Investigating cognitive processes in different item formats in reading tests through eye-tracking and verbal protocols

Hatice Akgün
School of Foreign Languages, Marmara University, Istanbul, Turkey

Aylin Ünaldı
School of Education and Professional Development, University of Huddersfield, Huddersfield, UK

This study has investigated the differences in cognitive processes that test-takers undergo while answering reading comprehension questions in multiple-choice and open-ended short answer formats. For this purpose, data were collected from a group of undergraduate students in an English medium university through eye-tracking technology, immediate retrospective verbal protocols, and short semi-structured interviews. The results showed that the participants used careful reading skills more and comprehended the text more thoroughly in the open-ended format. However, in the MC format, they read less carefully and used more test-taking strategies. These findings contribute to the ongoing discussion on how item format can alter the cognitive processes in a reading comprehension test and confirm the effectiveness of eye-tracking in unveiling cognitive processes in combination with qualitative methods. This study has implications for reading test development.

Key words*:* Eye-tracking in reading, reading processes, item format effect in reading tests, cognitive validity in reading tests

Introduction

Reading comprehension may involve a combination of a variety of skills such as

Email address for correspondence: akgunhatice88@gmail.com

expeditious and careful reading at global or local text spans with processes at lower (i.e. word recognition, lexical access, etc.) and higher cognitive levels (i.e. inferencing, building a mental model) (Khalifa & Weir, 2009). Ideally, in designing reading tests, items are developed to operationalize one or more of these reading skills. The congruence between these intended skills and the actual reading processes a test-taker uses in responding to items is a strong construct validity evidence supporting the inferences that can be made on the score from that test (Messick, 1989).

Regarding test validity, Cohen and Upton (2006) suggest that test-taking strategies should be a part of validity arguments as well. The use of test-taking strategies is generally induced by item format (Sarnaki,1979) because the features of a specific item format can trigger unintended or irrelevant skill or strategy use, which may undermine the construct validity of the test. This implies that cognitive processes need to be investigated with reference to item format. The most widely used item formats in reading comprehension tests are multiple-choice (MC) and short answer open-ended (OE) questions. Previous studies have indicated that MC questions in reading tests can alter normal reading processes, cause segmented reading or trigger unintended item-specific processes which threaten construct validity (Martinez, 1999; Rupp et al. 2006; Ozuru et al., 2013; Lim, 2014). Cognitive processes that are triggered by certain item types during reading tests have been investigated in certain studies (Rupp et al., 2006; Cohen & Upton, 2006; Bax & Weir, 2012; Bax, 2013; Lim, 2014). Verbal protocols and interviews have been used extensively in such research. While eye-tracking technology has also been used to investigate various forms of reading processes, its use in investigating differences in cognitive processes that result from different item formats is scarce. Using eye-tracking technology accompanied by qualitative data from verbal protocols can provide valuable in-depth data to analyse the different dynamics of MC and OE questions. However, such studies should be done with care by balancing the items in the type and range of cognitive processes they may activate. Items in both formats must be comparable to each other, not only in terms of question stems but also the depth of processing (e.g. factual versus inferencing questions).

The current study aims to contribute to the discussion on the item format effect in reading tests by making a systematic comparison of test-takers' reading processes in responding to carefully-balanced MC and OE items. The study combines data from

immediate retrospective verbal reports, semi-structured interviews, and also eye-tracking technology to seek a deeper insight into the cognitive operations triggered by different item formats.

Investigating cognitive validity in reading tests

A comprehensive cognitive processing model, which accounts for readers' use of skills, strategies, and other related processes, is suggested by Khalifa and Weir (2009) for the investigation of reading constructs. The framework characterizes reading as taking place at the global versus local level and as being careful or expeditious in nature. In careful reading, the goal is to get complete meanings from the presented material and the approach in this reading type is "slow, careful, linear, incremental reading for comprehension" (Khalifa & Weir, 2009, p.46). Expeditious reading includes quick, selective, and efficient reading in order to find out targeted information. This type of reading consists of scanning, skimming, and search reading. In this model, the central core consists of lower-level processes (i.e. word recognition, syntactic parsing, etc.) and higher-level processes (i.e. inferencing, building a mental model, etc).

Khalifa and Weir (2009) further underline that a reader's perception of task demands determines the purpose of reading for the reader and affects the type of skills, strategies, and processes through metacognitive mechanisms of goal setting and monitoring. The congruence between the processes elicited by reading test items and real-life tasks is seen as *cognitive validity* evidence (Weir, 2005; Field, 2012). Khalifa and Weir's (2009) cognitive processing model of reading is used as a reference in this study.

As discussed above, test-taking strategies should also be investigated in test validation as some test-taking strategies triggered by certain item formats can undermine validity in reading tests because the use of test-taking strategies, particularly test-wiseness strategies which include the use of testing format, can help test-takers respond to items without engaging in the intended processes of the item (Cohen, 2013). For this reason, test-taking strategies will be explored in this study as well.

Item format effect

The discussion on strategy use inevitably leads to the item format effect as the use of

test-taking strategies is evidently related to the item format effect (Sarnaki, 1979). MC and short answer OE item formats are the most widely used formats in the assessment of reading comprehension. In fact, the MC format is more commonly preferred for a number of reasons such as practicality in rating, rating reliability, and wide content coverage. However, its effectiveness over OE items has been questioned. For instance, Martinez (1999) asserts that the MC format elicits low-level cognitive processing while more complex thinking is required by OE formats. Rupp et al. (2006) underline that reading to respond to MC questions is never a linear process but more of a segmentalized and localised process including heavy scanning for keywords. Where we expect an intensive careful reading as a response to an item, this might heavily undermine the cognitive validity claims. Also, Ozuru et al. (2013) conclude that the MC and OE format measure different aspects of the reading comprehension process as in their study, test-takers' performance in MC correlated more strongly with topic-specific prior knowledge. Lim's (2014) study also shows that test-takers may try to cope with the demands of MC items through the use of extra mental, item-responding processes along with genuine reading processes that the task intends to elicit. These studies evidence extensive interest in investigating the reading processes activated by certain item formats. However, except for Lim (2014), who used a mixed-method design, other studies used only qualitative methods for data collection such as cognitive interviews, self-explanation protocols, and concurrent or retrospective verbal reports (e.g. Cohen & Upton, 2006; Ozuru et al., 2013; Rupp et al., 2006). Although the value of such qualitative methods cannot be denied, they have been criticised for relying on participants' memory too much as participants may omit, forget or add information. Such methods as concurrent verbal reports are also suspected of distorting the normal reading processes (Cohen, 1998). Therefore, format effect studies can benefit from the non-intrusive method of eye-tracking technology to have a deeper insight into real-time processing. The triangulation of such data with qualitative data enhances the validity of the studies into the reading processes triggered by different item formats.

Eye-tracking research in cognitive processes in reading

Staub and Rayner (2007) assert that eye movements can "provide a moment-to-moment indicator of the ease (or the difficulty) with which readers are able to comprehend the text that they read" (p.327). Eye movements are capable of giving

online information about human cognition as a strong link between eye movements and cognition is assumed to exist (Rayner et al., 2005). Eye-tracking technology has been extensively used in first language reading and second language reading assessment research.

For example, McCray and Brunfaut (2018) investigated 28 test-takers' cognitive processing while responding to six banked gap-fill tasks designed to measure reading ability using an eye tracker. The study examined types of processing undertaken by high and low achievers based on the reading model developed by Khalifa and Weir (2009). Nine measures were used to analyse eye-tracking data such as mean fixation time on task, test, gap, word bank and number of visits on word bank. The study showed that there was a difference in the processing of low and high achievers; low achievers' increased attention to the words surrounding the gaps attested localised reading and also lower-level processing. Low achievers were also found to visit the word bank more frequently, and this was taken as evidence of their non-linear reading process.

There are also studies that used eye-tracking technology and verbal reports in a complementary way to investigate reading processes. For example, Bax (2013) investigated test-takers' cognitive processes while completing onscreen IELTS reading test items through the use of eye-tracking and stimulated recall interviews with a focus on the difference between successful and unsuccessful test-takers' cognitive and metacognitive processing. The study focused specifically on the analysis of careful and expeditious local reading as defined in Khalifa and Weir's (2009) model. The results showed significant differences in terms of expeditious reading. Unsuccessful students could not effectively read expeditiously to locate information and they spent more time looking for information. However, successful students showed greater success in locating the correct paragraph and focusing on key information.

In another study, Brunfaut and McCray (2015) investigated the cognitive processing of 25 ESL test-takers on computer-based Aptis reading tasks through a combination of eye-tracking and retrospective verbal reports with eye-tracking traces used as recall-enhancing stimuli. Eye-tracking data were analysed according to 11 eye-tracking metrics related to fixations, saccades, and regressions, and three processing-type groups, i.e., global processing, text processing, and task processing. Based on Khalifa

and Weir's (2009) model, test-takers were found to engage in a wide range of cognitive processes including both higher- and lower-level processes, which proved that Aptis reading component sampled extensively from the reading construct. However, for B1 tasks, there was a discrepancy between intended cognitive processes and elicited processes; this was taken as a risk of construct-irrelevant variance. The authors concluded that data from the two methods mutually confirmed each other; however, eye-tracking data provided more insights into lower-level reading processes while verbal reports were more helpful in understanding higher-level processes such as inferences.

In a follow-up study, Brunfaut (2016) investigated 25 ESL test-takers' cognitive processing on B1 level opinion matching tasks. The data were collected and analysed in the same manner as in Brunfaut and McCray (2015). The results showed that this task elicited both higher- and lower-level processes based on Khalifa and Weir's (2009) reading model. There was an alignment between intended and elicited processes with no obvious risks for construct-irrelevant variance.

Of specific relevance to the present study in item format comparison, Lim (2014) focused on item format effects by combining verbal protocol and eye-tracking technology: Lim investigated the extent to which item format affects test-takers' scores and reading processes. Two comparable reading texts were chosen from the TOEFL iBT test practice volume and two open-ended versions were created for two multiple-choice testlets. The eye movements of the participants were recorded while they were answering questions in OE and MC formats. Afterward, a recall interview was administered for only three items in the MC format. The predetermined criteria (Bax & Weir, 2012) were used to identify reading types. Regarding the test format, it was found out that MC questions were easier than OE questions. Total fixation duration on question stem, first paragraph, key sentence, and key phrase was longer in the OE format than MC in the data of the test-takers who got full scores in both formats. Hence, it was concluded that test-takers paid more attention to the key information in the OE format. Also, vocabulary items were solved without looking at the text but rather with the prior knowledge of collocations or adjacent words in MC questions. It was concluded that these vocabulary items did not measure inference abilities as intended, but rather vocabulary size, which undermined the cognitive validity of this

item type. As to the effect of item format, Lim (2014) concluded that tapping into true reading ability was problematic with indirect testing methods as indirect items required extra mental processes. One limitation reported by the author was that stimulated recall interviews were not systematically incorporated in the study as verbal reports were elicited on only three items in the MC format. The author suggested that more qualitative data should complement eye-movement data for a better understanding.

In short, there have been studies which aim to elucidate the contribution of eye-tracking technology in understanding the cognitive processes activated by different tests in judging the cognitive validity of reading tests. In terms of comparing cognitive processes in different item formats, quantitative and qualitative data should systematically complement each other. However, as indicated by Lim (2014), there is a need for a more systematic analysis.

The present study aims to expand this line of research by equating items focusing on their comparability in terms of question stem, content, difficulty, and the type/range of cognitive processes required as well as the comparability of text complexity in the tests. Besides, verbal protocols were collected on all the items to complement eye-tracking data for a full systematic comparison.

Research questions

In order to find out the differences in cognitive processes test-takers use in responding to two different item formats (multiple-choice and open-ended) in reading comprehension tests, the following research questions have been formulated:

(1) Do the reading scores change across parallel item formats?

(2) How do the cognitive processes of participants differ in MC and OE formats at test level analysis?

(3) How do the cognitive processes of participants differ in MC and OE formats at item-level analysis?

Method

Participants

The study was conducted at an English-medium state university in Turkey following all the necessary ethical requirements required by Boğaziçi University. An invitation e-mail was sent to different departments at the university and all the participants took part in the study on a voluntary basis. The data were obtained from 34 participants (30 females and four males) in verbal reports; however, data from three participants had to be eliminated from the eye-tracking analysis due to technical problems. The participants consisted of undergraduate students (16 sophomores and 18 juniors) studying in different departments. All the participants who had passed the university's English proficiency exam reported to correspond to 79 on TOEFL iBT and 6.5 on IELTS Academic on the university's website. Therefore, the language level of the participants was minimum upper-intermediate (B2 on the CEFR).

Instruments

The study made use of two reading texts adapted from TOEFL iBT practice tests chosen from The Official Guide to TOEFL® Test Third Edition (2009). The features of these reading texts are given in Table 1.

Table 1. Text comparison analysis

	Text 1	**Text 2**
Genre	Expository	Expository
Title	The Origins of Cetaceans	Swimming Machines
Word Count	638	631
Flesch Reading Ease Score	56	60
Flesch-Kincaid Grade Level	9.2	8.5
Coh-metrix L2 Readability	11	11
SMOG	8.7	8
K1+K2 Word%	78.1	80.7
AWL Percentage	4.05	4.44
Type and Token Ratio	0.43	0.45
Lexical Density	0.6	0.58
Narrativity	14	22
Syntactic Simplicity	80.7	68.7

Word Concreteness	86	85
Referential Cohesion	22.3	20
Deep Cohesion	42	55

Two reading passages were chosen as comparable after various textual features were analysed through the automatic text analysis tool *Coh-Metrix*, the vocabulary analysis tool *Compleat Lexical Tutor* and readability statistics. As can be seen in Table 1, the texts were quite comparable to each other in terms of their topics, length, lexical features, syntactic features, and readability. The two texts from the TOEFL iBT practice volume had their MC questions. For each MC item, an OE version was created only for the MC questions that could be converted into an OE format. This was done by carefully keeping equal the question stem (stem equivalence), the intended type of reading process, and the difficulty level of the items across MC and OE versions. The questions that could not be transformed into the OE format such as negative factual information, sentence simplification questions, and insert text questions (i.e. cohesion questions) were eliminated. Using a cognitive and contextual proforma adapted from Wu (2011), five university English language teachers evaluated the reading texts in terms of genre, rhetorical organization, cultural and content specificity, and text abstractness. The teachers evaluated the items' comparability in terms of content, explicitness, difficulty, and the text span that may need to be processed for generating an answer. The texts and items were revised based on their feedback to ensure test comparability.

A pilot study of the OE test was conducted on a group of B2 level EAP students (Text 1: 62 students; Text 2: 55 students) at a public university. The items which had low reliability were eliminated in the OE test and their equivalents were also taken out from the MC version (Cronbach's Alpha was .97 for Text 1 and .72 for Text 2). Finally, there were two different texts (Text 1 and Text 2) and each text had OE and MC forms, yielding four different versions. In each test, there were six questions.

These *a priori* qualitative and quantitative analyses helped to establish text and item equivalence in a systematic way. The item types and their counterparts are summarized in Appendix 1. OE tests and their answer keys are provided in Appendices 2, 3, and 4 respectively.

Procedures

The eye-tracker available for this study was Tobii x1 Light. This eye-tracker has a variable data rate which is typically between 28 and 32 Hz, i.e. typically between 20 to 32 data samples are collected per second for each eye. The minimum fixation duration for the current study was set at 100 ms (Staub & Rayner, 2007), and the area up to 5 degrees of visual angle was accepted to be within the perception at the time of a fixation.

Data were collected individually at Boğaziçi University Vision Lab. First, the participants completed informed consent forms. Then, four forms of tasks were counterbalanced in each condition. Half of the participants took Text 1 MC and Text 2 OE and the rest of them were given Text 2 MC and Text 1 OE. There was no time restriction and the time used by the participants ranged from 60 to 90 minutes. In all the tests, the text appeared on the left side of the screen and the questions appeared on the right side (see Appendix 2).

Before the data collection, the participants performed a 5-point calibration and completed a training task on eye-tracking and verbal report processes. The experiment started with eye-tracking recording when the participants responded to the tests. After this, they were asked to give an item-by-item verbal report on how they read the text and answered each question.

This was followed by a semi-structured interview that lasted approximately five minutes. During the interviews, the participants were prompted to comment on the issues such as the difficulty of the tests, how much they comprehended, and their overall preferences for a specific item format. All verbal reports and interviews were audio-recorded and transcribed.

Data analysis

For the first research question, scoring was done using answer keys. As MC tests were taken from TOEFL practice volume, they had the answer keys. For OE items, the answer keys were prepared based on the piloting results and expert teacher opinions. There were no partial credits; correct answers were scored as “1”, and incorrect answers “0” by strictly adhering to the prepared key. Total scores were converted into

percentages. Descriptive statistics were calculated for the scores for both methods and both texts. In order to investigate whether the variance in the scores is affected by test method or text effect, a two-way analysis of variance (ANOVA) was conducted with score percentages as the dependent variable and the method (MC or OE) and the text (Text 1 and Text 2) as categorical independent variables.

The second research question was answered with the help of eye-tracking, verbal reports and interview data. In the eye-tracking process, careful reading was operationalized as a minimum of three fixations on a sentence as suggested in the predetermined criteria by Bax and Weir (2012). The percentages of careful reading in MC and OE formats were then compared by the Mann-Whitney U test. In addition, text-based *total reading time* (TRT: the sum of all fixations within the target area) (Rayner et al., 2006) and text-based *total fixation count* (TFC: how many times the target area was fixated on or visited) were calculated. TRT and TFC counts were compared between two formats by using the Mann-Whitney U test. In addition, four representative gaze plots from two participants illustrating the different reading processes in MC and OE were analyzed. In the verbal report data, three types of strategy use were investigated: reading strategies, test-management strategies, and test-wiseness strategies. Firstly, based on the literature on reading skills and test-taking strategy use (Cohen & Upton, 2006; Ünaldı, 2004; Lim, 2014), coding rubrics for each type of strategy use were developed. Following this, the audio recordings of verbal protocols from MC and OE test-taking sessions were transcribed and coded by the two researchers separately and discrepancies were resolved through discussion. The rubrics were revised recursively based on the new categories that emerged from the data until the final form was reached. This required four cycles of coding in total. In the last coding, there were 1096 codes in OE with 83% of consistency and 1140 codes in the MC format with 81% of consistency in coding (See Akgün, 2018 for the details). The frequencies for each strategy use in the coding rubrics were calculated and compared between the two formats through paired samples t-tests. Lastly, the interview data were qualitatively categorized into three emerging themes: (a) participants' perceptions on which format was more difficult, (b) participants' perceptions on which item format was more helpful in their comprehension of the texts, and (c) whether they had a specific preference for MC or OE questions and the reasons.

The third research question was also investigated with the help of eye-tracking, verbal reports and interview data. For the eye-movement data, areas of interest (AoI) were identified for each question. AoI was defined as "specific words or parts of the text that a participant did or did not read carefully" (Jarodzka & Brand-Gruwel, 2017, p.195). Then, the participants' eye movements on each MC item were compared to their OE counterparts in terms of TRT and TFC in AoIs and the question stems (QS) (see Appendix 1 for parallel items). The eye movement data were then submitted to Mann-Whitney U test for item-level comparison. For the verbal report data, the most frequently used strategies were reported for each item type in the MC format and then, they were compared with the most frequently used strategies in the parallel OE items. Lastly, from the interview data, the participants' own explanations and justifications were included to complement eye-tracking and verbal report data in the item-level analysis.

Results

RQ 1: Do the reading scores change across parallel item formats?

The results of the descriptive statistics for both methods and for both texts are presented in Table 2, indicating that the mean of all OE questions (75.41%) is higher than the mean of all MC questions (60.82%). The average performance on each text, however, is not so different. Text 1 has received a slightly lower mean (65.68%) than Text 2 (70.56%) when MC and OE questions are combined in the calculation.

Table 2. Descriptive statistics for method and text by percent

	MC (Text 1&2)	OE (Text 1&2)	Text 1 (MC&OE)	Text 2 (MC&OE)
Mean	60.8	75.4	65.6	70.5
Median	67	83	67	67
Std. Dev.	21.2	18	25.2	15.8
Skewness	-0.5	-0.6	-0.5	-0.3
Kurtosis	1.1	0	0	-0.3
Variance	451.4	338.7	637.4	250.1

The results of ANOVA in Table 3 show that the difference between the mean scores of the two tasks is due to the test method effect. The effect for the method is significant: $F\,(1, 64) = 5.394$, $p < .05$. The result is not affected by texts used in the tests.

Table 3. Results of two-way ANOVA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5993.059	3	1997.6	5.394	.002
Intercept	315520.9	1	315521	851.9	.000
Method	3617.8	1	3617.8	9.76	.003
Text	405.2	1	405.2	1.094	.299
Method*Text	1969.9	1	1969.9	5.31	.024
Error	23702	64	370.3		
Total	345216	68			
Corrected Total	29695.06	67			

RQ 2: How do the cognitive processes of participants differ in MC and OE formats at test level analysis?

The results of eye-tracking data show that there is a statistically significant difference between MC and OE formats in terms of careful reading time, total reading time, and total fixation count, as is summarized in Table 4. In the OE format, the whole text is read for a longer time; there are more fixations in the overall text and a greater portion of the text is processed through careful reading.

Table 4. *Results for text-based careful reading, TRT, and TFC*

		Text-based Careful Reading	Text-based TRT	Text-based TFC
MC (N=31)	Mean	75.7	200.4	503.3
	St. Dev.	9.5	64.1	114.9
OE (N=31)	Mean	80.3	243.5	613
	St. Dev.	12.2	78.5	158.4
Mann Whitney U		302.5	320	282.5
Z		-2.5	-2.2	-2.7
Sig (2-tailed)		.012	.024	.005

To illustrate text-based careful reading in gaze plots and to visualize how reading patterns might change for the same participant in different formats, four representative sample gaze plots are provided here from two different participants for both MC and OE formats (see Figures 1 to 4). First, each participant's gaze plots for MC

and OE are provided respectively, followed by an explanation. For Participant 12, as can be seen in Figures 1 and 2, the red circles represent eye fixations on each part of the text. Smaller ones show shorter fixations while larger ones display longer fixations and they give a picture of the areas to which the reader gave the most visual attention. Larger and continuous red circles in Figure 2 indicate the use of careful reading because there is a "slow, careful, linear, incremental reading for comprehension" suggested by Khalifa and Weir (2009, p.46). On the other hand, in Figure 1, there are non-continuous, fewer, and shorter fixations, which suggests the presence of expeditious reading which is "quick and selective" in nature (Khalifa & Weir, 2009). It should be noted that Participant 12 answered all OE questions correctly but the participant answered only half of the MC questions correctly.

Figure 1. Gaze plot from Participant 12 in MC

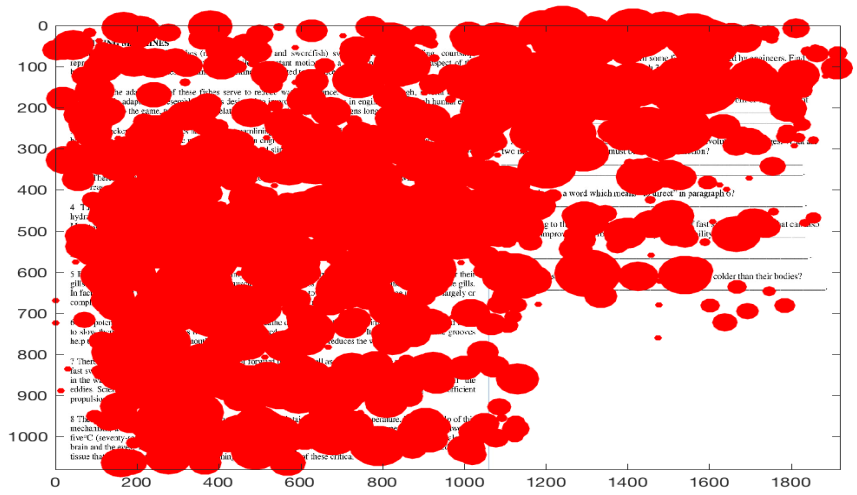

Figure 2. Gaze plot from Participant 12 in OE

When gaze plots from Participants 7 and 12 are compared, the same pattern can be observed in Figures 3 and 4 as in Figures 1 and 2 in terms of the reading skills in different item formats. Compared to Figure 3, Figure 4 suggests careful reading which is more linear and less fragmented than Figure 3. It should be noted that Participant 7 answered all questions correctly in both formats but with different reading types as implied by the gaze plots.

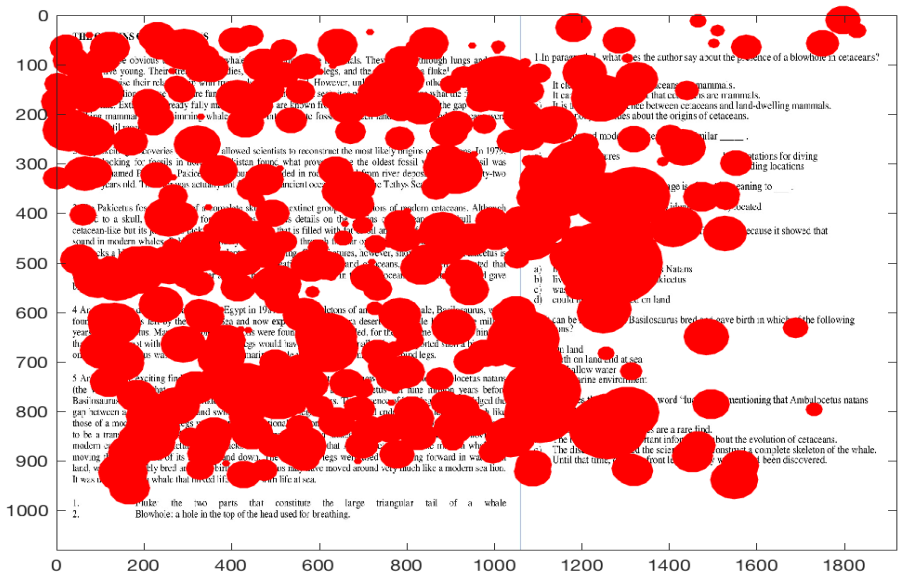

Figure 3. Gaze plot from Participant 7 in MC

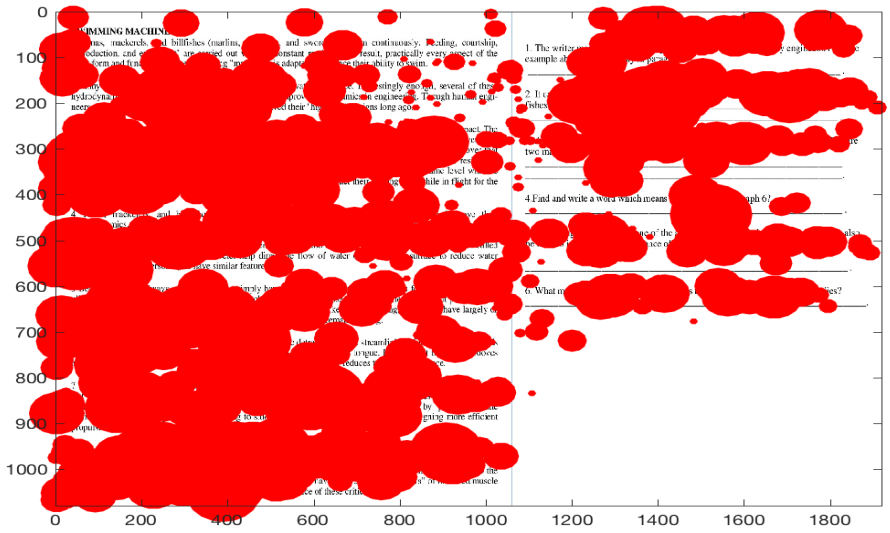

Figure 4. Gaze plot from Participant 7 in OE

In order to understand which overall reading skills and strategies are used in different formats, the frequencies from the verbal report data were calculated based on rubric categories, compared, and presented in Table 5.

Table 5. The frequency of overall reading skills and strategies

	MC	MC	OE	OE
	Count	**Percentage**	**Count**	**Percentage**
Prior to test-taking				
R1 Reads the whole text carefully before the test	22	3.0	24	2.7
R2 Reads the whole text quickly before the test	0		0	
Expeditious Reading Skills				
R3 Scanning	119	16.5	120	13.6
R4+R6 Search Reading	147	20.4	199	44.1
R5 Skimming	33	4.6	37	4.2
Total	299	41.6	356	40.4
Other reading strategies				
R8 Reading only parts that seem relevant to question	254	35.3	296	33.6
Careful reading Skills				
R9 Focusing on parts of a sentence	29	4.0	42	4.7
R10 Reading carefully across sentences	54	7.5	87	9.8
R11 Creating a textual representation	2	.2	3	.3
R14 Rereading important parts	29	4.0	30	3.4
R7 Making inferences based on the text	29	4.0	43	4.8
Total	143	19.9	205	23.2

First, it can be seen in Table 5 that more careful and expeditious reading skills are used in OE items. The results of paired samples t-tests which were conducted to compare the differences between the means of overall reading skill use also support that there is a significant difference in OE (*M*= 80.09, *SD*= 86.10) and MC (*M*= 65.27, SD= 77.18) conditions: *t*(10)= -3.306, *p*= .008. However, the effect size (*d* = .18) for this analysis was small. This finding indicates that these reading operations are only slightly more deployed in the OE format.

As to the test-management and test-wiseness strategies, the frequencies are summarized in Tables 6 and 7, respectively. Results showed that 43.49% of test-management strategies in MC items are option-related. The paired samples t-test indicated that there is a significant difference in the test-taking strategy use in OE (*M*=11.38, *SD*= 17.98) and MC (*M* = 21.10, *SD* = 20.96) conditions: *t* (20) = 2.159, *p* = .04. The effect size *(d* = .49) for this analysis was found to be moderate. Finally, it is clear in Table 7 that the participants attempt to use test-wiseness strategies more in MC than OE questions.

Table 6. Frequencies of test-management strategies

	MC	MC	OE	OE
	Count	**%**	**Count**	**%**
T1 Rereading question for clarification	42	9.7	66	28.0
T4 Reading question and options before the text	32	7.4	0	
T5 Skipping a difficult question	13	3.0	18	7.6
T6 Using the order of questions as a clue	64	14.8	53	22.5
T8 Producing answer after reading the text	42	9.7	22	9.3
T11 Identifying an option with unknown vocabulary	5	1.1	3	1.2
T14 Selecting preliminary options with uncertainty	34	7.9	4	1.7
T17 Eliminating similar options	2	.4	0	
T18 Wrestling with option meaning	3	.7	0	
T19 Making an educated guess	33	7.6	4	1.7
T22 Selecting options based on background knowledge	6	1.4	0	
T23 Selecting options based on paragraph meaning	70	16.2	0	
T29 Expressing uncertainty at the correctness of an answer	23	5.3	14	5.9
T30 Stopping reading the options when the answer is found	35	8.1	0	

T31 Receiving clues from other items	6	1.4	9	3.8
T35 Continuing to read the text when the answer is found	6	1.4	19	8.0
T36 Stopping reading the text when the answer is found	14	3.2	23	9.7

Table 7. Frequencies of test-wiseness strategies

	MC	OE
	Count	**Count**
TW1 Selecting an option out of a vague sense, even if it is not understood	3	0
TW2 Using clues in other items to answer the item under consideration	2	2
TW3 Selecting an option as it has a keyword/phrase from the passage	8	0
TW4 Chooses a phrase as an answer which is in the same sentence as the keyword	0	2

Based on the interview data, the participants' own overall comments on MC and OE tests were classified with frequencies presented in Table 8. During the interviews, the participants indicated that they comprehended the text in the OE format more because they had to read and comprehend it more carefully to produce their own answers with their own words. This made them read the text in a more careful and detailed way and they even had to reread some parts. Some participants claimed that even though the OE test was more difficult, they understood the texts better and answered the questions with more confidence. Most of the participants explained that answering OE was easier as that format did not confuse them during their reading process and they enjoyed it more in general. Some claimed that they preferred the OE format more because in the MC format, even if they understood the related part in the text and knew the answer, they had difficulty matching their own correct answer with the test writer's interpretation of the text in the correct option. In addition, when they thought they found the correct answer and stopped reading the options, they could choose the wrong answer as a result of very subtle differences among the options or because of the tricky nature of the options.

Table 8. Percentages of supporting comments from the interviews

	MC	OE
Difficulty	43.7	43.7
Comprehension	6.2	87.5
Overall Preference	37.5	53.1

The results from these three types of data confirmed and complemented each other. The eye-tracking data showed that the text in the OE format was read in a longer time, more carefully, and fixated on more. Similarly, the gaze plots illustrated that the reading process in the OE format was more linear and incremental, and it involved the processing of larger text spans compared to the processed text spans in the MC format. In the verbal report data, the participants reported to have used more reading strategies in the OE format and read the text more carefully while they made use of more test-taking strategies, especially option-related strategies in the MC format. The data from the interviews strongly supported the above findings of eye-tracking and verbal reports and indicated that in order to comprehend the text and find the right answers, the participants had to read the text in the OE format more carefully.

RQ 3: How do the cognitive processes of participants differ in MC and OE formats at item level analysis?

The detailed results of descriptive statistics of item-based eye movement data are given in Appendix 5, with the results of Mann-Whitney U test summarized in Table 9. This table shows which item format (OE or MC) in each item type (see Appendix 1 for item types) scores higher in terms of the eye-tracking measure category (TFC and TRT) by the text part (QS and AoI). For example, in Item 1 (factual-local), the combined scores from OE items from Text 1 and Text 2 (Q1 and Q6) are higher than those from MC items in terms of all measures.

Table 9. Summary of overall item type-based differences

ITEM TYPE	TEXT SPAN	QS TFC	QS TRT	AoI TFC	AoI TRT
1. Factual	Local	OE	OE	OE *	OE *
2. Factual	Global	OE	OE	OE *	OE *
3.Vocabulary	Global	OE *	OE	OE *	OE *
4. Factual	Global	MC	MC *	OE	OE
5. Inference	Global	OE *	OE *	MC	MC

6. Factual	Global	MC	MC	MC *	MC*

Notes. * The difference is statistically significant at $p < .05$.

The overall results in Table 9 show that the two measures, TFC and TRT, of areas of interest and question stems are higher in the OE format in 16 out of 24 cases in total, and in 9 cases, this difference is statistically significant. This indicates that the relevant areas in the text and the question stems are usually read for a longer time and visited more frequently in the OE format.

In the item-based analysis of the verbal report data, the percentages for the most frequently used skills and strategies were calculated. Tables 10 and 11 below list the top five reading skills utilised in each format on an item basis.

Table 10. Percentages of the most frequent skills and strategies for MC items

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6
R8 Reading only the parts that seem relevant to question	29	22	24	28	24	33
R6 Search reading	20	15		14	6	19
R3 Scanning	12	11		7	9	12
T19 Making an educated guess			20			
T23 Selecting options based on paragraph meaning	8			5		6

Table 11. Percentages of the most frequent skills and strategies for OE items

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6
R8 Reading only the parts that seem relevant to question	27	26	35	27	33	23
R6 Search reading	20	19	8	15	7	15
R10 Reading carefully across sentences	8	9	10	7	13	10
R3 Scanning	9	12		15		17
T1 Rereading question for clarification					7	7

As can be seen in Tables 10 and 11, the most frequently used reading skills in both formats were "*R8"* and "*R6"*. "*R3"* is the third most frequently used skill in MC, but it is "*R10"* in all OE items. As to the test-taking strategies, "*T19:making an educated guess"* was frequently used in the MC vocabulary question. Also, *"T23:option elimination based on an overall understanding of the paragraph"*, was among the top

test-taking strategies in the MC format. On the other hand, the most frequently used test-taking strategy in OE was *"T1:rereading questions for clarification"*.

When the results are examined in an item-based manner, it can be seen that in Item 1 (sentence-level factual reading) and Items 2, 4, and 6 (paragraph/across paragraphs factual reading), search reading, scanning, and option elimination were reported in the MC format while in the OE format, search reading and scanning were followed by careful reading. This implies that after reading expeditiously to locate the key information in both formats, the participants read the relevant area more carefully in the OE format while they spent more time eliminating the options in the MC format. This is also confirmed by the eye-tracking data as the participants spent more time reading the AoIs in the OE items except for Item 6. In line with the eye-tracking and verbal report data, the interview data also suggested that the participants spent more time on the options instead of the text. During the interviews, the participants who found the MC format difficult mostly mentioned that the distractors confused or misled them. Some also stated that they were more passive while answering MC questions as they could answer some questions without fully understanding the text thanks to the options. The participants who regarded the OE format as more difficult claimed that options in MC helped them have an overall idea about the question intent or they could find an answer by matching keywords in the text and options, so they could spend less time reading the text. The participants who preferred MC items generally considered this format as practical to answer as they could find clues or identify keywords from the options.

For Item 3, which is a vocabulary question, when the strategies used in MC and OE formats are compared, it is clear that the participants used the guessing strategy in the MC format while in the OE format, they reported search reading and careful reading. Eye-movement data confirm this finding as they spent more time reading the question stem and AoIs in the OE format.

For Item 5, which is an inference question, the participants reported to have read carefully and reread the question stem in the OE format along with search reading. The eye-movement data showed that the participants read the question stem for a longer time in the OE format, thus confirming the verbal reports. On the other hand, the participants reported to have used search reading and scanning in the MC format.

However, they read the AoIs in the text for a longer time in the MC format, but the difference is statistically non-significant.

Discussion & Conclusion

The three types of data are successfully consolidated in this study to show the different types of cognitive processing that have been triggered by reading comprehension item formats. Below we discuss our observations on OE and MC item formats with reference to these three types of data which confirm and complement each other.

Open-ended format

The findings of this study are in line with previous research done on format comparison (e.g. Lim, 2014). Both eye-tracking and the verbal report data show that in the OE format, the text is read more carefully and more time is devoted to reading the text. The gaze plots also show that for several question types, the text span that needs to be processed in an OE question is usually larger. The fact that the participants spent more time reading the text in OE as suggested by eye movement data might imply that either there is difficulty in comprehension, or greater attention is given to higher-level processes in reading (i.e. inferencing, building a mental model of the text, etc.). The interviews and higher scores in the OE format indicate that the participants did not have more difficulty with the OE format and longer time spent here implies careful reading and a focus on comprehension. While MC items were found to be easier in Lim (2014), in the current study, MC and OE items were perceived to be equally difficult by the participants but still, they comprehended the text in the OE format better due to careful reading processes.

As to the reading skills yielded by verbal report data, it is obvious that both formats triggered expeditious reading skills to find the answers, as is expected in a test-taking condition. However, the participants engaged in more careful reading at both local and global levels in the OE format. Recall that in this study the participants were allowed to read without time pressure and in the manner they preferred. The time they spent on OE items should be devoted to detailed reading and re-reading of larger text spans to be able to formulate an answer, which could not be done only through expeditious reading. On the other hand, although the participants were not under time pressure,

they still preferred fast and selective reading to answer MC items as this is what the MC format leads test-takers to do.

In addition, item-based verbal report results indicate that in every OE item, "*R10: making connections while reading carefully across sentences"* was one of the most frequently used skills and this implies that the participants engaged in more complex thinking in the OE format (Martinez, 1999) unlike in the MC one, which is found to be more conducive to expeditious reading skills in this study. As a result, it can be concluded that the OE test may demand deeper cognitive processing (Lim, 2014). The verbal report data also showed that certain test-management strategies are used differently across formats. In OE items, strategies that involved careful reading were used more frequently. These included rereading or translating the question for clarification and continuing reading to make sure that they had found the correct answer. Cohen (2013) asserts that not every test-management strategy undermines cognitive validity (i.e. rereading the question for clarification), as they might be different from the test-wiseness strategies which clearly make readers engage in unintended processes. These test-management strategies used in OE items are not disruptive in terms of altering normal reading processes or leading participants to wrong answers due to confusion. Therefore, test-taking strategies reported in OE items did not trigger any unintended cognitive processes and, in this sense, they cannot be claimed to undermine the cognitive validity of the test. The interview data support this argument in that some participants stated that the lack of options in the OE format made the reading process less confusing and more linear.

Multiple-choice format

On the other hand, eye movement data showed that MC tests triggered more expeditious reading skills and produced lower scores; total fixation counts were fewer in the interest areas of MC questions. While the participants reported to have used expeditious reading skills in both formats in an almost equal way, in the OE format, the participants read carefully after using expeditious reading skills, but the answers in the MC format could be located by mostly expeditious reading followed by test-taking strategies. MC questions seem to be encouraging search processes but discouraging deeper processing of the text when a likely answer is assumed to have been found. In the gaze plots, it can also be seen that in the MC format, the texts were

read in a more fragmented manner. The fixations focused on mostly different words or phrases, which might suggest the use of lower-level processes such as word recognition or keyword matching between the text and the questions. These processes may disrupt the extended linear and deeper reading process participants may engage in otherwise. If such lower-level processes are not intended by a specific item, then the cognitive validity can be at risk (Field, 2012).

A closer look at the use of test-management strategies reveals that in the MC format almost 43% of total test-management strategies are option-related. Based on our personal observations during data collection and as reported by the participants, options helped them find the correct answer for some questions. However, it is evident that they also misled them in others. This was caused by subtle differences and the tricky nature of the options, difficulty in matching their own comprehension with the test writer's interpretation, and assuming that they found the answer based on only keyword matching. Rupp et al. (2006) underline that option-related strategies such as matching key information in the text with the options can alter normal reading processes to a greater extent compared to what happens in the OE format. In addition, they assert that reading a text with MC questions is more like a problem-solving activity that requires unique reasoning when readers need to contemplate the subtle differences among options or the plausibility of the options instead of forming a coherent text representation. These claims resonate with the findings of the current study, which revealed that the participants resorted to additional test-taking strategies, especially option-related strategies, and read in a segmented manner due to the presence of options.

At the item level, the three sets of data complemented and confirmed each other, except for Items 5 and 6, which required processing of longer text spans. In these two items, although participants reported to have read more carefully in the OE format, the eye movement data showed that reading time was slightly higher in Item 5 and significantly higher in Item 6 for AoIs in the MC format. One reason for this longer reading time might be that as indicated during the interview sessions, the participants tried to answer this question by matching keywords in the text and question stem/options. This seems to have required frequent visits to the relevant AoIs through scanning and search reading, and hence increased the total reading time.

Another point worth mentioning is that MC vocabulary questions can be responded to through the use of background knowledge. In line with Lim (2014), in this study as well, the participants frequently reported using *"T19: making an educated guess"* strategy in answering vocabulary questions. The verbal report and gaze plot data in this study showed that the "guessing the meaning" type of vocabulary questions require the processing of larger text spans carefully in the OE format by nature, which we see as an additional virtue of this format.

Lastly, as mentioned in the results, the use of test-wiseness strategies was very limited in the current study as the participant group consisted of minimum upper-intermediate learners, a group of participants who strived to do their best. Had this not been the case, a more diversified group of participants might use more test-wiseness strategies and come up with correct answers. This would strengthen the arguments against the cognitive validity of MC items.

Conclusions and Limitations

A controlled and systematic investigation of the cognitive processes that participants went through in MC and OE reading comprehension items showed us that the differences in test method lead to differences in cognitive processes of reading and test-taking strategy use. The findings revealed and confirmed many issues relating to the MC item format. First of all, it is undeniable that MC items feature the practical advantages of scoring and marker reliability. However, it should be noted that the MC format can disrupt reading processes and have "an undue effect on measurement" (Weir, 1990, p. 44). Thus, as suggested by Field (2020), if possible, the OE format should be preferred in local testing settings where the ease of marking is not badly needed. The fact that MC questions lead test-takers to a number of test-taking strategies (i.e. option elimination or guessing the correct answer based only on the options) to cope with the item demands raises concerns about the validity of these item types. In addition, test developers should pay attention to the quality of each option. Options should not be based on absurdities, ambiguities, or subtle differences but rather on the comprehension or miscomprehension of the text. Item writers should also pay attention to the fact that the formation of text comprehension (micro or macro-proposition formation) can change from reader to reader. Therefore, an item writer's paraphrasing and summarisation of a text should be as objective and explicit

as possible and free from their personal interpretations so that a test-taker would not have any additional difficulty in matching their understanding with the one reflected in the correct option. Secondly, as Weir (1990) argues, answering MC items is an unreal task because, in real life, one expresses the understanding of what has been read by writing or speaking. Therefore, MC reading exercises should be used cautiously in the teaching of reading, too, as they do not reflect authentic extended reading processes.

This study has a few limitations. First, the data were collected in a lab, not in a real test environment, and this might have had an effect on the performance of participants. More importantly, the eye-tracker used in the study was a slow one in terms of the sampling rate. However, this study is based on the comparison of two item formats and if technical shortcomings affected the results, they should have affected the results from both item formats.

This means that even if there was a reduction in the quantity of recorded eye-movement data due to the quality of the eye-tracker, the general patterns we observed from two item formats could not be altered because of this. Despite the fact that the participants gave their verbal reports without being informed about their eye-tracking results, the eye movement data were confirmed by verbal reports and interviews. This suggests that although the data collected from the eye-tracker were crude, the findings are still meaningful and informative. In addition, future studies can include more eye-movement measures in the analysis such as analysis of regressions, thus extending the sensitivity of analysis. Lastly, the test-taking process was not investigated under the time limit, as it would be under normal test-taking conditions, and this might be seen as limiting the generalizability of the study. However, as we have mentioned, this helped us to see the behaviours of the participants better.

Despite the limitations, the study is exemplary in the sense that the research tools were carefully developed and piloted before being used in data collection. In conclusion, this study has confirmed the shortcomings of MC questions of reading comprehension and contributes to the existing literature by establishing systematic and controlled development and comparison of parallel item formats. It also offers guidance for further studies on item format effect on reading comprehension.

References

Akgün, H. (2018). *The effect of item format on the choice of reading and test-taking strategies* [Unpublished MA Thesis]. Boğaziçi University. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye tracking. *Language Testing*, *30*(3), 1–25. doi:10.1177/0265532212473244.

Bax, S., & Weir, C. (2012). Investigating learners' cognitive reading processes during a computer-based CAE reading test. *Research Notes, 47*, 3–14. http://hdl.handle.net/10547/337904

Brunfaut, T. (2016). *Looking into reading II: A follow-up study on test-takers' cognitive processes while completing APTIS B1 reading tasks*. British Council. https://eprints.lancs.ac.uk/id/eprint/82618/1/Brunfaut2016_LookingIntoReadingII.pdf

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. https://eprints.lancs.ac.uk/id/eprint/72005/1/Brunfaut_and_McCray_final_report_FINAL.pdf

Cohen, A. D. (1998). *Strategies in learning and using a second language*. Longman.

Cohen, A. D. (2013). Using test-wiseness strategy research in task development. In A. J. Kunnan (Ed.), *The companion to language assessment – Vol. 2: Approaches and development, Part 7: Assessment development*. Wiley and Sons.

Cohen, A. D. & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph Series Report No. 33). Educational Testing Service. http://www.ets.org/Media/Research/pdf/RR-06-06.pdf

ETS. (2009). *The official guide to the TOEFL® test*. McGraw-Hill.

Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In P. Thompson (Ed.), *IELTS Research Reports* (Vol. 9) (pp. 17–66). British Council. http://hdl.handle.net/10547/225496

Field, J. (2020). Cyril Weir and cognitive validity. In L. Taylor & N. Saville (Eds.), *Lessons and legacy: A tribute to professor Cyril J Weir (1950-2018).* (pp. 54-83). Cambridge University Press.

Jarodzka, H., & Brand-Gruwel, S. (2017). Tracking the reading eye: Towards a model of real-world reading. *Journal of Computer Assisted Learning, 33*(3), 193-201. doi:10.1111/jcal.12189.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Lim, H. J. (2014). *Exploring the validity evidence of the TOEFL iBT reading test from a cognitive perspective* [Unpublished PhD Thesis]. Michigan State University. https://www.proquest.com/docview/1648169236?pq-origsite=summon

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218. doi: 10.1207/s15326985ep3404_2.

McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing, 35*(1), 51-73. doi: 10.1177/0265532216677105

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13- 103). American Council on Education & Macmillan.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale, 67*(3), 215-227. doi:10.1037/a0032918.

Rayner, K., Chace, K.H., Slattery, T.J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10*, 241 - 255. doi: 10.1207/s1532799xssr1003_3.

Rayner, K., Reichle, E. D., & Pollatsek, A. (2005). Eye movement control in reading and the E-Z Reader model. In G. Underwood (Ed.), *Cognitive processes in eye guidance* (pp. 131 – 162). Oxford University Press.

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

Sarnaki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research, 49*(2), 252-279. doi: 10.3102/00346543049002252.

Staub, A., & Rayner, K. (2007). Eye *movements* and on-line comprehension processes. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327-342). Oxford University Press.

Ünaldı, A. (2004). *Construct validation of the reading subskills of the Boğaziçi University English Proficiency Test* [Unpublished PhD Thesis]. Boğaziçi University. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Weir, C. J. (1990). *Communicative language testing*. Prentice Hall.

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

Wu, R. Y. (2011). *Establishing the validity of the general English proficiency test reading component through a critical evaluation on alignment with the common European framework of reference* [Unpublished PhD Thesis]. University of Bedfordshire. http://hdl.handle.net/10547/223000

Appendix 1: Items of Text 1 MC & OE and Text 2 MC & OE*

TEXT 1: MC& OE	ITEM TYPE	TEXT SPAN (Where the answer was found)	TEXT 2: MC & OE
Q1	1. Factual	One sentence (Local)	Q6
Q2	2. Factual	One paragraph (Global)	Q3
Q3	3.Vocabulary	One paragraph (Global)	Q4
Q4	4. Factual	Across sentences (Global)	Q5
Q5	5. Inference	One paragraph (Global)	Q1
Q6	6. Factual	Across paragraphs (Global)	Q2

*: The items based on Text 1 and Text 2 are balanced with respect to cognitive difficulty, item focus and the amount of text that needs to be processed.

Appendix 2: Text 1 OE Format

THE ORIGINS OF CETACEANS

1 It should be obvious that cetaceans—whales and dolphins—are mammals. They breathe through lungs and give birth to live young. Their streamlined bodies, the absence of hind legs, and the presence of a fluke[1] and blowhole[2] cannot disguise their relationship with mammals living on land. However, unlike the cases of other animals such as seals and sea lions whose limbs are functional both on land and at sea), it is not easy to visualize what the first whales looked like. Extinct but already fully marine cetaceans are known from the fossil record. How was the gap between a walking mammal and a swimming whale bridged? Intermediate fossils between land mammals and cetaceans were missing until recently.

2 Very exciting discoveries have finally allowed scientists to reconstruct the most likely origins of cetaceans. In 1979, a team looking for fossils in northern Pakistan found what proved to be the oldest fossil whale. The fossil was officially named Pakicetus. Pakicetus was found embedded in rocks formed from river deposits that were fifty-two million years old. The river was actually not far from an ancient ocean known as the Tethys Sea.

3 The Pakicetus fossil consists of a complete skull of an extinct group of ancestors of modern cetaceans. Although limited to a skull, the Pakicetus fossil provides precious details on the origins of cetaceans. The skull shape is cetacean-like but its jawbones lack the enlarged space that is filled with fat or oil and used for receiving underwater sound in modern whales. Pakicetus probably detected sound through the ear opening as in land mammals. The skull also lacks a blowhole, another cetacean adaptation for diving. Other features, however, show experts that Pakicetus is a transitional form between a group of extinct flesh-eating mammals and cetaceans. It has been suggested that Pakicetus fed on fish in shallow water and was not yet adapted for life in the open ocean. It probably bred and gave birth on land.

4 Another major discovery was made in Egypt in 1989. Several skeletons of another early whale, Basilosaurus, were found in sediments left by the Tethys Sea and now exposed in the Sahara desert. This whale lived twelve million years after Pakicetus. Many incomplete skeletons were found but they included, for the first time a complete hind leg that features a foot with three tiny toes. Such legs would have been far too small to have supported such a big animal on land. Basilosaurus was undoubtedly a fully marine whale with possibly nonfunctional hind legs.

5 An even more exciting find was reported in 1994, also from Pakistan. The now extinct whale Ambulocetus natans (the walking whale that swam) lived around three million years after Pakicetus but nine million years before Basilosaurus. The fossil luckily includes a good portion of the hind legs. The presence of hind legs clearly bridged the gap between a walking mammal and swimming whale. The legs were strong and ended in long feet very much like those of a modern seal. The legs were certainly functional both on land and at sea. This clearly proved Ambulocetus to be a transition species. Moreover, the whale retained a tail and lacked a fluke, the major means of moving in modern cetaceans. The structure of the backbone shows, however, that Ambulocetus swam like modern whales by moving the rear portion of its body up and down. The large hind legs were used for moving forward in water. On land, where it probably bred and gave birth, Ambulocetus may have moved around very much like a modern sea lion. It was undoubtedly a whale that linked life on land with life at sea.

1. *Fluke: the two parts that constitute the large triangular tail of a whale*
2. *Blowhole: a hole in the top of the head used for breathing.*

1. In paragraph 1, what can the presence of a blowhole in cetaceans not conceal according to the author?

2. What do Pakicetus and modern cetaceans have in common?

3. Find and write a word which means "visible" in paragraph 4.

4. The hind legs of Basilosaurus show that it couldn't have walked on foot because

5. Find and write a sentence in paragraph 4 that supports the inference that "Basilosaurus bred and gave birth in water".

6. Why does the author use word "luckily" while mentioning that Ambulocetus natans fossil included hind legs?

Appendix 3: Text 2 OE Format

SWIMMING MACHINES

1 Tunas, mackerels, and billfishes (marlins, sailfishes, and swordfish) swim continuously. Feeding, courtship, reproduction, and even "rest" are carried out while in constant motion. As a result, practically every aspect of the body form and function of these swimming "machines" is adapted to enhance their ability to swim.

2 Many of the adaptations of these fishes serve to reduce water resistance. Interestingly enough, several of these hydrodynamic adaptations resemble features designed to improve aerodynamics in engineering. Though human engineers are new to the game, tunas and their relatives evolved their "high-tech" designs long ago.

3 Tunas, mackerels, and billfishes have made streamlining into an art form. Their bodies are sleek and compact. The body shapes of tunas, in fact, are nearly ideal from an engineering point of view. Most species lack scales over most of the body. This feature makes their bodies smooth and slippery. They also have a slick and transparent cover that reduces water resistance. The fins are stiff, smooth, and narrow. These qualities also help reduce water resistance. When not in use, the fins are tucked into special grooves or depressions so that they lie at the same level with the body. Therefore, they do not break up its smooth contours. Airplanes retract their landing gear while in flight for the same reason.

4 Tunas, mackerels, and billfishes have even more sophisticated adaptations than these to improve their hydrodynamics. The long bill of marlins, sailfishes, and swordfish probably helps them move smoothly the water. Many supersonic aircraft have a similar needle at the nose. Most tunas and billfishes have a series of keels and finlets near the tail. Although most of their scales have been lost, tunas and mackerels retain a patch of coarse scales called the corselet. The keels, finlets, and corselet help direct the flow of water over the body surface to reduce water resistance. Again, supersonic jets have similar features.

5 Because they are always swimming, tunas simply have to open their mouths and water is forced in and over their gills. Accordingly, they have lost most of the muscles that other fishes use to suck in water and push it past the gills. In fact, tunas must swim to breathe. They must also keep swimming to keep from sinking, since most have largely or completely lost the swim bladder. Swim bladder helps most other fish remain floating.

6 One potential problem is that opening the mouth to breathe detracts from the streamlining of these fishes and tends to slow them down. Some species of tuna have specialized grooves in their tongue. It is thought that these grooves help to channel water through the mouth and out the gill slits. This process reduces the water resistance.

7 There are adaptations that increase the amount of forward thrust as well as those that reduce water resistance among fast swimming fishes. Perhaps most important of all is their ability to make use of swirls and eddies (circular currents) in the water. They can glide past eddies that would slow them down and then gain extra thrust by "pushing off" the eddies. Scientists and engineers are beginning to study this ability of fishes in the hope of designing more efficient propulsion systems for ships.

8 These fishes also have a highly efficient mechanism that maintains a warm body temperature. With the help of this mechanism, a bluefin tuna in the water of seven°C (fortyfive°F) can maintain a core temperature of over twenty-five°C (seventy-seven°F). This warm body temperature may help not only the muscles to work better, but also the brain and the eyes. The billfishes have gone one step further. They have evolved special "heaters" of modified muscle tissue that warm the eyes and brain, maintaining peak performance of these critical organs.

1. The writer mentions that the design of some fish bodies is used by engineers. Find the example about this similarity in paragraph 3.
_____.

2. It can be understood that the adaptations such as fins and long bill of some types of fishes help them to swim smoothly by _____
_____.

3. According to the text, tunas have gone through major evolutionary changes. What are two major reasons why tunas must be in constant motion?
_____
_____.

4.Find and write a word which means "to direct" in paragraph 6?
_____.

5. According to the passage, one of the adaptations of fast swimming fishes that can also be used to improve the performance of ships is their ability to _____
_____.

6. What makes bluefin tunas swim in waters that are much colder than their bodies?
_____.

Appendix 4: Answer keys

TEXT 1

1. It cannot conceal their relation with mammals on land.
2. Skull shape
3. Exposed
4. Legs were too small to support it on land
5. Basilosaurus was undoubtedly a fully marine whale with possibly nonfunctional hind legs
6. Ambulocetus was a transition species / It bridged the gap between a walking mammal and a swimming whale

TEXT 2

1. Airplanes retract their landing gear while in flight for the same reason / the fins
2. Reducing water resistance
3. İ. To breathe ii. To keep from sinking/ they lost their swim bladder
4. To channel
5. Make use of circular currents/ glide past the eddies and pushing off eddies
6. A highly efficient mechanism

Appendix 5: Item-based Eye-tracking Descriptive Statistics

		MC		OE		MC		OE	
		AoI TFC	AoI TRT	AoI TFC	AoI TRT	QS TFC	QS TRT	QS TFC	QS TRT
ITEM 1	Mean	3.92	3.50	9.93	10.54	2.15	2.17	2.49	2.19
	St. Dev.	2.15	1.98	3.98	3.73	1.21	1.39	1.2	1.1
ITEM 2	Mean	6.52	6.64	9.23	9.63	2.06	1.95	2.67	2.63
	St. Dev.	2.45	2.61	3.34	4.29	0.94	1.13	2.1	2.23
ITEM 3	Mean	5.33	5.43	7.12	8.24	1.65	1.76	2.67	2.31
	St. Dev.	2.91	2.83	3.71	4.14	0.72	0.96	1.28	1.25
ITEM 4	Mean	7.13	7.45	10.01	11.68	2.6	2.96	2.09	2.06
	St. Dev.	3.03	3.19	6.92	8.05	1.03	1.02	1.32	1.24
ITEM 5	Mean	8.21	8.63	7.23	7.76	2.03	2.08	2.93	2.71
	St. Dev.	2.85	3.07	3.96	4.73	1.47	1.6	1.5	1.53
ITEM 6	Mean	7.91	7.81	4.64	4.59	1.69	1.81	1.63	1.54
	St. Dev.	4.18	4.08	3.7	3.71	0.85	1.19	0.75	0.82