



교육학석사학위논문

# The Effects of Response Format and Question Type on Reading Comprehension Test Performance

문항의 응답 형태와 질문 유형이 영어 독해 시험 수행에 미치는 영향

2023년 8월

서울대학교 대학원 외국어교육과 영어전공 성 수 빈

# The Effects of Response Format and Question Type on Reading Comprehension Test Performance

by

Subin Seong

A Thesis Submitted to the Department of Foreign Language Education in Partial Fulfillment of the Requirements for the Degree of Master of Arts in Education

At the Graduate School of Seoul National University English Major

August 2023

# The Effects of Response Format and Question Type on Reading Comprehension Test Performance

지도교수 소 영 순

이 논문을 교육학 석사 학위논문으로 제출함 2023년 5월

- 성수빈의 석사학위논문을 인준함 2023년 7월
- 위 원 장 <u>(인)</u> 부위원장 <u>(인)</u> 위 원 <u>(인)</u>

# The Effects of Response Format and Question Type on Reading Comprehension Test Performance

APPROVED BY THESIS COMMITTEE:

BYUNGMIN LEE, COMMITTEE CHAIR

JUNGEUN CHOI

YOUNGSOON SO

# ABSTRACT

The Effects of Response Format and Question Type on Reading Comprehension Test Performance

Subin Seong

English Major, Dept. of Foreign Language Education The Graduate School of Seoul National University

The purpose of the current study is to identify the effects of response format, question type and their interaction on measuring Korean mid-tolow proficiency level English learners' L2 reading comprehension ability. A correlation analysis was performed to investigate relationships between different item conditions. Total four conditions which consist of two different response formats (Multiple-choice questions, Open-ended questions) and different question types (asking literal comprehension, asking inferential comprehension). Results from the analysis revealed that items sharing the same response formats were more homogeneous than items sharing the same question type. A mixed-effects modeling was also performed and revealed fixed effects of response format on readers' performance on reading test items. But, no significant effects of question type or interaction between response format and question type were not found. Results from the study suggest that measuring method can be a factor that affects the construct validity of test items. For this reason, the response format should be carefully considered with the purpose of the

reading assessment.

Key Words: Response format, question type, reading assessment, reading

test, validity, Korean English learners

Student Number: 2019-26486

# **TABLE OF CONTENTS**

ABSTRACT
TABLE OF CONTENTSiii
LIST OF TABLESvi
LIST OF FIGURESvi
CHAPTER 1. INTRODUCTION1
1.1 The Background and Goal of the Study1
1.2 Research Questions
1.3. Organization of the Thesis
CHAPTER 2. LITERATURE REVIEW9
2.1 Reading Comprehension
2.1.1 The Simple View of Reading10
2.1.2 A Construction-Integration Model11
2.1.3 Components of Reading Comprehension12
2.2 Reading Assessment15
2.2.1 Response Format
2.2.1.1 Multiple-Choice Questions (MCQs)19
2.2.1.2 Open-Ended Questions (OEQs)21
2.2.2 Response Format Effects
2.2.3 Question Types25
2.2.4 Interaction between Response Format and Question Type 27
2.3 Summary of the Chapter

CHAPTER 3. METHODOLOGY	30
3.1 Participants	30
3.1.1 Participants	30
3.2.1 Item Reviewers	31
3.2 Instruments	32
3.2.1 Reading Test	32
3.2.1.1 Material-Text Selection	32
3.2.1.2 Item Development	32
3.3 Procedure	34
3.3.1 Data Collection	34
3.3.2 Data Analyses	35
CHAPTER 4. Results	41
4.1 Descriptive Statistics	41
4.2 A Correlation Analysis	42
4.3 Mixed-effects Model Analysis	43
CHAPTER 5. DISCUSSION	47
5.1 Effects of Response Formats on Construct Equivalence	47
5.1.1 Correlation of Scores from Multiple-Choice Questions an	nd
Open-Ended Questions	47
5.2 Effects of Response Format and Question Type on Reading	
Comprehension Test Performance	52
CHAPTER 6. CONCLUSION	55
6.1 Major Findings and Implications	55

6.2 Limitations and Suggestions	
REFERENCE	59
APPENDICES	70
ABSTRACT IN KOREAN	

# LIST OF TABLES

Table 3.1 Participants' Mock KCSAT grades for Test A group	.31
Table 3.2 Participants' Mock KCSAT grades for Test B group	.31
Table 3.3 Item Composition of Test A	.34
Table 3.4 Item Composition of Test B	.34
Table 3.5 Descriptions of Models	.39
Table 4.1. Descriptive Statistics for Students' Reading Test Scores	.41
Table 4.2. Results from the Bootstrapped Correlation Analysis	.43
Table 4.3 Estimation of Fixed Effects and Random Effects of Models	.44
Table 4.4. The AIC, BIC and the log-likelihood ratio test results	.45

# LIST OF FIGURES

Figure 2.1 A heuristic for thinking about reading comprehension	
(Snow, 2002, p.12)	13

# CHAPTER 1.

The goal of the study is to investigate the effects of response format on measuring the reading comprehension ability of Korean high school English learners. This chapter introduces the motivation and theoretical framework of the current study. Section 1.1 suggests the background and goal of the study. Section 1.2 presents the study's research questions and the last Section 1.3 offers the organization of the thesis.

# 1.1 The Background and Goal of the Study

Assessment is a crucial component of teaching and learning in that it helps quantify students' academic growth, identify and diagnose students with academic difficulties, plan classroom instruction and quantify the effects of instruction (Cain & Oakhill, 2006). In the context of teaching and learning, the goal of reading assessment is to provide immediate help to a reader with the reader's current developmental stage. Educators diagnose students' abilities and make pedagogical decisions on appropriate instruction and intervention. In the decision-making process, results from reading tests are important evidence for inferencing students' abilities. With the crucial role of assessment and tests in teaching reading, numerous studies have been conducted on the interpretation of reading comprehension test results and factors that affect the test results interpretation (Collins, Compton, Lindström, & Gilbert, 2020; Collins, Lindström, & Sandbank, 2021; Hua & Keenan, 2017; Keenan, Betjemann, & Olson, 2008;

Keenan, Hua, Meenan, Pennington, Willcutt, & Olson, 2014; Keenan & Meenan, 2014). Among the plethora of literature, some researchers emphasized the necessity of careful consideration on reading comprehension test results interpretation by focusing on several factors that affect reading test performance (Alderson, 2000; Collins et al., 2020; Collins et al., 2021; Keenan, Betjemann et al., 2014; Keenan, Hua et al., 2014)

Two major features of reading comprehension are often said to complicate constructing valid assessments and inferencing a reader's ability from reading comprehension tests: (a) a multifaceted-process of reading comprehension (Alderson, 2000; Grabe, 2009) and (b) the opacity of the process (Carpenter & Paris, 2005; Shohamy, 1984). The unobservable process of reading comprehension deepens the matter of reading comprehension test results interpretation. As a reader's reading process cannot be overtly observed, it remains obscure whether a reader's comprehension process corresponds to the intended process to measure.

Recent studies present various theoretical models that depict the dynamic interaction of numerous skills and knowledge in the reading comprehension process (e.g., Direct and indirect effects model of reading; Kim, 2020a, 2020b), indicating a complicated nature of reading comprehension. Some researchers demonstrated the complexity of reading comprehension attributes to considerable inconsistency in diagnosing students' ability with different reading comprehension tests (Collins, Lindström, & Compton, 2018; Collins et al., 2021; Keenan et al., 2008; Keenan & Meenan, 2014). For example, Keenan and Meenan (2014) compared different reading comprehension tests to identify whether the different reading comprehension tests diagnose a child with comprehension

deficiency in a similar way or not. The researchers revealed considerable inconsistency across different reading tests in diagnosing children with comprehension deficits, showing a low average correlation of .37.

The inconsistencies in test performances and decisions based on the test results imply the effects of other variables not intended to measure. Cohen and Upton (2006) stated that "While linguistic and background knowledge appears to be primary sources for individual differences in reading skills, there are numerous other variables" (p.2). The researchers emphasized that the interaction between such variables can also influence test scores and their interpretation. That is, other variables may affect inference on one's ability in target language use. For this reason, the effects of other variables on reading comprehension test performance lead researchers to delve into the sources of test performance variance (Collins et al., 2020; Collins et al., 2021; Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan & Meenan, 2014; Kulesz, Francia, Barnes, & Fletcher, 2016; Ozuru, Best, Bell, Wintherspoon & McNamara, 2007; Ozuru, Briner, Kurby & McNamara, 2013).

The researchers divided sources of reading comprehension test score variances into test characteristics (e.g., response format; see Collins et al., 2020; Collins et al., 2018; Collins et al., 2021; Shohamy, 1984) and reader characteristics (e.g., prior knowledge; see Kulesz et al., 2016; Ozuru et al., 2013). The researchers investigated their contributions on reading comprehension test performance. Among several factors that affect a reader's test performance, response format (Collins et al., 2020; Collins et al., 2021) has long been paid attention to (Cordon & Day, 1996; In'nami & Koizumi, 2009; Kobayashi, 2002; Shohamy, 1984).

Shohamy (1984) examined the effects of response format, which is one of the actively debated test characteristics, on measuring L2 reading comprehension. The researcher compared the effects of multiple-choice questions and open-ended questions, the language of questions (L1 or L2) on students' reading test scores. Results from the study revealed that items with multiple-choice response format were consistently easier than the items with open-ended response format. Also, items presented with L1 were easier than items presented with L2 in general. In addition, the researcher also suggested that the effects of different test characteristics vary across test takers' proficiency level by revealing low-level test takers were susceptible to response format and other test characteristics, while high-level test takers were hardly affected. The researchers attribute the effects of response format on items difficulty to different cognitive activities elicited by different response formats. To be specific, answering multiple-choice questions involves comprehension and selection process while answering open-ended question involves production process, which is regarded as a higher-level cognitive task than selection. For this reason, the researchers viewed production process accompanied by open-ended questions may have contributed to the relative difficulty of open-ended questions.

A recent study of Collins et al. (2021) estimated the contributions of test characteristics (text genre, response format) and reader characteristics (e.g., language knowledge). In their research, the researchers viewed language knowledge as a composite of academic knowledge, listening comprehension and vocabulary knowledge. The researchers showed that the greatest portion of score variance was explained by response format and its estimated effects increased from the emerging language knowledge group to the proficient language

knowledge group, showing the response format effects vary across different characteristics of readers. Results from their research presented response effects varying across first-language learners' language knowledge. The relationship between response format effects and first-language learners' language knowledge indicates interaction between test characteristics and a test-taker's individual characteristics. However, as the study was conducted with L1 children only, further research is required to identify if such interaction effects still exist in L2 reading comprehension test-takers' test-taking process. Furthermore, considering Collins and colleagues included general academic knowledge, listening comprehension and vocabulary in the definition of language knowledge, the definition of it should be different in studies with L2 students. In defining language knowledge for L2 adolescent readers, the feature of L2 adolescent readers that they have more academic knowledge than L1 children should be carefully considered. For this reason, additional studies are necessary to identify the relationship between response format effects and reader characteristics of second language (L2) learners. In'nami and Koizumi (2009) conducted a metaanalysis on the effects of multiple-choice and open-ended formats in L1 reading comprehension test, L1 listening comprehension test, L2 reading comprehension test and L2 listening comprehension test. Results from their research revealed that in L2 reading, there was no significant difference between the two formats in overall. Instead, significant differences were observed when the effects of the two formats were compared in between-subjects design, when subject groups were randomly assigned, when stem-equivalent items were given and when subjects' L2 proficiency level was high.

While many researchers shed light on the response format effects, some

researchers paid attention to the effects of question types (Eason, Goldberg, Young, Geist, & Cutting, 2012; Kulesz et al., 2016; Ozuru et al., 2013). Eason and colleagues (2012) compared the effects of three types of questions: questions examining literal understanding, questions examining the ability to make inferences and questions examining the ability to synthesize and evaluate extracted information from the text. Results from the research indicated that question type can be a source of performance variance, by showing students' performance variances on the three types of questions. Kuelsz et al. (2016) also investigated the question type effects and emphasized that the effects of question type and its interaction effects with other test components should be carefully considered. The importance of confirming the interaction effect is on the construct validity of a reading comprehension test. The interaction effects of test characteristics and question type refers to the interaction between how to measure and what to measure. That is, the interaction may reveal what a test targets to measure can be affected by how to measure the intended ability. However, as a few studies were conducted focusing on the interaction effects of question type and other test characteristics, further studies are required to explore interaction effects among various test-related factors.

To respond to the necessity of further study, the primary purpose of the current study is to examine the effects of response formats and question types on L2 reading comprehension test performance and to help educators or researchers with an in-depth interpretation of L2 students' reading comprehension ability from reading comprehension test results. The current study targets to investigate the effects of response format, question type and their interaction effects. In terms of response format, this study will focus on two representative response formats;

multiple-choice questions (MCQs) and open-ended questions (OEQs). The question type in the current study will include questions that ask literal comprehension and questions that ask inferential comprehension.

#### **1.2. Research Questions**

With the problems hitherto stated, the current study attempts to investigate the effects of response format and question types on L2 reading comprehension test performance. This research goal will be guided by the following research questions.

- What kind of relationships are there between response format effects and question type effects?
- 2. What are the effects of response format and question type on L2 reading comprehension test performance? Are the effects of response format and question type affected by individual differences?

## 1.3. Organization of the Thesis

The present thesis is organized as follows. Following this introduction, Chapter 2 presents the review of literature on models describing reading comprehension, reading assessment, response format effect and question type effects of a reader's performance on a reading test. Chapter 3 explains the methods and designs of the study focusing on participants, the design of test material, the procedure and data analysis. Chapter 4 will report descriptive statistics, results from a correlation analysis and mixed effects model analyses. In Chapter 5, the study's major findings will be discussed based on research questions. Finally, Chapter 6 concludes the study by summarizing the significant findings, proposing implications and addressing the current study's limitations.

# CHAPTER 2.

#### LITERATURE REVIEW

This chapter presents the theoretical framework for the research by reviewing previous research on reading comprehension and reading assessment. The first section 2.1 will explain two representative models of reading comprehension: the simple view of reading and Kintsch's construction-integration model. Section 2.2 will deal with reading assessment and related concerns on the validity Test characteristics and individual characteristics that affect reading comprehension test performance will also be discussed in Section 2.2.

## 2.1. Reading Comprehension

Reading comprehension is a complex process that accompanies multidimensional cognitive activities. The complexity of the comprehension process derives from the process of transforming textual information into mental representations (Alderson, 2000; Catts, 2018; Hoover & Tunmer, 2018; Høin-Tengesdal, 2010; Kintsch, 2012; Kintsch & van Dijk, 1978; Snow, 2002). To explain the reading comprehension process, various models of reading comprehension were suggested. The simple view of reading and Kintsch's Construction-Integration model are widely cited to explain the reading comprehension process.

## 2.1.1 The Simple View of Reading

The simple view of reading (SVR) considers decoding and language comprehension as two major components that drive reading comprehension (Gough&Tunmer, 1986; Hoover & Tunmer, 2018; Kim, 2017). Decoding (*D*) in SVR refers to 'the ability to recognize printed words accurately and quickly' while language comprehension (*C*) refers to 'the ability to extract and construct literal and inferred meaning from linguistic discourse represented in speech' (Hoover & Tunmer, 2018, p.1). The SVR represents reading comprehension (*R*) with a simple equation,  $R=D \times C$ , where each variable ranges from 0 to 1 (Gough & Tunmer, 1986; Hoover & Tunmer, 2018). Recent studies on SVR are also paying attention to the complexity of the reading comprehension process by revealing that both decoding and language comprehension are complicated processes (Hoover & Tunmer, 2018; Kim, 2017).

Some researchers have shown that the two major components of reading comprehension explain the reading comprehension variance of young readers (Catts 2018; Language and Reading Research Consortium & Chiu, 2018; Lonigan, Burgess, & Schatschneider, 2018). For example, Language and Reading Research Consortium & Chiu (2018) found that 94% of the reading comprehension variance of third-grade students is explained by word recognition and listening comprehension which refers to the understanding a text read aloud for beginner readers. Such results corroborate the usefulness of the SVR framework for understanding basic factors of reading comprehension. However, some studies have shown that the contributions of decoding and listening comprehension on reading comprehension vary across grade levels (Tilstra,

McMaster, van den Broek, Kendeou, & Rapp, 2009) and across different reading tests (Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan & Meenan, 2014). Results support the necessity of studies including various factors such as test characteristics of response format and text type to explain performance variances of students' reading comprehension test performance (Cutting & Scarborough, 2006).

# 2.1.2 A Construction-Integration Model

Numerous studies have been accumulated on construction-integration (CI) model, which postulates several different levels of mental representation that a reader constructs while reading through texts (Kintsch, 1988; 1998; 2012). The mental representation consists of three distinct levels: the linguistic surface structure, the textbase and the situation model. The linguistic surface structure refers to a representation of the actual words and syntactic relations between them. This level of representation is easily forgotten. The textbase is 'a representation of propositions of the text and their interrelationships' (Kintsch, 2012, p.22). Textbase is a verbal construction in that it requires knowledge of vocabulary, grammar and conventions of discourse to identify idea units of the text. Lastly, the situation model represents the integration of the textbase meaning, a reader's world knowledge and reading goals. That is, portions of the textbase meaning are reconstructed in relation to the reader's prior knowledge and reading goals while reading through a text. In the process of forming a situation model, new knowledge from the text is integrated with a reader's prior knowledge, constructing a coherent mental representation.

The construction and integration process of reading comprehension and different levels of mental representation remind the complex nature of reading comprehension. Kintsch (2012) stated "Comprehension is not a unitary process that can be measured with a single score. At minimum, it requires a distinction between superficial and deep comprehension—comprehension at the level of the textbase versus the level of the situation model" (Kintsch, 2012, p.33).

# 2.1.3. Components of reading comprehension: the reader, the text and the activity

Snow and RAND Reading Study Group (RRSG) (2002) also shed light on the construction of a coherent mental representation. The RRSG group defined reading comprehension as 'the process of simultaneously extracting and constructing meaning through interaction and involvement with written language' (Snow, 2002, p.11). The RRSG also emphasized that text is not a determinant of reading comprehension. Instead, they suggested three essential elements of reading comprehension which interact with each other under socio-cultural context: reader, text and activity. Figure 1 shows a heuristic for understanding reading comprehension (Snow, 2002). The heuristic is a commonly used framework in reading comprehension research (Collins et al., 2020; Snow, 2018).



Figure 2.1 A heuristic for thinking about reading comprehension (Snow, 2002, p.12)

'The reader' component denotes a subject who reads and comprehends texts and the reader's characteristics. This component embraces reader characteristics that directly or indirectly influence reading comprehension: cognitive capacities (e.g., attention, memory), motivation (e.g., an interest in the topic, reading self-efficacy) and various types of knowledge (e.g., linguistic knowledge, strategic knowledge). Snow (2002) emphasized that differences in the characteristics are sources of reading comprehension variability.

'The text' is an object of reading comprehension, which includes characteristics of texts. In comprehension processes, readers do not simply extract textual information but they also activate their own knowledge to construct different mental representations of the text (Kintsch, 1988, 2012). The effects of text on readers' comprehension have been actively debated and revealed that part of reading comprehension variance is attributable to text characteristics. To be specific, some researchers paid attention to the effects of text genre on reading comprehension (Collins et al., 2020; Eason et al., 2012; McNamara, Ozuru, & Floyd, 2011), effects of text cohesion (Best, Floyd, & Mcnamara, 2008;
McNamara et al., 2011) and effects of text length (Keenan & Meenan, 2014; Kim, 2022; Ozuru, Rowe, O'Reilly, & McNamara, 2008).

Following the reader and text components, 'the activity' is a broad term that embraces any act that a reader engages in while reading through a text. The term also covers purposes of reading, methods for text processing and reading comprehension consequences, which are all affected by reader capabilities and various text features (Collins et al., 2020; Snow, 2002). The current study will pay attention to the activity and investigate the effects of response format and question type which are important elements of test activity that influence reading test and test performance.

Researchers have been actively delving into the effects of the activity component (Collins et al., 2020; Collins et al., 2018; Collins et al., 2021; Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan & Meenan, 2014; Kulesz et al., 2016; Ozuru et al., 2007; Ozuru et al., 2013) and several issues were discussed regarding the activity characteristics. For example, time restriction during reading comprehension tests (Clemens, Davis, Simmons, Oslund, & Simmons, 2015) and reading methods of oral or silent reading (García & Cain, 2014) were investigated. The researchers found that differences in activity components produce different reading test results. Considering the effects of the activity on reading outcomes, there is a growing concern about components that affect reading comprehension test results and the test results interpretation (Collins et al., 2020; Collins et al., 2021).

## 2.2 Reading Assessment

A plethora of research has demonstrated the complexity of the meaning construction process and its multidimensionality. The complex nature of the reading comprehension process showed that reading comprehension ability is a multi-component construct and posed several issues in the field of reading assessment. One of the major issues in the field of reading assessment is the matter of validity.

Test is a measurement instrument designed to elicit a specific sample of an individual's behavior (Bachman, 1990), from which we draw inferences about learners' abilities (Mislevy, 1995). Results from the tests are evidence for inferencing what students know and can do (Mislevy, 1995). The matter of validity occurs when inferencing a reader's ability from test scores.

Validity refers to the appropriateness of interpreting a reader's ability from test scores (Haladyna & Rodriquez, 2013). *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014) defined validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p.11). Thus, the concept of validation can be regarded as "a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (p.11). Furthermore, Rodriguez (2003) accentuated the roles of items and tasks to gather evidence for specific decisions to be made based on the reading test results.

In the field of reading assessment, validity has been a critical issue and

researchers have been paying attention to constructs of reading tests and factors that threaten the validity of reading tests. Due to the complexity of reading comprehension, commonly used reading tests seem to measure slightly different parts of the construct (Collins et al., 2020; Cutting & Scarborough, 2006; Francis, Fletcher, Catts, & Tomblin, 2005; Keenan et al., 2008; Keenan & Meenan, 2014; Snow, 2002).

For example, Keenan, Betjemann and Olson (2008) instantiated that various reading comprehension tests measure different parts of multiple component skills. The researchers compared four different reading comprehension tests in the market: The Gray Oral Reading Test-3 (GORT-3), the Qualitative Reading Inventory-3 (QRI-3), the Woodcock-Johnson Passage Comprehension subtest (WJPC-3) and the Peabody Individual Achievement Test (PIAT).

All four reading comprehension tests have different test characteristics, measuring children's reading comprehension ability with different formats. The GORT-3 asks children to answer multiple-choice comprehension questions after reading aloud narrative and expository passages, the QRI-3 requires students to retell the passage and answer open-ended questions after reading aloud narrative and expository passages. On the other hand, WJPC-3 requires children to read one or two sentences silently and provide a missing word. The PIAT also asks children to silently read a single sentence, but it requires children to match a picture that best describes the sentence. The researchers found that contributions of children's decoding ability and listening comprehension (which refers to the comprehension of a text read aloud) varied across the four tests. That is, different characteristics of reading comprehension tests contribute to different parts of the complex reading comprehension ability.

Furthermore, Keenan and Meenan (2014) found that different reading tests elicit different decisions in diagnosing students with comprehension difficulties. The researchers used the same tests as Keenan et al. (2008) and revealed that different test activities affect the specific comprehension skills assessed, contributing to inconsistent diagnoses of children's reading ability. The researchers of the two studies acknowledged that contributions of reading component skills may vary across different test activities.

In line with the studies that showed the effects of different test activities, researchers have been actively examining sources of reading comprehension performance variance. Various research was conducted in terms of reader characteristics (Collins et al., 2020; Collins et al., 2021), text characteristics (Eason et al., 2012; McNamara, Ozuru, & Floyd, 2011; Kim, 2022) and test characteristics (Clemens et al., 2015; Collins et al., 2020; Collins et al, 2021; Ozuru et al., 2008).

Collins et al. (2020) investigated sources of reading comprehension scores in three components of reading comprehension: reader characteristics, text characteristics and test activity characteristics. The researchers focused on the contributions of a reader's ability (cognitive capacity, word reading fluency, word reading fluency, reading motivation and knowledge), genre of a text and response format (open-ended and multiple-choice questions) to reading comprehension score variance. Results from the study showed a general tendency that students were more likely to answer correctly in multiple-choice questions than in openended questions. The interaction effects between a reader's language knowledge (e.g., general academic knowledge, listening comprehension) and response format were also found, instantiating that students with low language knowledge scores

had less probability to answer correctly in open-ended questions.

In a similar vein, Collins et al. (2021) narrowed down their research interest to the effects of response format and text genre on reading comprehension scores and their interaction effects with language knowledge. The researchers viewed response format and text genre as sources of measurement error and found that response format is a major source of error variance. In addition, the contribution of response format varied with different language knowledge proficiency, posing that response format can be a factor that hampers the appropriate interpretation of reading test results.

To sum up, previous studies investigated sources of reading comprehension test score variance and posed a possibility that response format can be a validity-threatening factor of reading comprehension tests by affecting readers' performance on reading tests.

#### 2.2.1 Response formats

Response format, which is defined as a method for collecting answers from readers (Collins et al., 2020; Walker, 2017), has been regarded as a factor that affects readers' comprehension and the question-answering process (Cain & Oakhill, 2006). Various issues on response format have been actively debated in the field of reading assessment. In the mid of such discussion, there is a growing consensus on that "No one best method for testing reading" (Alderson, 2000, p.203). That is, there is no ubiquitous format that fulfills every intended use of test results. Instead, every response format has its own benefits and drawbacks. Cain and Oakhill (2006) also pointed out that different response formats impose

different processing demands on question answering process. The researchers viewed response format as a factor that affects the sensitivity to identify differences in reading ability.

Some researchers have shown the response format effects on the construct of reading tests. Results from various studies instantiated that different response formats measure slightly different parts of reading comprehension ability, eliciting different educational decisions on the same readers (Clemens et al., 2021; Collins et al., 2020; Collins et al., 2021; Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan & Meenan, 2014; Kobayashi, 2002).

Considering the different characteristics of response formats and their influence on reading assessment, it is necessary to delve into the strengths and weaknesses of each response format and choose the most appropriate one according to the intended purpose of the reading assessment.

Among the various types of response formats, the most commonly used response formats in reading tests are multiple-choice questions and open-ended questions. As the two response formats are widely used in classroom assessments and high-stake tests, numerous researchers shed light on the features of the two response formats and compared them in several ways.

#### **2.2.1.1 Multiple choice question (MCQ)**

Multiple-choice questions (MCQ) have been regarded as a dominant way of assessing reading (Alderson, 2000). Generally, multiple-choice questions ask students to choose one option from several answer options. Haladyna and Rodriguez (2013) recommended the application of MCQs for measuring mental skills and any knowledge of different cognitive demands. The potent merit of using this type of question is its efficiency in large-scale tests (Haladyna & Rodriguez, 2013). Responses collected through MCQs can even be scored by machine. For this reason, numerous large-scale tests apply this type of response format.

Another feature of MCQ is the role of distractors, however the effects of distractors are double-edged. Carefully designed distractors will promote readers' interaction with the given text and reveal the source of readers' comprehension errors. But, at the same time, distractors can be a factor that interferes with readers' natural comprehension process (Alderson, 2000; Rodriguez, 2003). That is, distractors can influence readers' thought processes, giving "possibilities they may not otherwise have thought of" (Alderson, 2000, p.211). The intervention effects of distractors will interfere with measuring students' meaning construction of the given text.

One of the major weaknesses of an MCQ is the ambiguity of the readers' question-answering process. Alderson (2000) pointed out that "the testers do not know why the candidate (test-taker) responded the way she/he did (p.212)". That is, although the test-taker answered correctly, it is unknown if the test-taker constructed the intended mental representation or not.

In line with the ambiguity of question answering process, there has been growing concern about the feasibility of MCQs (Alderson, 2000; Daneman & Hannon, 2001; Katz, Lautenschlager, Blackburn, & Harris, 1990; Rupp, Ferne, & Choi, 2006). For example, Katz et al. (1990) revealed that students performed better-than-chance levels with MCQs even if passages were not given to them. Results from Daneman and Hannon (2001) corroborated Katz et al. (1990) in that

test-takers could achieve better-than-chance level performance without reading given passages on the Verbal Scholastic Assessment Test (VSAT). However, the researchers also revealed that the availability of passages affects test-takers' strategy use. Also, the researchers stated the construct validity of MCQs and provided strategies used in passage available condition as a supporting evidence. In addition, results from another multiple-choice reading comprehension test had higher predictive power on VSAT scores when passage was given. That is, results from passage available condition and unavailable condition may look similar, but the accompanied processes are different. For these reason, the researchers emphasized the validity of MCQs.

# 2.2.1.2 Open-Ended Question (OEQ)

Although the MCQ type is dominantly used, the open-ended question (OEQ) type is also commonly employed in various educational contexts (Rodriguez, 2003). OEQs require test-takers to produce answers in their own language. The answers from OEQs are regarded as reflecting a test-taker's understanding, facilitating the interpretation of what the reader really understood. For that, the OEQ is often suggested as an alternative to the MCQ (Alderson, 2000) as it does not require justification for the answer choice.

One of the weaknesses of an OEQ is its scoring methods. When utilizing OEQs in reading comprehension tests, there should be much consideration on the objectivity of the scoring process. As test-takers produce answers reflecting their own thought, test writers should contemplate various possible answers when writing answer keys. The cost and efforts required for scoring are also weaknesses

 $2\ 1$ 

of OEQs in reading tests.

## 2.2.2 Response Format Effects

A number of literature has been accumulated regarding response format effects. Researchers delved into the features of different response formats and their effects on assessing reading comprehension ability.

As MCQs and OEQs are representative response formats in reading comprehension tests, there has been substantial literature comparing the two response formats (Kobayashi, 2002; In'nami & Koizumi, 2009; Kurby & Mcnamara, 2013; Ozuru et al., 2007; Ozuru et al., 2013; Rodriguez, 2003; Rupp et al., 2006; Shohamy, 1984). Ozuru (2007) classified the accumulated studies into two major streams according to the approach each study adopted: studies applied the statistical approach and studies applied the experimental-oriented approach. For example, the investigation of the effects of MCQs and OEQs on what an item measures has been actively debated by applying a statistical approach (Ozuru et al., 2007; Rodriguez, 2003), while different processes accompanied when answering MCQs and OEQs have been actively studied with the experimental-oriented approach (Campbell, 1999; Ozuru et al., 2013).

Studies on response format effects on L2 reading comprehension test performance have also been conducted (Shohamy, 1984; In'nami & Koizumi, 2009; Kobayashi, 2002; Park, 2017). For example, Shohamy (1984) examined response format effects on L2 test-takers' reading comprehension test. In the study, scores from MCQs and OEQs were compared and the results suggested MCQs are generally easier than OEQs. In addition, the effects of the response

format on L2 test-takers' reading comprehension test scores varied across their proficiency level. Test-takers in the lowest proficiency group were most affected by the response format while test-takers in the highest proficiency group showed little difference by response format.

With a number of studies, some researchers conducted meta-analyses on the response format effects of MCQs and OEQs. (In'nami & Koizumi, 2009; Rodriguez, 2003). However, the meta-analyses slightly differed in their research interests. Rodriguez's (2003) meta-analysis was conducted to identify if MCQs and OEQs measure the same ability (which is referred to as 'construct equivalence' in the research). By synthesizing correlation coefficients between scores from MCQs and OEQs, the researcher revealed substantial heterogeneity of correlations under the stem-equivalent condition. That is, there is a difference in what MCQs and OEQs really measure.

In'nami and Koizumi (2009) also conducted a meta-analysis comparing MCQs and OEQs. While Rodriguez (2003) paid attention to the effects of response format and its effects on what a question measures, In'nami and Koizumi (2009) shed light on the effects of response format on test scores and the relative difficulties of items. The researchers synthesized results from studies that compared mean scores of MCQs and OEQs and no significant response format effect was detected in L2 reading tests. Instead, if questions were given with stem-equivalent questions or if the test data was collected from high-proficiency test-takers, MCQs were generally easier than OEQs.

Although Hohesinn and Kubinger (2011) included both L1 subjects and L2 subjects in their study, the researchers offered supporting evidence for both Rodriguez (2003) and In'nami and Koizumi (2009). The researchers investigated

if MCQs and OEQs measure the same reading comprehension ability and if response format affects the difficulty of reading comprehension test items. Their IRT models indicated that questions with different response formats are measuring the same latent traits. Instead, different response formats significantly affected item difficulty by showing multiple-choice items tend to be easier than open-ended items.

To summarize, the studies reviewed above show incongruent views toward response format effects on what a question measures. Instead, the studies reached an agreement on the response format effects on item difficulty.

On the other hand, some researchers indicated response format effects on constructs of reading comprehension tests by focusing on processes accompanied when answering the two different response formats (Campbell, 1999; Ozuru et al., 2013). Campbell (1999) demonstrated that OEQs lead test-takers to more active engagement to the texts and elicit their higher-level cognitive processing. Also, Ozuru and colleagues (2013) found test-takers' scores from OEQs were positively correlated with self-explanation scores, while scores from MCQs were significantly correlated with their prior knowledge. Results from the study implied that answering to OEQs requires more active engagement in text comprehension processes.

Although previous studies have shown an agreement on response format effects on item difficulty only, recent studies are shedding light on its effects on what a test measures. Recent studies on reading comprehension assessment indicate the effects on response format by suggesting that different response formats measure slightly different parts of the multi-component construct (Collins et al., 2020; Collins et al., 2021; Cutting & Scarborough, 2006; Keenan et al.,
2008; Keenan, & Meenan, 2014; Kulesz et al., 2016; Ozuru et al., 2007; Ozuru et al., 2013). A recent research by Collins and colleagues (2020) showed the effects of response format on predicting correct responses on MCQs and OEQs. Collins and colleagues revealed that when text characteristics (e.g., text genre) and reader abilities (e.g., attentive behavior, language knowledge, learning motivation, nonverbal reasoning, word reading fluency, working memory) are controlled, the possibility of a correct response for MCQs was greater than OEQs. Such results imply that the same students' reading comprehension ability can be measured differently when they are given in different response formats. Furthermore, the researchers emphasize that measuring students' comprehension with only one response format may lead us to a limited estimation of their ability. In line with the study, Collins et al. (2021) also demonstrated that response format accounted for a comparable degree of variance in reading comprehension scores.

Although a plethora of literature has been accumulated on response format effects, there is a lack of congruence on the effects of response format on specific constructs of a reading comprehension test. To fill the gap, the current study aims to investigate the response format effects on what a reading test measures by performing a correlation analysis and applying mixed effects modeling. In addition, this study targets to identify the response format effects in relation to question type.

### 2.2.3 Question type

Question type, which is also a part of test activity characteristics, has been regarded as a factor that exerts an impact on the comprehension processes and

reading test performance (Andreassen & Bråten, 2010; Cerdán, Vidal-Abarca, E., Martínez, Gilabert, & Gil, 2009; Eason et al., 2012; Kulesz et al., 2016; Basaraba, Yovanoff, Alonzo, & Tindal, 2013). In the current study, question type is deeply related to what a question measures. The current study divided question type into literal questions and inferential questions. According to Eason (2012), literal questions are defined as questions that assess readers' recall of information explicitly stated in texts, while inferential questions are defined as questions that require readers to build their own mental representation.

Dividing question type into literal and inferential questions is a commonly used categorization in commercial tests (e.g., Qualitative Reading Inventory) and reading assessment studies (Basaraba et al., 2013; Best et al.,2012; Eason et al., 2012; Hua & Keenan, 2014; Leslie & Caldwell, 2001). Based on Kintsch's (1998) model, some researchers have been emphasizing the importance of questions measuring different levels of representation respectively to specify a reader's sources of reading difficulty (Basaraba et al., 2013; Eason et al., 2012; Hua & Keenan, 2014; Kintsch, 2012; Kulesz et al., 2016).

Some researchers paid attention to question types and the different cognitive demands they impose on readers (Eason et al., 2012; Kulesz et al., 2016). For example, Eason and colleagues (2012) examined effects of text and question types and their interaction effects on reader's performance. The researchers revealed readers' performance differences on different question types.

In terms of question type effects on test performance, the literal comprehension questions are typically regarded to be easier than inferential comprehension questions (Basaraba et al., 2013; Eason, 2012; Hua & Keenan, 2014). Basaraba and colleagues (2013) investigated the relationship between

question type and item difficulty. The researchers revealed that literal questions tend to be easier than inferential questions, but literal questions can be more challenging than inferential questions. Hua and Keenan (2014) also compared literal comprehension questions and inferential comprehension questions. Results from their study revealed that inferential questions require readers to remember more of the texts to answer correctly.

Kulesz and colleagues (2016) also conducted research on the role of question types and their processing demands. On the contrary to previous studies, the researchers found that the different question types were not highly predictive factors of item difficulty. The researchers also emphasized the necessity of investigating the effects of question type and their interaction effects with other characteristics of reader, text and test related characteristics to elicit appropriate decisions on what readers know and can do.

### 2.2.4 Interaction between response format and question type

As reviewed above, response format, question type and their interaction effects are deeply related to construct validity of a reading test. However, only a few studies were performed on the interaction effects between them (Ozuru et al., 2007; Ozuru et al., 2013). Ozuru et al. (2007) examined relationship between scores from MCQs and OEQs and their relationship varies across three different question types. The three question types refer to questions asking text-based comprehension, local bridging inference and global bridging inference. The study did not indicate significant correlation between MCQs and OEQs when a text is available. Ozuru et al. (2013) also investigated the effects of response format

(MCQs, OEQs) and question type (questions require text-based comprehension, questions require local bridging inference and questions require global bridging inference) on L1 participants' reading test performance. Results from the study suggested response format effects and question type effects on item difficulty. To be specific, the study corroborates the previous studies that revealed that MCQs are generally easier than OEQs. In terms of question type effects, questions asking global bridging inference were more difficult than text-based or local inference questions. In addition, the researcher also indicated that different response formats may affect what a question measures. However, the researchers did not examined the interaction between response format and question type.

In L2 context, some researchers investigated the interaction between question type and individual characteristics (e.g., working memory, language knowledge, vocabulary knowledge) (Kim, 2023, Lim, 2019). The researchers performed studies on the relationship between question type and required reading sub-skills. However, the interaction between question type and response method has rarely been studied.

### 2.3 Summary of this chapter

As discussed in the first section of this chapter, reading comprehension is a complex process that requires readers to construct multi-level mental representations. The complex nature of reading poses a difficulty in 1) designing reading tests and 2) inferencing a reader's ability from the test results.

For this reason, many researchers have been delving into the factors that affect test-takers' reading comprehension test performance. One of the actively

debated issues is the effects of response format. Numerous studies have been conducted on response format effects on reading assessment. Researchers revealed response format effects of MCQs and OEQs on item difficulty. However, response format effects on what a test measures remain inconclusive, requiring further studies on the issue. Another important issue is a question type, which is highly related to what a question measures. The current study divided question type into literal comprehension questions and inferential comprehension questions. Researchers have shown that different question type requires different cognitive demands, causing differences on item difficulty.

Previous studies have shown that response format and question type are important test components. Especially, response formats can be a validitythreatening factor by affecting the ability that a question measures. Despite their importance in reading comprehension tests, the relationship between the two components and their interaction effects were scarcely investigated in the L2 context. Considering that the interaction between response format and question type can affect the construct validity of a reading comprehension test, the current study targets to identify the effects of the response format, question type and their interaction on measuring reading comprehension ability of Korean L2 English learners.

# CHAPTER 3

### METHODOLOGY

This study applied a statistical approach to investigate construct equivalence between MCQs and OEQs and to identify the effects of response format and question type on L2 reading test performance. Section 3.1 explains the participants of the study and their characteristics and Section 3.2 introduces the instruments of the study. Section 3.3 describes the data collection process and data analysis procedures in detail.

### **3.1 Participants**

### 3.1.1 Participants

In this study, two groups of participants were recruited. One group was recruited for reviewing test items and the other group was the experimental group for the study. The first group was recruited to test the difficulty of reading test items. This group consisted of seven high school 2<sup>nd</sup>-grade students. The students' mock KCSAT (Korean College Scholastic Aptitude Test) English results ranged from 3-6 grades. Students in this group have similar proficiency with the students in the experimental group.

The experimental group was the test-takers group, consisting of 29 high school 2<sup>nd</sup>-grade students. The participants' mock KCSAT English grade scores were collected and referred to allocate them into two different groups. The two groups were consisted for comparing the effects of different response formats. One group took Test A and the other group took Test B. Items with MCQs in Test

A were provided with OEQs in Test B with the same item stems. Their mock KCSAT English grades ranged from 3 to 8. The specific distribution of test-takers' mock KCSAT English grades is presented in Table 3.1.

Mock KCSAT	Number of
grade	students
4	4
5	4
6	2
7	3
8	2
Total	15

Table 3.1 Participants' Mock KCSAT grades for Test A group

fable 3.2 Participants	' Mock KCSAT g	grades for Te	st B group
------------------------	----------------	---------------	------------

Mock KCSAT	Number of
grade	students
3	1
4	3
5	4
6	2
7	1
8	3
Total	14

## 3.1.2 Item Reviewers

In item writing process, four graduate students were recruited to judge whether the reading test items measure the intended literal comprehension ability or inferential comprehension. Two students were studying for a doctoral degree and the other two were studying for a master's degree. And one of them participated in the process of scoring participants' responses on OEQs. They were asked to confirm if reading comprehension test items require literal or inferential comprehension of reading comprehension.

### **3.2 Instruments**

### 3.2.1 Reading test

### **3.2.1.1 Material-Text selection**

To develop items for L2 reading test, two expository texts were excerpted from high school 2<sup>nd</sup>-grade English 1 text book (Donga, Kwon). None of the participant has read the given texts. The title of the texts were 'IoT: Connecting everything (Chapter five)' and 'Palm oil, the biggest threat to orangutans (Chapter three)'.

## **3.2.1.2 Item development**

Item development procedure consisted of three stages. In the first stage, the researcher made 12 questions for each text, six questions requiring literal comprehension to answer and the other six questions requiring inferential comprehension. In item writing processes, Leslie and Caldwell (2011) and Lane, Raymond and Haladyna (2015) were referred to in designing two types of reading comprehension questions. To be specific, commonly used question stems for literal and inferential questions from Leslie and Caldwell (2011) were reviewed. Item writing guidelines from Lane et al. (2015) were also referred in the process of writing item stem and options of test items.

To investigate the effects of response format and question type, two forms

of tests were developed in a counterbalanced way. Subjects were allocated to test A group or test B group and took test A or test B. The group allocation was conducted to control their proficiency. Students in test group A took test A while test group B took test B. Both test A and test B were composed of stem-equivalent questions. For example, if a question that requires literal comprehension was given with MCQ in Test A, the same question was given with OEQ in Test B. In the same way, questions given in an MCQ in Test A were given in OEQ in Test B. As a result, every item was given in both MCQ and OEQ.

In the second stage, a pilot study was conducted to test the difficulty of the test items. Seven high school 2<sup>nd</sup>-grade students were recruited for the pilot study through online postings. The recruited students' mock KCSAT grade score range was similar to that of test-taker group students. The students were instructed to read the two texts and answer 24 questions. As there were only 7 students in the pilot study, the difficulty of each item was calculated as a proportion of correct answers. Items with correct answer rates under 20% and above 90% were deleted from the test. Through this process, total of 20 questions were left.

In the following stage, 10 literal comprehension questions and 10 inferential comprehension questions were under investigation for their item validity. To ensure that the questions required literal or inferential comprehension as intended, item reviewers confirmed the required comprehension level of the questions. They read the texts and classified each question into a literal or inferential comprehension question. After the judgment process, the researcher compared their results of classification and excluded questions that showed incongruence in their judgment from the test. As a result, 8 questions for Text 1 and 8 questions for Text 2 were left. Every test taker was given 4 MCQs asking

literal comprehension, 4 MCQs asking inferential comprehension, 4 OEQs asking literal comprehension and 4 OEQs asking inferential comprehension. A detailed structure of the test design is summarized in Table 3.2. and Table 3.3.

Table 3.2 Item composition of Test A				
	Pasnonsa format	Level of	Item	
	Response format	comprehension	number	
	MCO	Literal	1, 5	
Toyt 1	MCQ	Inferential	4, 8	
OEQ		Literal	3, 6	
	UEQ	inferential	2, 7	
MCQ Text 2 OEQ	Literal	2, 5		
	MCQ	Inferential	6, 8	
	OFO	Literal	1, 4	
	Inferential	3, 7		

Table 3.3 Item composition of Test B Level of Item Response format comprehension number Literal 3, 6 MCQ Inferential 2,7 Text 1 Literal 1, 5 OEQ inferential 4,8 Literal 1,4 MCQ Inferential 3, 7 Text 2 2, 5 Literal OEQ Inferential 6, 8

### **3.3 Procedures**

# 3.3.1 Data Collection

High school 2<sup>nd</sup>-grade students could contact the researcher freely if they were interested in the experiment. Students who agreed to participate in the

experiment submitted a consent form to the researcher and scheduled a meeting for the reading comprehension test. In this stage, students sent their mock KCSAT (on March) English grades. Then, participants were assigned to a group for test A or test B. The group allocation was conducted to control students' English proficiency and examine the effects of response format and question type. The average mock KCSAT English grade for group A was M=5.67, SD=1.40 (N=15) and for group B was M=5.57, SD=1.59 (N=14).

After group allocation, students took the reading test at the scheduled time. All participants had 40 minutes to complete the test. Students' performance on MCQs was scored by using answer keys and their performance on OEQs was scored by two raters. The raters scored students' responses as correct if the student included important keywords in the answer. Both raters scored the whole OEQs and discussed when there was an incongruence between their ratings. Following the scoring process, the correct answer was rated either 1 or 0.

### 3.3.2 Data Analyses

To answer the research questions this study conducted two quantitative analyses. A correlation analysis and a mixed effects model analysis were performed to identify the effects of response format and question type. All the quantitative analyses were performed on R version 4.2.1., which is an open source statistical program.

### **3.3.2.1** Correlation Analysis

Campbell and Fiske (1959) emphasized that in the process of validation, two types of validation are required to establish construct validity: 1) convergent validation, 2) discriminant validation. Convergent validation indicates that independent measures of the same trait should converge while discriminant validation refers that tests intended to measure different traits should diverge.

Items in the current study consist of items measuring two different traits (literal comprehension, inferential comprehension) with two different measuring methods (MCQ, OEQ). In other words, different traits are measured in two different ways and the measurement features can also contribute to test score variance. For that, the correlation analysis will be performed to investigate relationships between items measuring two different reading comprehension abilities with two different response formats.

To be specific, the goal of the correlation analysis is to test convergence between items measuring the same reading ability and divergence between items measuring different reading abilities. If results from the correlation analysis show a higher correlation between items measuring the same reading ability, the results will indicate the convergence validity of items measuring the same reading ability and divergence validity of items measuring different reading abilities. Then, it can be said that the items are measuring the same reading ability regardless of different response formats. On the other hand, a higher correlation between items with different response format may represent stronger effects of response format than the effects of what an item measures. In that case, construct validity of items will be weakened by different response formats.

# 3.3.2.2 Mixed effects model

To answer the second research question that focused on the effects of response format and question type on L2 reading test performance, mixed effects models were applied in the current study.

A mixed-effects model is a statistical model that includes both fixed effects and random effects (Cunnings & Finlayson, 2015). In mixed-effects modeling, independent variables are often designed as fixed effects and possible effects that cause individual variation are designed as random effects. Another characteristic of mixed effects model is its robustness. It is often regarded as a robust method in analyzing data from repeated-measures design (Cunnings & Finlayson, 2015; Larson-Hall, 2015; Linck & Cunnings, 2015; Schielzeth, Dingemanse, Nakagawa, Westneat, Allegue, Teplitsky, Réale, Dochtermann, Garamszegi & Araya-Ajoy, 2020). Considering that subjects in this study answered literal comprehension and inferential questions repeatedly (with different response formats), the current study can be said to have a repeateddesign.

In the current study, mixed-effects modeling was conducted to identify the fixed effects of response format, question type and their interaction effects on second language learners' performance on L2 reading comprehension test and to examine random effects of individual differences.

Especially in second language studies, individual difference, which can easily be overlooked in some statistical analyses, is often designed as a random effect (Linck & Cunnings, 2015; Cunnings & Finlayson, 2015). By considering both fixed effects and random effects, mixed effects models in second language

studies are regarded as "potentially offering a fruitful way of examining how individual differences may affect L2 acquisition." (Linck & Cunnings, 2015, p.186). The goal of the current study is not only to identify the effects of response format, question type and their interaction effects and but also to provide further explanation on variance that the effects of response format or question type cannot explain. To achieve the goal, the current study included the effects of individual differences, which is mock KCSAT grade in this study, as a random effect. For this reason, mixed model analyses were performed to estimate the fixed effects of response format, question type and their interaction and the random effects of individual differences on reading comprehension test score variance.

The linear mixed-effects models included response format (multiple choice vs. open-ended questions), question type (literal question vs. inferential question) and their interaction as fixed effects. The fixed effect factors of response format and question type were recoded using treatment coding: multiple choice response format was coded as -0.5, while the open-ended format was coded as 0.5, the literal comprehension question was coded as -0.5 and the inferential comprehension question was coded as 0.5.

To account for the unexplained score variance by response format, question type and their interaction effects, three models with different random effect structures were compared. The specific structures of the two models are presented in Table 3.4.

	Fixed effects	Random effects		
		Random intercept	Random slope	
Model 1	response format, question type, and their interaction	Mock KCSAT English grade		
Model 2	response format, question type, and their interaction	Mock KCSAT English grade	Response format	
Model 3	response format, question type, and their interaction	Mock KCSAT English grade	Question type	

**Table 3.4 Descriptions of models** 

Model 1 includes mock KCSAT English grades as a random intercept to explain the score variance not identified by the effects of response format, question type and their interaction. Model 2 includes a random slope to confirm whether the response format effects vary across individuals' mock KCSAT grades, while Model 3 includes question type as a random slope to identify if question type effects vary across individuals' mock KCSAT English grades.

**Model 1.** The first model includes response format (multiple-choice vs. open-ended), question type (literal question vs. inferential question) and their interaction as fixed effects. In terms of random effects, the model included each subject's mock KCSAT English grade, which represents individual difference, as a random intercept.

**Model 2.** The second model also includes response format and question type as fixed effects. The only difference with Model 1 is its random effects. In the second model, a subject's mock KCSAT English grade is also included as a random intercept and response format effects varying by mock KCSAT English grade is added as a random slope. **Model 3.** The third model also has the same fixed effects design. The only difference is on its random effects design. While Model 2 includes response format effects varying by mock KCSAT English grade, Model 3 includes question type effect varying by mock KCSAT English grade is added as a random slope.

Linear mixed-effects models in the current study were analyzed using R version 4.2.1., which is an open source statistical program. In the process of modeling linear-mixed effects models, lmer () function of the 'lme4' package version 1.1.31 was applied with a restricted maximum likelihood estimation technique. And anova() function was employed to compare the models and identify the best-fitted model. In the model comparison process, maximum likelihood (ML) estimation technique was applied (Cunings & Finlayson, 2015, Hoffman, & Rovine, 2007).

# CHAPTER 4

### RESULTS

In this chapter, results from the quantitative analyses will be presented. This chapter includes descriptive statistics for the subjects' reading comprehension test results, results from a correlation analysis and mixed effects analysis.

### 4.1 Descriptive Statistics

Table 4.1 shows descriptive statistics for the means and standard deviations of students' reading comprehension test performance with different response formats and question types. The perfect score for each item condition was 4 points and the total score for the test was 16 points.

Table 4.1. Descriptive statistics for students Teading test scores					
Response	Question	Scores			
format	type	n	M	SD	95% CI
MC	Literal	29	2.17	1.42	[1.63, 2.71]
MC	Inferential	29	2.00	1.25	[1.52, 2.48]
OE	Literal	29	1.83	1.26	[1.35, 2.31]
OE	Inferential	29	1.00	1.34	[0.49, 1.51]

 Table 4.1. Descriptive statistics for students' reading test scores

Table 4.1 indicates literal comprehension questions with MC questions show the highest mean score of 2.17 out of 4 points while OE inferential questions show the lowest mean score of 1.00 out of 4 points. Considering the results give overlapping confidence intervals (95% CIs), further analysis is required to compare the effects of response format and question type. To be specific on 95% CIs from the four conditions, there are some overlapping among 95% CI ranges. The only exception is on 95% CIs of MC inferential and OE inferential. Although it is a slight difference, there is no 95% CI overlap between OE inferential questions and MC inferential questions.

## 4.2 A Correlation Analysis

To investigate the relationships between the effects of different response formats and question type, a correlation analysis was performed first. The goal of the analysis was to identify relationships among results from MCQs asking literal comprehension, MCQs asking inferential comprehension, OEQs asking literal comprehension and OEQs asking inferential questions and to discuss convergent validity and divergent validity of the items with different conditions.

Assumption-checking tests showed that data from the current study has a linear relationship between each pair of variables, non-normal distribution and unequal variance. For this reason, robust methods were considered (Larson-Hall, 2015) and a Pearson correlation analysis was performed with 1,000 bootstrapped samples.

As shown in Table 4.2, results from the analysis indicated significant correlation between MCQs asking literal comprehension and MCQs asking inferential comprehension (r=0.423, CI [.025, .704]), MCQs asking literal comprehension and OEQs asking literal comprehension (r=0.391, CI [.078, .650]), between OEQs asking literal comprehension and OEQs asking literal comprehension and OEQs asking literal comprehension and OEQs asking literal comprehension (r=0.391, CI [.078, .650]), between OEQs asking literal comprehension and OEQs asking literal comprehension asking literal comprehension and OEQs asking literal comprehension a

significant correlation was found between MCQs asking literal comprehension and OEQs asking inferential comprehension, between MCQs asking inferential comprehension and OEQs asking literal comprehension and between MCQs asking inferential comprehension and OEQs asking inferential comprehension. The highest correlation was found between OEQs asking literal comprehension and OEQs asking inferential comprehension, followed by the correlation between MCQs asking literal comprehension and MCQs asking inferential comprehension. The lowest correlation was found between MCQs asking inferential comprehension.

Table 4.2 Results from the bootstrapped Correlation Analysis					
	MC	MC	OE	OE	
	-literal	-inferential	-literal	-inferential	
MC	1	0.423*	0.391*	0.340	
-literal					
MC		1	0.000	-0.043	
-inferential					
OE			1	.527**	
-literal					
OE				1	
-inferential					

 Table 4.2 Results from the Bootstrapped Correlation Analysis

\*. Correlation is significant at the 0.05 level (2-tailed).

\*\*. Correlation is significant at the 0.01 level (2-tailed).

### 4.3 Mixed Effects Analysis

To answer the second research question, three mixed-effects models were compared. The three models had different random effects designs. Results from the analysis showed that only Model 1 and Model 2 have a significant model fit. Model 3 showed a singular convergence error, indicating that the estimated contribution of random effects is converging to zero (R Core Team, 2014). For this reason, only results from the Model 1 and Model 2 are reported in Table 4.3.

Fixed effects					Random effects
Parameters	Estimate	SE	t	р	SD
Model 1					
Intercept	5.716	0.222	25.692	0.000**	0.446
Response format	-0.655	0.228	-2.872	0.004**	_
Question type	-0.448	0.228	-1.965	0.051	_
Response format x Question type	-0.621	0.456	-1.361	0.176	_
Model 2					(by response format)
Intercept	5.731	0.226	25.344	0.000**	0.462
Response format	-0.529	0.387	-1.366	0.175	0.735
Question type	-0.448	0.222	-2.018	0.046*	_
Response format x Question type	-0.621	0.444	-1.397	0.165	_

 Table 4.3. Estimation of fixed effects and random effects of Models

Note. All factors were re-coded using contrast coding, as follows

: Response format (-0.5=multiple choice, 0.5= open-ended), Question format (-0.5=literal, 0.5=inferential).

Model 1 formula: Score~Response format\*Question type+(1 |mock KCSAT grade). Model 2 formula: Score~Response format\*Question type+(1+Response format |mock KCSAT grade)

\*. The effect is significant at the 0.05 level (2-tailed).

\*\*. The effect is significant at the 0.01 level (2-tailed).

Model 1 included fixed effects of response format effects, question type

effects and their interaction effects (Response format x Question type) on

comprehension scores. In the model, subjects' mock KCSAT grades were

included as a random effect to explain the variance not accounted by the effects of

response format, question type and their interaction. The model revealed

significant fixed effects of response format (estimate=-0.655, SE=0.228,

t=-2.872) only. The effects of question type nor interaction between response

format and question type did not indicate significant results, with the absolute value of the *t* statistics less than 2.0 (Link & Cunnings, 2015).

Model 2 also included the same fixed effects with Model 1. The only difference was on the design of random effects. In the Model 2, subjects' mock KCSAT grades were included as an intercept and the effect of response format varying by subjects' mock KCSAT grades was included as a random slope. Results from Model 2 indicated reversed results of Model 1, revealing only the question type effect (estimate=-0.448, SE=0.222, t=-2.018) was significant in Model 2. The response format effect and interaction effects between response format and question type were not observed. As the two models differed in their results, a model-fit comparison was conducted to choose the best-fitted model. Table 4.4 shows AIC, BIC and log-likelihood of Model 1 and Model 2.

,	8		
Number of parameters	AIC	BIC	Log-likelihood
6	395.23	411.75	-191.61
8	398.36	420.39	-191.18
	Number of parameters 6 8	Number of parametersAIC6395.238398.36	Number of parameters         AIC         BIC           6         395.23         411.75           8         398.36         420.39

Table 4.4. The AIC, BIC and the log-likelihood ratio test results

Considering the model-fit statistics of Akaike's (1974) information criterion (AIC) and Bayesian information criterion (BIC), Model 1 showed better model-fit than Model 2. For this reason, Model 1 was selected as the final model of the current study.

The final model indicates the estimate of 5.761 for intercept, indicating the mean of test score when there is no effect of response format, question type and their interaction. The estimate of response format -0.655 indicates that the mean score from OEQs is 0.655 lower than that of MCQs. No significant effect of question type nor the interaction between response format and question type was detected in the final best-fitted model.

# CHAPTER 5.

### DISCUSSION

In this chapter, summaries of the major findings of the study are presented and discussed regarding research questions.

### 5.1 Effects of Response Format on What an Item Measures

The ultimate goal of the study is to confirm the effects of measuring methods and measuring traits on measuring mid-to-low English proficiency level Korean English language learners' L2 reading comprehension ability. To achieve the goal, the first research question elicited for this study pertains to the relationship between response format effects and question type effects. The second research question of the current study paid attention to the effects of response format, question type and their interaction by applying mixed-effects modeling.

# 5.1.1 Correlation of Scores from Different Response Formats and Question Types

Many researchers have been delving into the effects of response format on the construct of an item, but the discussion is still inconclusive (Collins et al., 2020; Collins et al., 2021; In'nami & Koizumi, 2009; Kobayashi, 2002; Ozuru et al., 2007; Ozuru et al., 2013; Rodriguez, 2003) and fewer studies were performed on the relationship between response format and question type (Ozuru et al., 2007; Ozuru et al., 2013). To answer the first research question, a correlation analysis was conducted to identify the relationship between response format and question type on L2 reading comprehension test performance. Relationships between items sharing the same response format and between items measuring different levels of L2 reading comprehension. Scores from the four-item conditions were analyzed: literal comprehension questions with MCQs, literal comprehension questions with OEQs, inferential comprehension questions with MCQs and inferential comprehension questions with OEQs.

Before conducting the correlation analysis, two hypotheses were postulated. First, a higher correlation between items sharing the same response format will indicate that response format affects what an L2 reading comprehension test item measures and it will threaten the construct validity of items. Second, a higher correlation between items sharing the same question type will indicate the effects of question type exceed those of response format on what an L2 reading comprehension test item measures.

Considering that different question types in the current study indicate differences in the intended comprehension ability to measure, results from the correlation analysis pose a possibility that different response formats may affect what an item measures. The correlation analysis indicated higher correlations between items sharing the same response format than items measuring the same level of reading comprehension (sharing the same question type). The highest correlation was found between scores from literal comprehension questions with OEQs and inferential comprehension questions with OEQs (r=.527), followed by the correlation between literal comprehension questions with MCQs and inferential comprehension questions with MCQs (r=.423). Scores from literal

comprehension questions with MCQs and OEQs revealed a moderate correlation (r=0.391) while scores from inferential comprehension questions with MCQs and OEQs did not show a significant correlation between them. In other words, items sharing the same response format are more homogeneous than items measuring the same part of reading comprehension ability. Such results are supporting the first supposition that postulated the effects of response format on what an L2 reading comprehension test item measures. Although cautious interpretation is required with the small sample size of the current study, results from the current study pose a possibility that the effects of response format on measuring Korean L2 English learners' reading comprehension test overweigh the effects of what an item measures.

According to Campbell and Fiske (1959), construct validity is achieved when items measuring the same trait are homogeneous and items measuring different traits are heterogeneous. However, the current study revealed items measuring different levels of reading comprehension with the same response format can be more converging than items measuring the same level of reading comprehension with different response formats. That is, the construct validity of the items can be threatened by measuring methods, at least for measuring Korean mid-to-low proficiency level English language learners' reading comprehension ability. These findings cast doubt on the perspective that response format does not affect the construct of an item. While some researchers claimed that response format does not affect what a test item measures (In'nami & Koizumi, 2009) results from the current study corroborate studies that revealed different response format affect what an item truly measures (Collins et al., 2020; Collins et al., 2021; Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan, & Meenan,

2014; Kulesz et al., 2016; Ozuru et al., 2007; Ozuru et al., 2013; Shohamy, 1984).

Ozuru and colleagues (2007) also analyzed the correlation between scores from MCQs and OEQs and indicated a non-significant correlation between MCQs and OEQs in three different question types (text-based comprehension questions, local bridging inference questions and global bridging inference questions). The study applied different question type classifications, however, text-based comprehension is comparable to literal comprehension questions while local bridging inference questions and global bridging inference questions are included in inferential questions of the current study. Results from the correlation analysis of the study are on contrary to those of Ozuru et al. (2007). While Ozuru and colleagues showed a significant correlation between MC and OE local inference questions, no significant correlation was observed between MCQs asking inferential questions and OEQs asking inferential questions in the current study. To explain the gap between the results from the two research, the characteristics of test-takers can be referred to. Subjects of the Ozuru et al. (2007) were undergraduate students in the US. On the other hand, subjects in the current study are high school English learners who achieved mid-to-low grades in mock KCSAT. For that, the different characteristics of subjects, such as English proficiency and cognitive maturity, may have contributed to different results from the correlation analyses.

In addition, higher correlations between questions sharing the same response format indicate that different response formats may tap into different cognitive processes when readers are answering inferential questions. The differences in cognitive processes attributable to different test formats have also been emphasized by several researchers. For example, Cutting and Scarborough

(2006) indicated that different test formats resulted in measuring different cognitive processes of reading comprehension. Keenan et al. (2008) and Keenan & Meenan (2014) corroborated the test format effects on constructs of a reading comprehension test. In line with previous studies, the current study poses a possibility that response format, which is an important part of test characteristics, affects what an item measures. To be specific, the process of selecting the best option and producing the target answer may accompany different cognitive activities.

What is noteworthy is that while literal questions with MCQs and OEQs indicated a moderate correlation, there was no significant correlation between inferential questions with MCQs and with OEQs. A possible explanation can be made that answering inferential questions with OEQs may have been a burden for the subjects. The subjects of the current study are mid-to-low level proficiency level high school students. For that, some of the students could have struggled with identifying meanings of the words or sentences. That is, some of the students may have allocated much more cognitive resources to identify literal meanings of the text, leaving a smaller amount of cognitive resources for inferential comprehension.

Still, there is a need for further investigation on what features of inferential questions caused a non-significant correlation although they were intended to measure the same part of reading ability.

To sum up, the correlation analysis revealed that although questions were intended to measure different constructs, the response format of questions may affect the cognitive activities required for answering the comprehension question.

# 5.2 Effects of Response Format and Question Type on L2 Reading Comprehension Test

To examine the effects of response format and question type on L2 reading comprehension test scores, three mixed-effects models were compared and the final best-fitted model was selected. Model 1, which is the final bestfitting model of the current study, contains fixed effects of response format, question type, the interaction between them and subjects' mock KCSAT grade as a random effect.

There was a significant fixed effect of response, however, there was no significant effect of question type or interaction between them. The results are in line with previous studies shedding light on the effects of response format on reading comprehension tests (Collins et al., 2020; Collins et al., 2021; Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan, & Meenan, 2014; Kulesz et al., 2016; Ozuru et al., 2007; Ozuru et al., 2013).

The negative estimate of the response format effects on L2 reading comprehension test score indicates that the average scores from MCQs were higher at a rate of 0.655 than those of OEQs. That is, subjects scored higher in items with multiple choice response format and they were easier than the OEQs on average, corroborating the studies that revealed a tendency that MCQs to be easier than OEQs (Hohesinn & Kubinger, 2011; In'nami & Koizumi, 2009; Shohamy, 1984). In'nami & Koizumi's (2009) meta-analysis concluded that the accumulated literature on response format effects shows a tendency that MCQs are easier than OEQs in general. In line with the meta-analysis, the tendency was also detected in the current study. The score difference of 0.655 can be viewed as a slight difference, however, regarding the average scores from the four different item conditions ranging from 1 to 2.17, the difference of 0.655 is a considerable difference. The results are in line with the results from Collins et al. (2021) study where response format explained the greatest amount of variance in reading comprehension test performance.

The effects of question type and the interaction between response format and question type were not revealed in the final model of the current study. Previous research has shown the effects of question type on item difficulty by indicating that literal comprehension questions tend to be easier than inferential questions (Basaraba et al., 2013; Eason, 2012; Hua & Keenan, 2014). However, the effects of question type and the effects of interaction between response format and question type failed to indicate significant effects in the final model. It can be accounted for by the larger effects of response formats. For example, Participant 2 answered four MCQs asking literal comprehension, one OEQs asking literal comprehension and two MCQs asking inferential comprehension. The score variance was larger within literal comprehension questions than between literal comprehension questions and inferential comprehension questions. That is, at least for mid-to-low-levell Korean English learners, there was only a small difference between literal comprehension ability and inferential comprehension ability. Such results can be explained by the limited cognitive resources. The midto-low-level students may have struggled with constructing propositional meaning from the text. There is a possibility that the students may have difficulty identifying words used in the given texts or understanding the grammatical structure of the sentences. For that, the students could have allocated much more

cognitive resources to low-level processes, leaving a limited amount of cognitive resources available for connecting propositional meanings or connecting the meaning with their background knowledge.

Another explanation can be made by focusing on the overwhelming effects of the response format. Concerning the results from the best-fitted model from the current study, there is a possibility that the question type effect was shaded by the response format effect and did not indicate significant results in L2 reading comprehension test.

To sum up, considering the results from the current study, it can be said that how to measure had more impact than what to measure on measuring Korean mid-to-low English grade level proficiency students' reading comprehension ability. Although the small sample size of the mixed-model requires careful consideration on generalizing results from the current study, results from the current study confirmed that response format has a considerable impact on measuring L2 reading comprehension ability.

# CHAPTER 6.

### CONCLUSION

The present study delved into the effects of response format and question type on measuring L2 reading comprehension ability. This chapter summarizes the major findings of the current study and its pedagogical implications. In Section 6.1, a summary of the major findings will be briefly presented and its implication will be discussed. The chapter concludes with the limitations of the study and suggestions for future studies in Section 6.2.

### 6.1 Major Findings and Implications

The goal of the study was to investigate the effects of the testing method on measuring L2 reading comprehension ability. Specifically, the current study focused on the effects of response format and question type on L2 readers' reading test performance. The current study was designed to analyze reading comprehension test results in quantitative ways; a correlation analysis and a mixed-effects modeling.

The first research question of the study examined relationships between different response formats and different question types. Some studies have shown that response format affects reading comprehension test performance and the construct of reading tests. However, the specific relationship between response format and what specific reading ability an item measures have rarely been investigated. For this reason, the current study was conducted to identify the specific relationship between response format and an intended L2 reading

comprehension ability to measure. A correlation analysis was conducted and revealed higher correlations between items sharing the same response format, rather than items with the same question type. Such results indicate a possibility that different response formats may tap into slightly different part of L2 reading comprehension ability.

Regarding the second research question that studied the effects of response format, question type and their interaction on L2 reading comprehension test performance, results from mixed-effects modeling demonstrated the effects of response format on Korean mid-to-low level high school English learners' L2 reading comprehension test scores. But, no question type effect or interaction effect was found.

This study provides educators and researchers with some implications for designing a test and assessing one's L2 reading comprehension ability. The primary implication of the current study is that there is "no 'one best method' for testing reading" (Alderson, 2000, p.203). Our data has verified that response format which is a crucial component of reading comprehension test design, can affect reading test performance. That is, it is necessary to consider the effects of testing methods and their characteristics when designing a test and inferencing readers' abilities from the test.

Another implication is that features of the target test takers (e.g., L2 proficiency) should also be carefully taken into regard. As shown by our random-effects analysis, the effects of the response format vary across individuals' characteristics. For that, it is necessary to consider the features of test-takers when choosing which response format to use in test design and interpreting the results from questions with different response formats.

### 6.2 Limitations and Suggestions

There are several suggestions for future research. First, the sample size of the current study was not enough to reach a model fit convergence of more complex models. Further studies can be performed with a larger sample size to test the effects of response format and question type with various aspects.

Second, an investigation of testing methods can be conducted with a wider range of proficiencies will help sophisticate our insights. As suggested in this study, response format effects may vary across individual characteristics. In the current study, investigation on individual characteristics was limited to subjects' proficiency level and the recruited subjects' proficiency levels were skewed to mid-to-low grades in mock KCSAT English. The truncated proficiency sample attributed to the difficulty in generalizing the results. Future studies with a broader range of proficiency will help generalize the effects of response format and question type.

Third, the application of a qualitative method is recommended. The current study only focused on a quantitative analysis that did not shed light on the process of responding to given items. Considering some subjects showed evidently low response rates with open-ended questions, there is a need to investigate what made the subjects not answer the open-ended questions. Further studies on their process of answering questions will contribute to enhancing our understanding of the testing method's effects on measuring a test-taker's ability.

Despite these limitations, the current study adds to the previous studies of response format effects and provides implications for assessing mid-to-low proficiency L2 students' reading comprehension as to the response format affects

a test-takers' performance on reading tests consisting of literal comprehension questions and inferential comprehension questions and the degree of effects varies across different individual characteristics.

# Reference

- Alderson, J. (2000). Techniques for testing reading. In Assessing Reading (Cambridge Language Assessment, pp. 202-270). Cambridge: Cambridge University Press.
- 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.
- American Educational Research Association, The American Psychological
   Association, & National Council on Measurement in Education.
   (2014). Standards for Educational and Psychological Testing.
   Washington, DC: American Educational Research Association.
- Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford, UK: Oxford University Press.
- Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal inferential, and evaluative comprehension truly exist? *Reading and Writing: An Interdisciplinary Journal, 26*(3), 349-379.
- Best, R.M., Floyd, R.G., & McNamara, D.S. (2008). Differential competencies contributing to children's comprehension of narrative and expository tests. *Reading Psychology*, 29(2), 137-164.
- Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology,*

76(4), 697-708.

- Campbell, D. T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-104.
- Campbell, J.R. (1999). Cognitive process elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension. Unpublished doctoral dissertation. Temple University, Philadelphia.
- Carpenter, R.D., & Paris, S.G. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369-394). Mahwah, NJ: Erlbaum.
- Catts, H.W. (2018). The simple view of reading: advancements and false impressions. *Remedial and Special Education*, *39*(5), 317-323.
- Cerdán, R., Vical-Abarca, E., Martínez, T., Gilabert, R., & Gil, L. (2009). Impact of question-answering tasks on search processes and reading comprehension. *Learning and Instruction*, 19(1), 13-27.
- Clemens, N.H., Davis, J.L., Simmons, L.E., Oslund, E.L., & Simmons, D.C. (2015). Interpreting secondary student's performance on a timed, multiple-choice reading comprehension assessment: The prevalence and impact of non-attempted items. *Journal of Psychoeducational Assessment*, 33(2), 154-165.
- 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<sup>T</sup>1. *Language Testing*, *24*(2), 209-250.
- Collins, A.A., Compton, D.L., Lindström, E.R., & Gilbert, J.K. (2020).
   Performance variations across reading comprehension assessments:
   Examining the unique contributions of text, activity, and reader.
   *Reading and Writing*, 33(3), 605-634.
- Collins, A. A., Lindström, E. R., Compton, D. L. (2018). Comparing students with and without reading difficulties on reading comprehension assessment: A meta-analysis. *Journal of Learning Disabilities*, 51(2), 108-123.
- Collins, A.A., Lindström, E.R., & Sandbank, M. (2021). The influence of language knowledge and test components on reading comprehension scores. *Annals of Dyslexia*, 71(2), 238-259.
- Cordon, L.A., & Day, J.D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology*, 88(2), 288-295
- Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L.Plonsky (Ed.), Advancing quantitative methods in second language research. New York: ROUTLEDGE.
- Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how

comprehension is measured. *Scientific Studies of Reading*, 10(3), 277-299.

- Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General,* 130(2), 208-223.
- Dinsmore, D.L., & Alexander, P.A. (2016). A multidimensional investigation of deep-level and surface-level processing. *The Journal of Experimental Education*, 84(2), 213-244.
- Eason, S.H., Goldberg, L.F., Young, K.M., Geist, M.C., & Cutting, L.E. (2012).
   Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal* of Educational Psychology, 104(3), 515-528.
- Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369-394). Mahwah, NJ: Erlbaum.
- García, J.R., & Cain, K. (2013). Decoding and reading comprehension: A metaanalysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, 84(1), 1-38.
- Gough, P.B., & Tunmer, W.E. (1986). Decoding, reading and reading disability. *Remedial and Special Education*, 7(1), 6-10.

- Grabe, W. (2009). Reading in a second language: Moving from theory to practice. New York: Cambridge University Press.
- Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York: Routledge.
- Hoffman, L., Rovine, M.J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, 39(1), 101-117.
- Hohesinn, C., & Kubinger, K.D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychology*, 71(4), 732-746.
- Hoover, W.A., & Tunmer, W.E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education, 39*(5), 304-312.
- Hua, A.N., & Keenan, J.M. (2014). The role of text memory in inferencing and in comprehension deficits. *Scientific Studies of Reading*, 18(), 415-431.
- Hua, A.N., & Keenan, J.M. (2017). Interpreting reading comprehension test results: Quantile regression shows that explanatory factors can vary with performance level. *Scientific Studies of Reading*, 21(3), 225-238.
- Høin-Tengesdal, I. (2010). Is the simple view of reading too simple?Scandinavian *Journal of Educational Research*, 54(5), 451-469.

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.

- Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990).
   Answering reading comprehension items without passages on the SAT. *Psychological Science*, 1(2), 122-127.
- Keenan, J.M., Betjemann, R.S., Olson, R.K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281-300.
- Keenan, J.M., Hua, A.N., Meenan, C.E., Pennington, B.F., Willcutt, E., & Olson, R.K. (2014). Issues in identifying poor comprehenders. *L'Anne Psychologique*, 114(4), 753-777.
- Keenan, J.M., & Meenan, C.E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, 47(2), 125-135.
- Kim, M. (2023). Exploring literal and inferential reading comprehension among L2 adolescent learners: the roles of working memory capacity, syllogistic inference, and L2 linguistic knowledge. *Reading and Writing, 36*(2), 1085-1110.
- Kim, T. H. (2022). The effects of text length on EFL learners' reading comprehension of fill-in-the-blank texts in KCSAT. Unpublished master's thesis. Seoul National University, Seoul.

- Kim, Y. G. (2017). Why the simple view of reading is not simplistic: Unpacking the simple view of reading using a direct and indirect effect model of reading (DIER). *Scientific Studies of Reading*, 21(4), 310-333.
- Kim, Y. -S. G. (2020a). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology, 112*(4), 667-684.
- Kim, Y. -S. G. (2020b). Toward integrative reading science: The direct and indirect effects model of reading. *Journal of Learning Disabilities*, 53(6),469-491.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163-182.
- Kintsch, W. (1998). Comprehension: A paradigm for cognigition. Cambridge, MA: Cambridge University Press.
- Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J.P. Sabatini, E.R., Albro, & T. O'Reilly (Eds.), *Measuring up* (pp.21-38). Washington, DC: Rowman&Littlefield Education, UK.
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(1), 193-220.

- Kulesz, P.A., Francis, D.J., Barnes, M.A., & Fletcher, J.M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, *108*(8), 1078-1097.
- Lane, S., Raymond, M.R., & Haladyna, T.M. (Eds.). (2015). *Handbook of test development* (2nd ed.). New York, NY: Routledge.
- Language and Reading Research Consortium & Chiu, Y. D. (2018). The simple view of reading across development: Prediction of grade 3 reading comprehension from prekindergarten skills. *Remedial and Special Education, 39*(5), 289-303.
- Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R.* New York: Routledge.
- Leslie, L., & Caldwell, J. S. (2011). *Qualitative reading inventory-5<sup>th</sup> Edition*. Boston: Allyn & Bacon.
- Lim, H. (2019). Test format effects: A componential approach to second language reading. *Language Testing in Asia*, 9(1), 1-22.
- Linck, J.A., Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, 65(1), 185-207.
- Lonigan, C.J., Burgess, S.R., Schatschneider, C. (2018). Examining the simple viewof reading with elementary school children: Still simple after all these years. *Remedial and Special Education*, 39(5), 260-273.

- McNamara, D.S., Ozuru, Y., & Floyd, R.G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education, 4*(1), 229-257.
- Mislevy, R. J. (1995). Test theory and language-learning assessment. *Language Testing*, *12*(3), 341-369.
- Ozuru, Y., Best, R., Bell, C., Wintherspoon, A., & McNamara, D.S. (2007).
  Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, 25(4), 399-438.
- 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.
- Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D.S. (2008). Where's the difficulty in standardized reading tests: The passage of the question? *Behavior Research Methods, 40*(4), 1001-1015.
- Park, S. (2017). The effect of the test method on L2 reading and listening performance. *Journal of Pan-Pacific Association of Applied Linguistics*, 21(1), 45-63.
- R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

https://www.R-project.org/

Ricci, V. (2005). Fitting distribution with R (RELEASE 0.4-21).

- Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40(2), 163-184.
- 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.
- Schielzeth, H., Dingemanese, N.J., Nakagawa, S., Westneat, D.F., Allegue, H.,
  Teplitsky, C., Réale, D., Dochtermann, N.A., Garamszegi, L.Z., &
  Araya-Ajoy, Y.G. (2020). Robustness of linear mixed-effects models
  to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9), 1141-1152.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147-170.
- Snow, C.E. (2002). Reading for understanding: Toward a research and development program in reading comprehension. Santa Monica, CA: RAND.
- Snow, C.E. (2018). Simple and not-so-simple views of reading. *Remedial and Special Education*, *39*(5), 313-316.
- Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading

across grade levels. Journal of Research in Reading, 32(4), 383-401.

Walker, R. (2017). The effect of response format and presentation conditions on comprehension assessment for students with and without a reading disability. Unpublished doctoral dissertation. University of North Caroline, Greensboro.

## APPENDICES

APPENDIX 1	61
APPENDIX 2	67
APPENDIX 3	73

# APPENDIX 1. TEST A

#### Text 1

1

2

3

5

6

7

### **IOT:** Connecting everything

There is a car accident near your home while you are sleeping. Your phone receives the news and automatically sets your alarm earlier by 20 minutes. When the alarm rings, the lights turn on to wake you up and the shower selfadjusts the water temperature b your preference. When you leave the house all of the machines stop working automatically.

This situation may sound like something from a sci -fi movie. Guess what? This is one example of what the "Internet of Things," or IoT, can do. The IoT is becoming an increasingly popular topic. But what exactly is the Internet of Things, and how is it changing our lives?

What is the IoT? How does it work?

- The IoT is a network of machines, vehicles, buildings, and other devices that collect data anækchange information using electronics, software, and sensors. To put it simply, we can connect devices over the Internet, letting <u>them</u> talk to each other and to us. The "things" in the IoT include everything from smartphones, coffee makers, washing machines, lamps, cars, and entire buildings to almost any other object you can imagine.
- You might think that the IoT is nothing new. We already interact with machines. There are home heating systems that you can control with your phone while you're away. The IoT takes this interaction to a <u>higher level</u>. It does not require human control, and determines by itself when and how to turn the systems on and off. These home automation systems are able to find your location from your phone's GPS signal. They then determine when to turn themselves off when you leave the house and when to turn themselves on before your arrival. A smart fire alarm will not only alert your family in case of a fire, but also no tify the local fire and police stations. In such examples, through the IoT, machines exchange data and "talk" to each other, making our lives easier, safer and more efficient.

#### What is the IoT good for?

- At this point, you may wonder why you would want so many devices talking to each other. There are many examples of the IoT's potential value. Imagine that you have just finished the last bottle of milk from your refrigerator. Your fridge notices that there's no milk, so<u>it</u> automatically orders your favorite brand of milk from the store you frequently shop at. Or suppose you're meeting your friend outside when your favorite TV show comes on. Your TV would automatically start recording the show for you because it knows your preferences and viewing history. How convenient!
- While the IoT began with the concept of "smart homes," it is now moving toward the concept of "sm cities." One of the most popular IoT applications is the use of sensors built into the roads to find available parking spaces. Searching for parking spots in busy urban areas is frustrating for drivers, intensifies traffic jams, and increases pollution from circling cars. Sensors built into the roads can tell us how many parking spaces are free, and where they are located.
- Besides providing convenience, the IoT also has large -scale implications for public safety. In 2007, the Interstate 35W bridge in Minnesota collapsed, killing and injuring many people. The collapse was due to steel plates that were incapable of handling the bridge's load. Now, when we build bridges, we use smart cement that has sensors that monitor the bridge for cracks or other stresses. Also, these same road sensors could inform your car of driving hazards, such as icy roads, and warn your car to slow down. As such, the IoT makes us aware of problems, helping to prevent accidents and hopefully save countless lives.

#### What are the concerns of using the IoT?

8

Nothing is perfect in life, and the IoT is no exception. While there is an endless list of benefits associated with the IoT, there are also serious concerns about security and privacy. All of the devices and systems that make up the IoT collect a great deal of personal data. With an increasing number of home devices being connected over the Internet, wh at can people do to make sure that their personal information stays secure? Someone might access one device in your home, even just a coffee maker, and be able to gather information about your entire network. For example, someone could use information from an Internet-connected door lock to find out when you enter or leave your home. Smar TVs and child monitors could also be used to spy on you. With these concerns in mind, a serious discussion about the benefits and downsides of the IoT is needed.

- 1. What is the IoT?
  - 1) a network of exchange
  - (2) a network of only machines
  - (3) an imagined network of everything
  - (4) a network of things
- 2. What does the underlined 'higher level' in paragraph 4 mean? (answer within 10 words)
- 3. What does the sensors built into the roads do? (answer within 10 word)
- 4. What does the 'smart cities' imply?
  - ① cities with many schools
  - 2 cities with various concepts
  - ③ cities with IoT systems

- 1. How does the IoT contribute to public safety?
  - 1 by providing convenience
  - (2) by building bridges with light steel plates
  - (3) by handling the bridge's load
  - (4) by monitoring and preventing accidents
- 2. What does the smart cement do? (answer within 10 words)
- 3. What does 'spy on' in paragraph 8 mean? (answer within 10 words)
- 4. What is the expected impact of using the IoT?
  - ① IoT accelerates the technology improvement
  - ② IoT let us share a great amount of information
  - ③ IoT makes our life safer and more convenient.
  - ④ IoT will connect us with our neighbors.

#### Text 2

2

3

4

5

5

6

7

8

9

#### Palm oil, the biggest threat to orangutans

1 It was late evening. Into the Orangutan Rescue Center on the island of Borneo rushed a rescue worker holding a tiny bundle in his arms. Inside, with a face as small as a mouse's, was a 3-month-old orangutan Those who were there that day remember two things about their first meeting with the baby orangutan: her big bright eyes, and her left arm, half of which was missing.

"Baby orangutans, who ride on their mothers' stomachs by holding onto their long hair, have incredibly strong grips. They do not let go," said the director of the Orangutan Rescue Center. "The baby's hand was chopped off, probably to separate her from her mother."

Rescuers believe that the baby and her mom had been driven out of their forest home when it was destroyed to make way for a palm oil plantation. Hungry and weak, the mother was likely walking in the forest in search of food when she encountered some people working there. Luckily, the baby orangutan was rescued before more harm was done. Rescuers named the baby orangutan "Kesi," meaning "child born in difficult times." **That name is fitting for many orangutans living in the wild today.** 

A hundred years ago, about 230,000 orangutans lived on Earth. Today, fewer than 50,000 are left in the wild, living exclusively on the islands of Borneo and Sumatra. These islands happen to be ground zero for one of the world's most controversial and rapidly growing industries: palm oil.

The cultivation of palm oil is the biggest threat to orangutans. Although <u>their</u> survival gravely depends on rainforests, people have continuously destroyed rainforests to make room for the production of palm oil. Deforestation has reduced the habitat of the orangutans, which in turn has caused thousands of the animals to die. At this rate, orangutans could become extinct in less than 25 years.

So what exactly is palm oil? You may have never heard of palm oil, but you have almost certainly come into contact with it, probably multiple times every day. Palm oil is a type of vegetable oil that is derived from the palm fruit. It can be found in thousands of consumer goods, from packaged foods to hair products. Cheaper and more efficient to produce than other vegetable oils, palm oil is currently used in half of all packaged products.

The consumption of palm oil has rapidly increased in the past two decades. Although the United States and Europe are <u>its</u> top consumers, demand is growing throughout the world. Recently, the U.S. Food and Drug Administration (FDA) banned trans fat in foods, which will likely increase the consumption of palm oil. Though by no means healthier, palm oil is set to be the "clear alternative" for food producers experts say.

The production of palm oil is the number one cause of forest loss in Indonesia and Malaysia. If this continues, the extinction of orangutans will become a sad reality. <u>The future of Kesi's species is in our hands.</u>

Did you know that orangutans are one of our closest relatives? The name orangutan is derived from the Indonesian words *orang* meaning "person," and *hutan* meaning "forest." So orangutan means "person of the forest." They are so intelligent that they make umbrellas out of leaves when it rains.

Unfortunately, these smart animals are getting closer to extinction. I recently read an article saying that palm oil plantations are destroying the homes of orangutans. The sad thing is that most people are not familiar with palm oil, even though it is in many products that we use on a daily basis. This might surprise you, but palm oil is used in ice cream, cookies, bread, chocolate, soap, and cleaning agents!

So what can we do to help save the orangutans? Well, you can talk to your friends and family about

alm oil and the dangers orangutans are facing. You could also be an active consumer next time you sit the supermarket. Read the labels of products carefully to see if they contain palm oil, and try to avoid them if they do.

11

Please help spread the word about the dangers of the palm oil industry and name any products containing palm oil that you know in the comments below. I really hope you will join me in protecting these "**people of the forest**."

- 1. What is the feature do people remember about the baby orangutan? (answer within 10 words)
- 2. Why did people destroy the forest?
  - ① to send orangutans to a rescue center
  - ② to make an oil factory
  - ③ to make a way for palm oil plantation
  - ④ to make a highway
- 3. What does the sentence "<u>That name is fitting for many orangutans living in the wild today</u>." in paragraph 3 imply? (answer within 10 words)
- 4. Why does the cultivation of palm oil threaten orangutans? (answer within 10 words)
- 5. What is palm oil made from?
  - 1 from the palm fruit
  - 2 from the palm leaf
  - ③ from oil fruit
  - (4) from any vegetable

- 1. What does the sentence "The future of Kesi's species is in our hands." in paragraph 7 imply?
  - ① The production of palm oil depends on Kesi.
  - ② The forest loss is due to Kesi's species.
  - ③ We are responsible for reducing palm oil production.
  - ④ We are responsible for adopting orangutans.
- 2. What is the relationship between palm oil and orangutans? (answer within 10 words)
- 3. What does "people of the forest" in paragraph 11 refer to?
  - ① people living in the forest
  - (2) orangutans living in the forest
  - (3) active consumers in neighborhood
  - (4) people who are named forest

## APPENDIX 2. TEST B

#### Text 1

1

2

5

6

#### **IOT:** Connecting everything

There is a car accident near your home while you are sleeping. Your phone receives the news and automatically sets your alarm earlier by 20 minutes. When the alarm rings, the lights turn on to wake you up and the shower selfadjusts the water temperature b your preference. When you leave the house all of the machines stop working automatically.

This situation may sound like something from a sci -fi movie. Guess what? This is one example of what the "Internet of Things," or IoT, can do. The IoT is becoming an increasingly popular topic. But what exactly is the Internet of Things, and how is it changing our lives?

What is the IoT? How does it work?

The IoT is a network of machines, vehicles, buildings, and other devices that collect data and exchan information using electronics, software, and sensors. To put it simply, we can connect devicesover the Internet, letting <u>them</u> talk to each other and to us. The "things" in the IoT include everything from smartphones, coffee makers, washing machines, lamps, cars, and entire buildings to almost any other object you can imagine.

You might think that the IoT is nothing new. We already interact with machines. There are home heating systems that you can control with your phone while you're away. The IoT takes this interaction to a <u>higher level</u>. It does not require human control, and determines by itself when and how to turn the systems on and off. These home automation systems are able to find your location from your phone's GPS signal. They then determine when to turn themselves off when you leave the house and when to turn themselves on before your arrival. A smart fire alarm will not only alert your family in case of a fire, but also notify the local fire and police stations. In such examples, through the IoT, machines exchange data and "talk" to each other, making our lives easier, safer and more efficient.

#### What is the IoT good for?

At this point, you may wonder why you would want so many devices talking to each other. There are many examples of the IoT's potential value. Imagine that you have just finished the last bottle of milk from your refrigerator. Your fridge notices that there's no milk, so<u>it</u> automatically orders your favorite brand of milk from the store you frequently shop at. Or suppose you're meeting your friend outside when your favorite TV show comes on. Your TVwould automatically start recording the show for you because it knows your preferences and viewing history. How convenient!

While the IoT began with the concept of "smart homes," it is now moving toward the concept of "sm cities." One of the most popular IoT applications is the use of sensors built into the roads to find available parking spaces. Searching for parking spots in busy urban areas is frustrating for drivers, intensifies traffic jams, and increases pollution from circling cars. Sensors built into the roads can tell us how many parking spaces are free, and where they are located.

Besides providing convenience, the IoT also has large -scale implications for public safety. In 2007, the Interstate 35W bridge in Minnesota collapsed, killing and injuring many people. The collapse was due to steel plates that were incapable of handling the bridge's load. Now, when we build bridges, we use smart cement that has sensors that monitor the bridge for cracks or other stresses. Also, these same road sensors could inform your car of driving hazards, such as icy roads, and warn your car to slow down. As such, the IoT makes us aware of problems, helping to prevent accidents and hopefully save countless lives.

#### What are the concerns of using the IoT?

8

Nothing is perfect in life, and the IoT is no exception. While there is an endless list of benefits associated with the IoT, there are also serious concerns about security and privacy. All of the devices and systems that make up the IoT collect a g reat deal of personal data. With an increasing number of home devices being connected over the Internet, what can people do to make sure that their personal information stays secure? Someone might access one device in your home, even just a coffee maker, and be able to gather information about your entire network. For example, someone could use information from an Internet-connected door lock to find out when you enter or leave your home. Smar TVs and child monitors could also be used to <u>spv on</u> you. With these concerns in mind, a serious discussion about the benefits and downsides of the IoT is needed.

- 1. What is the IoT? (answer within 10 words)
- 2. What does the underlined 'higher level' in paragraph 4 mean?
  - 1) The interaction between people and machines
  - (2) The interaction between a smartphone and GPS signal
  - ③ The interaction between machines
  - ④ The interaction between automation systems and human control
  - 3. What does the sensors built into the roads do?
    - 1 help us find available parking spaces.
    - (2) help us searching for destination
    - (3) help us know the number of cars in urban areas
    - (4) help drivers to intensify traffic jams
  - 4. What does the 'smart cities' imply? (answer within 10 words)
  - 5. How does the IoT contribute to public safety? (answer within 10 words)

- 6. What does the smart cement do?
  - ① It handles the bridge's load.
  - (2) It monitors cracks and other stresses.
  - ③ If fills a crack on the bridge.
  - ④ It counts the number of cars on the bridge.
- 7. What does 'spy on' in paragraph 8 mean?
  - 1 to protect security
  - to protect privacy
  - ③ to monitor safety
  - 4 to monitor home devices
- 8. What is the expected impact of using the IoT? (answer within 10 words)

### Text 2

2

3

4

6

7

8

9

#### Palm oil, the biggest threat to orangutans

1 It was late evening. Into the Orangutan Rescue Center on the island of Borneo rushed a rescue worker holding a tiny bundle in his arms. Inside, with a face as small as a mouse's, was a 3-month-old orangutan Those who were there that day remember two things about their first meeting with the baby orangutan: her big bright eyes, and her left arm, half of which was missing.

"Baby orangutans, who ride on their mothers' stomachs by holding onto their long hair, have incredibly strong grips. They do not let go," said the director of the Orangutan Rescue Center. "The baby's hand was chopped off, probably to separate her from her mother."

Rescuers believe that the baby and her mom had been driven out of their forest home when it was destroyed to make way for a palm oil plantation. Hungry and weak, the mother was likely walking in the forest in search of food when she encountered some people working there. Luckily, the baby orangutan was rescued before more harm was done. Rescuers named the baby orangutan "Kesi," meaning "child born in difficult times." **That name is fitting for many orangutans living in the wild today.** 

A hundred years ago, about 230,000 orangutans lived on Earth. Today, fewer than 50,000 are left in the wild, living exclusively on the islands of Borneo and Sumatra. These islands happen to be ground zero for one of the world's most controversial and rapidly growing industries: palm oil.

The cultivation of palm oil is the biggest threat to orangutans. Although <u>their</u> survival gravely depends on rainforests, people have continuously destroyed rainforests to make room for the production of palm oil. Deforestation has reduced the habitat of the orangutans, which in turn has caused thousands of the animals to die. At this rate, orangutans could become extinct in less than 25 years.

So what exactly is palm oil? You may have never heard of palm oil, but you have almost certainly come into contact with it, probably multiple times every day. Palm oil is a type of vegetable oil that is derived from the palm fruit. It can be found in thousands of consumer goods, from packaged foods to hair products. Cheaper and more efficient to produce than other vegetable oils, palm oil is currently used in half of all packaged products.

The consumption of palm oil has rapidly increased in the past two decades. Although the United States and Europe are <u>its</u> top consumers, demand is growing throughout the world. Recently, the U.S. Food and Drug Administration (FDA) banned trans fat in foods, which will likely increase the consumption of palm oil. Though by no means healthier, palm oil is set to be the "clear alternative" for food producers experts say.

The production of palm oil is the number one cause of forest loss in Indonesia and Malaysia. If this continues, the extinction of orangutans will become a sad reality. <u>The future of Kesi's species is in our hands.</u>

Did you know that orangutans are one of our closest relatives? The name orangutan is derived from the Indonesian words *orang* meaning "person," and *hutan* meaning "forest." So orangutan means "person of the forest." They are so intelligent that they make umbrellas out of leaves when it rains.

Unfortunately, these smart animals are getting closer to extinction. I recently read an article saying that palm oil plantations are destroying the homes of orangutans. The sad thing is that most people are not familiar with palm oil, even though it is in many products that we use on a daily basis. This might surprise you, but palm oil is used in ice cream, cookies, bread, chocolate, soap, and cleaning agents!

10

So what can we do to help save the orangutans? Well, you can talk to your friends and family about palm oil and the dangers orangutans are facing. You could also be an active consumer next time you visit the supermarket. Read the labels of products carefully to see if they contain palm oil, and try to avoid them if they do.

11

Please help spread the word about the dangers of the palm oil industry and name any products containing palm oil that you know in the comments below. I really hope you will join me in protecting these "**people of the forest**."

- 1. What is the feature do people remember about the baby orangutan?
  - 1) her name of the baby orangutan
  - ② a bundle that she was holding
  - ③ her dark eyes
  - (4) injured left arm
- 2. Why did people destroy the forest? (answer within 10 words)
- 3. What does the sentence "<u>That name is fitting for many orangutans living in the wild today</u>." in paragraph 3 imply?
  - ① There are many orangutans named 'Kesi'.
  - ② There are many orangutans which are in danger.
  - ③ There are many children in the forest.
  - ④ There are many orangutans rescued by people.
- 4. Why does the cultivation of palm oil threaten orangutans?
  - ① It makes thousands of animals to live in the rainforest
  - 2 It causes deforestation
  - ③ It is toxic to orangutans
  - ④ It grows the number of orangutans
- 5. What is palm oil made from? (answer within 10 words)

- 6. What does the sentence <u>"The future of Kesi's species is in our hands.</u>" in paragraph 7 imply? (answer within 10 words)
- 7. What is the relationship between palm oil and orangutans?
  - ① Orangutans are exploited in palm oil plantations.
  - 2 Palm oil threatens orangutans' health.
  - ③ Palm oil plantation threatens our lives.
  - ④ Palm oil plantation threatens orangutans' lives.
- 8. What does "people of the forest" in paragraph 11 refer to? (answer within 10 words)

# APPENDIX 3. Answer Key for Open-Ended Questions

Test -A			
Text	Item	Sample answer and Keywords	
Number	Number	Sample answer and Keywords	
Text 1	2	Sample answer	
		: An interaction between machines	
		Key words: interaction, machines	
	3	Sample answer	
		: They help us find available parking spaces.	
		Key words: help, parking spaces	
	6	Sample answer	
		: It monitors cracks and other stresses.	
		Key words: monitor, cracks, stresses,	
	7	Sample answer	
		: to monitor home devices	
		Keywords	
		: monitor, invade privacy, home devices	
Text 2	1	Sample answer	
		: her injured left arm	
		Key words: injured, left arm,	
	3	Sample answer	
		: There are many orangutans which are in danger.	
		Key words: in danger, endangered, in crisis	
	4	Sample answer	
		: It causes deforestration.	
		Key words	
		: devastating, deforestration, destroy forest (or habitat)	
	7	Sample answer	
		: Palm oil plantation threatens orangutans' lives.	
		Key words: threaten, threat, in danger	

Test -B		
Text	Item	Sample answer and Keywords
Number	Number	Sample answer and Keywords
Text 1		Sample answer
	1	: A network of things
		Keywords
		: network, connected, things, devices
	4	Sample answer
		: Cities with IoT systems
		Keywords
		: IoT, smart homes, extended
	5	Sample answer
		: by monitoring and preventing accidents
		Key words
		: monitor, prevent, accident
		Sample answer
	8	: IoT makes our life safer and more convenient.
		Key words
		: safe, convenient, faster, safer
Text 2	2	Sample answer
		: to make a way for palm oil plantation
		Key words
		: make, build, way, road, palm oil plantation
	5	Sample answer
		: palm fruit
		Keywords
		: palm fruit, vegetable
	6	Sample answer
		: We are responsible for reducing palm oil production.
		Keywords
		: responsible, have to reduce, help reduce, we should
		reduce (palm oil consumption), protect orangutans
	8	Sample answer
		: Palm oil plantation threatens orangutans' lives.
		Keywords: threaten, threat, in danger

# 국문초록

글을 읽고 이해한다는 것은 텍스트에서 추출된 의미와 독자의 경험, 배경지식 등의 상호작용을 통해 일종의 정신적 표상을 구성하는 것이다. 이러한 정의에 따르면 독해 능력은 텍스트에서 의미를 추출할 수 있는 능력, 추출한 의미를 활용하여 개인의 독자적 정신적 표상을 구성하는 능력 등을 포함한다. 즉, 독해 능력이란 단일한 구인이 아닌, 다양한 요소들로 구성된 복합적인 구인인 것이다.

구인의 복합성은 독해 능력 평가에 있어 두 가지 주요 시사점을 제시한다. 첫째, 독해 능력을 평가하기 위한 다양한 시험들이 동일한 독해 능력을 측정한다고 볼 수 없다. 둘째, 독해 시험들이 측정하는 능력이 다르다는 점은 독해 능력 평가결과 해석 및 해석의 타당도에 영향을 미친다. 이에 따라, 선행 연구들 독해 시험의 구인에 영향을 미치는 요소들을 독자, 텍스트, 시험의 특성을 중심으로 살펴보았다.

본 연구에서는 시험 특성이 한국의 고등학생 영어 학습자들의 독해 능력을 측정하는 데 미치는 영향을 알아보고자 문항의 응답 형태와, 문항 질문의 종류가 학습자들의 독해 능력을 측정하는 데 미치는 영향에 대해 분석하였다. 본 연구에 활용된 분석 방법은 상관관계 분석과 혼합 효과 모형이다. 먼저, 상관관계 분석을 통해 동일한 문항 형태를 갖는 문항들의 동질성과 동일한 질문 형태를 갖는 문항들의 동질성을 분석하였다. 다음으로 혼합 효과 모형(Mixed effects model)을 적용하여, 문항의 응답 형태와 문항 질문 유형이 독자들의 문항 응답 수행에 미치는 영향을 알아보았다.

분석 결과, 동일한 응답 형태를 갖는 문항들은 동일한 질문 유형의 문항들 보다 높은 상관관계를 갖는 것으로 나타났다. 혼합효과 모형 분석에서 문항 응답 형태가 독해 시험 수행에 큰 영향을 미치는 것으로 나타났다. 이러한 결과는 시험의 특성이 독해 시험 결과를 근거로 한 피험자의 독해 능력 추론에 영향을 미친다는 것을 보여주며, 시험 구성 단계 및, 시험 결과 활용 단계에서 문항의 응답 형태와 같은 시험 특성에 대한 충분한 고려가 필요함을 시사한다.

주요어: 문항 응답 형태, 질문 유형, 독해 평가, 독해 시험, 타당도, 한국 영어학습자

학번: 2019-26486