

VALIDATING MULTIPLE-TEXT READING TASKS
IN FOREIGN LANGUAGE PROFICIENCY TESTS
THROUGH VERBAL PROTOCOLS AND EYE TRACKING

HATİCE YURTMAN KAÇAR

BOĞAZIÇI UNIVERSITY

2018

VALIDATING MULTIPLE-TEXT READING TASKS
IN FOREIGN LANGUAGE PROFICIENCY TESTS
THROUGH VERBAL PROTOCOLS AND EYE TRACKING

Thesis submitted to the
Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements for the degree of

Master of Arts in
English Language Education

by

Hatice Yurtman Kaçar

Boğaziçi University

2018

DECLARATION OF ORIGINALITY

I, Hatice Yurtman Kaçar, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date25.04.2018.....

ABSTRACT

Validating Multiple-Text Reading Tasks in Foreign Language Proficiency Tests Through Verbal Protocols and Eye Tracking

The purpose of this study is to investigate how multiple texts reading skill in tasks in existing English proficiency tests are operationalized, whether the subskills and strategies specified in these exams match the theoretical explanations, and whether the actual use of skills and strategies reflects a sufficient and accurate coverage of theoretically designated multiple texts reading skill. ISE II, MET, and ECCE have been found to aim at assessing multiple-text reading comprehension. The tasks purportedly measuring multiple-text reading comprehension in these proficiency exams were administered to 10 participants of varying nationalities. Data were collected through eye tracking and retrospective think aloud method. The results revealed that ISE II does not attempt to operationalize multiple texts reading skill representatively, while MET specifications are not specific enough on multiple-text reading skill. ECCE specifications show that this task attempts to operationalize these skills representatively. When it comes to the operationalization of the specified multiple texts reading skill in each task, ISE II and MET do not sufficiently operationalize multiple-text reading skill, while ECCE is found to operationalize these skills to some extent. The findings also have implications for the design of multiple texts reading comprehension test tasks.

ÖZET

Yabancı Dil Yeterlilik Sınavlarındaki Çoklu-Metin Okuma Becerileri Ödevlerinin Sesli Düşünme Tekniği ve Göz Hareketi Takibi ile Doğrulanması

Bu çalışmanın amacı mevcut dil yeterlilik sınavlarında çoklu metin okuma becerilerini ve stratejilerini test eden soruların, bu becerileri nasıl işlevselleştirmeyi hedeflediğini, bu becerilerin literatürde teorik olarak tanımlanan beceri ve stratejilerle ne derece örtüştüğünü, ve bu becerilerin fiili kullanımının teorik olarak tanımlanan çoklu metin okuma becerilerini yeterli ve doğru bir biçimde kapsayıp kapsamadığını incelemektir. ISE II, MET ve ECCE gibi dil yeterliliğini ölçen sınavlarda bulunan çoklu metin okuma becerilerini ölçmeyi hedefleyen sorular, farklı anadillere sahip 10 öğrenciye uygulanmıştır. Veri, göz hareketi takip teknolojisi, ve ardıl sesli düşünme tekniği aracılığıyla toplanmıştır. Sonuçlar ISE II çoklu metin okuma becerilerini yeterli ve doğru bir biçimde ölçmeyi hedeflemediğini, MET'in beceri tanımları işlevselleştirilebilecek kadar spesifik olmadığını, ve ECCE beceri tanımlarının çoklu metin ölçme becerilerini yeterli ve kapsamlı bir biçimde ölçmeyi hedeflediğini gösteriyor. Ayrıca, çoklu metin okuma becerilerini bu mevcut sınavların nasıl işlevselleştirdiğine bakıldığında, ISE II ve MET'in, bu becerileri yeterli ve doğru bir biçimde işlevselleştiremediği, ECCE'nin belirli bir ölçüde işlevselleştirdiği gözlemlenmiştir. Bu çalışma ayrıca çoklu metin okuma becerileri sınav ödevleri tasarımlarında tavsiyeler de sunmaktadır.

ACKNOWLEDGEMENTS

I would like to first thank my thesis advisor, Assist. Prof. Aylin Ünalđı, for her guidance and invaluable suggestions on the research design, implementation, and her comments and suggestions on the manuscript. Without her constructive criticism and enlightening guidance, this thesis would not possibly be finalized. I also want to express my thanks to the members of my thesis committee, Assist. Prof Nur Yiđitođlu and Assoc. Prof. Zeynep Koçođlu for their precious comments and suggestions. I also would like to thank Assist. Prof. İnci Ayhan for providing us with the chance to carry out the study in the Vision Laboratory at the Psychology Department, and Emre Oral for his technical support on the use of the eye tracker. If it weren't for him, this research would never be complete. I would like to send my deepest gratitude to my parents, Necla Yurtman and Mehmet Yurtman, my sisters, Hülya Yurtman and Betül Yurtman, for always believing in me and supporting me without a question. Finally, I am grateful to my awesome husband, Hakan Kaçar, as I am aware that all the time devoted to this thesis was actually taken from his time.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
1.1 Introduction	1
1.2 Aims of the study.....	2
1.3 Overview of methodology.....	3
1.4 Significance of the study	3
1.5 Research questions	5
1.6 Overview of the thesis.....	5
CHAPTER 2: LITERATURE REVIEW	6
2.1 Introduction.....	6
2.2 Validity.....	6
2.3 Theories of reading	10
2.4 Foreign language assessment frameworks	21
2.5 Construct validation research through eye tracking.....	24
2.6 Conclusion to the literature review chapter	26
CHAPTER 3: METHODOLOGY	28
3.1 Introduction.....	28
3.2 Research questions	28
3.3 Participants.....	29
3.4 Multiple-text reading tests	30
3.5 Instruments.....	32
3.6 Procedure.....	33
3.7 Data analysis	35
3.8 Conclusion to the methodology chapter.....	36
CHAPTER 4: RESULTS	38
4.1 Introduction.....	38
4.2 RQ1: Do multiple texts reading (MTR) tasks used in language proficiency tests attempt to operationalize MTR skill and subskills as defined in theory representatively?	38
4.3 RQ2: Do test takers use substantial MTR skill as defined by theory and as specified in the test specifications in responding to the MTR tasks in tests where such tasks are available?.....	44
4.4 Conclusion to the results section.....	79
CHAPTER 5: DISCUSSION.....	80
5.1 Introduction.....	80

5.2 RQ1: Do multiple texts reading (MTR) tasks used in language proficiency tests attempt to operationalize MTR skill and subskills as defined in theory representatively?	81
5.3 RQ2: Do test takers use substantial MTR skills as defined by theory and as specified in the test specifications in responding to the MTR tasks in tests where such tasks are available?.....	85
5.4 Conclusion to the discussion chapter	97
CHAPTER 6: CONCLUSION.....	98
6.1 Introduction.....	98
6.2 Overview of the findings.....	98
6.3 Implications on test design.....	101
6.4 The limitations of the study.....	102
6.5 Suggestions for further research.....	103
6.6 Conclusion	103
APPENDIX A: ISE II	104
APPENDIX B: MET.....	106
APPENDIX C: ECCE.....	107
APPENDIX D: TRAINING TASK	107
APPENDIX E: READING STRATEGY CODING RUBRIC	109
APPENDIX F: ISE II (1) AREAS OF INTEREST	111
APPENDIX G: ISE II (2) AREAS OF INTEREST.....	114
APPENDIX H: MET AREAS OF INTEREST	116
APPENDIX I: ECCE AREAS OF INTEREST	119
APPENDIX J: ISE II (1) EYE-MOVEMENT SEQUENCE OF PARTICIPANT 7 AND PARTICIPANT 10	121
APPENDIX K: ISE II (2) EYE- MOVEMENT SEQUENCE OF PARTICIPANT 7 AND PARTICIPANT 10	122
APPENDIX L: MET-EYE-MOVEMENT SEQUENCE OF PARTICIPANT 7 AND PARTICIPANT 10.....	123
APPENDIX M: ECCE- EYE-MOVEMENT SEQUENCE OF PARTICIPANT 7 AND PARTICIPANT 10	124
REFERENCES.....	125

LIST OF TABLES

Table 1. Facets of Validity by Messick.....	8
Table 2. The Participants' Scores on ISE II Multi-Text Reading Task	44
Table 3. ISE II Task 1 Frequency of Accuracy for Each Item	45
Table 4. The Participants' Performance on MET.....	55
Table 5. The Participants' Performance in ECCE.....	60
Table 6. Average Fixation Count, Fixation Duration, and Careful Reading in ISE, MET, and ECCE.....	66
Table 7. Fixation Count, Fixation Duration, and Careful Reading in ISE II (1).....	67
Table 8. Fixation Count, Fixation Duration, and Careful Reading in ISE II (2).....	68
Table 9. Fixation Count, Fixation Duration, and Careful Reading in MET	69
Table 10. Fixation Count, Fixation Duration, and Careful Reading in ECCE.....	69
Table 11. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in ISE II (I)	71
Table 12. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in ISE II (I)	71
Table 13. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in ISE II (2)	73
Table 14. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in ISE II (2)	73
Table 15. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in MET	76
Table 16. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in MET	76

Table 17. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in ECCE	78
Table 18. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in ECCE	78

LIST OF FIGURES

Figure 1. Cognitive processing in reading by Khalifa and Weir.....	22
Figure 2. ISE II Task 1 overall strategy use.....	45
Figure 3. ISE II Task 1 reading operations	46
Figure 4. ISE II Task 2 overall strategy use.....	47
Figure 5. ISE II Task 2 reading operations	48
Figure 6. ISE II Task 3 overall strategy use.....	48
Figure 7. ISE II Task 3 reading operations	49
Figure 8. MET Item 1 overall strategy use	56
Figure 9. MET Item 1 reading operations.....	57
Figure 10. MET Item 2 overall strategy use	57
Figure 11. MET Item 2 reading operations.....	58
Figure 12. ECCE Item 1 overall strategy use.....	61
Figure 13. ECCE Item 1 reading operations	61
Figure 14. ECCE Item 2 overall strategy use.....	62
Figure 15. ECCE Item 2 reading operations	63

CHAPTER 1

INTRODUCTION

1.1 Introduction

Data collection, analysis and interpretation of these data are integral processes of all testing and assessment endeavors to give meaning to the results. It is highly necessary that decisions and interpretations based on the scores obtained from tests be indicative of the actual performance in real life. For a test to predict actual performance, it is of grave importance for the test to be proven valid and reliable. Reliability is concerned with the consistency of the scores produced by a certain test. Validity is about whether a test measures what it attempts to measure; namely, the meaningful interpretations of the relationship between the score produced by the test and the observed performance in real life. The content and the importance of skills and subskills to be included in an exam depend on the context the exam is used in. Therefore, in testing and assessment, the first step in devising tests is to start with skills definitions to be included in test specifications. These definitions are rooted in theory. Consequently, testing and theory are closely connected because devising tests that operationalize the necessary skills requires a grasp of theory that defines the constructs underlying tests. Khalifa and Weir (2009) suggest different cognitive levels to reading comprehension. Their model starts with decoding of words and syntactic analysis, and inferencing, and textual and intertextual comprehension are conceptualized as the highest levels. Therefore, a reading test intending to follow this model needs to operationalize reading skills at different cognitive levels depending on the context. Higher-level skills such as textual and intertextual comprehension are necessary at tertiary level of education (Ünaldı, 2010; Goldman, 2011), and as

suggested by Goldman (2011), locating, evaluating, and integrating information are vital skills of reading and understanding, which are described as subskills of intertextual comprehension. These skills are also very significant at tertiary level of education since employment of skills as such facilitate deeper learning from texts (Cerdán & Vidal-Abarca, 2008; Bråten, Ferguson, Anmarkrud, & Strømsø, 2013; Hagen, Braasch, and Bråten, 2014), and this is the type of learning expected in academic contexts. Hence, it is paramount that tests in academic contexts sample from skills operationalized at higher cognitive levels as well. At present, only three international language proficiency exams aim to assess reading comprehension at intertextual level. However, there is no study on the investigation of the validity of these exams. Therefore, these three international language proficiency exams, which often act as gatekeepers to universities, need to be examined in terms of construct or cognitive validity through the comparison of the construct definitions specified in the test specifications with the construct definitions in theory and the collection of process based data.

1.2 Aims of the study

This study aims to investigate whether international language proficiency exams aiming to assess multiple texts reading skill (intertextual comprehension) are valid by comparing the construct operations in their test specifications to the construct operations defined in theory, and reveal what sort of cognitive process are employed by test takers while completing the multiple texts reading comprehension tasks in these exams.

1.3 Overview of methodology

As the study aims to investigate how multiple texts reading skill is operationalized in the language proficiency tests available in the field, all English as a foreign language proficiency exams were scanned, and the ones aiming to assess multiple texts reading skill as a reading construct per se, rather than integrating multiple texts reading skill with writing, have been identified. The tests including multiple texts tasks were analyzed using verbal protocol and eye tracking methods in terms of the use of related skills by the test takers. To this end, the formats of these exam tasks were modified to be compatible with the eye tracker screen, and high quality images of these tasks were produced. An individual session with each participants was arranged. The participants completed tasks one by one, and upon completion of a task, they were required to think aloud on what strategies they used, and what order they followed etc. A reading strategy coding rubric was developed based on the construct definitions of reading skills in theory together with an Applied Linguistics expert. The verbal accounts of the participants were coded by two raters at the same time upon discussion of the use of each strategy. Basic descriptive statistical analyzes were carried on the reported strategies. In addition, eye tracking data were analyzed using descriptive statistics. Fixation duration, fixation count and sequence of the eye movements were analyzed. These provided information on the strategy use of the participants and the extent of multiple texts reading skill use.

1.4 Significance of the study

Firstly, there is no study on cognitive validation of exams aiming at operationalizing multiple texts reading skill. It is important to note that this skill has been tested recently in a few exams. However, without undergoing cognitive validation

processes, these exams may not be guaranteed to operationalize the construct they aim to operationalize. These processes are quite significant because testing and assessment involve several stakeholders from test takers to their parents, teachers, and test writers and users. Unless a test is valid, people's lives might be affected negatively if their abilities are judged using these tests. For example, if universities admit students based on the scores from tests inadequately representing academic reading skills, both the student, not having the necessary skills and the instruction at university may suffer. Namely, interpretations and decisions based on the scores of these exams may not reflect actual performance if these exams are not valid. It is also known that tests have a washback effect on curriculum and materials design. Therefore, when a test successfully and representatively operationalizes multiple texts reading skill, which is the highest level of comprehension, then this may be reflected on language teaching contexts, and more emphasis might be placed on teaching this skill. This study will reflect a comprehensive investigation on the use of multiple texts reading skill in certain exams therefore exemplify an important validation process for the quality of test design.

In addition, data were collected through a very innovative method, which provides moment by moment information on the strategies test takers apply. This study is a first that employed eye tracking in collecting validity evidence for multiple texts reading skill in present reading tests.

Finally, through the detailed analysis of think aloud data, it provides valuable insight on the thought process of test takers, which may have implications for test writers.

1.5 Research questions

In order to attain the aims mentioned above, the study proposed two research questions. The first research question aims to investigate the construct validity of international language proficiency exams from a theoretical point of view. The second question aims to investigate the construct (or cognitive) validity of these exams through deeper look into the processes used to complete the test tasks.

RQ1: Do multiple texts reading (MTR) tasks used in language proficiency tests attempt to operationalize MTR skill and subskills as defined in theory representatively?

RQ2: Do test takers use substantial MTR skill as defined by theory and as specified in the test specifications in responding to the MTR tasks in tests where such tasks are available?

1.6 Overview of the thesis

Chapter 1 is an introductory chapter. Chapter 2 is the review of literature on test validity, reading theories and frameworks, and the results of relevant research studies. Chapter 3 outlines the methods used in this study in detail. Chapter 4 presents the results under two research questions. It presents the analysis of the test specifications of each test, think aloud and eye movement data, as well as the detailed analysis of two specific cases. Chapter 5 presents the discussion of the findings. Chapter 6 presents an overview of the study together with conclusions, implications and the limitations.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents an overview of the relevant literature on validation of language tests, specifically reading skills. It starts with the theory of testing validation because the core of language assessment lays in mainly collection of evidence, analysis and interpretation of data in language use. Based on these analyzes and interpretations, inferences, and consequent decisions are made. There are two important considerations to arrive at accurate interpretations and decisions. First, a test must be reliable; that is a test must produce consistent results across different observations of the target behavior or skill so that trust is instilled on the score awarded (Bachman, 1996). Secondly, the use of a test and interpretation of its scores must be a valid judgement. That is, a test must measure the knowledge and abilities that are desired to be measured. Validation could be satisfied through the collection of different types of validity evidence. This second concern needs to be examined in detail in any inquiry on language tests.

This chapter also presents reading theories as validation and theory are closely related. Finally, after outlining reading theory, language assessment frameworks and validation studies carried out using eye trackers will be presented.

2.2 Validity

In the literature, various types of validity have been put forward. Cronbach and Meehl (1955) divided validity into three main types, which were described as

criterion-oriented validity, content validity, and construct validity. Criterion-oriented validity is concerned with the relationship between a particular test and a criterion against which predictions are made. Namely, validity evidence is the strength of correlation between the test score and the performance on the criterion. Content validity is concerned with showing that the content of the test is representative and comprehensive enough to sample target behavior effectively. Finally, construct validity is involved when what is to be operationalized matches with the operational definition of the certain ability to be assessed as the ones in the literature. Namely, it is the extent of similarity between the interpretations of the scores and the actual operations in real life.

2.2.1 Bachman and Palmer's test usefulness

Bachman and Palmer (1996) assert that construct validity is about building a validity argument, and meaningfulness and appropriateness of the decisions based on test scores (p. 21). Any decision or interpretation must be justifiable. Justifying test scores and interpretations requires validity evidence showing that the score awarded capture the areas of language ability to be measured. The first step to provide this evidence is to define the construct. Bachman and Palmer (1996) considers the construct as the definition of an ability that is used as a basis for a test or task and for the interpretations of scores (p.21).

Construct validity also deals with the domain of generalizations (Bachman, 1996, p.21), which is the array of tasks in the Target Language Use (TLU) domain. Therefore, the choice of tasks to be included in a test are of crucial importance. Due to that, the abilities tested by the tasks must match those in the TLU. Besides, a variety of tasks testing different abilities must be included to make the scores and

interpretations generalizable to the TLU. These show us that abilities or operations of a skill must be derived from research and theory, and these abilities must be tested representatively to make accurate interpretations of the scores.

2.2.2 Messick's facets of validity

Messick (1989, p.6) criticizes the views that conceptualize validity comprising of different types because he asserts that all these types of validity try to justify the valid interpretation and the use of scores, and that they must be seen as supplementing one another by providing evidence, on which interpretations are based on. He sees validity as a unified concept under construct validity, and places construct validity at the heart of the validation process including different facets. An overview of the facets of validity is presented in Table 1.

Table 1. Facets of validity (Messick 1989, p.20)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV + Relevance/Utility(R/U)
Consequential Basis	CV+R/U	CV+R/U +VI
	Value Implications (VI)	Social Consequences (SC)

In this framework, different types of validity evidence contribute to the formulization of construct validity. Construct validity is not concerned with the validity of a test, rather with the degree to which an interpretation or use of a test is justified based on a test score. Thus, it can be said that validity is not a quality of a test, but how and with what purpose a test is used for. In Messick's (1989) model, evidential basis for test interpretation is about the construct, namely how the construct is defined as operations in the theory. Evidential basis for test use is about the purpose the test is used for. Here, the context the test will be utilized is highly

important because certain aspects of a construct might be emphasized while others are left aside considering the purpose the test is used for. Consequential basis for test interpretation is the value implications, which are concerned with what label is given to a construct. Namely, how test writers and users believe and define the construct to be. These beliefs and definitions by test writers and users is undoubtedly shaped by the context, and affect the interpretation of scores in return. Finally, the last label, social consequences, become known when the test is in use. Social consequences deal with interpretations and inferences made using the test scores, and how these interpretations affect different stakeholders in question. This model is quite significant in that social consequences were mentioned explicitly for the first time. Furthermore, one suggestion of this model is that validity cannot be proved, but evidence from different sources is gathered to make luminary interpretations and decisions of test scores. Weir (2005) describes these sources in more detail in his socio-cognitive framework.

2.2.3 Weir's socio-cognitive framework

According to the framework by Weir and Shaw (2005), validation process consists of two stages: *a priori* validation, comprising of theory based validity and context validity and *a posteriori* validation, which takes place after an exam is administered and includes scoring, consequential, and criterion-oriented validity (in Zainal, 2012). Theory based validity covers both *a priori* evidence collected before the test and *a posteriori* evidence collected after the test is administered (Weir 1988a in Weir 2005, p. 17). *A priori* evidence is the match between operations defined in the theory and test whereas *a posteriori* evidence is gathered through statistical analysis of the data to reveal underlying commonalities as well as through criterion referenced studies to

compare the results produced by similar tests. The other important *a priori* evidence component, context validity deals with the social aspect of the language. A test targets to assess certain skills, abilities, or knowledge, and it is necessary for it to ensure that it complies with the specifications, which are designed considering the context. Bachman and Palmer name (1996, p. 23) this concept as authenticity. A task must be authentic in that it must elicit similar behaviors to the TLU. It is evident that validation starts with theory-based validity because the construct to be measured is defined in theory, and is followed by context validity as context determines the relative importance of the subskills of the construct depending on the context. In addition, O’Sullivan and Weir (2011) emphasize cognitive validity in their recent model of test development and validation. Cognitive validity is the type of evidence collected during or immediately after a test regarding the actual mental processes a certain task operationalizes, and whether there is a match between these operationalizations and the operations defined in the theory. Whether it is called construct validity or cognitive validity, the crucial point is that the operations to be tested must be based on theory.

Therefore, as all validation models sees construct validity as very significant, we can conclude that to be able to make judgements regarding the validity of a test, it is necessary to construe how a construct is defined in theory. In the next section, a review of reading theories and models will be presented.

2.3 Theories of reading

2.3.1 Process models of reading

Urquhart and Weir (1998) and Taylor (2013) categorize reading theories into two: process models and componential models. In process models, the focus is on the

process, which means what happens in each stage and how these stages follow one another is of crucial importance. Process models are categorized into two: Bottom-up and Top-down Models. Bottom-up Models aim to explain reading comprehension as a sequential process consisting of several stages. The most important component of reading comprehension is the text, and comprehension starts with the letters decoded into words, and move up. Top-down models, on the other hand, focus on reader expectations which govern the reading ability (Grabe & Stroller, 2002, p. 32). Namely, readers are assumed an important role in reading. In short, componential models deal with separate set of skills or knowledge areas that are used during reading (Urquhart & Weir, 1998). Process models focus on what actually happens as a reading act goes on. Both lines of theories are significant because it is necessary to define the knowledge domains involved and understand how these come into play as reading unfolds. Therefore, interactive models, which combine the useful components of Bottom-up and Top-down models (Grabe & Stoller, 2002, p.33) emerged. Unlike Top-down and Bottom-up models, there is no sequential order (Urquhart & Weir, 1998, p.44). Interactive Models propose a parallel processing approach (Taylor, 2013, p.20). There are two significant interactive models conceptualized by Rumelhart (1977 in Urquhart & Weir, 1998, p.44) and Stanovich (1980 in Taylor, 2013, p. 21). Rumelhart suggested that information from all levels such as orthographic, lexical, syntactic, semantic as well as visual input interact at the same time for a reader to reach a meaningful interpretation (1977 in Taylor, 2013, p. 21). Stanovich's (1980 in Taylor, 2013) supporting Rumelhart's model, added that if there is any deficiency in any of the stages of reading comprehension, heavier reliance on another information source irrespective of each stage's position in the hierarchy may compensate for that deficiency while reading. For this reason,

he named his model as an Interactive Compensatory Model (Stanovich, 1980 in Taylor, 2013).

2.3.2 Componential models of reading

While the focus is on precisely what actually happens in the mind of a reader during reading in process models, for componential models, the focus is on what subskills and knowledge sources form or guide the ability to read. Thus, these models conceptualize the reading ability to be decomposed into subskills and knowledge types instead of being comprised of a group of psychological processes (Taylor, 2013, p.21). Hoover and Tunmer (1993, p3) proposed a componential model called the Simple View of Reading, in which reading can be deconstructed into two components, which are decoding and comprehension. This view does not assert that reading ability is just a simple task; rather it divides the complexities into two headings. In the absence of one, reading cannot take place. First, a text must be decoded, and then the message must be comprehended. If there is just decoding, it will not be reading, but just word calling as put forth by Hoover and Tunmer (1993, p3). Again, for comprehension to take place, input, which in the case of reading has to be derived by decoding, is necessary as well. Hoover and Tunmer (1993) exclude background knowledge, for they aim to explain reading ability not as performance with the rationale that background knowledge is constant and does not differ for reading and listening; as a result, it cannot be used to differentiate between the two (Urquhart & Weir, 1998, p.62). In addition, Coady (1979 in Urquhart and Weir, 1998, p.50) formulated a three-component model, which are Conceptual Abilities, Process Strategies and Background knowledge. Conceptual abilities are defined as the intellectual abilities. As for Process Strategies, Coady covers both competence and

production. In other words, knowledge of the language system as well as how to use this knowledge is crucial (Urquhart & Weir, 1998, p.50). Finally, Background Knowledge is what learners bring with them. It is important for two main reasons; first, reading a text is not complete unless a reader supplements it with the knowledge they have about the world, and second, which is more crucial for L2 readers, is that any deficiency in any other component may be supported and compensated by background knowledge (Urquhart & Weir, 1998, p.63). It might be concluded that background knowledge is a vital component for both L1 and L2 reading. Similarly, Bernhardt (1991) argued for a three-component model comprising of Language, Literacy and World Knowledge. Bernhardt's model is quite similar to Coady's. World Knowledge captures background Knowledge, Language is the knowledge of morphology, syntax, and semantics. Literacy is about knowing how to handle a text, knowing why to read it and what to do with it, which are all operational. Her model made reference to both higher and lower level skills (Kurt, 2015).

On one hand, component models describe what skills and knowledge is necessary for reading to take place. On the other hand, process models explain what happens during decoding. The two are invaluable to understand the ability to read. However, as Goldman et al. (2013) argues that these simpler views of reading comprehension cannot sufficiently explain and guide the literacy skills necessary in the 21st century because of the abundance of information requiring people to combine information from various sources in their personal, academic, and professional lives (Britt, Rouet, & Braasch, 2013; Coiro, 2011; Coiro & Dobler, 2007; Goldman, 2004; Goldman et al., 2012; Wiley et al., 2009 in Goldman et al., 2013). Readers are required to move beyond decoding and comprehension process

focusing on word and sentence level (Goldman et al., 2013). Doing this effectively necessitates to assess a document's in its entirety, which requires the analysis of the features of the documents such as the author, the date it was written in as well as the strength of the arguments in the text. Next section, proposition based approach, which focuses on the potential product of reading, rather than the process (Taylor, 2013, p. 22) will be explained.

2.3.3 Proposition based (text base) approach

With the developments in cognitive psychology, new approaches emerged as regards to reading theory. Kintsch and van Dijk (1983) proposed a multi-layered text processing theory consisting of surface level, propositional level, and situation model. Surface level represents syntax, morphology and lexicon, which help encode as well as decode ideas. Propositional level explains the relations between propositions that comprise a predicate and an argument (Urquhart & Weir, 1998, p.79). In principle, each proposition is an idea unit carrying varying levels of prominence within a text. For instance, a claim argued by the author carries more weight or prominence in the mental representation of the text compared to justifications s/he presents to support his/her claim. The relations between propositions at the local level (i.e. within a paragraph) is called microstructure. Some relations between propositions are made explicit via argument structure or linkers such as yet, however, but etc. while others are relatively implicit and might require the initiation of inferencing on the part of the reader. A collection of microstructures forges a macrostructure, which is also called text base. These microstructure and macrostructure levels are connected by macro rules, which are a group of distinct semantic mapping rules. Therefore, it is conceivable to conclude that text base is the

coherence graph of a whole text (Urquhart & Weir, 1998, p.79). As mentioned earlier, the linkages of propositions and microstructures are on occasion implicit. In the case that they are implicit, readers are required to infer the missing links, which requires consultation to comprehenders' background knowledge.

The earlier model by van Dijk and Kintsch (1978) had been criticized on two main grounds: Firstly, Brown and Yule (1983) argued that this approach sees reading as something that is vested chiefly in text, and is found lacking to emphasize the significance of reader interpretation as well as the intended meaning of an author. However, these two issues are truly vital during comprehension because the act of reading is goal driven (Gil, Bråten, Vidal-Abarca, & Stromso, 2010; McCrudden, Magliano, & Schraw, 2010) since readers as well as writers have a purpose. Secondly, Brown and Yule (1893 in Taylor, 2013, p.23) found this model lacking in that it cannot account for the different interpretations reached by different readers even when an identical text is read by them, and this model cannot particularly assert which interpretation is the best. With the addition of situational model, van Dijk and Kintsch enhanced their 1978 model by acknowledging that text representation not only involves text elements, but also knowledge elements (1983, p.336). Texts provide propositions, and relations between propositions, which help create a mental image of the texts on the minds of readers. However, without an existing background knowledge, the creation of a mental image would be beyond possible. van Dijk and Kintsch (1983, p.337) articulate that the presence of background knowledge enables readers to form situational models. It is, therefore, reasonable to infer that situational level is the interpretation of the information presented in a text in the light of the already existing background knowledge to reach a mental representation of the situation that is being described in the text.

The model does not prescribe what happens during decoding or parsing, yet focuses on what relationships between propositions in a text are, and what readers get out of a text kneading the text base with the background knowledge. Not the surface level representation but the gist (prominent propositions) is retained since the input is transformed into some other conceptual form.

According to the premises of this model, which is called Construction-Integration Model, text base, namely the bottom-up processes of decoding the propositions, which are generally incoherently represented ideas or concepts as well as the elements that become activated by those concepts and ideas are the construction component (Goldman, 2005). Integration component, on the other hand, lays in the relevance and strength of links and nodes that are activated during the construction stage. Those ideas and concepts that are highly linked are brought forward, while less connected ones are marked as irrelevant and ignored. Nodes with more connections are associated with the core meaning of a text. Therefore, ideas that are highly connected are integrated with the ideas derived from the text, and fill in the information gap if there is one (Goldman, 2005). From an assessment perspective, the end result representations of this process closely correlates with performance in comprehension tasks (Goldman, 2005). However, Lacroix (1999) claims that how a mental representation is formed is not very clear. Therefore, she postulates that comprehending a text comprises of two distinct levels of macrostructuring: Macrostructure Construction and Macrostructure Organization. Macrostructure Construction requires the identification and hierarchisation of information extracted from the texts in the abstract form. On the other hand, Macrostructure Organisation is where readers form a coherent mental representation by connecting several text representations. Still, formation of a mental representation

is the ultimate goal, and requires global level careful reading for the abstraction of the propositions in a text. Therefore, textual level representation leads to deeper learning. To illustrate, Ozuru, Best, and McNamara (2004) studied learner aptitudes when learning from text at secondary level. Skilled readers employed a more constructivist approach by combining information from different parts of an expository text “to see the larger picture” in Ozuru et al.’s terms, involved in more elaboration and inferencing. Citing Chi (2000), they also support that use of more elaboration strategies are associated with deeper learning.

2.3.4 The documents model

All aforementioned models set out to explain how readers process a single piece of text by either focusing on the components or processes involved. However, not any single model can satisfactorily account for the reading skills needed by students in academic contexts, where students must do more than knowledge-telling, which is telling what they know with text, and be involved in knowledge transformation, which is knowledge construction, reasoning, and argument with the text (Bereiter & Scardamalia, 1987). It is also acknowledged that when students in academic contexts are to solve a problem or make a decision, reference to multiple texts and sources is pivotal (Coiro, 2011). Therefore, Perfetti et al. (1999 in Strømsø, Bråten, Britt, & Ferguson, 2013) moved beyond the concepts of text base and situation model, and presented the Documents Model. The nature of textual and intertextual comprehension differ in that an extra layer is added to explicate the process of combining the information between each separate text and forming an intertextual net of relations between documents. Different rhetorical relations may be present between documents such as contradicting, exemplifying, supporting, and

complementing (Britt & Rouet, 2012). Information from each document “updates” the situation model. Documents are connected through nodes that demonstrate links and relationships between each document. As a result, intertext model is created (Perfetti, 1999 in *ibid*) as well as a situations model, which is an integrated mental representation of all the situations that are being described in each text (Strømsø & Bråten, 2014). As these layers of representation are formed, readers form a coherent and deep understanding of the information presented, but at the same time, they need to keep a record of “who says what” according to Hagen et. al (2014). In addition, Hagen et al. (2014) assert that “relations between ideas and concepts are often complex and implicit”; therefore, while making connections across texts, readers need to transform the information by making inferences.

Multiple texts reading is commonly practiced in the field of history (Anmarkrud, Bråten, & Strømsø, 2014). Providing an account for an historical event necessitates consultation to various texts and documents because a certain historical event may well be interpreted or reflected discrepantly by different parties considering the context at the time. Thus, reaching an accurate interpretation of an historical event is a demanding job involving sourcing, contextualizing, and corroboration, all of which are document level skills (Britt & Aglinskas, 2002; Gil, Bråten, Vidal-Abarca, & Strømsø, 2010a). These are the skills necessary in the 21st century considering the vast amount of information available through books, articles, and various online sources, which may handle an issue from very discrepant positions. Therefore, sourcing, contextualization, and corroboration are skills necessary especially in academic contexts. Sourcing encapsulates noticing the source of a document, using source information in the prediction and interpretation of a document’s content or evaluating its reliability, or making reference to the source

when utilizing a text's content (Rouet, 2006; Wineburg, 1991). Contextualization requires consideration of time and place, basically, spatial temporal context, when evaluating the relevance and trustworthiness of information presented in texts (Wineburg, 1991). Finally, corroboration “involves a systematic comparison of content across documents to examine potential contradictions or discrepancies among them” (Gil, Bråten, Vidal-Abarca, & Strømsø, 2010b). However, it may not be possible to include all these skills in reading comprehension tests. There is need to clarify intertextual reading construct as an operation that could be tested. For this reason, Goldman, Lawless, and Manning (2013) attempted to conceptually clarify the construct for multiple text comprehension through the use of Evidence-Centered Design approach. Goldman et al. (2013) defined the construct of multiple texts reading comprehension for answering an inquiry question. There are two sides to their assessment approach. The Domain Model and the Student Model. The Domain Model is shaped in the light of theory and research (Goldman et al., 2013), and based on this domain model, the student model, which consists of claims and evidence for the operations, analysis, synthesis, and integration, is developed. Even though this construct is defined for an inquiry task, it can still be applied to reading comprehension because analysis, synthesis, and integration can all be accomplished through the analysis of single texts as well as multiple texts (Goldman et al., 2013) without the need of manifestation of these operations through a productive skill. These analysis, synthesis, and integration components will be explained in detail while presenting the results of RQ1.

In addition, several studies investigated multiple texts reading comprehension and its effects on learning gains, which is an important consideration for students in academic contexts.

Cerdán and Vidal-Abarca (2008) investigated the effects of tasks on the level integration across sources through two different tasks: an intra-text question and an intertextual essay question. The results revealed that intertextual essay task led to more slow and incremental reading. In addition, this task resulted in more learning. They conclude that information integration is highly influenced by the task, and establishment of intertextual links predicts and increases learning outcomes.

Bråten, Ferguson, Anmarkrud, and Strømsø (2013) investigated the impact of word level processing, strategic approach and reading motivation on learning from multiple texts of adolescents. The results indicate that word recognition skills positively correlated with learning gains. However, for strategic approach and reading motivation, no unique variance is found. In addition, what is more important is that the results showed that good performance on multiple texts task led learners to elaborate information and integrate perspectives across texts. Another important suggestion of their study is that lower level skills such as word recognition still play a role in multiple-text reading comprehension.

Hagen et al. (2014) investigated note-taking while processing multiple texts in different task conditions. It was found that more elaborate, intertextual notes led to deeper comprehensions as opposed to summarization. In addition, when low performers and high performers were compared, it was observed from the notes of the participants that the ones establishing more intertextual connections ended up with better retention.

In conclusion, as stated by Anmarkrud et al. (2014), “multiple-documents comprehension, therefore, generally seems to require deliberate, goal- directed attentional, transformative, and integrative processing”, all of which are higher level

processes, and it is necessary that they be exhibited by students in academic contexts. The fact that university students need these skills is not just a product of observation, but proven through research.

2.4 Foreign language assessment frameworks

Rosenfeld, Leung, and Oltman (2001) investigated the skills undergraduate and graduate students must have in order to better design TOEFL (Test of English as a Foreign Language) to in a way to representatively sample the target behavior; that is, to create a framework. Data were collected from students and experts in the field on the task statements designed by framework teams. Experts and students gave feedback regarding whether these task statements reflect academic performance. For reading, the results revealed that basic comprehension is highly valued. In the study, an example task which was believed to operationalize the reading skills in academic contexts by the experts and students was presented. When operationalizations of this task are examined, it is seen that basic comprehension, learning and integration are all included. Task statements under integration are comparing and contrasting ideas in a single text and/or across texts and synthesizing ideas in a single text and/or across texts (Rosenfeld et al., 2001). This finding suggests that both students and experts in the field see higher level reading such as textual and intertextual comprehension a part of academic studies. In order to provide a general framework for assessing reading, Khalifa and Weir (2009) formulated a reading assessment framework.

Khalifa and Weir (2009) believe that reading purpose determines the type and level of reading, and the sources of knowledge to be used. In this model (See Figure 1), there is a goal setter, which determines the type of reading based on the task. If

expeditions reading to be employed, it could be employed at local (scanning, search reading) and global level (skimming). If careful reading, which could also be employed both at local and global levels, is to be employed, it is divided into other levels. Careful reading is a bottom-up process starting with word recognition, lexical access, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, and creating a text level structure. There is a hierarchy and as it moves up, the type of ability requires higher level reading skills. In this model, there is also emphasis on background knowledge because when creating propositional, textual

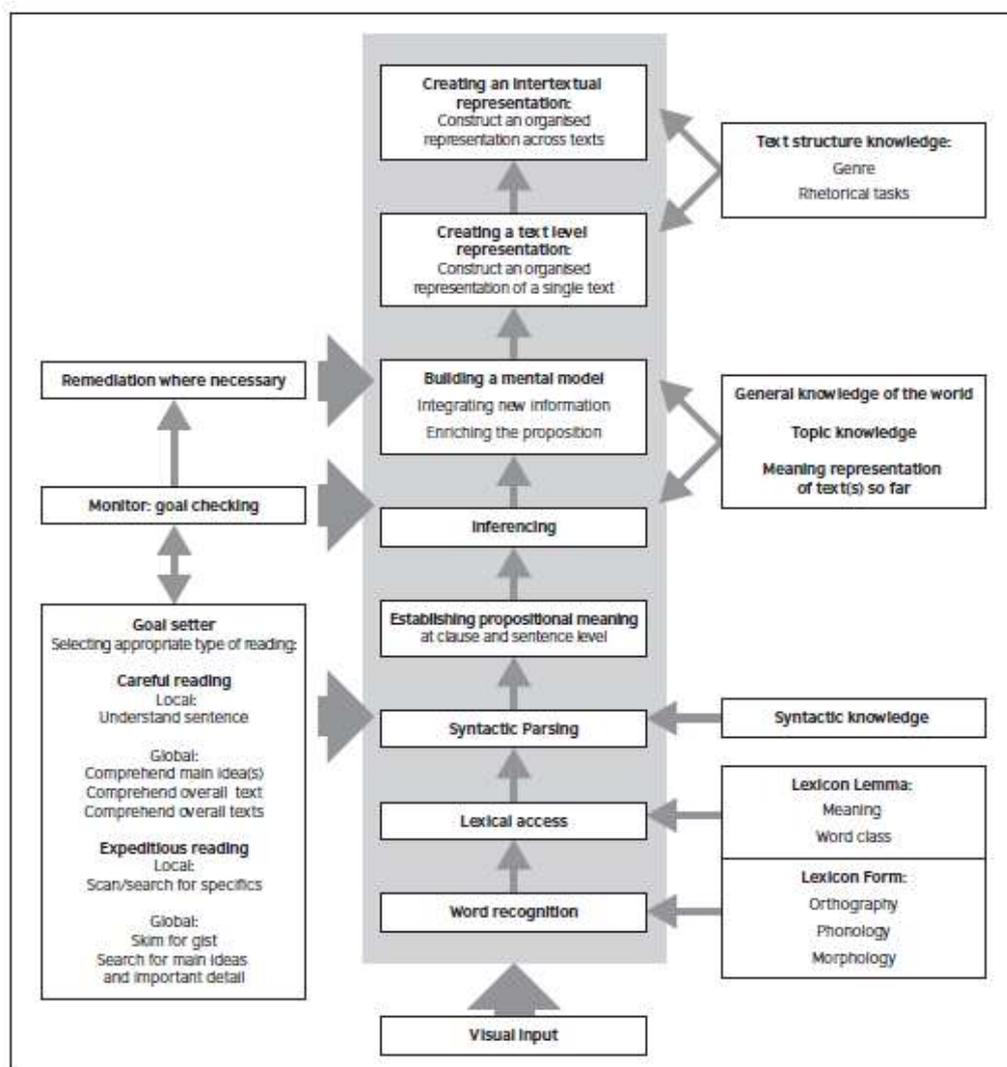


Figure 1. Cognitive processing in reading by Khalifa and Weir (2009, p.43)

and intertextual meaning, it is necessary to consult background knowledge. As mentioned earlier, otherwise, it would not be possible to create a situational representation.

Based on this framework, Ünalı (2010) investigated the types of reading students in academic contexts entertain at a British University. Data were collected on the nature of reading prevalent among university students through a questionnaire, reading diaries and interviews. It was revealed that reader goals determined the reading processes. Depending on the task, the students employed expeditious strategies to locate information, and careful reading for establishing propositional meaning, inferencing, textual and intertextual representations. This study is invaluable in that process based data were collected in a naturalistic academic setting, and it shows us that tests assessing reading at academic contexts need to operationalize both lower and higher levels of reading.

The studies reviewed so far show us that multiple texts reading skill is necessary in academic contexts. As it is shown by several studies, multiple texts reading comprehension, or the use of certain subskills of it improves retention and result in deeper learning, consequently better performance. Therefore, it is important to assess whether prospective university students do have the necessary skills before starting their studies. For this reason, the exams attempting to measure multiple texts comprehension skill need to be scrutinized for their validity as accurate and sufficient representation of multiple texts reading skill in EFL tests is crucial. In this study, this will be done through two means: retrospective think aloud and eye tracking. Thus, studies on cognitive validation of exam tasks, which guide this study will be presented next.

2.5 Construct validation research through eye tracking

Bax and Weir (2012) investigated readers' cognitive processes as they read a computer-based CEA (Cambridge English: Advanced) test. Data were collected through eye tracking and a questionnaire. The exam tasks consisted of 13 items in total. As participants took the exam, eye-movements were recorded, and after each item, retrospective questionnaire on reading strategy use appeared on the screen. The eye movement data were analyzed with four different tools: visual analysis of the eye-movement –data, showing the sequence of the movements, gaze plot data, heat maps and statistical analysis of the fixations. If there were three fixations on a question or option, it was considered to have been read. The findings suggested that readers employed a range of strategies in the framework of Khalifa and Weir (2009) from the lower level skills to whole text comprehension. Bax and Weir's (2012) study is valuable because it demonstrated new directions using a very innovative method as eye tracking despite the fact that it is necessary but difficult to find ways for systematically analyzing the eye movement data.

Bax (2013) aimed to examine the cognitive validity of IELTS (International English Language Testing System) by using 11 items from IELTS practice tests, which they asserted to be representative of an average IELTS test. In this study, Bax targeted only local level careful reading and local level expeditious reading. There is no explanation regarding this choice of scope. Data were collected through eye tracking and stimulated recalls on the video recordings of the eye movements of the participants. This method seems quite reliable, as the participants' verbal accounts will not suffer from forgetting. The results suggested that IELTS operationalizes lower level reading skills, but not on the levels of inferencing, building a mental model and understanding text function. The study did not set out to assess the

presence of those higher level skills anyway. It was revealed that careful and expeditious reading strategy use differed among low achievers and high achievers. Low achievers were not found to use expeditious reading strategies effectively, had difficulty locating relevant information, and had to process longer chunks of texts. Bax (2013) suggests that test writers may include lower level reading skills to some extent in their test design. However, it is imperative to include items aiming at different cognitive levels in Khalifa and Weir (2009) to reach a greater cognitive validity.

Brunfaut and McCray (2015) investigated the cognitive validity of Aptis Reading test following a similar methodology to Bax (2013). The purpose of the study was to identify whether items aiming at different CEFR levels trigger different cognitive processes. It was observed that a wide range of cognitive processes were activated by Aptis reading test except the intertextual representations which shows that Aptis reading test samples cognitive processes of reading representatively. Certain differences between different CEFR levels in terms of operationalisation of the cognitive processes were observed; however, these were attributed to task type rather than the CEFR level. This finding suggests us that task type plays a big part on the cognitive processes employed while reading.

These three studies employed a rather innovative methodology, which can help us be informed of the actual cognitive processes readers go through during reading test/task completion. As a result, through studies as such, the data gathered on online processes provide evidence on the cognitive validity of exams and exam tasks.

2.6 Conclusion to the literature review chapter

Language assessment validation is an ongoing process which starts with the definition of the skills and subskills to be tested considering the context the test will be used in and the purpose the test will be used for. These definitions must be derived from theory that is accumulated through years of research and observation to increase the chance of operationalizing the required skills and subskills accurately. Khalifa and Weir (2009) conceptualize reading ability as a multifaceted skill that involves skills operationalized at different cognitive levels. In their reading framework, multiple texts reading skill is defined as requiring an operation at the highest cognitive level. Previous research suggest that in academic contexts multiple texts reading skill is a requirement (Ünaldı, 2010; Rosenfeld et. al, 2001). Therefore, multiple texts reading skill needs to be operationalized in tests to be used in academic contexts besides the lower level comprehension skills to be valid. Otherwise, the scores obtained in a test do not reflect actual performance, which may have ramifications on different parties from test takers to test users, which Messick (1989) calls Social Consequences in his validity framework. Therefore, it is conceivable to conclude that validation of any sort of assessment is an essential process where there are tests to be used. In line with this, several high stakes exams such as CEA, IELTS, and Aptis have undergone such a validation process (See Bax & Weir, 2012; Bax, 2013; Brunfaut and McCray, 2015) in the light of validation frameworks such as Messick's (1989) and Weir's (2005). Recently, certain tests such as ISE II, MET, and ECCE have been found to aim at operationalizing multiple texts reading skill after the investigation of all the available English language proficiency exams. It is imperative to ensure that these exams representatively and accurately operationalize multiple texts comprehension skill due to the fact that these exams

aim at determining the language proficiency levels of students pursuing university education. Therefore, with the aim of providing evidence for the cognitive validation of these tests, two research questions were formulated. The research questions and details regarding the methodology is presented in Chapter 3.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The aim of this study is to investigate what reading processes the present multiple texts reading skill tasks in present language proficiency exams operationalize, and whether those processes match the ones defined in multiple texts reading theory and the specifications publicized by the institutions offering these proficiency exams. To achieve the aims of this study, a mixed-method was followed. Data were collected through retrospective think aloud protocol and eye tracking technique. The details regarding the participants, instruments, and procedure and data analysis are presented in the next section. The chapter is concluded with the detailed explanation of the data analysis procedure.

3.2 Research questions

As mentioned earlier, a priori validation is concerned with theory based validity and context validity. Therefore, it is necessary to investigate what sorts of reading processes the multiple texts reading tasks in the present study trigger. In other words, collection of cognitive validity evidence is crucial by investigating the task types and mental processes of test takers. For this reason, this study aims to answer the following research questions:

RQ1: Do multiple texts reading (MTR) tasks used in language proficiency tests attempt to operationalize MTR skill and subskills as defined in theory representatively?

RQ2: Do test takers use substantial MTR skill as defined by theory and as specified in the test specifications in responding to the MTR tasks in tests where such tasks are available?

3.3 Participants

10 young adult participants who are enrolled in a foundation university in Turkey as well as students studying at a public university voluntarily took part in the study.

They were contacted during class hours through announcements. A brief information was given about the study, and an individual experiment session was arranged with each participant based on their school schedule and availability.

In this study, minimum required proficiency level is set to B1 so that the level of the students would match the level of the tasks to be given. Students had varying nationalities as Turkish, Syrian, and Lebanese. Majority stated that they had been learning English for at least 10 years. The participants had normal or corrected to normal vision. Those wearing glasses, if possible, were kindly asked to wear contact lenses; otherwise, the eye tracker could not record eye-movements. Therefore, the participants who had vision problems and could not wear contact lenses were excluded from the experiment.

3.3.1 Ethical procedures and consent

Ethical consent was obtained from Boğaziçi University, Institute of Social Sciences Ethics Committee before data collection. In addition, all the participants have signed a consent form on which all the details of the study were outlined before initiating the data collection session.

3.4 Multiple-text reading tests

Four tasks purportedly measuring multi-text reading comprehension from three different exams were selected after scanning all the English language proficiency exams available. These tests are international language proficiency exams aimed at non-native speakers of English. Below is a description of each exam and task.

3.4.1 Integrated Skills in English (ISE) II

ISE II is a B2 level English proficiency exam taken by adolescents, young adults and adults with occupational and educational purposes. It tests reading, writing, listening and speaking skills. Reading and Writing section consists of three parts, one being long single text reading, and one multi-text reading, and an extended writing task.

The section in concern, multi-text reading, has four different tasks. The first task (5 items in total) requires matching questions with the text that accommodates the information to those questions. The second section (5 items in total) requires candidates to locate specific information in any text and decide whether the statements given are true or false. The third task, which again comprises 5 items, is an outline summary task, in which candidates need to fill in the gaps in the summary in an outline form by writing maximum three words extracting words and numbers from the information in all the texts. The last task is a reading into writing task, which requires candidates to compose an essay based on the information presented in the four texts. Based on the specifications, the following are the abilities tested:

- The ability to understand the main idea or purpose of each text,
- the ability to understand specific, factual information at the sentence level,
- the ability to understand specific, factual information at the word and/or phrase level across the texts. (Integrated Skills Examination, 2017)

3.4.2 Michigan English Test (MET)

MET is part of Cambridge Michigan Language Assessment, which is an exam measuring proficiency between A2 to C1 levels. It is taken by adults and adolescents for educational, occupational, or promotion-related purposes. In the test, there are four reading tasks each including three passages followed by a few questions focusing on individual texts, and then one or two questions that are to be answered based on the information from all the texts. In the test specifications, it is stated that the following are the abilities tested.

- at global level,
 - understanding main idea/gist,
 - understanding author's purpose/opinion/attitude,
 - making connections across texts;
- at the local level,
 - understanding vocabulary in context,
 - identifying referents;
- at inferential level,
 - understanding implicit ideas,
 - drawing conclusions
 - identifying rhetorical function. (CaMLA, 2017a)

3.4.3 Examination for Certificate of Competency in English (ECCE)

ECCE is a B2 level English Language Proficiency test taken by teenagers, young adults, and adults with academic and occupational purposes. It consists of two parts. The first part consists of single texts followed by multiple-choice questions, and the second section comprises of two sets, each of which includes four thematically

linked texts followed by 10 questions. Only three of these 20 questions require information from two or more texts to be answered. In the test specifications, it is stated that the following are the abilities tested:

- at global level,
 - comparing/contrasting features of one or more texts,
- at the local level,
 - understanding explicitly stated ideas (detail) from one or more texts,
- at the inferential level,
 - drawing an inference/conclusion from one or more texts. (CaMLA, 2017b)

All these tests are available online and can be seen in Appendix A, B, and C.

3.5 Instruments

3.5.1 Retrospective think aloud protocol

Verbal reports have been in use in cognitive science, education, and psychology for a long time (Leow & Morgan-Short, 2017). The affordance of this methodology such as the provision of online disclosure of mental processes makes it an invaluable tool (Cohen & Cavalcanti, 1987). In multi-text reading comprehension studies, this procedure is commonly employed (See Bråten & Strømsø, 2003; Cerdán & Vidal-Abarca, 2008; Ferguson, Bråten, & Strømsø, 2012; Strømsø et al., 2013). In testing, since validation is an ongoing process, collecting evidence about what processes each task and item yield, and whether those match the processes defined in theory is quite significant (Green, 1998); therefore, think-aloud protocol is commonly used as well. However, Van Den Haak, De Jong, and Jan Schellens (2003) state that concurrent think aloud protocols may lead to reactivity and may impact on task

performance because test takers have to verbalize what they think, which may create burden on the working memory. Therefore, in this study, retrospective think aloud protocol will be used. However, not a single data collection method is without any limitations as is the case for retrospective think aloud protocol. For instance, there is the risk of forgetting since participants do not concurrently report (Tai, Loehr, & Brigham, 2006). Therefore, in this study, retrospective think aloud protocol and eye tracking techniques were paired for data collection to accurately tap into the cognitive processes employed during reading considering the complexities of cognitive processes reading triggers.

3.5.2 Eye tracking

Tobii Eye Tracker x1 Light in Vision Lab at Psychology Department at Boğaziçi University was used during the experiment to collect process based data. This eye tracker recorded binocular eye movements at a rate of 30 Hz / 1000 ms. A chin rest was used to get more accurate results. The participants were sat at a distance of 55 centimeters from the computer screen. The exam tasks were presented on a screen with a 1980x1080 resolution.

For the purposes of this study, total fixation count, total fixation duration, careful reading percentage was calculated for each text in each exam task.

3.6 Procedure

Each participant took the tasks alone with the researcher in one session. No time limit was set for completing the tasks in order to avoid any anxiety and affective factors (Dolgunsöz & Sariçoban, 2016). Thus, the durations of the sessions ranged between 35 minutes to 85 minutes. First, the participants were explained the stages in

the study. Then, they were asked to sit in front of a computer placing their head on a chin rest, and asked to sit at a comfortable position by arranging the height of their seat as they would not be allowed to change position throughout the experiment. Experiment stage began with calibration where the participants were asked to look at the blue dots appearing and disappearing on the computer screen. The purpose of the calibration was to check whether eye movements were being accurately recorded.

After the calibration, the first task (See Appendix D) presented to candidates was a training task comprising of two sections. First, the participants read the text on the left of the screen and answered the questions on the right by telling their answers aloud. At the same time, their eye movements were recorded. Then, in the next stage, the participants were asked to verbalize what they thought or did while completing the task step by step. If they failed to give a detailed account of what they did, the researcher guided them by modelling think aloud on the same task. This was done to exemplify what was meant by thought processes.

The experiment stage consisted of three exam sets. MET and ECCE tasks included two multiple choice questions each while ISE II consisted of three sections each of which included 5 items. Therefore, the tasks of ISE II were presented on two different pages. ISE II Task I was presented on the first page, and ISE II Tasks 2 and 3 were presented on the second page together. For each task, the participants read the texts and answered the questions, during which eye movements were recorded. Then, in the second stage, they reported what they did to reach the response they decided on. The order of the exams for each student was randomized to prevent any fatigue effect on a certain task. All the sessions were audio recorded with Olympus VN-541PC recorder.

3.7 Data analysis

A reading strategy coding rubric (See Appendix E) was developed together with an Applied Linguistics expert based on Cohen and Upton (2006) and Goldman, Lawless, and Manning, (2013). Overlapping strategies were identified and eliminated. Reading strategy coding rubric was expanded as new strategies emerged from the data. As suggested by Grounded Theory, researchers can contribute to the theory through rigorous analysis of empirical data (Charmaz & Belgrave, 2007). In this case, strategies defined in theory did not enclose the strategies emerged in the data; thus new ones were added.

The verbal accounts the participants provided on what they did while reaching at the responses in each task were analyzed in terms of the strategies employed. Each recording was listened by two raters at the same time, and the strategies the participants used while taking the tasks were identified by consensus using the aforementioned rubric. This was because the rubrics were revised to fit the purpose during the coding, which also meant recoding certain data in a recursive manner. So, the two raters listened to the recordings simultaneously and did the coding which was immediately followed by the discussion on strategy used. For MET and ECCE, each question was analyzed separately. For ISE II, as there are three types of questions in three different sets, the questions in each set were analyzed together. For each item, the frequency count of the strategies was calculated, and analyzed through descriptive statistics using Microsoft Excel 2016.

The eye movement records obtained from 10 participants were analyzed through MATLAB and Excel. An expert on computer sciences cleaned the data based on the validity values, which means only the data obtained from both the eyes

at the “0” validity value were included in the analysis. Then, the consecutive data points were accepted within the same fixation if they fell within 5cm radius following the first data point. The eye tracker used in this study records data at 30 Hz. Namely, each data point was approximately $1000/30$ ms. Therefore, the fixation duration for each fixation was calculated using the following formula : $1000/30 * \text{number of fixations}$. The data points whose fixation duration was longer than 100 ms and above was accepted as fixations. Instructions, text headings, titles, every sentence in each text as well as the questions and options were labelled as Areas of Interests (AOIs). Each AOI was assigned a number for further analysis (See Appendices F, G, H and I). Then, these AOIs were defined as boxes using the pixel values. For each participant, the fixations calculated before were matched with the AOIs through a code run on MATLAB. After this analysis, a set of data showing the fixation and fixation duration on each AOI was obtained. Based on these data, fixation counts, fixation durations, and the presence of careful reading on each key area were calculated using Excel descriptive statistics. Minimum fixation count was designated as three for a sentence to be considered carefully read, which was a predetermined criterion in Bax and Weir (2012).

3.8 Conclusion to the methodology chapter

This study aimed at investigating the cognitive validity of language proficiency exams such as ISE II, MET, and ECCE, aiming to operationalize multiple texts reading comprehension. Therefore, the study for which data were collected through retrospective think aloud and eye tracking, followed a mixed method to calculate the frequencies of the strategies employed by test takers to gain insight on the thought processes of the test takers and the quantitative analysis of the

reading behavior to determine the nature of reading processes used in doing the tasks. The reported strategies were coded using a reading strategy coding rubric, which was developed in the light of reading theory. Eye movement records were used to calculate fixation count, fixation duration, and careful reading proportion, which was done with the aim of understanding whether tasks lead the texts to be read equally and carefully to what extent. In addition, two individual cases were further analyzed to identify the sequence of the reading behavior multiple texts reading tasks in the tests in the present study required. The results of the analyses are presented in Chapter 4.

CHAPTER 4

RESULTS

4.1 Introduction

This study aims to examine the current multiple-text reading comprehension tasks in international English language proficiency exams available in the market in terms of the reading strategies they trigger, and whether the strategies triggered by these tasks match the strategies outlined in the test specifications of these exams and the strategies defined in theory. In other terms, this study aims to assess the construct validity of these exams. To this end, these tasks were administered to the participants and data were collected through two means: eye movements and retrospective think aloud protocol. The verbal accounts of the participants were analyzed by coding each strategy used by the participants when they were taking the tasks using Reading Strategy Coding Rubric (See Appendix E) developed with the help of an Applied Linguistics expert. After the coding, frequency counts were calculated and analyzed through descriptive statistics. The eye movement data were analyzed using MATLAB and Excel. The results of the analysis are presented under the relevant research question for each exam task.

4.2 RQ1: Do multiple texts reading (MTR) tasks used in language proficiency tests attempt to operationalize MTR skill and subskills as defined in theory representatively?

Multiple-text reading skills is considered as the highest level of reading ability (Khalifa & Weir, 2009). These skills are rooted in The Documents Model by Perfetti

et al. (1999). The Documents Model is defined as consisting of two components: Intertext Model and Situations Model (Perfetti et al., 1999). Intertext Model is the representations of the connections between texts whereas Situations Model is formed based on the mental representations of the situations being described in each text (Perfetti et al., 1999). Therefore, both the content of the texts and their relations to each other such as complementary or conflicting is quite important in multiple texts reading comprehension. In addition, when reading multiple sources to solve a problem, sourcing, corroboration and contextualization in the literature are listed as document level skills (Britt & Aglinskias, 2002; Gil, Bråten, Vidal-Abarca, & Strømsø, 2010). However, sourcing is not an ability that could potentially be tested in reading exams because sourcing involves the investigation of the trustworthiness and relevance of a document when completing problem based tasks through searching for information. It is known that in reading exams, texts are provided to the test takers, so they are not required to evaluate sources' relevance and reliability. The second skill, contextualization, is "to situate document content in a broad spatial-temporal context" (Wineburg, 1991 in Anmarkrud, Bråten, & Strømsø, 2014). This is also an ability that cannot be easily tested in comprehension tests. On the other hand, corroboration, which involves checking consistency of claims and evidence across texts (Rouet, & Britt, 2011), requires comparing and contrasting information across texts and could be an important ability to be tested in comprehension tests. Goldman et al. (2013) defines the multiple texts reading domain as comprising of six components. They separate the comprehension component into analysis, synthesis, and integration subcomponents for the ease of assessment. As for the analysis subcomponent, it is necessary to "determine the relevance of information to the task, and identify claims and evidence in each text"

(Goldman et al., 2013). The synthesis subcomponent requires the comparison of claims across texts for consistency and relevance whereas the integration subcomponent is the combination of similar claims, and organization of complimentary claims, and relation of evidence to claims irrespective of the way they were introduced in the texts (Goldman et al., 2013). Therefore, for Goldman et al, (2013) a task testing multiple-text reading comprehension must operationalize the following skills:

- determining the relevance of information to the task,
- identifying claims and evidence in each text,
- comparing claims across texts for consistency and relevance
- comparing evidence from different sources,
- combining similar claims; organizing complementary claims
- relating evidence to claims regardless of how they were introduced in the texts. (Goldman et al., 2013)

Considering the definition explicated above regarding multiple texts reading comprehension, the exam tasks used in this study will be examined and the multiple texts reading skill operations specified in the test specifications will be compared against the multiple texts reading skill operations in the literature as cited above.

4.2.1 ISE II

Multi-text reading task in ISE II comprises three different tasks involving five questions each. The examination body delivering this exam introduces the operationalized skills in the test specifications as:

- Task 1: the ability to understand the main idea or purpose of each text,

- Task 2: the ability to understand specific, factual information at the sentence level,
- Task 3: the ability to understand specific, factual information at the word and/or phrase level across the texts. (Trinity College London, 2017).

Based on the test specifications, ISE II does not seem to attempt to operationalize multiple texts reading skill sufficiently and representatively. Task I seemingly operationalizes an ability that is at the textual level, as “to understand the main idea or purpose of each text”, requires careful reading at the global level, and the formation of macro structures of texts but not necessarily corroboration of information across texts. Task 2 attempts to operationalize an ability at the sentential level as stated in the specifications, which requires careful reading at the local level. Only ISE II Task 3 focuses on multiple texts and this task attempts to operationalize the ability to understand specific, factual information at the word and/or phrase level across the texts. However, a test measuring multiple texts reading comprehension must go beyond word and phrase level, and should test claims at least at propositional level because in the literature, multiple texts comprehension is defined as the ability to form a coherent mental representation based on the information gathered from different sources which provide information on the issue in question from different perspectives (Britt, Perfetti, Sandak, & Rouet, 1999; Goldman, 2004). It is evident from this definition that textual comprehension is a prerequisite for multiple text comprehension. However, this task does not attempt to operationalize this ability, rather a lower level ability. Still, as test takers are required to complete the tasks based on four different texts, what operationalizations are achieved could be identified through the analysis of the operations the participants employed.

4.2.2 MET

MET is presented as operationalizing the following reading skills:

- at global level,
 - understanding main idea/gist,
 - understanding author's purpose/opinion/attitude,
 - making connections across texts;
- at the local level,
 - understanding vocabulary in context,
 - identifying referents;
- at inferential level,
 - understanding implicit ideas,
 - drawing conclusions
 - identifying rhetorical function. (CaMLA, 2017a).

The MET task seems to sample a large array of lower and higher level reading abilities. When operations to be tested are examined in detail, at global level, the task attempts to operationalize making connections across texts. However, what sort of connections to be tested is not specified. This operation needs to be specified in a way to reflect the observable behavior test takers may present. Therefore, based on the specifications, it is beyond possible to make comments on the representativeness of the abilities tested in MET as far as multiple texts reading comprehension is concerned. As stated above, Goldman et al. (2013) specifies the subcomponents of multiple source comprehension as analysis, synthesis, and integration; however, as mentioned earlier, these MET specifications do not outline the sorts of connections to be made across texts. In addition, there is no specific reference to different sections of the exam regarding what skills each aims to operationalize. Therefore, it is

necessary to scrutinize the actual operations this task achieves through the analysis of the participant reports and eye tracking data, which will inform us about the construct validity of this task.

4.2.3 ECCE

Based on the test specifications, ECCE aims to operationalize the following skills:

- at the global level,
 - comparing/contrasting features of one or more texts,
- at the local level,
 - understanding explicitly stated ideas (detail) from one or more texts,
- at the inferential level,
 - drawing an inference/conclusion from one or more texts. (CaMLA, 2017b).

When specifications are scrutinized, it is seen that all the abilities tested are abilities across texts or they could be operationalized across texts. The ability at the global level, comparing/contrasting features of one or more text, is an important skill in source evaluation (Bråten, Strømsø, & Britt, 2009), which is one of the key elements of multiple texts reading comprehension (Perfetti et al., 1999). The other two abilities are listed in Goldman et al. (2013). The subskill, understanding explicitly stated ideas (detail) from one or more texts, attempts to operationalize the analysis component, drawing an inference/conclusion from one or more texts, does so the synthesis and integration components. Therefore, ECCE could be considered as attempting to operationalize multiple texts reading skill representatively. However, in the test specifications, these operations are constructed in a way that the same ability may be tested at the textual level or intertextual level. The inclusion of

the phrase “from one or more texts” prevents the task to be interpreted as a multiple-text reading comprehension task. All in all, it could be concluded that ECCE attempts to operationalize multiple texts reading skill based on the specifications if we assume that ‘or’ option would be used in favor of multiple texts reading comprehension. It is necessary to investigate how these operations listed in the specifications are operationalized during the actual reading process. An investigation into the participants’ reported strategies will reveal the strategies these two items operationalize.

4.3 RQ2: Do test takers use substantial MTR skill as defined by theory and as specified in the test specifications in responding to the MTR tasks in tests where such tasks are available?

4.3.1 Reading strategy use in ISE II

To begin with, in total, there were 15 items in this task, and Table 2 shows the participants’ scores (See Table 2).

Table 2. The Participants’ Scores on ISE II Multi-text Reading Task

Participant Number	Task 1 Score	Task 2 Score	Task 3 Score	Total Score/15	Total percentage
P01	5	4	0	9	60%
P02	2	3	1	6	40%
P03	4	3	1	8	53%
P04	5	3	2	10	67%
P05	5	4	1	10	67%
P06	5	3	2	10	67%
P07	4	4	3	11	73%
P08	5	3	1	9	60%
P09	3	3	2	8	53%
P10	4	2	1	7	47%
Mean	4.2	3.2	1.4	8.8	59%

It is seen that 80% of the participants had at least 50% success rate overall. A closer look at the individual task scores reveals that participants performed better in Task 1; however, the scores in Task 3 is rather low compared to the other two.

Task 1

ISE II Task 1 is a matching task where candidates are expected to match questions to the texts that accommodate the answer to those questions (See Appendix A)

Table 3 shows the accurate responses each item attracted in Task 1. As can be seen, 90 % of the participants accurately responded to items 1.3, 1.4, and 1.5.

Table 3. ISE II Task 1 Frequency of Accuracy for Each Item

Item number	Frequency	Percentage
1.1	7	70%
1.2	8	80%
1.3	9	90%
1.4	9	90%
1.5	9	90%

When the participants were responding to the items in this task, they reported 108 strategies in total. Of all these, majority (50%, n=54) was expeditious reading strategies, followed by careful reading strategies (42%, n=45). Multiple texts reading strategies used in this task was only 2 % (n=2) (See Figure 2).

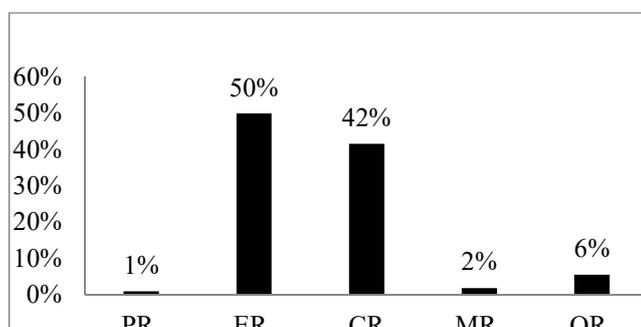


Figure 2. ISE II Task 1 overall strategy use

When reading strategies are closely examined (See Figure 3), among the expeditious reading strategies, ER5, which is “searching for/identifying key words/ideas in the text related to the question”, was the most commonly used strategy (18%, n=19). The second most commonly employed strategy (13%, n=14) was ER6, which is “searching for/identifying key words/ideas in the texts related to the question”. ER7, which is “based on the prior knowledge of texts and visuals, identifying/ trying to identify the relevant information related to a task” was reported as the third common expeditious reading strategy (8%, n=9) for this task. When careful reading strategies reported in the study are scrutinized, the results indicate that CR7, which is “reading only the part of the text which seems related to specific question/s” is the most common strategy (14%, n=15). CR8, “rereading the difficult and/or relevant parts of the text”, was the second mostly applied strategy (6%, n=6).

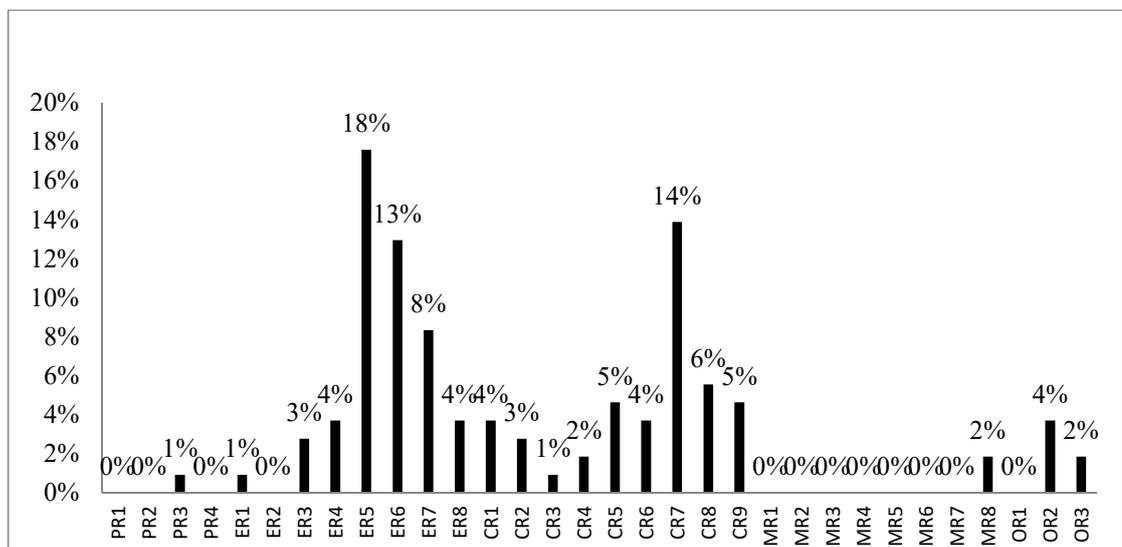


Figure 3. ISE II Task 1 reading operations

Task 2

ISE II Task 2 questions require the participants to choose five items that are true out of eight based on the information provided in four different texts (See Appendix A)

For these items, overall reading strategy use is presented in Figure 4, and the

distribution of reading strategies operationalized through this task are presented in Figure 5.

In total, 114 strategies were reported for this task. Majority of these strategies were careful reading strategies (51%, n=58), which was followed by expeditious reading strategies (42%, n=48). Only 2 % of the strategies reported were multiple texts reading strategies (n=2). For this task, no pre reading strategy was reported (See Figure 4).

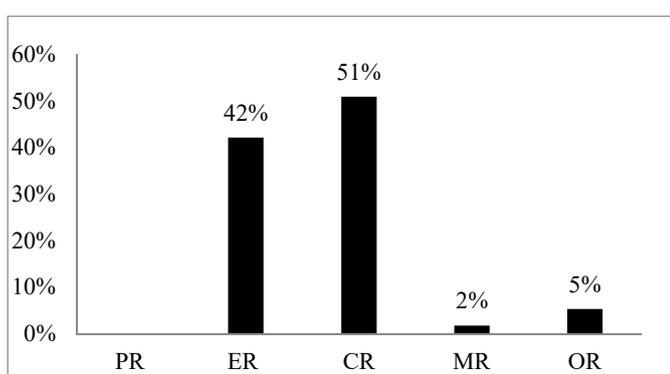


Figure 4. ISE II Task 2 overall strategy use

When reading strategies are analyzed in detail (See Figure 5), the findings reveal that, in terms of expeditious reading, ER5, “searching for/identifying key words/ideas in the text related to the question” is the most common strategy (13%, n=15). It was followed by ER7, “based on the prior knowledge of texts and visuals, identifying/ trying to identify the relevant information related to a task” (11%, n=13) and ER8, “choosing one text which seems related to a specific question depending on prior skimming, (8%, n=9). Concerning the careful reading strategies, CR9, “choosing one text which seems related to a specific question or option depending on prior careful reading”, is the mostly reported strategy (13%, n=15) among careful reading strategies, and it is used as frequently as ER5. CR9 was followed by CR3, “reading carefully across sentences (to establish the connections of ideas between sentences or parts of the text by identifying relationships such cause and effect, claim

and supports etc.)”, and CR7, “reading only the part of the text which seems related to specific questions”, both of which were reported 10% of the time (n=11). As regards to multiple texts reading skill, MR8, “comparing the gists of different texts” was the only strategy pronounced (n=2).

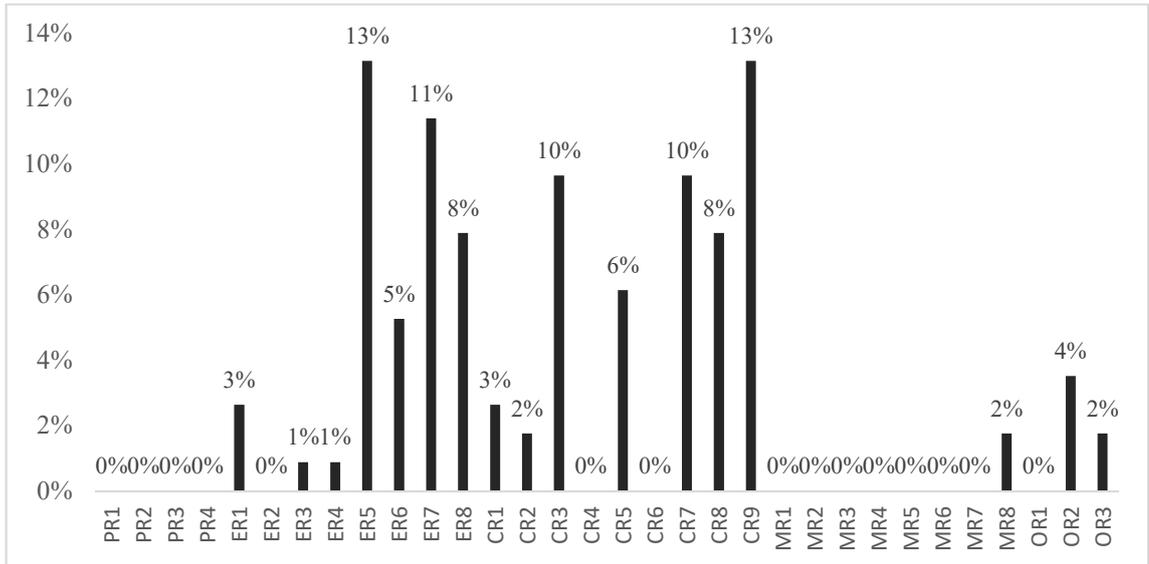


Figure 5. ISE II Task 2 reading operations

Task 3

ISE II Multi-text Reading Task 3 requires test takers to complete a summary by filling in the blanks in maximum three words with the missing information by extracting words and numbers from the texts (See Appendix A).

As it is clear from Figure 6, for this task, of all the total 93 reported strategies, majority were expeditious (44%, n=41). The second most common type of strategy

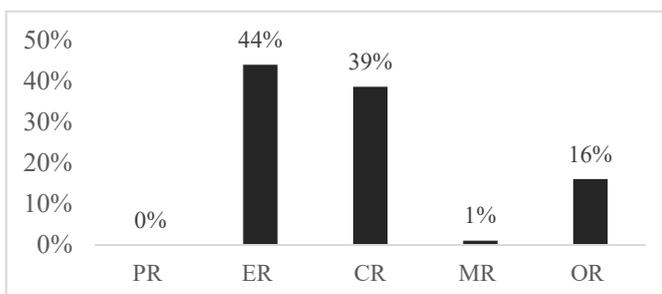


Figure 6. ISE II Task 3 overall strategy use

was careful reading strategies (39%, n=36). Unlike the previous two tasks, the

proportion of other strategies is rather high for this task (6%, n=16). Similar to Task 2, no pre reading strategy was reported.

When the distribution of the strategies is analyzed (See Figure 7), it is seen that ER5 “searching for/identifying key words/ideas in the text related to the question” (15%, n=14) is the mostly reported strategy followed by ER6 “searching for/identifying key words/ideas in the texts related to the question” (14%, n=13), and ER8 (10%, n=9), “choosing one text which seems related to a specific question depending on prior skimming”. The proportion of OR2, “using background knowledge to support understanding / guess or interpret meaning” (9%, n=8) and OR3, “answering the question based on the information gathered up to that point without going back to the text/ or only to confirm” (8%, n=7) is relatively high as well. Only one instance of multiple texts reading strategy, MR2, “identifying claims that agree, disagree and complement one another in different texts” use was reported (1%).

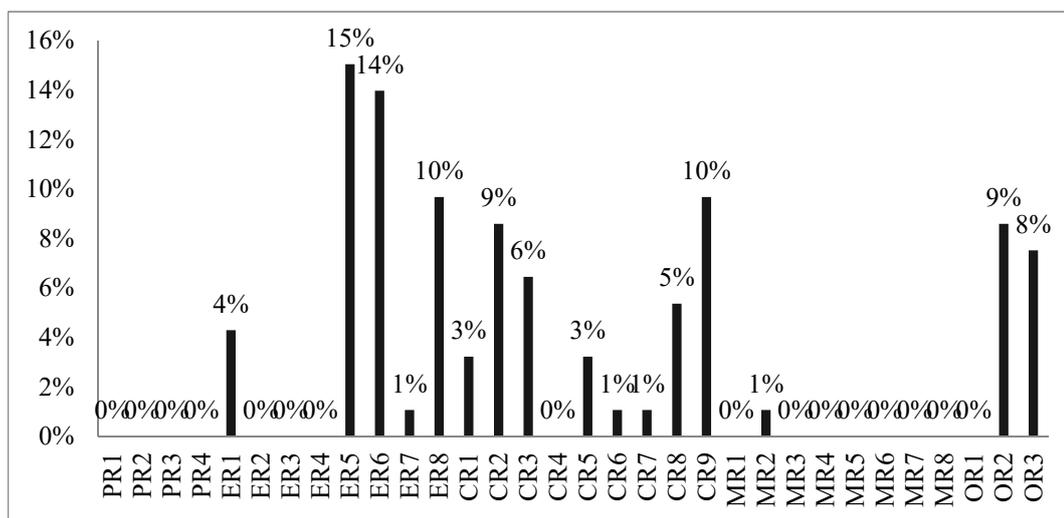


Figure 7. ISE II Task 3 reading operations

4.3.1.1 Task characteristics of ISE II multiple texts reading skill

Apart from the analysis of the strategies employed, the verbal protocols of the students provided insight into the thought processes of test takers, which may have implications on the embedded characteristics of these tasks. In the next section, several instances regarding certain items will be reported.

In Task I, Item 1, was responded accurately by 70% of the participants (n=7) as mentioned earlier. For the participants whose response was inaccurate, there was a general trend for the way they justified their responses. Two of these three participants matched Item 1 with Text C because Text C contained information regarding the nutrient content of locally produced food, and that vitamin levels drop quite soon after picking, which is not the case for locally grown food. In addition, the person composing Text C included the details of her well-being since she started eating “such good” food. Although implicit, it is possible to deduce from the context that since it is possible to eat locally produced food directly after picking, the food will be fresher, so will taste better. The participants making such an inference, which is a higher-level reading skill than simply matching key words, through which the correct answer could be reached for this item, are at a disadvantage. Others who simply matched the key words accurately answered this item.

The justification the participants provided is below:

I looked at Text C. Here, a person talks about her opinion, so you expect her to say something like that (referring to the item), and she says, “I have been eating such good food, and I feel fantastic”. So good food, I thought tastes better. (Participant 7)

In Text C, the sentence starting with “I always used to”, she says, “I have been eating such good food and I fell fantastic”. So, I matched it with Text C. (Participant 8)

In Task I, Item 2, accuracy rate was 80% (n=8). The justifications of the two students who responded inaccurately to these items are parallel and presented below. For this item, the participants simply matched key words (return to-in the item, and shift-back-in the text), which led them to an inaccurate response.

In Text A, the question says “criticizes the idea that people could return to”...yeah... it says here (Text A) that local food movement wants a shift-back to small scale farming. (Participant 8)

I think I found it in “Nowadays in such areas local food movement wants a shift-back to small scale farming. (Participant 3)

In task I, Item 3, accuracy rate was 90% (n=9). For this item, majority of the participants reached the answer through key word matching. Some examples from the verbal protocol regarding how they reached the answer are presented below:

Jane was talking about vitamins I guess. Where is it? “Locally grown food is better for us. That’s another reason why people should buy it”... “The change has been incredible” That’s why...I just looked at the text and saw the word vitamin, I matched it with feeling better. (Participant 3)

I first skimmed the texts. In text C, something caught my attention. It was about health, and I remembered there was something about health in the questions, so I matched them. (Participant 1)

Here (Text C) it talks about vitamins, chemicals and her feeling better. So, I thought these are better for well-being. (Participant 4)

Task I, Item 4 was responded accurately by 90% of the participants as mentioned above (n=9). All of these participants reported that it was the first item that they responded to because when they read the word “stage” in the question stem, they directly matched it with the text which provided a diagram with arrows showing the stages of a process. For instance:

First, I read the questions, and I looked at the texts, in Text B different stages are shown in a picture, so there was not even need to read it, and I matched it with 4. (Participant 4)

It is clear I think, Farming, Storing, Transporting... (Participant 6)

Different stages... It is obvious. There are arrows. I didn't even read it. (Participant 4)

Task 1, Item 5 was matched with the correct text by the majority of the students (90%, n=9). Item 5 was “Which text compares the farming in the last century with the popularity of farming nowadays?” Only Text A included numbers/dates. Therefore, it was easy to reach the correct answer. The justifications the participants provided were parallel. Some examples from their verbal protocols as regards to their justification of their response are presented below.

It just talks about that 90s and then 20s like they just comparing the last century and 21 century. That's why, I chose it (Text A). They are just all statistics. (Participant 3)

In the question it says last century. When I saw 90s (in Text A), I directly chose it. (Participant 4)

In this text (Text A) it says 90s, and now. Actually, I didn't read the text. I just saw the dates and matched it. (Participant 7)

When it comes to Task 2, there are 8 items in total, out of which five are true based on the information provided in all four texts. It is important to note that although majority of the students could comprehend the texts, which was evident through their summarization when talking about their thought process, they could not always accurately respond to the Items 1, 5, and 6 (respectively A, E, and F) in Task 2 (See Appendix A). Below are the extracts from their verbal protocol, which may enlighten us about the reason behind this case.

A is an item that is false but mostly chosen as true because the participants attempted to answer it just through key word matching. Item A is "US local food supporters want a return to farming levels of the 1900s." The answer lies in Text A; however, the text has information regarding farming style, not level.

When I read the statement A, I thought the answer was probably in Text A as it talks about faming levels. I read that part, and found the answer. (Participant 7)

I chose Text A because they want a shift back to small scale farming.

(Participant 6)

E is an item that is true but mostly was decided to be false. Item E is "Small farms sometimes use chemicals". Text C contained this information: "Large farms often use more chemicals than smaller ones" With this item, one of the main

propositions of the text is not tested, but what is tested is rather a small detail, which frequently misled the participants. Below is an extract:

In Text C, she mentions that she has been eating good food and feels fantastic. Also, here it says large farms often use more chemicals than smaller ones. So, small farms don't use chemicals but large ones do. E is false.
(Participant 7)

F is an item despite being false considered as true. Item F is "Jane believes there has been a slight improvement in her health and mood." Text C included the following information "The change has been incredible. I always used to get colds, now I never do since I have been eating such good food." What makes this statement incorrect is only the presence of the adverb "slight". Even though, the participants could make the connection between good food and health, which is one of the main propositions of Text C, they could not accurately respond to this item because of the oversight of the word 'slight'.

Last sentence, I have been eating good food and I feel fantastic. (In Text C).
(Participant 6)

I did not feel the need to go back and check. I believe the answer was in Text C. (Participant 7)

4.3.2 Reading strategy use in MET

Two items from MET were administered to the participants in this study. Item 1 required the participants to identify what the authors of Text B and C agreed on. Item 2 tested the ability to guess the meaning of a phrase from the contexts presented in Text A and C (See Appendix B).

Table 4 shows the participants' overall performance in MET task consisting of 2 items in total. 80 % of the participants accurately responded to both of the questions.

Table 4. The Participants' Performance on MET

Participant No	Item 1	Item 2	Total correct answers	Percentage
P01	1	1	2	100%
P02	1	0	1	50%
P03	1	1	2	100%
P04	1	1	2	100%
P05	1	1	2	100%
P06	1	1	2	100%
P07	1	1	2	100%
P08	0	0	0	0%
P09	1	1	2	100%
P10	1	1	2	100%

Item 1

Item 1 requires candidates to find the point given in the options that explains the issue two authors agree on based on the information from two texts (See Appendix B).

Figure 8 shows the participants' overall strategy use when responding to MET Item 1. The findings demonstrate that majority of the strategies reported were careful reading strategies (44%, n=18). The second most commonly reported strategy belonged to other strategies category (22%, n=9), which was followed by expeditious reading strategies (20%, n=8). In addition, multiple texts reading strategies was reported 10% of the time (n=4). Figure 9 presents the distribution of reading strategies in detail.

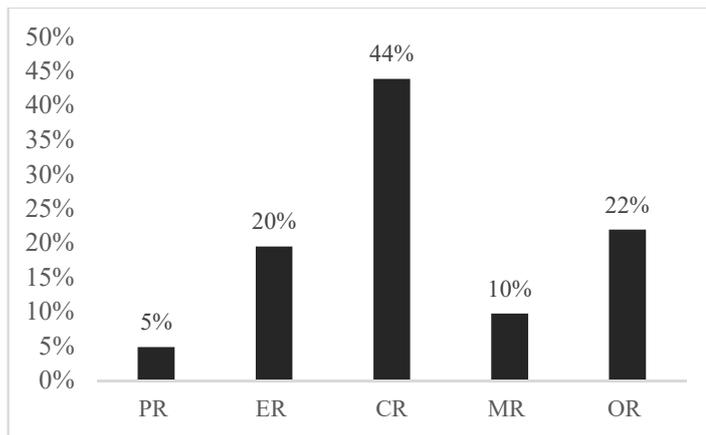


Figure 8. MET Item 1 overall strategy use

Figure 9 shows that of all the careful reading strategies, CR6, “reading the texts linearly from the beginning to the end carefully” (12%, n=5), was the most common, followed by CR5, “reading the text linearly from beginning to the end carefully” (10%, n=4). Among the expeditious reading strategies, the most commonly reported strategy was ER3, “trying to understand the information in the text quickly (by using the title, subtitles, section headings, first and last sentences) through skimming” (7%, n=3). It was followed by ER1, “rapidly looking for/matching figures, dates, names, specific words, etc. in the text” (5%, n=2) and ER5, “searching for/identifying key words/ideas in the text related to the question” (5%, n=2). From the multiple texts reading category, MR8, “Comparing the gists of different texts”, (7%, n=3) and MR2, “identifying claims that agree, disagree and complement one another in different texts”, (2%, n=1) are the only strategies reported. Besides, from the other strategies category, OR2, “using background knowledge to support understanding / guess or interpret meaning”, was also used (10%, n=4).

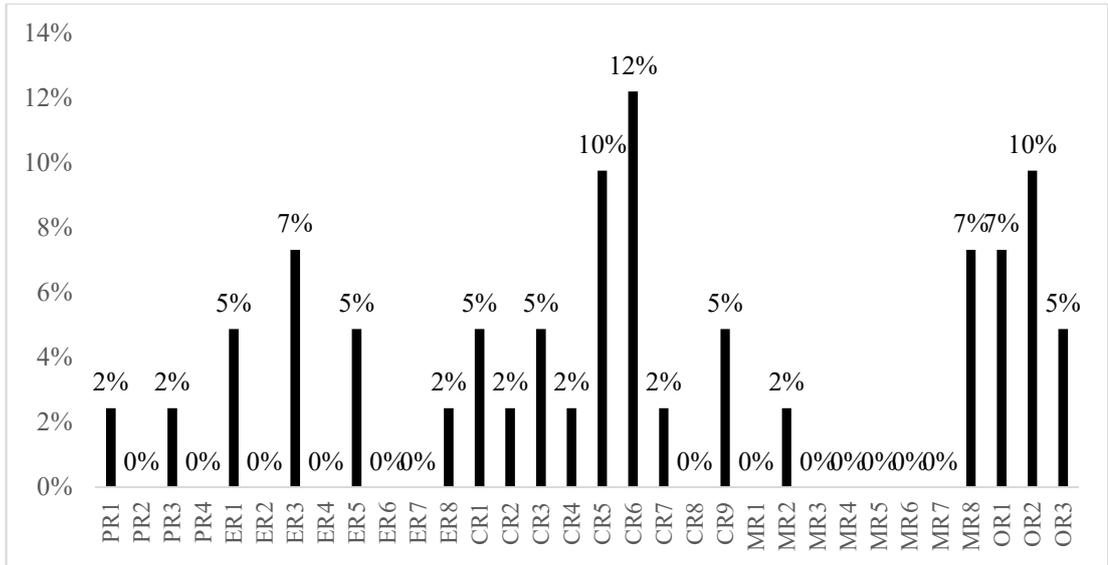


Figure 9. MET Item 1 reading operations

Item 2

In the task used in the experiment, Item 2 tests guessing meaning from context, which is provided through two texts (See Appendix B). Figure 10 below presents the overall strategy use for this item.

It is clear from Figure 10 that the most commonly reported strategies are careful reading strategies (63%, n=20) followed by expeditious reading strategies (34%, n=11). No pre reading and multiple texts reading strategies were reported.

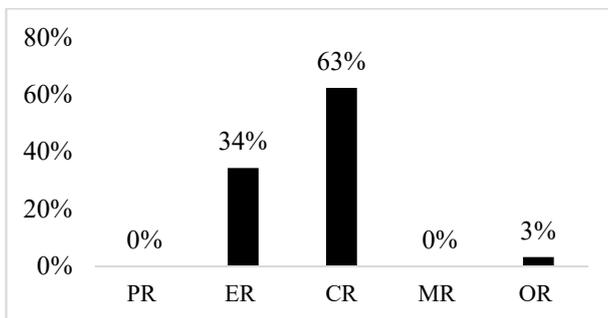


Figure 10. MET Item 2 overall strategy use

Figure 11 shows the detailed analysis of the strategies operationalized in these exams. As for careful reading strategies, CR3, “reading carefully across sentences (to

establish the connections of ideas between sentences or parts of the text by identifying relationships such cause and effect, claim and supports etc.)” and CR8, “rereading the difficult and/ or relevant parts of the text” (16%, n= 5) are the most common strategies reported, and CR2, “focusing on one sentence (and/or its parts) to understand it clearly” (13%, n=4) follows them. ER5, (19%, n= 6), “searching for/identifying key words/ideas in the text related to the question, is the most common expeditious reading strategy followed by ER1, “rapidly looking for/matching figures, dates, names, specific words, etc. in the text” (16%, n= 5).

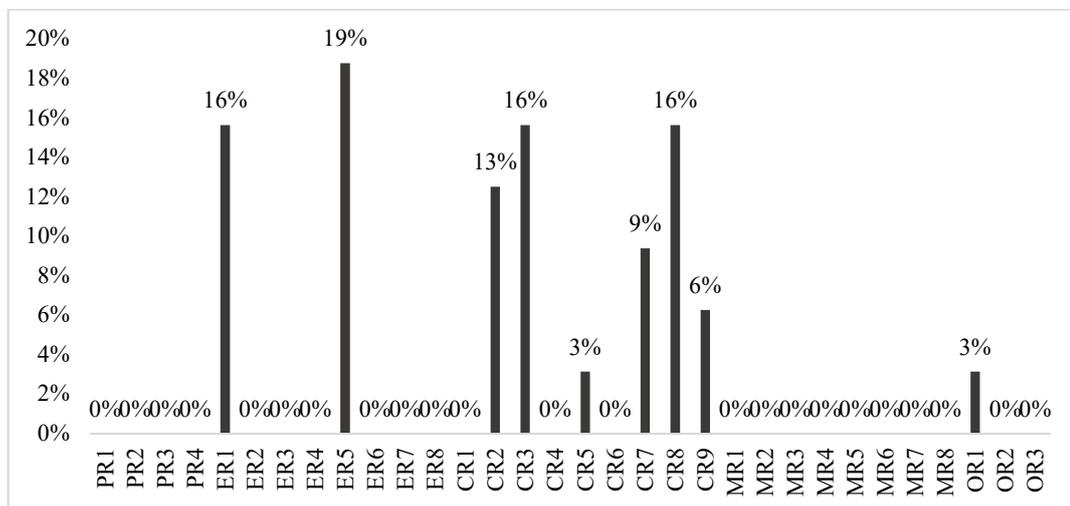


Figure 11. MET Item 2 reading operations

4.3.2.1 The task characteristics of MET based on the participants’ comments

Item 1 required the participants to identify what options two authors agree on based on the information in two different texts (B and C), and the accurate response was chosen by the 90% of the participants (n=9). Again, their thinking processes reveal a few points to consider in terms of the design of the exam tasks. The extracts from the participants’ verbal protocols are introduced below.

I read all the texts in order. I eliminated option A because I thought it will not say anything bad about Copy Pro (considering the information looks like an

advertisement). I eliminated option B because again, when introducing a product, it should not say that it should not be used in draft-mode. Then, I had two options, C and D. Option D is mentioned in Text A. I thought D is a better option, so I chose D. (Participant 5)

I looked at the options, and I thought “home and office” in option D is easier to scan in the texts, so I scanned the texts and find the options. There was no need to consider the other options. (Participant 4)

Item 2 required the participants to guess the meaning of a phrase based on the contexts presented in two different texts. 80% of the participants accurately responded to this question. The students who chose the wrong option did so not because they could not guess the meaning of the phrase, but because they did not know the equivalent word in the correct option. One sample extract from their verbal protocol is presented below.

The ink didn't run, it stayed there. It doesn't change color. Maybe it (the answer) is “b” or “c” (options). Just by luck, I chose b (the correct answer) (Researcher asked: what do you think the meaning of “smudge proof” is? What did you understand?). It is permanent, it doesn't go away or it will not be damaged. (Participant 10)

4.3.3 Reading strategy use in ECCE

Two items from ECCE purportedly measuring multiple texts reading skill were administered to the participants. Item 1 requires the participants to identify how text C differs from the other texts. Item 2 necessitates the participants to find out what all the four texts imply.

Table 5 illustrates the correct and incorrect responses of the participants', and the total score each participant received. It is seen that all the participants responded to Item 2 accurately whereas Item 1 was accurately responded by 60% of the participants (n=6). Strategies they employed while responding to these items are demonstrated in Table 5.

Table 5. The Participants' Performance in ECCE

Participant No	Item 1	Item 2	Total correct answers	Percentage
P01	1	1	2	100%
P02	0	1	1	50%
P03	0	1	1	50%
P04	1	1	2	100%
P05	1	1	2	100%
P06	0	1	1	50%
P07	1	1	2	100%
P08	1	1	2	100%
P09	1	1	2	100%
P10	0	1	1	50%

Item 1

Item 1 asks readers to identify how Text C differs from the other four texts.

Strategies the participants reported after the completion of the task is presented in Figure 12.

Of a total of 27 strategies reported for this item, careful reading strategies were the most commonly employed strategy type (44%, n=12). Careful reading strategies were followed by expeditious reading strategies (26%, n=7) and multiple texts reading strategies (26%, n=7).

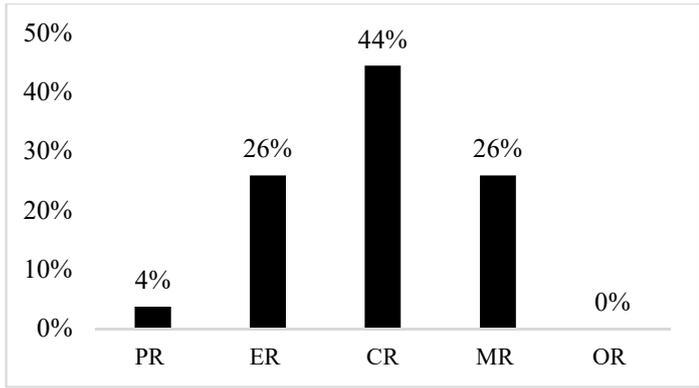


Figure 12. ECCE Item 1 overall strategy use

As it can be seen in Figure 13, the analysis of the operationalized strategies reveals that CR6, “reading the texts linearly from beginning to the end carefully”, (26%, n=7) and CR5, “reading the text linearly from beginning to the end carefully”, (15%, n= 4) are the mostly employed ones. As for expeditious reading, ER6, “searching for/identifying key words/ideas in the texts related to the question”, (11%, n= 3) and ER5, “searching for/identifying key words/ideas in the text related to the question”, (7%, n=3) were the most common expeditious reading strategies. Furthermore, MR8, “comparing the gists of different texts” was the only multiple texts reading strategy reported (26%, n=7).

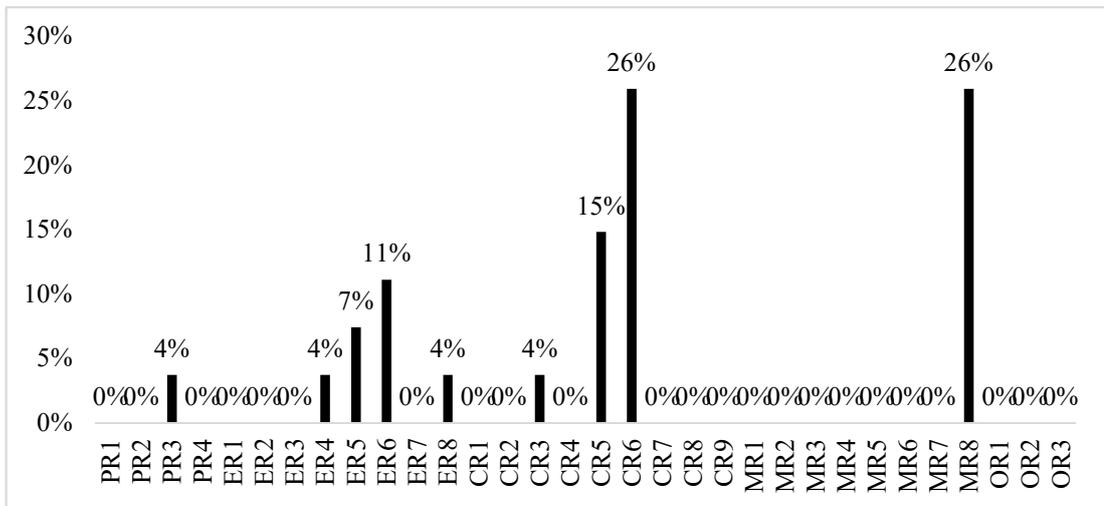


Figure 13. ECCE Item 1 reading operations

Item 2

Item 2 requires the participants to identify what all the four texts imply (See Appendix C). As mentioned above, all the items were responded accurately by the participants. Details regarding the overall strategy use and individual strategy use proportion are presented respectively in Figure 14 and Figure 15.

For this item, 30 strategies were reported by the participants. Figure 14 shows that Item 2 yielded mostly expeditious reading strategies (70%, n=21). For the second question other strategies (17%, n=5) is reported as the second common, which was followed by careful reading strategies (13%, n=4). No pre reading and multiple texts reading strategies were reported.

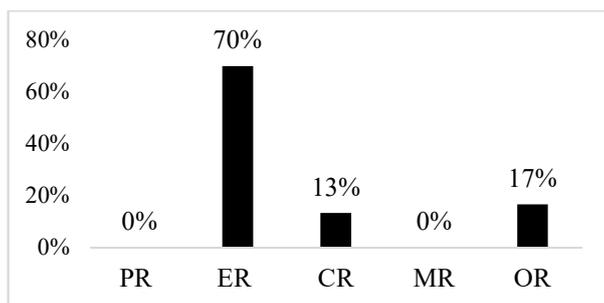


Figure 14. ECCE Item 2 overall strategy use

Figure 15 shows the individual strategies triggered by Item 2 as reported by the participants. ER6, “searching for/identifying key words/ideas in the texts related to the question”, is the most common strategy (53%, n=16), and followed by ER5, “searching for/identifying key words/ideas in the text related to the question”, (n=3). OR1, “using knowledge of the text: Notes the discourse structure of the text (cause/effect, compare/contrast, etc.)”, is reported 10% of the time as well (10%, n=3). Among the careful reading strategies, CR6, “reading the texts linearly from beginning to the end carefully”, was also reported (7%, n=2).

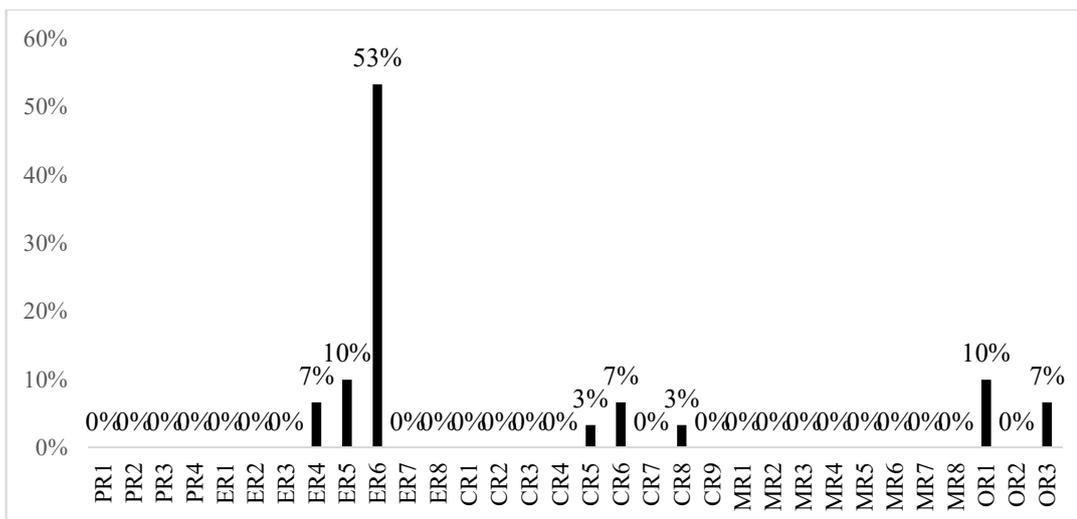


Figure 15. ECCE Item 2 reading operations

4.3.3.1 Task characteristics based on the participants' comments

ECCE, Item 1, required the participants to identify how one text (Text C, See Appendix C) differs from the other four texts. 40% of the participants could not find the correct answer even though they could comprehend the text which was evident based on their verbal accounts. The reason may be because the participants did not chose the correct answer, which was about language learning because, they believed, only a small proportion of Text C is about language learning. Therefore, they concluded that the answer cannot be related to language learning. In other words, this item did not test comprehension at the textual level. However, the participants expected the answer not to lay in the details, but in the text as a whole. The extracts from their verbal accounts is presented below.

For the first question, nearly all the options seemed plausible. But when I look at the C (Text), it was about language...right ..but it is not just about learning language. That's why it is not B (option related to language learning). It also talks about specific cooking techniques and tips, but it is not only about that as well. It (Text C) offers options when you are travelling and

cooking. In option D, it says it discusses traveling. Yes, but it is not just about travelling. So I chose option A (It is written for a specific audience).

(Participant 3)

Text C is the only paragraph talking about learning Spanish. There, most of the talk was about cooking tips and stuff. That's why, I changed my answer to option A (from option C, the correct answer). (Participant 10)

ECCE, Item 2, which requires test takers to find the option that is implied by all the four texts was accurately responded by all the participants in this study. When the participants were trying to find the answer, the way they eliminated the other options was worth taking into consideration because it shows that the participants could reach the answer through matching the key words in the options, rather than directly identifying what is implied by all the texts. Here, the participants employed a test-taking strategy, and the answer could be reached without having the necessary abilities, which indicates that the quality and the discriminating power of the options may impact on the validity of the exam.

Before going back to the texts, I read the options because I had already read the texts, and I wanted to see if I could answer this question with what I remembered. I eliminated option D because Basque food is not mentioned in all the texts. It is just mentioned in C and D, and it was not a common point in all the texts. Another option mentions ethnic food. But ethnic food was not mentioned in all the texts. I eliminated that too. Also, not all the texts mention cooking, just Text B, so I chose option B which is about international food.

(Participant 7)

Actually, I looked at the options. I pondered on this question a little bit. I proceeded by eliminating the options. I eliminated option A because it does not have anything in common with Text C. Then, I looked at option D. The word, “Basque”, was mentioned in only D, so I eliminated that. The other option... ethnic restaurants are expanding their businesses... There is not much information about this. I directly chose option B. (Participant 9)

I eliminated option D because, in Texts A and B, there was no mention of Basque. Option C was not mentioned in Text C. I was in between Option A and B. I was thinking B must be the correct answer. So, in all the texts, I searched for it. (Participant 1)

In this section of the chapter, the results regarding each exam task and the reading operations they have triggered are presented. The insights these results provide will be discussed in the Chapter 5 after the presentation of the eye movement data in this chapter.

4.3.4 The results of the eye movement data

The eye movement records of the 10 participants were analyzed using descriptive statistics. Fixation durations, fixation counts, the presence and proportion of careful reading on AOIs were calculated using descriptive statistics for each task and participant. In addition, the sequence of the eye movements of two participants as reading encloses is also analyzed. The findings are presented below.

4.3.4.1 Overall fixation time, fixation count and careful reading proportion in each exam task

Table 6 demonstrates the average fixation count, the average fixation time and the average proportion of careful reading in each task. Since ISE II was presented to the participants in two consecutive pages, the results are presented separately ISE II (1) includes Task 1 whereas ISE II (2) includes Task 2 and 3.

Based on the figures, ISE II (2) attracted the highest number of fixations (M=627.6), and it was followed by ISE II (1) (M=279.6). The items with the lowest number of fixations was ECCE (M=208.9). There seems to be a correlation between total fixation count means and total fixation duration means of the tasks except MET. Although the mean fixation count of MET (M=261.7) was lower than ISE II (1) (M=279.6), the average total fixation duration of MET (M=48068) was higher compared to ISE II (1) (M=37350). The lowest total fixation duration was observed in ECCE (M=34255). When the careful reading proportions are examined, on average, ISE II (2) led readers to do more careful reading compared with other tasks (74%). The tasks that necessitated the least amount of careful reading was ECCE (48%), which is in line with total fixation count (M=208.9) and total fixation duration of this task (M=34255).

Table 6. Average Fixation Count, Fixation Duration, and Careful Reading in ISE, MET, and ECCE

Exams	Total Fixation Count (Mean)	Total Fixation Duration (ms)(Mean)	Careful Reading (Percentage)
ISE II (1)	279.6	37350	63%
ISE II (2)	627.6	93135	74%
MET	261.7	48068	68%
ECCE	208.9	34255	48%

Considering the purpose of this study, it is of great importance to investigate how much of each text in each exam task is processed, or to put it in other words,

carefully read because careful reading is considered to be a key component of multiple texts reading comprehension. Therefore, in the next section, detailed analysis of each exam will be elucidated.

ISE II (1)

ISE II (1) included the Task 1 of ISE II, where the participants were required to match questions to a text that bears the information to answer those questions.

Should this task to be completed using multiple texts reading skill, it is necessary that a considerable proportion of each text is carefully read. Consequently, the total fixation count and fixation duration in each text must be relatively high. Table 7 below shows the total fixation count, fixation duration and careful reading proportion presented for each text in ISE II Task I. In addition, the cognitive processes of the two participants will be scrutinized in detail.

Text A attracted the highest number of fixations ($M=70.9$), and the total fixation time is the highest of all the four texts ($M=11315$). It seems that cognitive processing of Texts B, C, and D was very similar as the average total fixation count and fixation time as well as careful reading proportion of all are close to one another. It is also seen that except Text A almost 40% of the texts are not carefully read.

Table 7. Fixation Count, Fixation Duration, and Careful Reading in ISE II (1)

Texts in ISE II	Total Fixation Count (n=10)	Total Fixation Time (ms) (n=10)	Careful reading proportion (n=10)
Text A	70.9	11315	78%
Text B	36.5	5385	58%
Text C	40.3	5330	58%
Text D	48.5	6460	54%

ISE II (2)

The data here were obtained for ISE II Task 2 and Task 3. Task 2 requires test takers to identify five statements that are true out of eight based on the information in four texts. Task 3 is a summary completion task with five gap fill items. Table 8 shows the part of the data including the participants' average fixation count and duration as well as the proportion of careful reading they employed for the same four texts in ISE II (1) above, but for this time for Task 2 and Task 3.

For the Task 2 and Task 3, it is seen that the participants fixated on Text A the most ($M=91.2$), the total fixation time for Text A is the highest among the four ($M=17735$). As in Task 1, this text yielded the highest proportion of careful reading. On the other hand, the participants had the least number of fixations ($M=50$) on Text B, which is supported by visuals, and where the information is presented in phrases rather than sentences. In addition, total fixation time for Text C was the lowest ($M=6625$).

Table 8. Fixation Count, Fixation Duration, and Careful Reading in ISE II (2)

Texts	Total Fixation Count	Total Fixation Time	Careful Reading Proportion
Text A	91.2	17735	81%
Text B	50	7685	63%
Text C	52.6	6625	64%
Text D	82.8	11080	67%

MET

Two items from MET measuring multiple texts reading skill by definition were administered. Item 1 required the participants to identify what the authors of Text B and C agree on, and Item 2 required the participants to guess the meaning of a phrase based on the contexts in Text A and Text C. Below are the details of the eye movement data.

Table 9 demonstrates that Text C was fixated the most ($M=102.7$), 78% of the text was read carefully. In contrast, Text A was fixated the least ($M=48.5$), and 49% of the Text A was read carefully.

Table 9. Fixation Count, Fixation Duration, and Careful Reading in MET

Texts	Total Fixation Count	Total Fixation Time	Careful Reading Proportion
Text A	48.5	10543.329	49%
Text B	54.2	10338.345	59%
Text C	102.7	17040.008	78%

ECCE

Two items purportedly measuring multiple texts reading comprehension in ECCE was given to the participants. Item 1 required the participants to identify how Text C was different from the other texts, and Item 2 required the participants to find out what all the texts implied. Below are the details of the eye movement data.

Table 10 demonstrates that there is almost an equal distribution of the fixation counts across all the texts. The same case is valid for the total fixation durations except Text A, which was fixated a little longer ($M=8195$). Besides, 67% of the Text A is carefully read, which is higher than the other texts. The proportion of careful reading was similar in Text B, C, and D.

Table 10. Fixation Count, Fixation Duration, and Careful Reading in ECCE

Texts	Total Fixation Count	Total Fixation Time	Careful Reading Proportion
Text A	40.6	8195	67%
Text B	30	5305	49%
Text C	42.1	4562.5	51%
Text D	39.1	6600	50%

Apart from the cognitive processing details of each individual text in each task, it is imperative to investigate switches from the questions to the texts and

between the texts for the purposes of this study because it is important to explicate which texts were processed with which order when responding to the questions in each task. Therefore, in the next section two participant cases will be presented. Participant 7 and Participant 10 were chosen because based on the analysis of the verbal protocol and of the eye movement data, Participant 7 was observed to do more expeditious reading while Participant 10 did more careful reading.

4.3.5 Cognitive processing data considering Participant 7 and Participant 10 in ISE II, MET, and ECCE.

ISE II (1)

In this section, only data from ISE II Task I is presented. ISE Task I included five questions and four texts, and required the participants to match the questions to the text where the answer lies. Table 11 shows the fixation time and duration, and the amount of careful reading of Participant 7 and Table 12 does so for Participant 10.

The findings indicate that overall, Participant 10 (60%) was engaged in more careful reading than Participant 7 (41%). There is also a significant difference between overall the total fixation count of Participant 7 (169) and Participant 10 (418), which is also reflected in total fixation duration. 89% of Text A was carefully read by both of the participants. However, while Participant 7 fixated 52 times with a duration of 6850 ms, Participant 10 fixated 117 times with a total duration of 24300 ms. Besides, it is conceivable to conclude that Participant 7 read Text B, Text C, and Text D relatively expeditiously considering the carefully read proportion of the texts. The fact that fixation counts for these texts are rather low also supports this.

Table 11. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in ISE II (I)

	Total Fixation Count	Total Fixation Duration	Careful Reading Proportion
Overall processing	169	17550	41%
TEXT A	52	6850	89%
TEXT B	12	950	25%
TEXT C	20	1950	33%
TEXT D	25	2050	31%

Table 12. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in ISE II (I)

	Total Fixation Count	Total Fixation Duration	Careful Reading Proportion
Overall processing	418	76800	60%
TEXT A	117	24300	89%
TEXT B	44	10450	63%
TEXT C	65	10000	89%
TEXT D	80	15750	56%

In addition, the sequence of the participants' eye movements were analyzed. It was observed that there were a few differences between Participant 7 and Participant 10 (See Appendix J).

Participant 7 started with reading the instructions and the first three questions despite these visits' being short. Then, the participant fixated on Text A and Text C, Item I and Text A and C again, where s/he seeks the answer to Item I. Her verbal protocol also supports this. After Item 2, the participant fixated on Text D twice, and then moved on to Item 3. After Item 3, she switched back and forth between Texts A, B, and D, and the item. The participant fixated back on Item 2 and Text D. After she fixated on Item 4, she only fixated on Text B. For Item 5, she fixated mostly on Text A, and once on Text D (See Appendix J). However, it could be concluded that the participant was reading rather expeditiously based on that the fixation count for each text was rather low and is not adequate for these texts to be read carefully.

Participant 10 preferred skimming the texts, which could be concluded from the fact that he made short visits to Text A, Text B and D, then Item 1 and Item 2. This pattern was followed during the beginning. The participant paid short visits to the questions and texts. This type of reading behavior indicates that he was trying to locate the relevant information by search reading. Then, it is observed that the participant read each text carefully with occasional switches to the items. For instance, after reading Item 4, the participant switched back and forth between Text A and Text B. Then, after s/he fixated on Item 3, Text A was fixated around 30 times. This shows that this participant was reading the text linearly and incrementally. The same pattern was observed for the rest of the task (See Appendix J). The participant tried to identify the text that included the information to a specific question and then read that text carefully linearly.

It is plausible to conclude that ISE II Task 1 required both participants to do search reading. Participant 7 did not read the texts from the beginning to the end linearly while Participant 10 did so to some extent. Considering the performances of the participants for this task (Participant 7 80%, Participant 10 80%), it is reasonable to conclude that ISE II Task I could be completed through expeditious reading paired with proportional careful reading, and that there is no need to read the texts from the beginning to the end to form micro and macro structures of the texts.

ISE II (2)

This section included the eye movement data of both Task 2 and Task 3. Task 2 required the participants identify five statements that are true out of eight based on the information from all the texts. Task 3 is a summary completion task with gap fill

items. The details of the eye movement records of Participant 7 and Participant 10 are presented respectively in Table 13 and Table 14 for these tasks.

Overall, Participant 10 carefully read 88% of the texts and the questions as well as the instructions whereas Participant 7 read 83% carefully. The total fixation duration of Participant 7 and Participant 10 for these tasks were 60950 and 147050 respectively. It could be concluded that Participant 10 spend more than twice as much time on this task. Furthermore, the difference between the two participants on the processing of Text A is significant. Although Participant 7 fixated only 17 times, 89% of the text is carefully read. However, Participant 10 fixated 135 times on the same text, and 78% of the text is carefully read.

Table 13. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in ISE II (II)

	Total Fixation Count	Total Fixation	
		Duration	Careful Reading Proportion
Overall processing	530	60950	83%
TEXT A	17	15000	89%
TEXT B	32	4550	88%
TEXT C	51	4700	67%
TEXT D	54	7400	69%

Table 14. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in ISE II (II)

	Total Fixation Count	Total Fixation	
		Duration	Careful Reading Proportion
Overall processing	722	147050	88%
TEXT A	135	32850	78%
TEXT B	49	8000	88%
TEXT C	91	19750	89%
TEXT D	94	10900	81%

The sequence and location of fixations also gives us valuable information regarding the cognitive processes readers go through. Participant 7 and Participant 10

both started with reading the instructions and questions, which is expected as they were already familiar with texts (See Appendix K).

Participant 7 was observed to have aligned fixations on Text A after fixating on Item 2.1. Then, Items 2.1, 2.7, and 2.3 followed by Text A were fixated on. All the items except 2.1 and 2.2 were fixated, and again followed by fixations on Text A. Probably, the participant was trying to identify which item she could answer based on Text A. There are several instances of fixating on the items and going back to the texts during this task, which may be because there are eight items. The number of aligned fixations were low, which indicates that the participant was engaged in expeditious reading followed by linear reading. For Task 3, the participant started with the items. First, there was aligned fixations on Item 3.1 with one fixation on 3.2, which was followed by aligned fixations on Text A. This pattern was followed again. After aligned fixations on 3.2, the participant fixated on Texts A and C. This was also repeated. After aligned fixations on 3.3, the participant is observed to have fixated on very briefly on Text A and C. For Item 3.4, all the texts were fixated with the following order: Text C, D, C, D, A, B, A, B, A. For Item 3.5, all the texts were fixated again, however, Text D and C attracted more fixations.

Participant 10 started with reading all the questions with occasional brief switches to texts. After fixating on 2.4, the participant is observed to have a brief visit to Text A, and then have aligned fixations on Text C, which includes the information to answer that question. He fixated on items 2.4, 2.5, and 2.6 a few times with occasional brief switches to Text A and B. The pattern Participant 10 follows indicate that while responding to the items except 2.4, there are not many aligned fixations on texts, only a few fixations across two or three texts. For this reason, there are many brief switches between texts and items. This is probably due to the

fact that the texts had already been processed for the previous task, only locating the relevant information for confirmation of the response would suffice. Then, the participant fixated on 2.4, which was followed by aligned fixations on Text A, which is also the text including the relevant information. When the eye movements of Participant 7 are examined for Task 3, it is clearly seen that this task engaged Participant 10 in more linear reading compared to Task 2 because there are more aligned fixations overall. This shows that for this task, just through search reading, the participant could not easily find the answer or could not answer the questions with the information s/he gathered thus far. After 3.1, the participant fixated on Texts A and D. After 3.2, the participant fixated on Texts D, A, B, C. Item 3.3 yielded aligned fixations on Text A, and C, with brief fixations on Text D and B, followed by aligned fixations on Text D. A similar pattern was followed for the rest of the items.

Task 2 engaged the participants in search reading as there was a high number of items. However, Task 3 required more careful reading which is reflected as more aligned fixations after Task 3 questions for both the participants.

MET

Two items from MET was used in this study. Item 1 required the participants to find out what the authors of Text B and Text C agree on (See Appendix B). The analysis of the eye movement records of Participant 7 and Participant 10 are presented in Table 15 and Table 16.

Table 15 shows that Participant 7 read 37% of the task carefully, while in Table 16, Participant 10 is seen to have read 63% of the task. The total fixation count and total fixation time is in line with this finding for both participants. There is a

positive correlation between total fixation count and time, and careful reading proportion. One significant difference between Participant 7 and 23 is that Participant 7 fixated only 20 times on Text B while Participant 10 did so 92 times.

Table 15. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in MET

	Total Fixation Count	Total fixation Duration	Careful Reading Proportion
Overall processing	130	15150	37%
TEXT A	34	5300	45%
TEXT B	20	1700	25%
TEXT C	32	3100	27%

Table 16. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in MET

	Total fixation Count	Total fixation Duration	Careful Reading Proportion
Overall processing	401	88700	63%
TEXT A	63	20850	64%
TEXT B	92	20500	67%
TEXT C	84	12800	47%

As far as the sequence of reading is concerned (See Appendix L), Participant 7 started by skimming the texts, which is indicated by the brief fixations on each text. Then, the participant fixated on Item 2, which requires guessing the meaning of a phrase from context. This fixation on this item was the only fixation, and then the participant read certain parts of Text C and Text A linearly with occasional fixations on Text B. It is also seen that this item was not challenging for the participant because s/he could respond to it with few fixations. Item 1, on the other hand, attracted more fixations. While responding to Item 1, after fixating on the question, aligned fixations on Text B, the item, and Text C could be seen. There were very brief switches to Text A, which is expected considering that the item requires the comparison of the ideas of the authors of Text B and Text C.

Participant 10 started by reading both the questions (First item 2, then Item 1), and a high number of aligned fixation on Text A is observed with occasional switches to Text C. After fixating on Item 2 again, Text B is highly fixated with occasional switches to Text C. It is sensible to conclude here that Text C is not processed carefully from the beginning to the end (See Appendix L).

To complete this task, it was necessary for both the participants to consult all the three texts while responding to both the questions. It was observed that while responding to Item 1, Texts B and C attracted more aligned fixations by both participants. For Item 1, both participants had more aligned fixations on Texts A and C, but it is seen that there is no need to process them in the same depth. Item 2, was not very challenging for Participant 7, and did not require many fixations, which could result from the fact that this vocabulary item was known by the participant, and s/he did not feel the need to consult the context to guess the meaning of it.

ECCE

Two items from ECCE were used in this study. Item 1 required the participants to identify how Text C differed from the others, while Item 2 required the participants to find out what all texts imply (See Appendix C). Below are the results of the eye movement data for Participant 7 and 23 (See Appendix M).

Table 18 shows that overall, Participant 10 was engaged in careful reading more. The general tendency of Participant 7 was to carefully read only half the texts except Text A (See Table 17). In addition, when we look at Text B, only 40 % of it was carefully read by both the participants. Furthermore, the findings suggest that Participant 10 read 86% of Text A and 78% of Text C, which was followed by 77% of Text D.

Table 17. Participant 7- Fixation Count, Fixation Duration, and Careful Reading in ECCE

	Total Fixation Count	Total Fixation Duration	Careful Reading Proportion
Overall processing	179	18600	45%
Text A	39	5150	57%
Text B	23	1450	40%
Text C	35	2250	44%
Text D	22	2950	38%

Table 18. Participant 10- Fixation Count, Fixation Duration, and Careful Reading in ECCE

	Total Fixation Count	Total Fixation Duration	Careful Reading Proportion
Overall processing	325	81550	77%
Text A	65	26100	86%
Text B	26	10600	40%
Text C	87	11200	78%
Text D	52	12700	77%

The sequence of reading while completing the task provides insightful information. When completing the ECCE task, Participant 7 first started with skimming the texts, which is indicated by few fixations on each text. It is also seen that the participant fixated on Texts B and D a little more. Then, s/he fixated on both the questions. After fixating on Item 2, the participant fixated once on Texts A, B and C. Then, she switched back and forth between Texts A and C with very few fixations on Texts B and D.

Participant 10 started by skimming the texts and the task. Then, after a long fixation on Item 2, the participant started reading the texts linearly in the order of Texts A, B, and D. Following that, the participant fixated on Item 1, and fixated on Text C linearly with occasional switches to Text A. This was followed by consecutive fixations on Item 1. Then, the participant was observed to have brief fixations on each text with occasional switches to the item. This was followed by aligned fixations on Text C and Item 1. Aligned fixations on Item 2 followed, but

afterwards only Text B was fixated twice, which could be an indicator that the participant accumulated the necessary information while responding to Item 1.

Two items from ECCE required the participants to consult all the texts with varying levels of depth to complete the task. Participant 7 was observed to fixate more on Texts A and C, and fixations on Text C is rather low for both participants. Participant 10 observed to have aligned fixations in all the texts, which is an indicator that he did more careful reading. It is reasonable to conclude that for the completion of this task, processing all the texts at a level to form macro structures of these texts was not necessary for participant 7.

4.4 Conclusion to the results section

The results indicate that based on the test specifications, ISE II does not attempt to operationalize multiple texts skill, MET specifications are not clear enough to make judgements on whether it attempts to operationalize multiple texts reading skill, and ECCE seems to attempt to operationalize multiple texts reading skill depending on the fact that test writers make a decision in favor of using multiple texts.

When the actual operations triggered by these exams are investigated through process based data, it is seen that ISE does not operationalize multiple texts reading skill, and the design of the task poses unnecessary challenge on the part of the readers. MET and ECCE found to operationalize multiple texts reading skill to some extent. The results are discussed in Chapter 5.

CHAPTER 5

DISCUSSION

5.1 Introduction

As Khalifa and Weir (2009) suggest reading is a multifaceted ability comprising of lower and higher level abilities. As far as the multiple texts (intertextual) reading skills are concerned, they are labelled as the highest level in the hierarchy because multiple texts reading comprehension requires establishing connections between different texts after the formation of micro and macro structures of individual texts. Multiple texts reading comprehension requires readers to go beyond the information presented in a text to form a stance based on all the information gathered from different documents. Therefore, multiple texts reading skill is essential for students in academic contexts considering the assignments to be completed and the abundance of information available. Then, should a language proficiency exam be taken for educational purposes, it must adequately sample from the target language use domain to be considered valid because a test must provide accurate information to different stakeholders regarding what a candidate can and cannot do. In the case of an academic context, this target language use domain comprises both higher and lower level reading abilities. Thus, this study set out to explore whether multiple texts reading tasks in present international language proficiency exams are cognitively valid by investigating firstly their designation in test specifications and then the processes test takers go through while completing these tasks.

The data were collected through two different means: retrospective think aloud and eye tracking method. The analysis of the data produced valuable information

about the validity of these tasks. The results will be discussed below for each research question separately.

5.2 RQ1: Do multiple texts reading (MTR) tasks used in language proficiency tests attempt to operationalize MTR skill and subskills as defined in theory representatively?

When compared with the multiple texts reading skill definitions in the literature, it was observed that ISE II does not representatively operationalize multiple texts reading skill. When the specifications of ISE II are examined, it is seen that only Task 3 attempts to operationalize multiple texts reading skill. However, a closer analysis reveals that this task attempts to operationalize “the ability to understand specific, factual information at the word and/or phrase level across the texts”. Therefore, we can understand that this is a local level careful reading task. However, considering that multiple texts reading skill require going beyond the information in a text, and necessitates the formation of a mental representation of the each situation being described in each text with a layer of the representation of the connections between these texts (Perfetti et al., 1999). Therefore, as Goldman et al. (2013) suggest, a multiple texts reading task must operationalize analysis, synthesis and integration. Therefore, it could be concluded that ISE II does not attempt to operationalize multiple texts reading skill representatively. Tasks aiming to operationalize multiple texts reading skill must not aim to operationalize reading comprehension at lower cognitive levels because multiple texts reading tasks require a global level comprehension of information firstly, in a single, then in multiple texts.

MET specifications make reference to multiple texts reading skill in very generic terms. The ability is listed in the specifications as “making connection across texts”. This ability could be considered to be a part of analysis, integration, and integration component of the assessment model of Goldman et al. (2013). Depending on the way this operation is interpreted, this task could well be attempting to operationalize multiple texts reading skill representatively. However, it is necessary to define what sorts of connections are expected.

ECCE makes reference to multiple texts reading skill in all the listed operations in the specifications. However, all the abilities are defined in way that the interpretations of the operations will determine whether the task will operationalize these skills as multiple texts reading skill or not since all the abilities are added “from one or more texts”. When these operations are considered as multiple texts reading operations, it is seen that ECCE attempts to operationalize a range of multiple texts reading skill as source evaluation, analysis, synthesis and integration by definition.

Construct validity is concerned with the extent of interpretations that can be made from the operationalizations specified in a test to the theoretical constructs where these operationalizations are derived, which inherently suggest that a test must base the operationalizations it aims to assess on the theory (Bachman & Palmer 1996). Mesick (1989) places construct validity in the center of the validation process. In addition, in his validity framework, he emphasizes value implications. Value implications are concerned with how different stakeholders, test writers and users, define the construct. This construct definition must still be rooted in theory, but based on the purpose and the use of a test. Certain aspects of a construct could be emphasized more or less. The way different stakeholders define the construct also affects the interpretations and inferences made based on the test scores. When we

turn to the theory of reading, Khalifa and Weir's (2009) reading model depicts reading as a construct comprising of various levels starting with word recognition, lexical access to textual and intertextual representations. Ünalı (2010) reaches the conclusion that reading tests in academic contexts must measure comprehension at sentential, textual, and intertextual levels after careful examination of the reading behaviors of students at a British university. Considering that the exams used in this study are aimed at young adults and are taken for educational purposes, it is necessary that these exams operationalize reading skills at all levels. In addition, as a common practice in testing field, these definitions of how a construct will be operationalized are included in the test specifications, which are used by different stakeholders such as test writers, test users and test takers. Test writers make use of test specifications to devise new exam tasks, test users analyze the specifications to assess whether the test fits their purpose and context, and test takers benefit from them to shape their studies. Therefore, it is highly necessary that test specifications be very clear and define operationalizations as observable behaviors, and there should not be any room for different interpretations. However, although MET and ECCE attempt to operationalize multiple texts reading skill, MET and ECCE specifications are formulated in a very ambiguous way, and whether multiple texts reading skill will be operationalized is based on interpretation. In addition, what is expected of test takers as observable behaviors is lacking, making it difficult to make interpretations and inferences based on the exam specifications regarding what test takers can and cannot do. The underlying reason behind this problem is that the operationalizations in the specifications are not based on the operational definitions of the construct. When it comes to ISE II, it attempts to measure a global level careful reading ability, summarization, through local level careful reading. It is clear

that summarization requires the formation of a mental representation of the information presented in a text, and this ability is placed on the higher end of the reading model by Khalifa and Weir (2009). In addition, as Goldman et al., (2013) suggest, a multiple texts reading comprehension task must operationalize the abilities specified under the subcomponents of multiple texts reading comprehension as analysis, synthesis, and integration, and all of these abilities are formulated as observable behaviors in this assessment model. However, the ability to be operationalized in ISE II Task 3, is placed towards the bottom of the model by Weir and Khalifa (2003). Thus, it is not expected that this task operationalizes multiple texts reading skill representatively, and it will definitely suffer from construct underrepresentation. Any interpretations and decisions based on the results of this exam will be misleading as well because it does not attempt to measure multiple texts reading skill as defined in the theory.

In the light of this information, it could be concluded that during the design phase of an exam, to increase the theory based validity, specifications must include specific operations that are necessary in a specific context, and these operations must be rooted in theory. If not, the test may fail to assess what it has set out to assess successfully. In addition, operations in test specifications must be worded clearly without leaving any room for misunderstanding and misguidance. Vague specifications may result in tasks measuring different skills and subskills when used to develop test items by different test writers. In addition, test users might be misinformed of the performance of test takers, and finally test takers may be misguided on what is expected of them in a certain test, which may impact on their performance poorly. Therefore, as can be seen, it is crucial to design test specifications that are clear and that include operations clearly defined in theory.

This is the first stage of the validation process that seriously affect the quality of a test. Still, it is not adequate to just design test specifications including operations that are rooted in theory. It is necessary that test tasks written based on these test specifications be validated through the collection of process based data. The actual operations that are triggered by tasks could only be determined when the test is in use. For that reason, in order to collect validity evidence for the exam tasks in question the second research question provided information on these tasks validity.

5.3 RQ2: Do test takers use substantial MTR skills as defined by theory and as specified in the test specifications in responding to the MTR tasks in tests where such tasks are available?

The discussion of the findings will be presented below for each exam task.

ISE II

All three tasks operationalized mostly expeditious reading strategies followed by careful reading strategies based on the reported strategies of the participants. It is seen that almost 40% of each texts was not carefully read. Interestingly, in Task 1 and Task 2, a few instances of multiple texts reading skill were reported. However, as mentioned earlier, ISE II Task 1 and Task 2 do not attempt to operationalize multiple texts reading skill. This finding suggests that when four texts are presented together, some participants read the fours texts and formed a gist of the texts irrespective of the tasks because some test takers prefer to read all the texts from the beginning to the end just to be on the safe side. In addition, the presence of multiple texts in a task forced the participants to do search reading across texts. That is why; ER6 was the second common expeditious reading strategy. Whether reading

expeditiously across texts should be considered as multiple texts reading skill is a question in concern. It is important to note here that this type of search reading may not be defined differently from search reading in a single text with multiple paragraphs.

For Task 1, the most common careful reading strategy was “reading only the part of the text which seems related to specific questions”. However, ISE II Task 1 attempts to operationalize “the identification of the main idea or the purpose of the text”. For the identification of the main idea or the purpose, it is expected that global level expeditious reading strategies are used for this task, which is not the case for this task. In addition, the verbal accounts of the participants’ support this finding, Items 1.1, 1.3, 1.4, 1.5 could be accurately responded just by key word matching or with the help of the visual. Especially 1.4 asking about the stages in food production was easily matched with the text, which included a diagram with arrows, and where information is presented in phrases not even sentences. This was also evident in strategies reported as ER7 was the second common expeditious reading strategy. A similar case was valid for Item 1.5, which included the key word “last century”. It was easily matched with Text A because it was the only text with numbers and dates. These findings suggest that Task 1 cannot successfully operationalize the ability specified in the test specifications. Furthermore, the eye movement data indicate that among the four texts, a higher proportion of Text A (78%) was carefully read, whereas other texts were carefully read below 60%. In addition, Text B was fixated on the least. The reason behind this finding is that Text A is propositionally denser; consequently required more careful reading, and finding the gist or the main idea just by simply key word matching was not possible. On the other hand, Text B did not include any claims or opinion, but just information in phrases supported by visuals.

Therefore, it was easier to process, and required less fixations. When two individual cases are analyzed, we see that both participants read a large proportion of Text A carefully (89%), which is also supported by aligned fixations on Text A. Unlike Participant 10, Participant 7 carefully read less than 50% of the other texts, which is supported by occasional switches between texts and questions.

Task 2, which attempts to operationalize “the ability to understand specific, factual information at the sentence level” operationalized mostly local careful reading strategies, and expeditious reading strategies. This was expected as the participants had already read the texts for Task 1. Therefore, they located the relevant information through search reading and read the relevant part carefully. When we look at the proportions of each type of strategy, careful reading strategies were reported 51 % and expeditious reading strategies were reported 48%. The detailed analysis of the reading behavior of the two participants based on the eye movement data also showed that in Task 2, there were occasional switches between questions and texts followed by aligned fixations, meaning that the participants completed the task through careful and search reading. All in all, this task was observed to operationalize the ability it aims to operationalize because the results show that the participants mostly did careful reading paired with search reading as the task only aims to test comprehension at the sentential level.

When it comes to Task 3, this was the most challenging task for the participants. Despite the fact that their verbal protocol indicated that they were able to summarize the tasks, majority could not respond to this task accurately, and some even gave up after spending a considerable amount of time. It was also clear from the eye tracking data of the two participants that Task 3 operationalized more careful reading, which is indicated by the more aligned fixations on the texts. As stated in

the test specifications, this task aims to operationalize “the ability to understand specific, factual information at the word and/or phrase level across the texts”. However, the nature of a summary task requires more global level careful reading because it requires the understanding the structure of a text (Khalifa & Weir, 2009). Therefore, a task that requires sentence level syntactic analysis, although these sentences are located in different texts, probably challenged the participants. It was also found that Task 3 directed the participants to consult their background knowledge more. It could be inferred that when test takers are challenged and cannot find the answers in the texts, they refer to their background knowledge to produce a response. Overall, this task may operationalize the ability specified in the specifications; however, there is no need to integrate information across texts, and only few participants reported that they compared the gists of texts and claims in different texts to complete this task. Therefore, it is possible to conclude this task do not operationalize multiple texts reading skill substantially and representatively.

When all the sections are taken into consideration, ISE II samples from both the lower and higher level abilities. However, ISE II cannot be claimed to operationalize multiple texts skill adequately and successfully. As mentioned earlier, this was expected just by looking at the test specifications. This exam did not base the skill definition or operation of multiple texts reading skill on theory. It attempted to measure a global level reading skill through syntactic analysis, which is a lower level careful reading ability. The fact that it attempted to operationalize this ability in such a way created unfair challenge, as a result construct irrelevant variance.

When these results are considered, it is revealed that, test specifications, if not based on theory, will not help to operationalize the ability that is aimed to be tested. Tasks that operationalize comprehension at lower cognitive levels cannot possibly

test multiple texts reading skill. If multiple texts reading skill is to be tested, operations that trigger global level careful reading must be included in the test specifications. When a task aims to operationalize a higher level reading ability through lower level abilities, this creates unrealistic expectations on the part of the test taker and may negatively affect test takers' performance. As Khalifa and Weir (2009) suggest, the purpose of reading determines the operations to be employed. When there is a mismatch between the purpose of the task, and the actual operations it requires, this creates an unfair challenge, and influence test takers' performance negatively. This is not desired because unnecessary difficulty leads to construct irrelevant variance; therefore invalidity because the true performance of the test taker cannot be revealed.

MET

Two items from MET were administered to the participants. Item 1 required the participants to identify what the authors of Texts B and C agree on. Item 2 tested the ability to guess the meaning of a phrase from the contexts provided in Texts A and C. Both items were responded accurately by 80% of the participants, which shows that this task was not challenging for the participants. Item 1 and 2, both were reported to be completed mostly through careful reading strategies. "Reading the text/s linearly from the beginning to the end carefully" was the mostly employed strategy for Item 1. Skimming and search reading was also reported to be used. For this item, multiple texts reading skill, as "identifying claims that agree, disagree and complement one another in different texts" and "comparing the gists of different texts" was also reported only by a small proportion of the participants. Another interesting finding emerged from the think aloud data is that, while responding to this item, one

participant could answer this question accurately just based on the information in Text B. This is surprising considering that the item asks what the authors of the both texts agree on. However, it was revealed that Text B included information regarding only the correct option, not the others. This suggests test writers that while designing multiple texts reading tasks, utmost care should be given to devise items that accurately operationalize the ability in question, and eliminate all the construct irrelevant variance.

As for Item 2, the participants reported mostly careful reading strategies, and expeditious reading strategies. However, no multiple texts reading strategies was reported. This finding is not surprising considering that this item tested the ability to guess the meaning of a word from the contexts presented in two texts. However, if the context in one text is adequate, then there should not be a need to consult the other. Even though there emerges a need to consult both texts, this cannot still be considered multiple texts reading skill, as only a few sentences in relevant texts would be adequate. As expected for this task, the most common careful reading strategy reported is CR3, “reading carefully across sentences (to establish the connections of ideas between sentences or parts of the text by identifying relationships such cause and effect, claim and supports etc.), and CR8, “Rereading the important or difficult / relevant parts of the text” was the second common strategy. It is expected for such items to operationalize expeditious reading to locate the words or phrase and careful reading strategies to process the relevant information, which is supported by the findings of this study.

When the eye movement data is analyzed for both items, it is seen that Text C attracted more fixations and 78% of it was carefully read. However, when two individual cases are examined, it is seen that Text C was carefully read the least. This

finding is interesting because Text C was the longest with five paragraphs among the three texts, and both questions directed the participants to Text C. However, without carefully processing it, the two participants could answer both the questions accurately. This is probably because the information to answer the two items was located in the first two paragraphs. Actually, there was not a need to process the rest of the text. Nevertheless, as suggested by Perfetti et al. (1999), it is known that forming a situational representation of a text is the first step to form intertextual representations. The formation of a situational representation requires the formation of macro structures, which is achieved through linear and incremental careful reading. Therefore, it could be concluded that MET does not operationalize substantial multiple texts reading skill. Item 1 does so to some extent based on the reported strategies of the participants.

In addition, the verbal accounts of the participants indicated that test takers may sometimes employ certain test taking strategies to easily reach the correct answers. These are worth mentioning because these provide valuable information to test writers to understand the thought process of test takers.

While responding to Item 1, one participant stated that he started with reading the options, and then in the options, he looked for a word that he could easily scan across texts, and the first option he scanned for happened to be the correct answer. This shows that options for tasks that attempt to operationalize careful reading skills must be written in a way to eliminate the operationalization of search reading skills. For the same item, another student eliminated all the negative options because he stated that the texts looked like an advertisement of a product; therefore, probably no negative points regarding the product would be raised. This strategy may not always

work. However, while designing tasks, it is necessary not to include options that could be eliminated through common sense.

With regard to Item 2, a few participants failed to respond to it accurately, not because they could not comprehend the texts to guess the meaning but because they did not know the equivalent in the options. This resulted from the fact that the correct option was a phrasal verb, and it is known that phrasal verbs are acquired later. This item failed to accurately test an ability, guessing meaning from context, the participants had. This shows us that when designing such items, it is necessary to place the less frequent word within context and the more frequent equivalent in the option to successfully assess this ability.

All in all, the results indicate the presence of multiple texts reading skill, but the indication of the eye movement data is that not the majority of the texts were carefully processed. This shows that the task required the participants to compare information across texts. However, for the successful completion of the task, it was adequate to read only certain parts of the texts carefully. This finding entails two things: first, if Perfetti et al.'s (1999) model of intertextual representation is to be followed, then items as MET Item 1, do not successfully operationalize multiple texts reading skill because the entire texts are not carefully processed. On the other hand, this finding may indicate that multiple texts reading skill is also multi layered within itself and there are lower level multiple texts reading skills such as comparing specific information across texts, and a more comprehensive multiple texts reading skill definition is necessary. Goldman et al. (2013), in their multiple texts comprehension assessment model, present subskills of multiple texts reading skill. However, they do not indicate how many of these subskills need to be employed for a test taker to be considered using multiple texts reading skill. This is an important

point of concern as especially the analysis component of their model could well be employed with only a single text. May be, for a task to be classified as testing multiple texts reading skill, the criteria should be that it operationalizes from each subcomponent of this assessment model; namely, synthesis and integration as well.

In addition, results from the MET administration show us that, similar to the results of ISE II, items that require comprehension at lower cognitive levels do not operationalize multiple texts reading skill. The item in question was a guessing meaning from context task. Therefore, we can conclude that items such as the ones requiring syntactic analysis are not the ideal tasks to assess multiple texts reading comprehension.

ECCE

Two items from ECCE were used in this study. Item 1 asked the participants to identify how Text C differed from the others. Item 2 asked what is implied by all the texts. These are the abilities tested based on the specifications “comparing / contrasting features of one or more texts, understanding explicitly stated ideas (detail) from one or more texts, and drawing an inference/conclusion from one or more texts” (CaMLA, 2017b).

The verbal protocols of the participants indicate that for Item 1, mostly careful reading strategies were employed and a high proportion of the participants reported that they read the texts from the beginning to the end. For this item, multiple texts reading skill use, “comparing the gists of different texts” was reported by 27% of the participants as well. It was also observed that two participants had realized the different information Text C presented. They did not choose the option with that information because they expected the difference Text C bears must be related to the

main idea of the text. However, the item in question focused on a specific detail. Therefore, despite comprehending the information presented in the text, they failed to answer this question accurately. This suggests that when the intertextual links between texts are assessed, it is necessary to focus on the gist, or the main propositions rather than specific details because as suggested by Perfetti et al., (1999) a more global understanding to form situations based on texts is necessary for multiple texts reading.

For Item 2, expeditious reading strategies were reported as the most common, which was “searching for/identifies key words/ideas in the text/s related to the question”. It was followed by “reading the texts linearly from the beginning to the end carefully”. No multiple texts reading skill were reported for this task. Considering the nature of the question, this was not expected. However, the participants responded to this item by doing mostly search reading and looking for the keywords in the options across (53%). The proportion of careful reading is quite high as well, which could be an indicator of multiple texts comprehension, but, although the careful reading was operationalized across texts, it does not necessarily show us that multiple reading comprehension is achieved because the task does not give us information regarding what a candidate can and cannot do with multiple texts. The verbal protocols of the participants also indicate that they searched for key words. Especially Option d (See Appendix C) was easy to eliminate because the key word in the option was a proper noun (Basque), which was easy to scan in the texts. When designing multiple texts reading tasks, it is imperative to include the key words in the options in all the texts, consequently eliminate the responses reached through expeditious reading.

When eye movement data is examined, except Text A (67%), around 50% of the other texts were carefully read. The results suggest that almost equal proportion of each texts is carefully processed. It is difficult to comment on whether this finding is in line with the reported strategies as Item 1 triggered more careful reading strategies, and Item 2 did so expeditious reading strategies. The analysis of the two individual cases may provide more insight.

As far as the two individual cases are concerned, although the participants differed in their overall preference towards expeditious (Participant 7) and careful reading (Participant 10), the least proportion of Text B was processed by the both. This might result from the fact that the information was presented in bullet points, and before the beginning of each section, the main topic was introduced. Therefore, it was easier to process, so there was no need for higher number of and longer fixations. In addition, one participant could answer both the items accurately, only by carefully reading around 50% of all the texts. This finding might indicate that the task could be successfully completed only with partial reading. However, multiple texts reading requires more global level careful reading as it is necessary to form mental representations of the situations formed in the minds of the reader based on the information presented in different texts (Perfetti et al., 1999).

Taking into consideration the test specifications and multiple texts reading theory, we can say that this task operationalizes multiple texts reading skill to some extent based on the reported strategies. However, this skill might have been employed just because four texts are presented at the same time irrespective of the task. It might also be claimed that Item 2 operationalizes making an inference across texts; still, the participants reach the correct answer through option elimination, which is not a reading strategy but a test taking strategy (Cohen & Upton, 2007).

Nevertheless, while eliminating the options, the participants had to do a substantial amount of search reading across texts. As mentioned earlier, this task triggers the use of all the texts while completing the task, and it is not different from search reading a text comprising several paragraphs. This is also a skill not defined in multiple texts reading theory. However, the fact that this skill necessitates the use of different texts might lead it to be considered a lower level multiple texts reading skill.

The findings of ECCE test administration demonstrate that a task aiming at global level careful reading comprehension actually triggered careful reading at local level, consequently, creating construct irrelevant variance by misguiding the students while determining what type of reading and level of reading they were supposed to do. As mentioned earlier, the purpose of reading determines the type and level of reading (Khalifa and Weir, 2009). Such construct irrelevant variance impact on the validity of an exam negatively as the candidates possessed the actual ability but could not show it due to task design, and such a test fails to provide a precise account of what a test taker can or cannot do. Besides, even though items such as Item 1 could be successfully responded to just through local level careful reading, it required test takers to consult or visit all the four texts. This could be because the presence of four texts created a need for the participants to read all the texts. Some test takers, to be on the safe side, or just as a matter of preference, read the whole text/s first, and then attempted the task. For this task, neither could it be claimed that the task actually required multiple texts reading skill, nor it did not. It is necessary that the participants be asked about their specific reading strategies. Finally, a task measuring multiple texts reading skill should not include options that could just be eliminated through expeditious reading. Therefore, all the options must include similar key words, and reaching the accurate answer must necessitate careful reading

of at least the majority of all the texts given. However, one can argue that the test takers did search reading across texts, which is different from search reading a single text. Then, it is also possible to suggest that such a lower level multiple texts reading skill must be conceptualized. But, this could only be done by showing how search reading across multiple texts differs from search reading a single text in terms of cognitive operations they both trigger.

5.4 Conclusion to the discussion chapter

To conclude, revisiting construct validity is necessary here. Designing exams and developing exam tasks are major responsibilities because decisions are made based on exam scores, which affect people's lives substantially. Therefore, designing exams and tasks is a rigorous and laborious task, which must be done intricately considering all the stakeholders, time, and, money involved. Making accurate interpretations, inferences, and consequent decisions depends on the construct validity of an exam. Therefore, firstly, the operationalizations of the construct to be assessed must be rooted in theory. Secondly, how these operationalizations are performed by test takers must be investigated through research to ensure that the abilities to be assessed match and reflect those defined in theory through the collection of process based data. This sort of data collection is also crucial to gather evidence whether the task used do actually trigger those abilities. As also revealed by this study, test takers may develop certain test taking strategies without entertaining the ability that a task aims to assess. Collecting ongoing process based data enlighten test writers on this regard providing them with the chance to devise tasks that would eliminate or decrease construct irrelevant variance, which in return will lead to more accurate interpretations on the exam scores.

CHAPTER 6

CONCLUSION

6.1 Introduction

This study aimed at investigating the cognitive validity of multiple texts reading skill test in the present international language proficiency exams. This chapter will present the overview of the findings, the implications on test design, the limitations of the present study, and suggestions for further research.

6.2 Overview of the findings

Firstly, when the reading operations in the specifications of these exams are compared against the literature, it was revealed that ISE II do not attempt to sample representatively from the target language use domain, as the task on multiple texts reading skill only attempts to operationalize reading across texts at word and phrase level. In addition, MET specifications for reading operations include multiple texts reading skill at a very generic sense. The operationalized ability is specified as “making connections across texts”. It is significant to point here that what sort of connections are to be made is not clear. The sort of connections to be made should be specified as abilities that could easily be observed in the items. Finally, ECCE specifications on reading show that all the operations listed in the specifications may or may not test multiple texts reading skill considering the way they are formulated. Therefore, it is not certain whether any reading exam that is written using those specifications will operationalize substantial multiple texts reading skill. When the operations listed in the specifications are examined in detail, it is seen that, ECCE

attempts to sample from both higher and lower level abilities because comprehension is tested on reading for specific details, making inferences and drawing conclusions as well as comparing features of one or more texts. Therefore, these specifications attempt to operationalize multiple texts reading skill representatively although the integration component of the multiple texts comprehension model by Goldman et al. (2013) is lacking.

Secondly, this study investigated whether actual cognitive processes test takers entertain match the ones defined in theory and the test specifications of these exams. ISE II could be considered to operationalize the majority of the operations listed in the specifications. Task 1 was observed to operationalize mostly expeditious reading even though the specifications mention that the task focus here is to test the ability to understand the main idea or purpose of each text. This seems to be done in our sample through key word matching between the questions and the text in the majority of the items. Task 2, which aimed to operationalize careful reading at the local level, actually was observed to operationalize this ability. Task 3, which is the only multiple texts reading skill task could be considered to operationalize the abilities specified in the specifications. However, it can be confidently uttered that this task did not require the participants to analyze, synthesize, and integrate information across text. It required local careful reading at the sentential level. It must also be acknowledged that this task required expeditious reading across texts. However, whether this is a multiple texts reading skill is questionable. A summary task, which normally should necessitate global level careful reading at text or intertextual level was operationalized here through local level careful reading. Therefore, even though the participants could comprehend the content, they could not frequently reach the accurate answer in this section, which raises concerns

regarding the task design. MET also does not substantially operationalize multiple texts reading skill because it was found that the participants did not need to process all the texts carefully, and there is no need to process all the texts incrementally to create a mental representation of the situations being described in the texts. However, it must be acknowledged that items required comparison/consultation across texts. Still, it was revealed that certain items could be accurately responded only based on the information in one text. Therefore, MET items in question neither substantially operationalize the multiple texts reading skill defined in theory nor operations specified in the test specifications.

It was also observed that ECCE operationalize multiple texts reading skill as defined in theory to some extent. The participants reported multiple texts reading strategy use in their verbal protocol. However, the results also revealed that only through partial careful reading, this task could also be successfully completed. This was also supported by the eye movement records, which revealed that three of the four texts were carefully read only around 50 %. In addition, investigation of the eye movements of two participants' while both consulted all the texts, one did more careful reading, which was clear from the aligned fixations. In addition, an item which requires the comparison of all the texts was inaccurately responded by a few participants even though comprehension was present in their verbal accounts. This was because when making comparisons across texts even though a more global level of comprehension is expected, and used by a few participants, this item focused on a specific detail, which turned out to be a confusing a case. Furthermore, an item seemingly requiring the comparison of all the texts could be answered just through option elimination, which is either realized through key word matching or local level careful reading. Consequently, it is necessary that the design of such tasks require

textual level comprehension, not just search reading paired with careful reading to eliminate certain options.

In conclusion, based on the findings of the present study, it is conceivable to reach that present multiple texts reading skill tasks in international English language proficiency exams do not substantially operationalize these skills. Even though there is an attempt, the design of the tasks prevents the achievement of the aim. This study also suggests implications on the task design which is presented below.

6.3 Implications on test design

Firstly, when designing multiple texts reading skill tests, it is imperative to include relatively long and propositionally dense texts which present an issue from different perspectives. Only through these sorts of texts can test takers be required to identify claims and evidence across texts on the same issue.

Secondly, multiple texts reading tasks must require test takers to synthesize information across texts after analysis, which means the task must not be completed only by using the information in one text. It is also imperative to mention that a multiple texts reading skill task must require the careful comprehension of the majority of the ideas in a text/ and even better the whole text and all the texts in the same manner, for multiple texts reading skill is a higher level ability requiring the establishment of links based on the relations of each text to one another.

In addition, items requiring local level careful reading must not be included in multiple texts reading skill tasks because such an ability is not included in the operational definition of multiple texts reading skill. In addition, guessing meaning of a vocabulary from contexts across two texts might not appropriately operationalize

multiple texts reading skill since if one context presented in a text is adequate to guess the meaning, there will not be a need to consult another.

Finally, the operationalizations in the test specifications of an exam must be based on the theory to reflect the operations defined in the theory regarding the construct to be assessed. In addition, test specifications must be clear to different stakeholders such as test writers to devise tasks measuring the abilities precisely, test users to decide on attest to fit their purpose and context and test takers to learn about the expectations and study accordingly.

6.4 The limitations of the study

One limitation of this study is the sample size as this was a small scale study. A larger participant size would definitely increase the reliability of the findings.

Another limitation of this study is that for ISE II two tasks were presented on the same page during eye tracking, which prevented collecting eye movement data for Task 2 and Task 3 separately.

Besides, due to time limitations the coding of the reading strategies was done by two raters at the same time. If the raters coded the strategies separately, and then disagreed items were discussed together, the reliability of the coding could be higher.

Lastly, the eye tracking methodology used in the study was very suggestive but because the amount of data produced was huge and included the distinctions based on minute details, it has not been possible to harness all the data and convert them into information.

6.5 Suggestions for further research

The first suggestion is that the same study be carried out with a larger sample size to be able to derive more reliable and generalizable results.

In addition, process based cognitive validity data on a task operationalizing multiple texts reading skill must also be collected through eye tracking and think aloud to accurately compare and contrast multiple texts reading skill operations with the ones the exams in question tap on.

Innovative methods of data analysis must be developed for eye tracking research that will better reflect the intricacies of the reading operations and sequences so that we can formulate better definitions of reading skills. At present, it is not possible to convert eye tracking data into a form that presents the sequence of movements and switches between texts and items in a way to be reliably interpreted. If new methods to be developed, eye tracking records, which produce immense data, cognitive processes of test takers could be investigated more accurately.

6.6 Conclusion

This study has exemplified breach of an important rule in assessment such that tests should accurately and comprehensively operationalize the skills that they claim to assess in line with the theoretical explanations of those constructs. Otherwise, they are risking the validity of the decisions to be made on the interpretation of the scores that they produce. Thus, test producing institutions should strive to warrant that this is not the case. This study presented a case to be thought upon, therefore contributed to our understanding of the validation of reading tests.

APPENDIX A

ISE II

<p>TEXT A</p> <p>Some countries are significant producers of local food, others less so. The local food movement is a campaign started in America, for example, in the 1900s over 40 per cent of the population lived on farms, whereas in 2000 the figure was 1 per cent. Nowadays, in such areas, the local food movement wants a shift back towards small-scale farming and locally-supplied food. This is an alternative to imported food, where producers are separated from consumers by 'food miles', resulting in long journey times. Although some big supermarkets stock local food, this is not the main trend as customers still want a wide choice of foods all year round. With local growing, the buyer can purchase food from the farmer in person or online, or from local shops. The farmer retains more money, which has a positive impact on local economies as money is kept within a region.</p>	<p>TEXT B</p> <p>TEXT D</p> <p>Robert: Going back to small-scale farming is incredibly unrealistic.</p> <p>Joseph: I disagree! I'm a farmer in Kenya, in Africa, and my family has always grown its own food.</p> <p>Robert: And do you export food, too?</p> <p>Joseph: Yes, I grow beans, corn and bananas for export. The money helps my family and the local and national economies.</p> <p>Robert: I'm sure. We'd have a very limited choice in Northern Scotland if we didn't import food. Local farmers couldn't produce enough for everyone in the area, so we couldn't do without food from abroad.</p> <p>Joseph: Aren't people worried about the effect transporting food has on the environment?</p> <p>Robert: Yes, but the environmental effect of transportation is actually not that high. In fact, the amount of greenhouse gases emitted in producing food locally is more than in the transportation of food. Apparently, cattle on open land produce more greenhouse gas than cows kept inside on large-scale farms.</p> <p>Joseph: Well, sending our produce abroad is great for us.</p> <p>Robert: And for us!</p>
<p>TEXT C</p> <p><i>I interviewed Jane Gold, a supporter of local food, for Green Magazine:</i></p> <p>Why do you support the local food movement, Jane?</p> <p>"Well, some countries rely too much on imported food. The effect of transporting food long distances obviously damages the environment, so eating local food is something we should all do to tackle the problem of greenhouse gases. Locally grown food is also better for us. That's another reason why people should buy it. Vitamin levels in food fall quite soon after picking, and large farms often use more chemicals than smaller ones. The change has been incredible. I always used to get colds and now I never do since I've been eating such good food — I feel fantastic!</p>	<p>Read questions 1-5 first and then read texts A, B, C and D. As you read each text, decide which text each question refers to. Choose one letter — A, B, C or D — and tell it out loud. You can use any letter more than once.</p> <p>Which text</p> <ol style="list-style-type: none"> suggests people buy local food because it tastes better? criticises the idea that people could return to producing food on small farms? suggests that local food production can reduce illness and increase well-being? explains that the different stages in the production of local food are connected? compares farming in the last century with the popularity of farming nowadays?

TEXT A

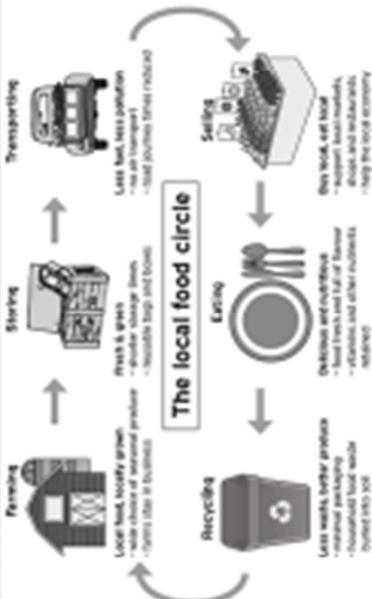
Some countries are significant producers of local food, others less so. The local food movement is a campaign started in countries which import more food than in the past. In America, for example, in the 1900s over 40 per cent of the population lived on farms, whereas in 2000 the figure was 1 per cent. Nowadays, in such areas, the local food movement wants a shift back towards small-scale farming and locally-supplied food. This is an alternative to imported food, where producers are separated from consumers by 'food miles', resulting in long journey times. Although some big supermarkets stock local food, this is not the main trend as customers still want a wide choice of foods all year round. With local growing, the buyer can purchase food from the farmer in person or online, or from local shops. The farmer retains more money, which has a positive impact on local economies as money is kept within a region.

TEXT C

I interviewed Jane Gold, a supporter of local food, for Green Magazine:
Why do you support the local food movement, Jane?

"Well, some countries rely too much on imported food. The effect of transporting food long distances obviously damages the environment, so eating local food is something we should all do to tackle the problem of greenhouse gases. Locally grown food is also better for us. That's another reason why people should buy it. Vitamin levels in food fall quite soon after picking, and large farms often use more chemicals than smaller ones. The change has been incredible. I always used to get colds and now I never do since I've been eating such good food — I feel fantastic!

TEXT B



TEXT D

Robert: Going back to small-scale farming is incredibly unrealistic.
Joseph: I disagree! I'm a farmer in Kenya, in Africa, and my family has always grown its own food.
Robert: And do you export food, too?
Joseph: Yes, I grow beans, corn and bananas for export. The money helps my family and the local and national economies.
Robert: I'm sure. We'd have a very limited choice in Northern Scotland if we didn't import food. Local farmers couldn't produce enough for everyone in the area, so we couldn't do without food from abroad.
Joseph: Aren't people worried about the effect transporting food has on the environment?
Robert: Yes, but the environmental effect of transportation is actually not that high. In fact, the amount of greenhouse gases emitted in producing food locally is more than in the transportation of food. Apparently, cattle on open land produce more greenhouse gas than cows kept inside on large-scale farms.
Joseph: Well, sending our produce abroad is great for us.
Robert: And for us!

Task 2: Choose the five statements from A–H below that are **TRUE** according to the information given in the texts on the left. **Tick** the letters of the **TRUE** statements (in any order).

- A US local food supporters want a return to farming levels of the 1900s.
- B Supermarkets generally support the local food movement.
- C Local farmers may use technology to help sell their food directly.
- D Storage times and the amount of packaging decrease with local farming.
- E Small farms sometimes use chemicals when producing their food.
- F Jane believes there's been a slight improvement in her health and mood.
- G The transportation of food damages the environment less than food production.
- H Both Robert and Joseph agree that exporting food to other countries is a good idea.

Task 3: The summary notes below contain information from the texts on pages 4 and 5. **Find an exact number, word or phrase (maximum three words)** from texts A–D to complete the missing information in gaps 1–5.

Summary notes

- Aims of local food movement:**
- to raise levels of production and sales of local food
 - a return to (1.) _____ and delivery of local food imported vs local food:
 - imported food: increased food miles between farmers and customers leads to (2.) _____
 - local food: bought direct from farmers
 - less time in storage after picking means higher (3.) _____
- Local food:**
- fresher and tastier
 - fewer food miles by (4.) _____ and road
- But:**
- greenhouse gases emitted in food production
 - insufficient locally farmed food: people in remote areas are unable to (5.) _____ imported food

APPENDIX B

MET

<p>A Introducing the New CopyPro</p> <p>The CopyPro's full-featured scanning, copying, and printing capabilities make it perfect for all your home office needs.</p> <ul style="list-style-type: none"> • Print images directly from your camera's memory card. No computer required! • Scan your photos and print them out in many sizes. • Replace ink cartridges only as colors run out with the special individual ink cartridge system. Four different color cartridges allow you to replace only the colors needed. • No need to worry about handling photos or other printed material. CopyPro uses quick-drying, smudge proof inks. • Edit and fix photos and images with CopyPro's instant Photo Expert software. 	<p>Regular Reviews: Honest Reviews by Ordinary People <i>Review of the CopyPro</i> by Steve Wilson, Philadelphia, PA</p> <p>I am quite pleased with this machine, and I think it offers tremendous value. One of the things I particularly liked about the CopyPro is that it prints at a normal speed with decent quality, which is unusual for printers in this price category. It has five levels of quality, although the draft mode is not recommended—pages are very light and dotty.</p> <p>CopyPro claims its ink is both water resistant and smudge proof. I tasted these claims by putting some color pages under running water; the ink did not run, and when the pages dried, the ink did not come off, even with rough handling, which supports CopyPro's claims. This is important for business users who make mailing labels and are concerned about exposure to the weather, and for home users worried about the durability of the photos they print.</p> <p>The CopyPro comes with four separate ink cartridges, meaning users can replace the colors as they run out. This is convenient, and it's cheaper in the long run than using a single cartridge for all colors that has to be replaced more often.</p> <p>The CopyPro has two memory card slots that can accommodate most types of camera memory cards. I find this to be very convenient—I can plug in my camera's card and print, without connecting my computer. However, the CopyPro Instant Photo Expert software was disappointing. It has minimal features and is not a replacement for full-featured photo editing software—the software that came with my digital camera is much better. Still, CopyPro Instant Photo Expert does let you resize your photos, rotate them, do basic color correcting, and some other things.</p> <p>In short, I think this is a good machine, and the low price makes it a good value.</p>	<p>B</p> <p>MEMO</p> <p>Jane,</p> <p>Last week when we discussed purchasing a new copier, you asked me to look into them and to give you my recommendation. I've looked at about ten different models so far. Here's one that I think will be perfect for our office: CopyPro. It has all the features that we discussed, and it is within the budget you mentioned. I looked online and found some product reviews. Most of the reviews for the CopyPro have been favorable—in fact, several computing websites have named it their top pick. Even though it's aimed at the home-user market (people who want to print photos, for example), its print speed, scan resolution, and copying capabilities are all things that we would take advantage of here in the office. Look at the attached product description and let me know what you think. If you like this, I'll be happy to take care of ordering one. If you don't, I'll continue looking at other models.</p> <p>Alan</p>
<p>Read the three texts on the left and answer the questions below.</p> <p>1. What do the authors of the memo and the review agree on about CopyPro?</p> <p>a. It should not replace a full-sized machine. b. It should not be used in draft mode. c. It should be used for photographs only. d. It is suitable for both home and business use.</p> <p>2. Which phrase is closest to "smudge proof" as it is used in the review and the advertisement?</p> <p>a. will not change color b. will not rub off c. will not be permanent d. will not dry on some kinds of paper</p>		

APPENDIX C

ECCE

<p>A A World of Cooking</p> <p>Pablo's Restaurant hosts a series of two-hour cooking workshops.</p> <p>Our award-winning chef Emily Winters cooks dishes from food cultures around the world, including European, Asian, and African cuisines. And of course, there will be a lesson on how to make Pablo's most popular Spanish dishes! Find details and sign up online, by phone, or ask next time you're at the restaurant. Discounted package rate for all 12 classes of the series.</p> <p>Learn from the best!</p>	<p>C A Recipe for Success</p> <p>Have you always wanted to learn Spanish? Or visit the beautiful Spanish countryside? Or maybe you really love traditional Spanish cuisine? If any or all of these apply to you, the Taste of Spain Study Tour is a perfect opportunity to realize your dreams!</p> <p>Our three-week itinerary provides a unique combination of Spanish-language instruction, travel to the most beautiful areas of Spain, and cooking lessons that cover traditional Spanish and Basque techniques and cuisines. Small group instruction and one-on-one feedback aids student learning.</p> <p>Don't wait any longer. At our affordable rates, there's no reason to put off the trip of a lifetime!</p>	<p>Read the four texts on the left and answer the questions below.</p> <ol style="list-style-type: none"> How does text C differ from the other sections? <ol style="list-style-type: none"> It was written for a specific audience. It mentions language learning. It offers specific cooking techniques and tips. It discusses traveling to different countries. What do all four sections imply? <ol style="list-style-type: none"> Increasing numbers of people are learning to cook. International foods are popular with many people. Ethnic restaurants are expanding their business. Basque food is difficult to make.
<p>B This Week's Shopping List</p> <p>Love to cook cuisines from other countries but can't find all the items on your grocery list? Check out my tips for getting the ingredients you need!</p> <ul style="list-style-type: none"> The International section. Big grocery stores often devote an aisle to foods from around the world. Where the locals go. If your community includes neighborhoods with strong ties to other countries, visit those stores to find great foreign foods. The Internet. You can find nearly any type of food online and have it delivered right to your door. Don't forget to visit my blog next week for more food-lover tips and for ideas for dishes you never imagined trying to make! 	<p>D Basque in It</p> <p>Spain is known worldwide for its food. And no small part of that recognition is thanks to Basque cuisine. The Basque are an ethnic group whose traditional territory is primarily in northern Spain but also extends into southern France. Basque cuisine is a distinct and important part of the Basque culture, and luckily for the rest of the world, it's delicious. Basque cuisine leans heavily on what's in season and what's local: fish straight from the ocean, mushrooms from the woods, vegetables from Basque farms. The highlight of Basque cuisine is its focus on using the highest quality ingredients and combining them in original, flavorful recipes. If you visit San Sebastian, Spain (also known as Donostia), a Basque food hotspot, you can hop from restaurant to restaurant like the locals do. Be sure to sample small dishes called pintos. Wherever you visit, you'll likely see someone asking the chef for whatever's best that day, rather than requesting specific menu items. Basque cuisine is popular in several parts of the world, with many restaurants serving pintos or traditional Basque dinners. We owe this to several notable chefs who've taken an interest in Basque cuisine and to Basque emigrants in other countries who have shared a taste of their homeland.</p>	

APPENDIX D

TRAINING TASK

<p>Good News for Chocolate Lovers!</p> <p>Recent research shows that eating moderate amounts of chocolate may be good for you. Several studies published in the last few months point to the health benefits in cocoa and other chocolates. These include keeping hearts healthy by lowering high blood pressure and maintaining healthy blood flow. Cocoa contains a substance that seems to help the body regulate nitric oxide levels, which are crucial to controlling blood flow and blood pressure. Cocoa beans also contain large amounts of compounds called flavanols. These plant compounds offer strong antioxidant properties and can prevent fats in the bloodstream from oxidizing. This helps reduce the potential for clogged arteries—a major contributor to heart disease. Dark chocolate contains more flavanols than milk chocolate or other kinds of processed chocolate, such as chocolate syrups or cocoa powder. This is because flavanols are destroyed or removed in processing. Dark chocolate is a less-refined product, therefore retaining more flavanols than other kinds of chocolates.</p>	<p>Read the text on the left and answer the question below.</p> <p>1. What is the main purpose of the article?</p> <ul style="list-style-type: none">a. to advertise a new chocolate-flavored productb. to explain possible health benefits of chocolatec. to compare dark chocolate to milk chocolated. to explain how much chocolate people should eat
--	--

APPENDIX E

READING STRATEGY CODING RUBRIC

Strategy	Description
Reading Strategies prior to text taking	
PR1	reading the text first carefully before attempting the task
PR2	reading the texts first carefully before attempting the task
PR3	reading the text expeditiously to have a general idea before attempting the task
PR4	reading the texts expeditiously to have a general idea before attempting the task
Expeditious Reading	
ER1	rapidly looking for/matches figures, dates, names, specific words, etc in the text.
ER2	Looking for markers of meaning in the text (e.g. definitions, examples, guides to paragraph development such as connectors)
ER3	Trying to understand the information in the text quickly by (using the title, subtitles, section headings, first and last sentences) through skimming.
ER4	Trying to understand the information in the texts quickly by (using the title, subtitles, section headings, first and last sentences) through skimming.
ER5	Searching for/identifies key words/ideas in the text related to the question
ER6	Searching for/identifies key words/ideas in the texts related to the question
ER7	Based on the prior knowledge of texts and visuals, identifying/ trying to identify the relevant information related to a task.
ER8	Choosing one text which seems related to a specific question depending on prior skimming
Careful Reading Strategies	
CR1	identifying the similarities between words / phrases in the text and the question (Through CR but without processing the whole sentence)
CR2	Focusing on one sentence (and/or its parts) to understand it clearly.
CR3	Reading carefully across sentences (to establish the connections of ideas between sentences or parts of the text by identifying relationships such cause and effect, claim and supports etc.)
CR4	Reading a proportion of a text by establishing connections between paragraphs

Strategy	Description
CR5	Reading the text linearly from the beginning to the end carefully
CR6	Reading the texts linearly from the beginning to the end carefully
CR7	Reading only the part of the text which seems related to specific questions.
CR8	Rereading the important or difficult / relevant parts of the text
CR9	Choosing one text which seems related to a specific question or option depending on prior careful reading

Multi-text Reading Strategies

MR1	Identifying evidence in a text that can be used to support a claim in another text.
MR2	Identifying claims that agree, disagree and complement one another in different texts.
MR3	Determining which evidence is consistent and inconsistent across texts.
MR4	Forming a unified idea by combining several claims from different texts.
MR5	Understanding how each text relates to one another as a document taking into account document characteristics such as genre, author, date, and context.
MR6	Mentioning the necessity of additional information
MR7	Evaluating the final representation of information that is being created as a result of multiple texts reading.
MR8	Comparing the gists of different texts

Other Strategies

OR1	Using knowledge of the text: Noting the discourse structure of the text (cause/effect, compare/contrast, etc).
OR2	Using background knowledge to support understanding / guess or interpret meaning
OR3	Answering the question based on the information gathered up to that point without going back to the text/ or only to confirm

APPENDIX F

ISE II (1) AREAS OF INTEREST

1 Text A

2 Some countries are significant procedures of local food, others less so.

3 The local food movement is a campaign started in countries which import more food than in the past.

4 In America, for example, in the 1900s over 40 per cent of the population lived on farms, whereas in 2000 the figure was 1 per cent.

5 Nowadays, in such areas, the local food movement wants a shift back towards small-scale farming and locally-supplied food.

6 This is an alternative to imported food, where procedures are separated from consumers by 'food miles', resulting in long journey times.

7 Although some big supermarkets stock local food, this is not the main trends as customers still want a wide choice of foods all year round.

8 With local growing, the buyer can purchase food from the farmer in person or online, or from local shops.

9 The farmer retains more money, which has a positive impact on local economies as money is kept within a region.

10 TextB

11 Farming and the pictures and the information below

12 Storing and the pictures and the information below

13 Transporting and the pictures and the information below

14 The local food circle

15 Recycling and the pictures and the information below

16 Eating and the pictures and the information below

17 Selling and the pictures and the information below

18 Text C

19 I interviewed Jane Gold, a supporter of local food, for Green Magazine: Why do you support the local food movement, Jane?

- 20 'Well, some countries rely too much on imported food.
- 21 The effect of transporting food long distances obviously damages the environment, so eating local food is something we should all do to tackle the problem of greenhouse gases.
- 22 Locally grown food is better for us.
- 23 That's another reason why people should buy it.
- 24 Vitamin levels in food fall quite soon after picking, and large farms often use more chemicals than smaller ones.
- 25 The change has been incredible.
- 26 I always used to get colds and now I never do since I've been eating such good food – I feel fantastic!
- 27 Text D
- 28 Robert: Going back to small-scale farming is incredibly unrealistic.
- 29 Joseph: I disagree!
- 30 I'm a farmer in Kenya, in Africa, and my family has always grown its own food.
- 31 Robert: And do you export food, too?
- 32 Joseph: Yes, I grow beans, corn and bananas for export.
- 33 The money helps my family and the local and national economies.
- 34 Robert: I'm sure.
- 35 We'd have a very limited choice in Northern Scotland if we didn't import food.
- 36 Local farmers couldn't produce enough for everyone in the area, so we couldn't do without food from abroad.
- 37 Joseph: Aren't people worried about the effect transporting food has on the environment?
- 38 Yes, but the environmental effect of transporting is actually not that high.
- 39 In fact, the amount of greenhouse gases emitted in producing food locally is more than in transportation of food.
- 40 Apparently, cattle on open land produce more greenhouse gas than cows kept inside on large-scale farms.
- 41 Joseph: Well, sending our produce abroad is great for us.

42 Robert: And for us!

43 Read questions 1-5 first and then read texts A,B,C, and D.

44 As you read each text, decide which text each question refers to.

45 Choose one letter – A,B,C, or D – and tell it outloud.

46 You can use any letter more than once.

47 Question 1

48 Question 2

49 Question 3

50 Question 4

51 Question 5

APPENDIX G

ISE II (2) AREAS OF INTEREST

1 Text A

2 Some countries are significant producers of local food, others less so.

3 The local food movement is a campaign started in countries which import more food than in the past.

4 In America, for example, in the 1900s over 40 per cent of the population lived on farms, whereas in 2000 the figure was 1 per cent.

5 Nowadays, in such areas, the local food movement wants a shift back towards small-scale farming and locally-supplied food.

6 This is an alternative to imported food, where producers are separated from consumers by 'food miles', resulting in long journey times.

7 Although some big supermarkets stock local food, this is not the main trend as customers still want a wide choice of foods all year round.

8 With local growing, the buyer can purchase food from the farmer in person or online, or from local shops.

9 The farmer retains more money, which has a positive impact on local economies as money is kept within a region.

10 Text B

11 Farming and the pictures and the information below

12 Storing and the pictures and the information below

13 Transporting and the pictures and the information below

14 The local food circle

15 Recycling and the pictures and the information below

16 Eating and the pictures and the information below

17 Selling and the pictures and the information below

18 Text C

19 I interviewed Jane Gold, a supporter of local food, for Green Magazine: Why do you support the local food movement, Jane?

20 'Well, some countries rely too much on imported food.

21 The effect of transporting food long distances obviously damages the environment, so eating local food is something we should all do to tackle the problem of greenhouse gases.

22 Locally grown food is better for us.

23 That's another reason why people should buy it.

24 Vitamin levels in food fall quite soon after picking, and large farms often use more chemicals than smaller ones.

25 The change has been incredible.

26 I always used to get colds and now I never do since I've been eating such good food – I feel fantastic!

27 Text D

28 Robert: Going back to small-scale farming is incredibly unrealistic.

29 Joseph: I disagree!

30 I'm a farmer in Kenya, in Africa, and my family has always grown its own food.

31 Robert: And do you export food, too?

32 Joseph: Yes, I grow beans, corn and bananas for export.

33 The money helps my family and the local and national economies.

34 Robert: I'm sure.

35 We'd have a very limited choice in Northern Scotland if we didn't import food.

36 Local farmers couldn't produce enough for everyone in the area, so we couldn't do without food from abroad.

37 Joseph: Aren't people worried about the effect transporting food has on the environment?

38 Yes, but the environmental effect of transporting is actually not that high.

39 In fact, the amount of greenhouse gases emitted in producing food locally is more than in transportation of food.

40 Apparently, cattle on open land produce more greenhouse gas than cows kept inside on large-scale farms.

41 Joseph: Well, sending our produce abroad is great for us.

42 Robert: And for us!

43 Question 1

44 A (question 1)

45 B (question 1)

46 C (question 1)

47 D (question 1)

48 E (question 1)

49 F (question 1)

50 G (question 1)

51 H (question 1)

52 Question 2

53 Summary Notes (Title)

54 Gap 1

55 Gap 2

56 Gap 3

57 Gap 4

58 Gap 5

APPENDIX H

MET AREAS OF INTEREST

1 Introducing the New CopyPro

2 The CopyPro's full-featured scanning, copying, and printing capabilities make it perfect for all your home office needs.

3 Print images directly from your camera's memory card.

4 No computer required!

5 Scan your photos and print them out in many sizes.

6 Replace ink cartridges system.

7 Four different color cartridges allow you to replace only the colors needed.

8 No need to worry about handling photos or other printed material.

9 CopyPro uses quick-drying, smudge proof inks.

10 Edit and fix photos and images with CopyPro's Instant Photo Expert software.

11 Call to order yours today!

12 MEMO

13 Jane, Last week when we discussed purchasing a new copier, you asked me to look into them and to give you my recommendation.

14 I've looked at about ten different models so far.

15 Here's one that I think will be perfect for our office: CopyPro.

16 It has all the features that we discussed, and it is within the budget you mentioned.

17 I looked online and found some product reviews.

18 Most of the reviews for the CopyPro have been favorable – in fact, several computing websites have named it their top pick.

19 Even though it's aimed at the home-user market (people who want to print photos, for example), its print speed, scan resolution, and copying capabilities are all things that we would take advantage of here in the office.

20 Look at the attached product description and let me know what you think.

21 If you like this, I'll be happy to take care of ordering one.

22 If you don't, I'll continue looking at other models.

23 Alan

24 Regular Reviews: Honest Reviews by Ordinary People Review of the CopyPro by Steve Wilson, Philadelphia, PA

25 I am quite pleased with this machine, and I think it offers tremendous value.

26 One of the things I particularly liked about the CopyPro is that it prints at a normal speed with decent quality, which is unusual for printers in this price category.

27 It has five levels of quality, although the draft mode is not recommended – pages are very light and dotty.

28 CopyPro claims its ink is both water resistant and smudge proof.

29 I tested these claims by putting some color pages under running water; the ink did not run, and when the pages dried, the ink did not come off, even with rough handling, which supports CopyPro's claims.

30 This is important for business users who make mailing labels and are concerned about exposure to the weather, and for home users worried about the durability of the photos they print.

31 The CopyPro comes with four separate ink cartridges, meaning users can replace the colors as they run out.

32 This is convenient, and it is cheaper in the long run than using a single cartridge for all colors that has to be replaced more often.

33 The CopyPro has two memory card slots that can accommodate most types of camera memory cards.

34 I find this to be very convenient – I can plug in my camera's card and print, without connecting my computer.

35 However, the CopyPro Instant Photo Expert software was disappointing.

36 It has minimal features and is not a replacement for full-featured photo editing software – the software that came with my digital camera is much better.

37 Still, CopyPro Instant Photo Expert does let you resize your photos, rotate them, do basic color correcting, and some other things.

38 In short, I think this is a good machine, and the low price makes it a good value.

39 Question1

40 Question1 a,b,c,d

41 Question2

42 Question2 a,b,c,d

43 Read the three texts on the left and answer the questions below.

APPENDIX I

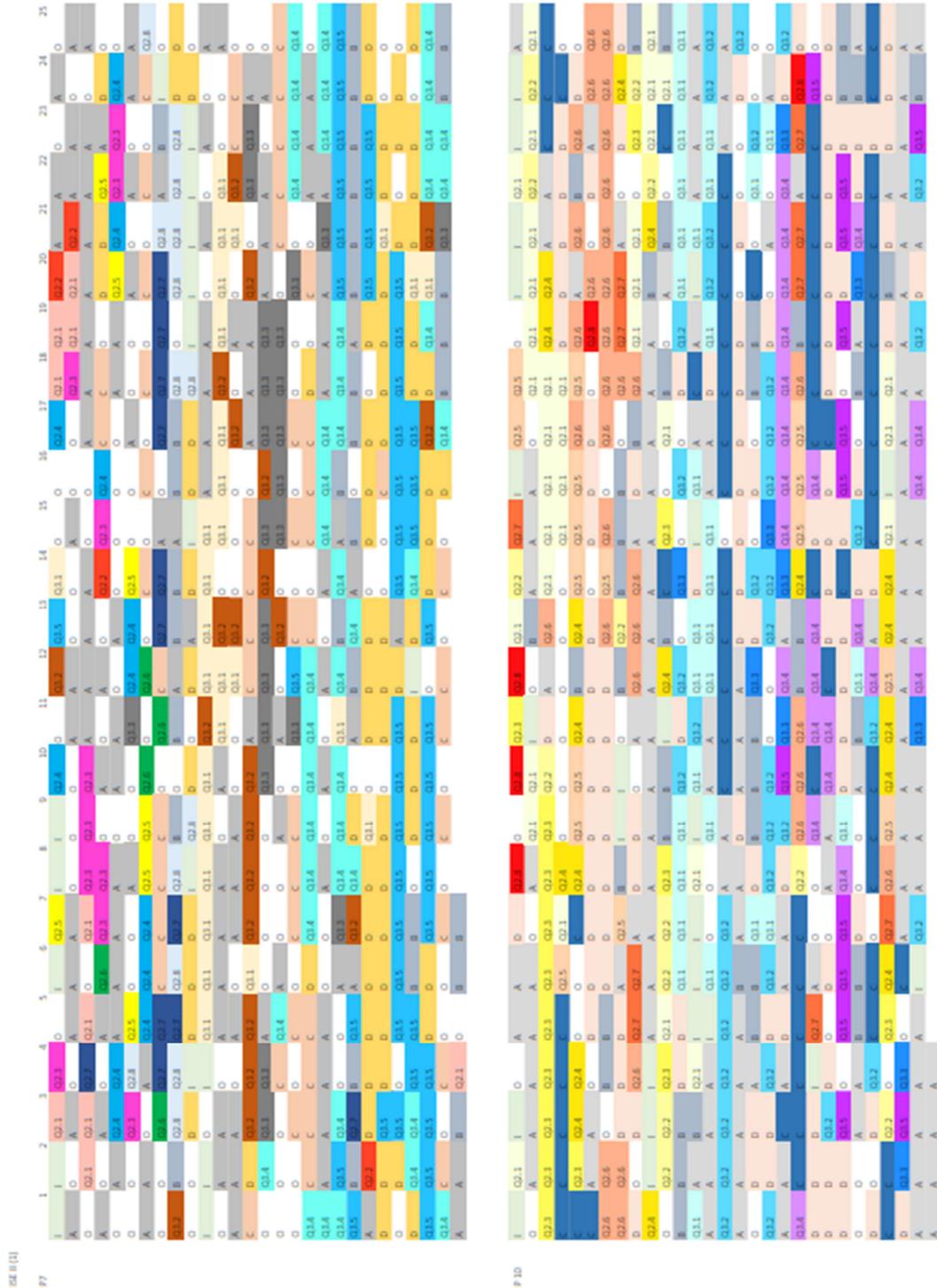
ECCE AREAS OF INTEREST

- 1 A world of cooking
- 2 Pablo's restaurant hosts a series of two-hour cooking workshops
- 3 Our award-winning chef Emily Winters cooks dishes from food cultures around the world, including European, Asian, and African cuisines.
- 4 And of course, there will be a lesson on how to make Pablo's most popular Spanish dishes!
- 5 Find details and sign up online, by phone, or ask next time you're at the restaurant.
- 6 Discounted package rate for all 12 classes of the series
- 7 Learn from the best!
- 8 This Week's Shopping List
- 9 Love to cook cuisines from other countries but can't find all the items on your grocery list?
- 10 Check out my tips for getting the ingredients you need!
- 11 The International section.
- 12 Big grocery stores often devote an aisle to foods from around the world.
- 13 Where the locals go.
- 14 If your community includes neighborhoods with strong ties to other countries, visit those stores to find great foreign foods.
- 15 The Internet.
- 16 You can find nearly any type of food online and have it delivered right to your door.
- 17 Don't forget to visit my blog next week for more food-lover tips and for ideas for dishes you never imagined trying to make!
- 18 A recipe for success
- 19 Have you always wanted to learn Spanish?
- 20 Or visit the beautiful Spanish countryside?
- 21 Or maybe you really love traditional Spanish cuisine?
- 22 If any or all of these apply to you, the Taste of Spain Study Tour is a perfect opportunity to realize your dreams!
- 23 Our three-week itinerary provides a unique combination of Spanish-language instruction, travel to the most beautiful areas of Spain, and cooking lessons that cover traditional Spanish and Basque techniques and cuisines.
- 24 Small group instruction and one-on-one feedbacks aids student learning.
- 25 Don't wait any longer.

- 26 At our affordable rates, there's no reason to put off the trip of a lifetime!
- 27 Basque in It
- 28 Spain is known worldwide for its food.
- 29 And so small part of that recognition is thanks to Basque cuisine.
- 30 The Basque are an ethnic group whose traditional territory is primarily in northern Spain but also extends into southern France.
- 31 Basque cuisine is a distinct and important part of the Basque culture, and luckily for the rest of the world, it's delicious.
- 32 Basque cuisine leans heavily on what's in season and what's local: fish straight from the ocean, mushrooms from the woods, vegetables from Basque farms.
- 33 The highlight of Basque cuisine is its focus on using the highest quality ingredients and combining them in original, flavorful recipes.
- 34 If you visit San Sebastian, Spain (also known as Donostia), a Basque food hotspot, you can hop from restaurant to restaurant like locals do.
- 35 Be sure to sample small dishes called pintxos.
- 36 Wherever you visit, you'll likely see someone asking the chef for whatever's best that
- 37 day, rather than requesting specific menu items.
- 38 Basque cuisine is popular in several parts of the world, with many restaurants serving pintxos or traditional Basque dinners.
- 39 We owe this to several notable chefs who've taken an interest in Basque cuisine and to Basque emigrants in other countries who have shared a taste of their homeland.
- 40 Question1
- 41 Question 1 a,b,c,d
- 42 Question 2
- 43 Question2 a,b,c,d
- 44 Read the four texts on the left and answer the questions below.

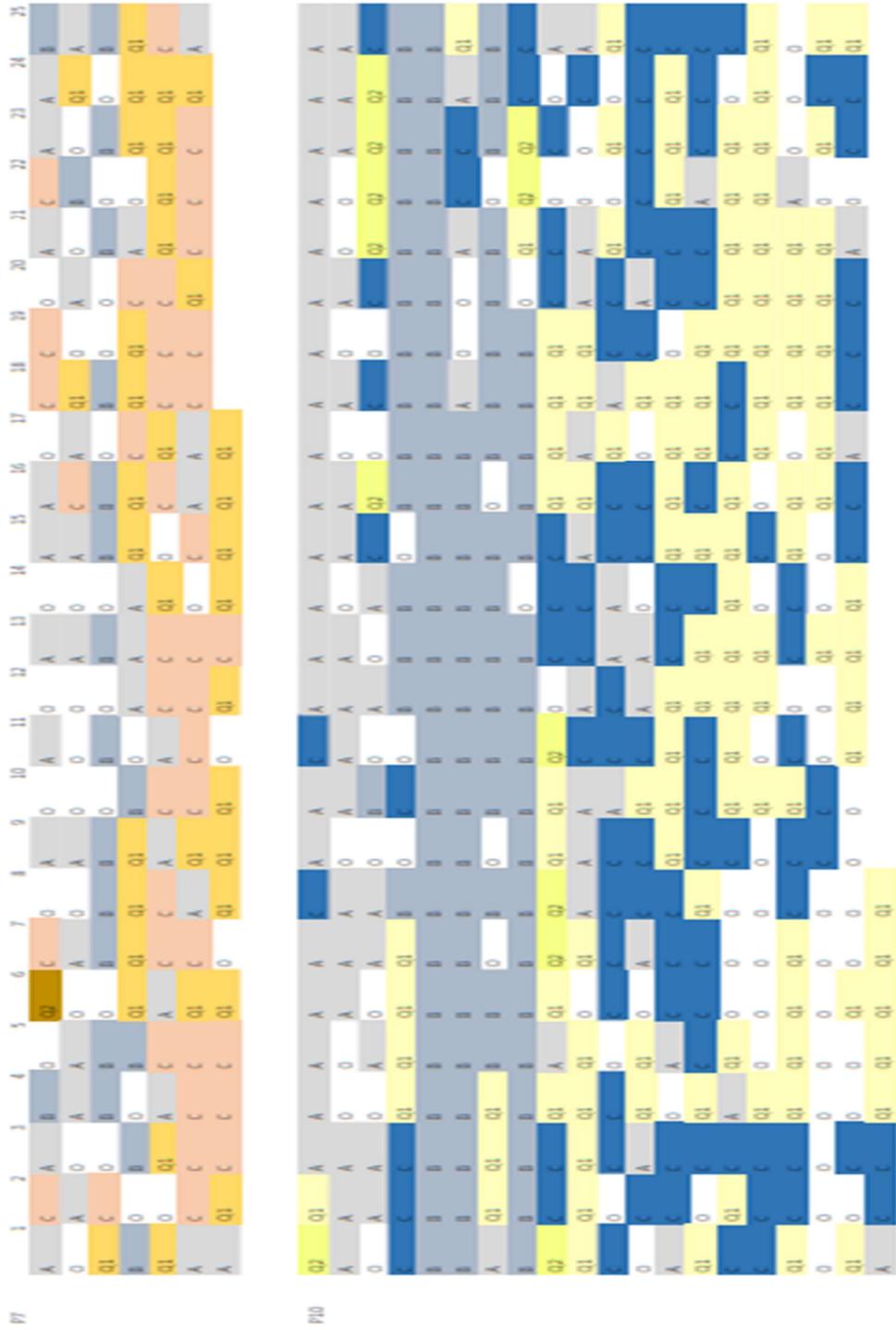
APPENDIX K

ISE II (2) EYE- MOVEMENT SEQUENCE
OF PARTICIPANT 7 AND PARTICIPANT 10



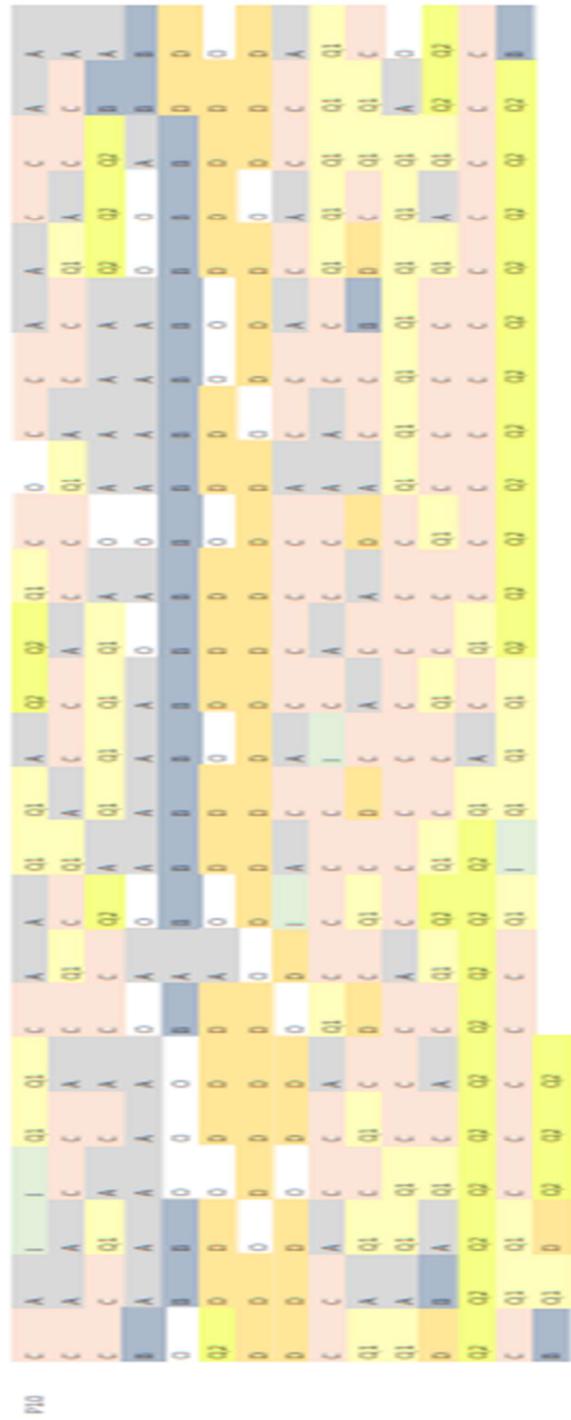
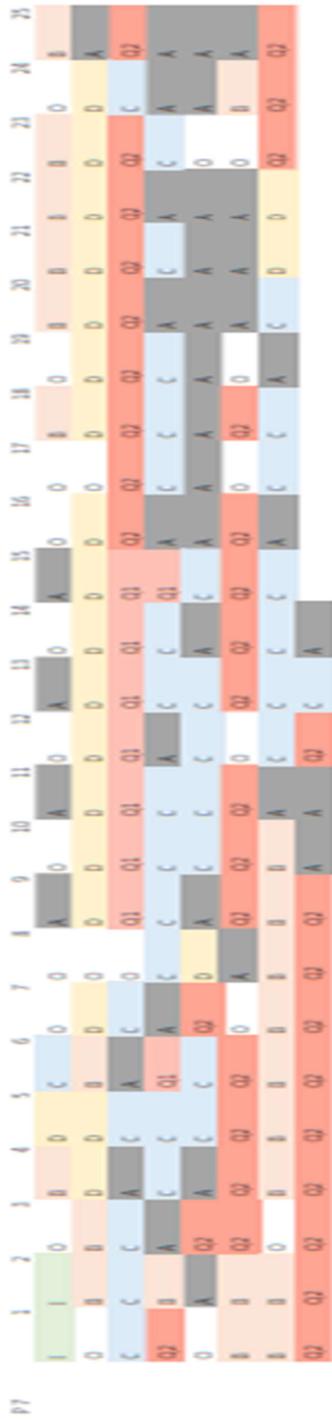
APPENDIX L

MET-EYE-MOVEMENT SEQUENCE
OF PARTICIPANT 7 AND PARTICIPANT 10



APPENDIX M

ECCE- EYE-MOVEMENT SEQUENCE
OF PARTICIPANT 7 AND PARTICIPANT 10



REFERENCES

- Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences, 30*, 64–76.
<http://dx.doi.org/10.1016/j.lindif.2013.01.007>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. (1st ed.). Oxford: Oxford University Press.
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye tracking. *Language Testing, 30*(4), 441–465.
<https://doi.org/10.1177/0265532212473244>
- Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE Reading test. *Cambridge ESOL Research Notes, 47*, 3–14.
- Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.
- Bråten, I., Ferguson, L. E., Anmarkrud, Ø., & Strømsø, H. I. (2013). Prediction of learning and comprehension when adolescents read multiple texts: The roles of word-level processing, strategic approach, and reading motivation. *Reading and Writing, 26*(3), 321–348.
<https://doi.org/10.1007/s11145-012-9371-x>
- Bråten, I., & Strømsø, H. I. (2003). A longitudinal think-aloud study of spontaneous strategic processing during the reading of multiple expository texts. *Reading and Writing, 16*(3), 195–218.
- Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*(1), 6–28.
<https://dx.doi.org/10.1598/RRQ.41.1.1>
- Britt, M.A., & Aglinskias, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*(4), 485–522.
http://dx.doi.org/10.1207/S1532690XCI2004_2
- Britt, M.A., Perfetti, C.A., Sandak, R., & Rouet, J.F. (1999). Content integration and source separation in learning from multiple texts. In S.R. Goldman, A.C. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 209–233). Mahwah, NJ: Erlbaum.
- Britt, M. A., & Rouet, J. F. (2012). Learning with multiple documents. *The Quality of Learning, 276*–314.

- Brunfaut, T. & McCray, G. (2015). Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye tracking and stimulated recall study. *British Council Assessment Research Awards and Grands: Research Reports*, 1-55. London: The British Council.
- Cambridge Michigan Language Assessment (2017a). *Michigan English Test*. Michigan, the USA: Cambridge Michigan Language Assessment. Retrieved from <http://cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/met/>
- Cambridge Michigan Language Assessment (2017b). *The Examination for the Certificate of Competency in English*. Michigan, the USA: Cambridge Michigan Language Assessment. Retrieved from <http://cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/ecce/>
- Cerdán, R., & Vidal-Abarca, E. (2008). The effects of tasks on integrating information from multiple documents. *Journal of Educational Psychology*, *100*(1), 209–222.
<https://doi.org/10.1037/0022-0663.100.1.209>
- Charmaz, K. (2008). Grounded theory. In Smith, J. A. (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 81–110). London: Sage.
- Cohen, A. D., & Cavalcanti, M. C. (1987). Giving and getting feedback on compositions: A comparison of teacher and student verbal report. *Evaluation and Research in Education*, *1*(2), 63-73.
- Cohen, A. D., & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. *ETS Research Report Series*, 2006(1).
- 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. *Language Testing*, *24*(2), 209-250.
<https://doi.org/10.1177/0265532207076364>
- Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research*, *43*(4), 352-392.
<https://doi.org/10.1177/1086296X11421979>
- Coiro, J., & Dobler, E. (2007). Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly*, *42*(2), 214-257.
- Cronbach, L. E. E. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.
<http://dx.doi.org/10.1037/h0040957>
- Dolgunsöz, E., & Sariçoban, A. (2016). CEFR and eye movement characteristics during EFL reading: the case of intermediate readers. *Journal of Language and Linguistic Studies*, *12*(2), 238-252.

- Ferguson, L. E., Bråten, I., & Strømsø, H. I. (2012). Epistemic cognition when students read multiple documents containing conflicting scientific evidence: A think-aloud study. *Learning and Instruction, 22*(2), 103–120.
<https://doi.10.1016/j.learninstruc.2011.08.002>
- Gil, L., Bråten, I., Vidal-Abarca, E., & Stromso, H. I. (2010a). Understanding and integrating multiple science texts: Summary tasks are sometimes better than argument tasks. *Reading Psychology, 31*(1), 30-68.
<https://doi.10.1080/02702710902733600>
- Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. (2010b). Summary versus argument tasks when working with multiple documents: Which is better for whom?. *Contemporary Educational Psychology, 35*(3), 157-173.
<http://dx.doi.org/10.1016/j.cedpsych.2009.11.002>
- Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Ferris & D. M. Bloome (Eds.), *Uses of intertextuality in classroom and educational research* (pp. 313–347). Greenwich, CT: Information Age Publishing.
- Goldman, S. R. (2011). Choosing and using multiple information sources: Some new findings and emergent issues. *Learning and Instruction, 21*(2), 238–242.
<https://doi.10.1016/j.learninstruc.2010.02.006>
- Goldman, S. R., & Bloome, D. M. (2005). Learning to Construct and Integrate. In A. F. Healy (Ed.), *Experimental Cognitive Psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. Washington, D. C.: American Psychological Association.
- Goldman, S. R., Lawless, K., & Manning, F. (2013). Research and development of multiple source comprehension assessment. In A. Britt, S. Goldman, J-F. Rouet (Eds.), *Reading: From words to multiple texts*. (pp. 180-197). New York: Routledge.
- Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. London: Longman.
- Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.
- Hagen, Å. M., Braasch, J. L., & Bråten, I. (2014). Relationships between spontaneous note-taking, self-reported strategies and comprehension when reading multiple texts in different task conditions. *Journal of Research in Reading, 37*(S1), 141-157.
<https://doi/abs/10.1111/j.1467-9817.2012.01536.x>
- Hoover, W. A., & Tunmer, W. E. (1993). The components of reading. In G. B. Thompson, W. E. Tunmer, & T. Nicholson (Eds.), *Language and education library, 4. Reading acquisition processes* (pp. 1-19). Clevedon, England: Multilingual Matters.

- Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading: Studies in Language Testing 29*. Cambridge: UCLES/Cambridge University Press.
- Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363.
<http://dx.doi.org/10.1037/0033-295X.85.5.363>
- Kurt, Y., (2015). *Development of a reading test for second language learners of Turkish*. (Unpublished master's thesis). Boğaziçi University, İstanbul, Turkey.
- Lacroix, N. (1999). Macrostructure construction and organization in the processing of multiple text passages. *Instructional Science*, 27(3-4), 221-233.
<https://doi.org/10.1007/BF00897320>
- Leow, R.P. & Morgan-Short, K. 2004: To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition 26*(1), 35–57.
<http://www.jstor.org/stable/44486713>
- McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: A mixed methods study. *Contemporary Educational Psychology*, 35(4), 229–241.
<https://doi.org/10.1016/j.cedpsych.2009.12.001>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13-103). New York: Macmillan.
- O'Sullivan, B., & Weir, C. J. (2011). *Test development and validation. Language testing: theories and practices*. London: Palgrave Macmillan.
- Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Towards a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels. *TOEFL Monograph (21)*. Princeton, NJ: Educational Testing Service.
- Rouet, J.-F. (2006). *The skills of document use: From text comprehension to web-based learning*. Mahwah, NJ: Erlbaum.
- Rouet, J. F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In G. Schraw, M. T. McCrudden, & J. P. Magliano (Eds), *Text Relevance and Learning from Text* (pp. 19-52). Charlotte, NC: Information Age Publishing, Inc.

- Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Cambridge monographs and texts in applied psycholinguistics. Advances in applied psycholinguistics, Vol. 1. Disorders of first-language development; Vol. 2. Reading, writing, and language learning* (pp. 142-175). New York, NY, US: Cambridge University Press.
- Strømsø, H. I., & Bråten, I. (2014). Students' sourcing while reading and writing from multiple web documents. *Nordic Journal of Digital Literacy, 2014(2)*, 92–111.
- Strømsø, H. I., Bråten, I., Britt, M. A., & Ferguson, L. E. (2013). Spontaneous sourcing among students reading multiple documents. *Cognition and Instruction, 31(2)*, 176–203.
<http://dx.doi.org/10.1080/07370008.2013.769994>
- Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research & Method in Education, 29(2)*, 185-208.
<https://doi.org/10.1080/17437270600891614>
- Taylor, L. (2013). *Testing reading through summary: Investigating summary completion tasks for assessing reading comprehension ability*. Cambridge: UCLES/Cambridge University Press.
- Trinity College London. (2017) *Integrated Skills Examination II*. Croydon, England: Trinity College London. Retrieved from
<http://www.trinitycollege.com/site/?id=3195>
- Urquhart, S., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. New York: Longman.
- Ünaldi, A. (2010). *Investigating reading for academic purposes: sentence, text, and multiple texts*. (Unpublished doctoral dissertation). University of Bedfordshire, Bedfordshire, England.
- Van Den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behavior & Information Technology, 22(5)*, 339-351.
<http://dx.doi.org/10.1080/0044929031000>
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York : Academic Press.
- Ozuru, Y., Best, R., & McNamara, D. S. (2004). Contribution of reading skill to learning from expository texts. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp.1071-1076) Mahwah, NJ: Lawrence Erlbaum Associates.

- Weir, C., & Shaw, S., (2005). Establishing the validity of Cambridge ESOL writing tests: towards the implementation of a socio-cognitive model for test validation. *University of Cambridge ESOL Examinations Research Notes (21)*, 10-14.
- Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C., Ash, I. K., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal*, 46(4), 1060–1106. <https://doi.org/10.3102/0002831209333183>
- Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, 83, 73–87. <http://dx.doi.org/10.1037/0022-0663.83.1.73>
- Zainal, A. (2012). Validation of an ESL writing test in a Malaysian secondary school context. *Assessing Writing*, 17(1), 1-17. <https://doi.org/10.1016/j.asw.2011.08.002>