to lack the empirical data that would substantiate the claim. Such a discrepancy immediately raises the question of the validity of the measurement—what the sentence processing task actually measures—rather than depreciating the role of sentence processing in L2 reading. The current study utilized CVs, instead of RTs, collected from the sentence construction task (adopted from Lim & Godfroid, 2015) as an index of the varying degree of learners' automatic sentence processing skills. A considerable amount of literature has validated the use of CVs to measure the development of learners' automatic lexical processing skills (e.g., Segalowitz & Segalowitz, 1993). In contrast, the attempt to extend the use of CVs to the sentence level has recently begun (e.g., Hulstijn, Van Gelderen, & Schoonen, 2009; Lim & Godfroid, 2015); thus, a logical step would be to first validate the testing method before using it for a secondary purpose. Admittedly, the contribution of sentence processing skills could cancel out after grammar knowledge and word recognition skills are controlled.

Finally, the learners' use of strategies, whether reading or test-taking strategies, failed to predict either MC or OE test scores. This finding might confirm Jeon and Yamashita's (2014) meta-analysis results, in which metacognition, including learners' metacognitive knowledge of people, task and strategies, and cognitive/affective experiences, was most weakly correlated with L2 reading comprehension. Nonetheless, we cannot rule out the possibility that reading strategies could have contributed to L2 reading ability to some extent but not enough to bring about individual differences, especially for advanced learners. The Chinese ESL students in this study were about to enter or had already started studying at a US university; thus, they were familiar with the reading genres of high-stakes academic reading tests. Given their level of proficiency and educational backgrounds, the relatively insignificant role of reading strategies may not be surprising; either all participants could be well equipped with the reading strategies specific to the expository reading passage of a TOEFL test or variations in the sample might not have been sufficient to attain statistical significance.

The test-taking strategies were broken down into two types: test management and test-wise strategies. Both are construct-irrelevant, and neither affected the test scores in any form. In the simple regression analysis, the participants scored higher on the MC test than on the OE test, which may hint at the potential impact of test-taking strategies on MC test performance. The subsequent stepwise regression revealed, however, that learners' use of test-taking strategies did not contribute to test scores. Another intriguing finding is that learners' use of test-wise strategies was negatively correlated with their MC test scores. A likely explanation for this result is either that ignorance of the advancement of item writing in language testing may have caused failure in those who learned testing skills at a cram school or that poor readers with relatively low L2 proficiency are prone to rely on their test-wise strategies, with no useful effect. To be able to pinpoint the reason, however, further qualitative investigations (e.g., interviews, think-aloud, eye-tracking) would be needed.

Why the MC test is not the best L2 reading test

The MC test has received much criticism, as it pursues test practicality and reliability at the expanse of test validity (e.g., Currie & Chiramanee, 2010; Farr, Pritchard, & Smitten, 1990). A major critique has been that actual reading does not take place in the MC test, while students likely view the MC reading comprehension questions as a problem set to solve, thus applying various construct-irrelevant strategies. Due to the item-response process, it is inevitable for the MC test to provoke unintended cognitive skills and thus divert test takers from natural reading. The findings of this study provide another reason for why the MC test is not the best L2 reading test. As opposed to the OE reading test, the MC test fails to properly approximate the construct to be tested, not because the elicited construct-irrelevant skills and knowledge contaminates the test scores (learners' use of test-taking strategies had a marginal effect on the test scores), but because the MC test items only evoke a limited range of construct-relevant skills and knowledge. In this study, vocabulary knowledge was the only significant predictor of the MC test scores, whereas grammar knowledge and word recognition skills together contributed to the OE test scores. In this regard, the construct validity of the MC test is probably not good enough, albeit not necessarily violated. This also accords well with Rauch and Hartig's (2010) claim that the OE test assesses higher linguistic processes than the MC test.

Additional attention needs to be given to the fact that the MC test failed to measure learners' processing skills at either the sentential or the lexical level. A number of reading researchers, practitioners, and language testers have underscored the importance of fast reading, especially at the post-secondary English-medium university (Grabe, 2010; Koda, 2007; Weir, Hawkey, Green, & Devi, 2006). Language learners are often slow readers (Fraser, 2007), which becomes a real obstacle in an academic setting where university students are asked to read a sheer number of texts for knowledge gain in a limited time. Fluency in reading, as in other language skills, entails accuracy; that is, an accurate comprehension is nested in the notion of fluent reading (Lim & Godfroid, 2015). As proposed in the skill acquisition theory (Anderson, 1982) and the ACTFL proficiency guidelines (American Council on the Teaching of Foreign Languages, 2012), language learners are likely to achieve fluency in the later stage of acquisition, after accuracy. In this regard, a valid L2 reading test, especially beyond the secondary school level, should be able to accurately assess learners' fast reading skills (Weir et al., 2006). In the present study, the OE test seems to gain superiority to the MC test in terms of assessing the fluency aspect of reading.

Conclusion

This study aimed to examine the format effect in L2 reading tests by means of componential analysis. The participants scored higher on the MC test than on the stem-equivalent OE test. More importantly, an L2 reading test in a different format involved different sub-reading components; vocabulary knowledge was the only significant predictor of MC test scores, whereas for the OE reading test, grammar knowledge, word recognition skills, and possibly inferencing strategies were found to be significant predictors. Such findings seem consistent with previous studies (e.g., Martinez, 1999; Ozuru et al., 2013; Rauch & Hartig, 2010) that have reported that OE items involve a

wider range of cognitive processes, which therefore require a greater number of knowledge sources, than the MC items.

Admittedly, this study has several limitations. First, the study relies heavily on correlation-based statistics, using only two reading passages, two topics, and one homogeneous group. A note of caution is due here because of the small sample size and the limited instruments. More importantly, such a quantitative approach appears insufficient in that it considers only ex post facto analysis and disregards test-taking processes. Thus, the results do not inform us of what is occurring in the test takers' minds during the test or how a test taker handled a key option or a distractor. Without insight into the test takers' minds, the argument for test validity or format effects would be only half-complete. Future studies should therefore shed light on such qualitative aspects by using interviews or eye-trackers. In particular, eye-tracking research will allow us to examine the reading processes or skills (e.g., skimming) that test takers actually use in real time.

Another problem lies in designing and administering the strategy questionnaires. Test takers' responses to the strategy questionnaires were collected and calculated cumulatively. However, students' reading behavior may change according to the coverage of the text and item types. Test takers would adopt careful local reading for details, expeditious global reading for gist, and presumably careful global reading for inferencing questions. Therefore, if learners' strategy use has not been investigated on an item basis, then aggregate numerical values based on questionnaire responses, such as means, would fail to disclose crucial information, including the actual contribution of reading strategies to L2 reading comprehension. Therefore, the finding that learners' use of strategies had little or no predictive power must be interpreted with caution.

Finally, the multicollinearity issue cannot be disregarded; vocabulary knowledge was highly correlated with grammar knowledge ($r = .77$) in both the MC and OE tests. In other words, although vocabulary knowledge failed to predict the OE test scores, the result does not minimize the important role of vocabulary knowledge in reading comprehension. Given the multicollinearity issue among component skills combined with the issue of insufficient information in the sample, it is challenging to identify the true determinants of the reading construct. Note, however, that uncovering the true determinant is not necessary for testing the existence of a test format effect, which is the main objective of this paper. Statistical evidence of non-identical predictors for MC and OE is sufficient for empirically establishing the test format effect.

Despite the limitations, the value of this study lies in the effort to empirically test format effects by taking a componential approach to reading. This study has tried to answer the question of *whether* the role of sub-reading skills varies systematically depending on the test format. A related but separated question would be *how* the role of sub-reading skills differs depending on the test format. In this respect, the predictive regression based on the data-driven model adopted in this study provided reliable evidence that format effects in L2 reading do exist, as sub-reading components play different roles in each format. Although it will never substitute for full-scale componential analysis or in-depth qualitative investigations of test takers' thought processes, the predictive regression results can still contribute to forming a consensus on the "big picture," which can be a valuable set of empirical facts that should be matched and explained by subsequent large-scale, in-depth research projects. Future studies with

more emphasis on the *how* question are therefore suggested. Further research should be undertaken to explore the underlying structure of the reading construct being tested in MC and OE tests, to identify the determinants of each construct, and to inspect test takers' cognition during a test. Although the current findings do not provide full answers regarding the test format effect in L2 reading and relevant validity issues, they will certainly serve as a stepping stone for future studies and alert language testers, practitioners, and teachers to the format effect when interpreting reading test scores.

Endnotes

[1]Because participants are randomly assigned to different groups, individual participants' characteristics are aggregated in the error term. Therefore, most variation in the data is naturally explained by the error term (low R^2), and this does not harm the validity of the analysis.

Abbreviations

EFL: English as a foreign language; ESL: English as a second language; L2: Second language; MC: Multiple-choice questions; OE: Open-ended questions; TOEFL iBT: Test of English as a Foreign Language (Internet-based test); WM: Working memory

Acknowledgements

I would like to thank anonymous reviewers for their insightful comments on an earlier draft of this paper.

Funding

The study was supported by a grant (DDG 2012) from the International Research Foundation for English Language Education.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

HL is the sole contributor to this research paper. The author read and approved the final manuscript.

Competing interests

The author declares that she has no competing interests.

Publisher's Note

Received: 17 January 2019 Accepted: 3 April 2019
Published online: 25 April 2019

References

Abbasian, G. R., & Hartoonian, A. (2014). Using self-regulated learning strategies in enhancing language proficiency with a focus on reading comprehension. *English Language Teaching, 7*(6), 160–167.

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing, 9*(2), 101–119. https://doi.org/10.1177/026553229200900201.

Alptekin, C., & Erçetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *Journal of Research in Reading, 33*(2), 206–219. https://doi.org/10.1111/j.1467-9817.2009.01412.x.

Alptekin, C., & Erçetin, G. (2011). The effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly, 45*(2), 235–266. https://doi.org/10.5054/tq.2011.247705.

American Council on the Teaching of Foreign Languages (2012). ACTFL proficiency guidelines 2012, Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369–406.

Andreassen, R., & Bråten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33*(3), 263–283. https://doi.org/10.1111/j.1467-9817.2009.01413.x.

Aryadoust, V., & Baghaei, P. (2016). Does EFL readers' lexical and grammatical knowledge predict their reading ability? Insights from a perceptron artificial neural network study. *Educational Assessment, 21*(2), 135–156.

Barnett, M. A. (1986). Syntactic and lexical/semantic skill in foreign language reading: Importance and interaction. *The Modern Language Journal, 70*(4), 343–349.

Calvo, M. G. (2001). Working memory and inferences: Evidence from eye fixations during reading. *Memory (Hove, England), 9*(4), 365–381. https://doi.org/10.1080/09658210143000083.

Carrell, P. L., & Eisterhold, J. C. (1983). Schema theory and ESL reading pedagogy. *TESOL Quarterly, 17*(4), 553–573. https://doi.org/10.2307/3586613.
Carretti, B., Cornoldi, C., De Beni, R., & Romanò, M. (2005). Updating in working memory: A comparison of good and poor comprehenders. *Journal of Experimental Child Psychology, 91*(1), 45–66. https://doi.org/10.1016/j.jecp.2005.01.005.
Cohen, A. D., & Upton, T. A. (2007). I want to go back to the text: Response strategies on the reading subtest of the new TOEFL(R). *Language Testing, 24*(2), 209–250. https://doi.org/10.1177/0265532207076364.
Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471–491.
Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450–466.
Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309–331.
Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*(3), 209–226.
Field, J. (2011). Into the mind of the academic listener. *Journal of English for Specific Purposes, 10*, 102–112.
Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906–911. https://doi.org/10.1037/0003-066X.34.10.90.
Fraser, C. A. (2007). Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks. *The Modern Language Journal,* 91(3), 372–394.
Grabe, W. (2010). Fluency in reading — Thirty-five years later. *Reading in a Foreign Language, 22*(1), 71–83.
Gui, M. (2013). Relative significance of component skills to EFL reading: Implications for diagnosis and instruction. Paper presented at the Langague Teting Research Colloquium (LTRC), Seoul, Korea.
Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied PsychoLinguistics, 30*(4), 555. https://doi.org/10.1017/S0142716409990014.
In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219–244. https://doi.org/10.1177/0265532208101006.
Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160–212. https://doi.org/10.1111/lang.12034.
Jiang, X., Sawaki, Y., & Sabatini, J. (2012). Word reading efficiency, text reading fluency, and reading comprehension among Chinese learners of English. *Reading Psychology, 33*(4), 323–349. https://doi.org/10.1080/02702711.2010.526051.
Khalifa, H., & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.
Kim, J., & Cho, Y. (2015). Proficiency effects on relative roles of vocabulary and grammar knowledge in second language reading. *English Teaching, 70*(1), 75–96. https://doi.org/10.15858/engtea.70.1.201503.75.
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge university press.
Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. New York: Cambridge University Press.
Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning, 57*, 1–44.
Kremmel, B., Brunfaut, T., & Alderson, J. C. (2015). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, 1–24. https://doi.org/10.1093/applin/amv070.
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.
Lim, H. (2016). Exploring the cognitive validity of the iBT TOEFL reading test. *Studies in Foreign Language Education, 30*(3), 231–258.
Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied PsychoLinguistics, 36*(5), 1247–1282. https://doi.org/10.1017/S0142716414000137.
Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology, 102*(3), 701.
Martinez, M. E. (1999). Cognition and the question of test item format cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2.
Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *Modern Language Journal, 87*(2), 261–276.
Nergis, A. (2013). Exploring the factors that affect reading comprehension of EAP learners. *Journal of English for Academic Purposes, 12*(1), 1–9.
Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology, 67*(3), 215–227. https://doi.org/10.1037/a0032918.
Paulson, E. J., & Henry, J. (2002). Does the degrees of reading power assessment reflect the reading process? An eye-movement examination. *Journal of Adolescent & Adult Literacy, 46*(3), 234–244.
Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*(1), 25–56.
Prince, P. (2014). Listening comprehension: Processing demands and assessment issues. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 93–108). Bristol: Multilingual matters.
Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items : A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354–379.
Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28*(3), 164–171. https://doi.org/10.1027/1015-5759/a000123.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234–247. https://doi.org/10.1177/00131649921969820.
Rupp, 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.
Sarnaki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research, 49*(2), 252–279.
Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x.
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing, 18*(1), 55–88. https://doi.org/10.1177/026553220101800103.
Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Metacognitive and language-specific knowledge in native and foreign language reading comprehension: An empirical study among Dutch students in grades 6, 8, and 10. *Language Learning, 48*, 71–106. https://doi.org/10.1111/1467-9922.00033.
Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice and the differentiation of speedup from automatization effects. *Applied Linguistics, 14*(3), 369–385. https://doi.org/10.1017/S0142716400010845.
Shiotsu, T. (2003): Linguistic knowledge and processing efficiency as predictors of L2 reading ability: A component skills analysis. Unpublished PhD thesis, The University of Reading.
Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99–128. https://doi.org/10.1177/0265532207071513.
van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology, 96*(1), 19–30. https://doi.org/10.1037/0022-0663.96.1.19.
Weir, C., Hawkey, R., Green, A., & Devi, S. (2006). The cognitive processes underlying the academic reading construct as measured by IELTS. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (Vol. 9, pp. 212–269). Cambridge: Cambridge University Press.
Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.
Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *Modern Language Journal, 96*(4), 558–575. https://doi.org/10.1111/j.1540-4781.2012.01398.x.
Zhang, D., & Koda, K. (2012). Contribution of morphological awareness and lexical inferencing ability to L2 vocabulary knowledge and reading comprehension among advanced EFL learners: Testing direct and indirect effects. *Reading and Writing, 25*(5), 1195–1216. https://doi.org/10.1007/s11145-011-9313-z.