International Journal of Assessment Tools in Education
2020, Vol. 7, No. 1, 98–113
https://dx.doi.org/10.21449/ijate.699494

Published at http://www.ijate.net http://dergipark.org.tr/en/pub/ijate *Research Article*

What You might not be Assessing through a Multiple Choice Test Task

Burcu Kayarkaya [1,*], **Aylin Unaldi** [2]

[1]School of Foreign Languages, Yıldız Technical University, Davutpaşa Campus, 34220, İstanbul, Turkey
[2]School of Education and Professional Development, University of Huddersfield, Huddersfield, HD1 3DH, UK

ARTICLE HISTORY

Received: 04 November 2019
Revised: 14 February 2020
Accepted: 06 March 2020

KEYWORDS

Textual reading comprehension,
Macrostructure formation,
Reading operations,
Multiple choice task,
Summary task

Abstract: Comprehending a text involves constructing a coherent mental representation of it and deep comprehension of a text in its entirety is a critical skill in academic contexts. Interpretations on test takers' ability to comprehend texts are made on the basis of performance in test tasks but the extent to which test tasks are effective in directing test takers towards reading a text to understand the whole of it is questionable. In the current study, tests based on multiple choice items are investigated in terms of their potential to facilitate or preclude cognitive processes that lead to higher level reading processes necessary for text level macrostructure formation. Participants' performance in macrostructure formation after completing a multiple choice test and a summarization task were quantitatively and qualitatively analyzed. Task performances were compared and retrospective verbal protocol data were analyzed to categorize the reading processes the participants went through while dealing with both tasks. Analyses showed that participants' performance in macrostructure formation of the texts they read for multiple choice test completion and summarization task differed significantly and that they were less successful in comprehending the text in its entirety when they were asked to read to answer multiple choice questions that followed the text. The findings provided substantial evidence of the inefficacy of the multiple choice test technique in facilitating test takers' macrostructure formation and thus pointed at yet another threat to the validity of this test technique.

1. INTRODUCTION

One requirement a second language (L2) reader in an academic context has to meet is to process a text thoroughly and carefully to extract complete meanings from the written material. Careful reading of extended texts is the basic skill for "learning" in academic environments (Weir, Hawkey, Green, & Devi, 2009). In tests of English for academic purposes (EAP), careful reading at whole text level should thus be assessed to ensure adequate construct representation. It is a general contention that through designing several test items on main ideas in a text, text level comprehension can be achieved even in multiple choice (MC) tests. However, there is research that generally points out that scores obtained in MC tests measuring reading comprehension may not truly represent test takers' understanding of the written material (Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008). It is also known that test

CONTACT: Burcu Kayarkaya ✉ burcudurak@hotmail.com School of Foreign Languages, Yıldız Technical University, Davutpaşa Campus, 34220, İstanbul, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

formats may determine how readers perform reading activities and test tasks in different formats can invoke different reading skills and strategies (Lee, 1986; Shohamy, 1984; Wolf, 1991). Studies conducted to discover whether task format has an effect on the extent and depth of reading comprehension have produced empirical evidence for the existence of inconsistency in reader performance due to task format (Kobayashi, 2002; Pearson, Garavaglia, Lycke, Roberts, Danridge, & Hamm, 1999). Thus, whether all task formats can facilitate the assessment of targeted skills and whether a test with certain tasks can operationalize the relevant reading skills with adequate coverage is an important issue for discussion, namely, a discussion on the construct validity of the test (ALTE, 2011).

Albeit, to our knowledge, there are no studies that focus on whether extensively used MC items can facilitate text level comprehension. To what extent this commonly used assessment technique can be instrumental in enabling readers to form a coherent mental representation of the text they read for test taking purposes is a question yet to be answered. This study aims at providing evidence for the claim that an MC reading test may be assessing comprehension of certain parts of a text but this may not necessarily mean that high scores from such a test reflect a complete understanding of the test text. This is yet another issue that challenges the validity of the use of MC format in reading assessment and it is important that this question be probed.

Reading in a foreign language is a complex process with many underlying cognitive components (Gernsbacher, 1997; Graesser, Singer & Trabasso, 1994; Kintsch, 1998; Myers & O'Brien, 1998). Examining these cognitive components is necessary to understand reading comprehension processes that take place in a reader's mind. Explanations on how a reader comprehends texts generally point to a series of processes to eventually arrive at constructing textual meaning. In the bottom-up processing, the reader starts with decoding linguistic structures or units to unfold propositions of the text one by one (Gough, 1972), or the top-down processing suggests more global processes of activating background knowledge to predict the content of the text and confirmation of the prediction takes place as the reading goes along (Goodman, 1967). In modern views of reading, reading process is seen as an interaction between bottom-up and top-down processes: Readers go along a continuum of selection of processes while reading, changing their focus from linguistic units towards textual clues or the other way round, making use of top-down and bottom-up processes in different quantities and sequences (Grabe, 1991).

Khalifa and Weir (2009) hypothesized that difficulty in reading is a function of the level of processing required by reading purpose and the complexity of the text. Reading is conceptualized as having several types; expeditious versus careful and local versus global reading. Moreover, careful reading is further divided into four levels including within-sentence (propositional meaning), across sentences (mental model; ongoing meaning making as the reader proceeds in the text), text (text model) and texts (documents model) models. Careful reading is predominantly a bottom-up process, starting with linguistic processing of the elements of a sentence and establishing propositional meaning (the literal interpretation of what is printed on the page). Through inferencing, the reader relates the message to the context. Inferencing is also functional in establishing coherence, or meaning between propositions, as the reader integrates new information into a mental representation of the text so far. This is the stage at which the reader starts to identify main ideas and impose a hierarchical structure on the information in the text. According to Kintsch and van Dijk (1978), this is the stage where microstructure rules are at work to link the textual pieces and reduce the content to higher propositions to be stored in working memory. Background knowledge on the content of the text and the meaning formed on the text so far facilitate inferencing and control of coherence and consistency in the text. At the text level, micropropositions are collapsed into macropropositions. Macropropositions are derived from a text through the application of macrorules: less important portions of the text are deleted, instances are generalized, and

summaries of events are constructed (van Dijk,1980). Macroproposition of a text is the skeleton that makes up the text body which can also be regarded as an organised form of the most important portions of the text in an hierarchical order. Recognition of the hierarchical structure of the text is of crucial importance in forming a unified understanding at the text level. The new information presented in the text is combined with what the reader already has in supply and eventually a situation model is produced (Kintsch & Kintsch, 2005; Kintsch & van Dijk,1978).

Whether a reader will construct a situation model of interpretation depends on what purposes the reader is engaged in the text for. Urquhart and Weir (1998) categorize types of reading that serve for different purposes as follows: a) expeditious reading (quick, selective and efficient reading to access desired information in a text- scanning, skimming and search reading), and b) careful reading (processing a text thoroughly with the intention of extracting complete meaning from presented material). They further make distinctions between global and local comprehension gains from reading texts. Global comprehension refers to reaching an understanding of the explicit information available in a text, including main ideas and the links between these ideas, through integrating and synthesizing information. The reader is then able to build logical relationships between ideas. Local comprehension is more related to an understanding of propositions within the sentence and is a process that involves word recognition, lexical access and syntactic parsing and maintaining meaning at the phrase, clause and sentence level (Bax, 2013, Khalifa & Weir, 2009, Weir & Bax, 2012). A reader, for example, looking for specific information in a text may favor search reading or expeditious reading to access the necessary information quickly (Guthrie & Kirsch, 1987). However, a reader comparing the arguments of a writer with those of another writer may embrace a global, careful style that would enable him/her to arrive at a deeper understanding of the text.

Elaborating on the purposes of reading, Enright, Grabe, Koda, Mulcahy-Ernt, and Schedl (2000) put forward four purposes for reading in L2: a) reading to find information (search reading), b) reading for basic comprehension, c) reading to learn, and d) reading to integrate information across multiple texts. Grabe (2009) built upon Enright's four purposes and added two further purposes: e) reading for quick understanding (skimming), f) reading to evaluate, critique and use information. For some of the purposes listed above (search reading or skimming), a reader does not necessarily form a text or situation model as deriving the macrostructure of a text is not the aim, but for some models of comprehension (reading to learn or reading for basic comprehension), this is required. Enright et al. (2000) explain that in the reader purpose perspective, a reader approaches a text depending on what he/she is supposed to do with the text, assuming that all readers read for a reason in certain contexts, be it for an exam purpose or orientation in real life. A reader's standard of coherence affects the depth of their comprehension and alters how a text is processed. Requirements of the reading task itself or the goals a reader sets for reading a text change the reading processes a reader goes through while reading. That is, readers shape and reshape their reading behavior, or the processes, to fit the requirements of the model of the task in mind (Britt, Rouet & Durik, 2017).

For L2 learners of English, reading ability in an academic context means that they can successfully perform a variety of reading skills including the ability to read and understand a text in its entirety with the purpose of learning from it. As mentioned before, this is usually referred to as "text/situation model formation" (Kintsch, 1998), "reading to learn" (Enright et al., 2000), and "reading at the whole text level" (Khalifa & Weir, 2009). The basic principle in this process is the formation of the macrostructure and thus text/situation model. Whether this can be adequately assessed in tests of academic reading and whether certain item formats are conducive or non-conducive to the assessment of text level comprehension is a worthy matter of inquiry in the field of language testing.

Reading comprehension is a multi-facetted complex activity in which features of input text, the

reader's competences and motivation and task demands interact with each other to determine the characteristics of reading behavior (Bachman & Palmer, 2010; Khalifa & Weir, 2009). As included in Bachman's (1990) framework, "the nature of the expected response to the input (the test format)", and its relationship with the input (the interaction between the written material and the test format)" determine the extent and depth of reading.

In test development, test items are aimed at operationalizing certain sub-skills of reading, presumably in adequate coverage. There is certainly no best method for testing reading since no single test method can fulfill all the varied purposes for which one might be testing (Alderson, 2000). Convenience, practicality and efficiency may become primary considerations while deciding on the most suitable method for assessment (Bachman, 2000). More often than not, objectively and economically scorable item formats such as MC and matching items are chosen instead of open-ended, extended response items (Prapphal, 2008; Watson-Todd, 2008). However, this brings up the issue of whether different item formats can measure the same ability or not – such as whether the MC and open-ended items assess equivalent reading skills. Besides, there is the question of whether test items can invoke reading skills at macro levels as well as they do at micro levels; and if they do, whether they can cover all the skills relevant to the test purpose. Pearson, Garavaglia, Lycke, Roberts, Danridge and Hamm (1999) investigate whether there are differences in the cognitive processes readers execute to complete an MC and a constructed response task. The results indicate that the MC task activates a significantly lower proportion of skills at the macro level (e.g. intertextuality). However, as Kintsch and Kintsch (2005) underline, questions requiring the readers to set off macro operations and questions targeting the macrostructure of a text are more instrumental in reflecting reading comprehension ability of the readers.

Obviously, some task types are more conducive to assessing text level comprehension processes whereas others may only assess it at local levels. MC items are widely used in reading assessment; therefore, it is important to understand the characteristics of these items types, and find out whether tests formed of MC items can tap into text level reading comprehension processes (Sheehan & Ginther, 2001).

There are several reasons why the MC technique is widely used in tests. First, relatively more content can be covered in MC tests when compared to other test formats (Haladyna & Downing, 2009). That is, MC technique provides more flexibility in covering larger bits of information. Second, it makes scoring easy and effortless – human raters or machine raters can be in charge and the answer key is fixed in that in a carefully planned MC test, there is usually only one correct option (Fuhrman, 1996). MC test format also economically represents whether and to what extent the reader can read and understand parts of a text.

However, while responding to MC questions, readers usually have to choose the best option from alternatives rather than verbalizing or producing answers themselves. This means that the question format may limit the understanding of the reader to the ideas as they are worded in the options by the item writer. MC questions may mostly encourage memorization and factual recall and may not promote high-level cognitive processes (Airasian, 1994; Scouller, 1998). Another serious problem concerning the MC format is that the rater simply does not know why readers respond to the questions the way they do (Lau, S. Lau, Hong, & Usop, 2011). There is also the risk of guessing effect, especially if the distractors are not written carefully (Kurz, 1999).

Research also suggests that test takers use various strategies which are not necessarily comprehension-based to answer MC items and critics claim that such test-taking strategies are executed to get an acceptable answer to the question rather than to understand the text (Anderson, Hiebert, Scott, & Wilkinson, 1985; Cohen, 1984). Shohamy (1984) and Wolf (1991) investigate the format effects comparing MC questions and open-ended questions and

conclude that MC items are easier to deal with for readers. Rupp, Ferne and Choi (2006) provide empirical evidence for the hypothesis that when readers respond to texts followed by MC questions, they go through different processes than they would while reading in non-testing contexts. Rupp et al. (2006) state that readers tend to approach reading tasks with MC questions as a problem solving task, rather than a comprehension task. That is, readers, as strategic test takers, use a number of techniques to "solve" the problems that appear as questions in tests and this makes the activity less similar to real-life reading. Similarly, the study by Cerdan, Vidal-Abarca, Martinez, Gilabert, and Gil (2009) point at the likelihood of readers' following a question-to-text sequence when answering MC questions.

Remembering that cognitive (construct) validity examines the relationship between what a test aims to measure and what it actually elicits from test takers (Weir, 2005), if the cognitive demands of a task do not adequately represent the demands of the skills in the target domain, the cognitive validity of such a task would be questionable. If specific task characteristics in a test affect the cognitive processes employed by test takers, then our inferences based on the scores from that test would be undermined (Smith, 2017). MC test format has been scrutinized from several aspects as discussed above (Martinez, 1999; Martinez & Katz, 1995; Rupp et al., 2006). However, although it is widely used in reading assessment, there is no study focusing on whether it facilitates comprehension at whole text level. With several MC items in a test, it is possible to cover all or most of the main ideas in a text. However, if test takers cannot form a successful macrostructure of the text they have dealt with during the test, then we can assume that processes leading to coherent mental representation formation are not facilitated or even precluded in reading tests with MC questions.

2. METHOD

This study aims at providing evidence as to whether MC items can assess text level reading by comparing the performances of test takers in an MC test and an oral summary task by addressing two research questions:

RQ1: To what extent can textual level comprehension be attained upon the completion of multiple choice and oral summary reading tasks?

RQ2: How do test takers' reading styles and preferences differ according to multiple choice and oral summary tasks?

2.1. Participants

A total of 32 (15 female, 17 male) students were selected for the study throughconvenience sampling. In selection of the students, a number of factors were taken into consideration. As the materials used in the study target learners of English above a certain level of proficiency (at B2 level), participants were chosen from a pool of almost 300 students taking an Advanced English mass course at a state university based in İstanbul, Turkey. When forming the participant group, the grades students got from the midterm exams were checked and those with scores at or above 80 out of 100 were shortlisted and a further elimination was made according to the performance those had in the reading section of the exam. Eventually, the 32 students who were regarded as eligible for the study were asked for consent and all agreed to participate in the study.

2.2. Instruments

2.2.1. *Tasks and texts used in the study*

Reading comprehension performance of the participants is measured by multiple choice (MC) and summarization tasks. Summarization is taken as a strong indicator of text level comprehension because summarizing a text requires the ability to identify main ideas in the text, integrate them into a text model, and develop a proper situation model of interpretation

(Grabe, 2009; Taylor, 2013). In order to understand main ideas, readers need to have a large receptive vocabulary, basic grammar, effective comprehension strategies, and strategic processing abilities to maintain a high level of comprehension and an awareness of discourse structure (Grabe, 2009; Pressley, 2002).

The summary technique provides a solid representation of how mental processes operate in the reader's mind, how they prioritize, construct and organize information as well as the retrieval strategies they use (Bernhardt, 1983). Khalifa and Weir (2009) state that global (text level) careful reading at the highest level requires the reader to understand the micro and macropropositions in a text and how these are interconnected, while integrating new information into a mental model to create a discourse level structure that is appropriate to their purpose. We can say that the cognitive processes a reader has to follow to summarize a text are text level macrostructure formation processes.

The participants in this study were asked to summarize the text they read to answer MC questions right after completing the task. A combined term, multiple choice summary (MCSUM) is used to refer to this task. MCSUM aimed to measure the extent of textual level comprehension that surfaced upon the completion of MC questions, and the performance in MCSUM is compared to that in oral summary (SUMONLY) task. The SUMONLY task is used as a baseline to assess the general summarization abilities of the participants so that we can identify whether the success or failure in summarization in the MCSUM task is due to what is comprehended after the text is processed or to the general comprehension abilities of the participants. The participants completed the tasks (MC and SUMONLY) designed on two different texts (Text A and Text B).

The texts and the set of MC questions accompanying them were taken from TOEFL preparation materials. TOEFL is a strong representative of MC EAP tests students take in college and university settings. To ensure comparability of the texts in different tests (Text A and Text B) in terms of textual features (vocabulary, topic, language use and level, cohesion, coherence, syntactic simplicity, narrativity, genre and interest), automatic text analysis tools were used along with ideas and suggestions from expert judgement.

For the MC test, there were eight questions in both versions intended to elicit a variety of reading subskills based on TOEFL (2014) test specifications. These questions were carefully matched in terms of subskills and question types across Texts A and B by the researchers (see Table 1). The SUMONLY task was a verbal instruction for the participants, informing that what they had to do with the text was to read it to summarize. The study counter-balanced task and text order in a four-way distinction. Table 2 describes how and in what order the tasks and texts were assigned to the participants.

Table 1. Question Types and the Reading Subskills the Question Types Measured.

Text A Q2	Text insertion/Cohesion formation	Text B Q5
Text A Q3	Sentence simplification	Text B Q8
Text A Q4	Factual information	Text B Q2
Text A Q5	Inference	Text B Q7
Text A Q6	Rhetorical purpose	Text B Q6
Text A Q7	Reference	Text B Q3
Text A Q8	Vocabulary	Text B Q1

Table 2. The Distribution of Tasks in Four Groups

Group	First session	Second session
Group I N=8	Text A – MC MCSUM VP (How did the participant read for the MC task?)	Text B – SUMONLY VP (How did the participant read for the SUMONLY task?)
Group II N=8	Text B – SUMONLY VP (How did the participant read for the SUMONLY task?)	Text A – MC MCSUM VP (How did the participant read for the MC task?)
Group III N=8	Text A – SUMONLY VP (How did the participant read for the SUMONLY task?)	Text B – MC MCSUM VP (How did the participant read for the MC task?)
Group IV N=8	Text B – MC MCSUM VP (How did the participant read for the MC task?)	Text A – SUMONLY VP (How did the participant read for the SUMONLY task?)

2.2.2. The scoring of summaries

The researchers worked with three instructors working at the School of Foreign Languages of the university mentioned above for expert judgement both in the formation of the tests and in the scoring of the summaries. The instructors read the texts to evaluate their appropriateness for the participant profile. They were also asked to identify the parts of the texts that were essential to include in an accurate summary of the texts. After having them make their own lists of relevant text parts, a meeting was organized with all the instructors to compare the summary lists including one of the researchers' and to discuss discrepancies. Thus, a consensus rubric for the scoring of the summaries was formed by the instructors and one of the researchers. These rubrics contained seven statements which carried primary information; i.e. main ideas and topic sentences for each text; Text A and Text B. While scoring, one point is given to each statement in the participants' summary that matched a statement in the rubric. The participants were free to summarize the text either in their L1, Turkish, or in L2, English. Each summary was scored twice by one of the researchers using the rubrics: during the sessions when the participants completed the tasks and upon completion of data collection. The time interval between the two scorings was 15 days. The final scores were turned into percentages. The scores obtained from the first and second scoring of the summaries for both conditions, MCSUM and SUMONLY, were compared and the intra-rater agreement, Cohen's kappa, was found to be 0.76 and 0.73, respectively. The difference between the scores was analyzed through one-way ANOVA and the effects of test methods and texts were analyzed through two-way ANOVA between subjects.

2.2.3. Retrospective verbal protocols

In order to investigate the cognitive processes the participants went through while reading to answer MC questions and summarizing the text, retrospective verbal protocol (VP) technique was used. Following the summarization process both after MC task (MCSUM) and in SUMONLY condition, the participants were asked to reflect on their reading behavior, explaining how they read the texts to answer MC tasks and whether or not reading for such a

purpose affected their reading style. The participants were also asked to describe their reading processes when they read to summarize the text. All VP sessions were video-recorded. When necessary, prompt questions were asked to help the participants express how they handled the task. The questions were in participants' L1, that is to say, in Turkish.

2.2.4. *Coding verbal protocols*

In order to develop a coding scheme to classify the reading processes emerging from VPs, several coding schemes were examined (Cohen & Upton, 2007; Lim, 2014; Unaldi, 2004; Weir et al., 2009). As a result, a list of reading operations that fit the purpose of the current study was compiled. Two more operations that emerged from the VP data were added to the reading operations. A customized scheme of 11 reading operations (RO) that the participants stated they executed while accomplishing the tasks was formed.

The VPs for each participant (following MC and SUMONLY tasks) were coded and scored by one of the researchers using the coding scheme, identifying each reading operation as test takers verbalized them. Following this, a second coder coded both tasks using the coding scheme and watching the video-recorded sessions. The two coders' results were compared and inter-rater agreement on the reading operations in MC task was found to be 81% and for the SUMONLY task, the agreement on the reading processes were calculated as 73%. A paired-sample t-test was run to determine the statistical significance of the difference between the reading operations used in MC and SUMONLY tasks.

3. RESULT / FINDINGS

3.1. Research Question 1

RQ1: To what extent can textual level comprehension be attained upon the completion of multiple choice and oral summary tasks?

To provide an answer for this question, the researcher used the summary rubric to count how many of the sentences in the rubric were produced by the participants while they summarized in two conditions. By doing so, the extent to which the macrostructure of the text had been successfully formed by the participant in both conditions was assessed. The scores for the two summaries were then compared to find out whether theypointed to a statistical difference in the formation of macrostructures.

A paired samples t-test was used to test the null hypothesis that there is no difference between the mean scores from two summaries (MCSUM and SUMONLY). In addition, a two-way analysis of variance (ANOVA) was used to investigate whether the texts or the tasks accounted for the main variance. Table 3 shows the distribution of the scores in the tasks. The mean scores (converted into percentages) the participants obtained showed that they had the highest scores in the SUMONLY task, when they read the text to make a summary of it. The lowest scores were observed in MCSUM task, when the participants summarized the text they had read to answer the MC questions (MC task). In other words, the participants received lower scores when they summarized the text upon the completion of the MC task in comparison to the case when they read a text to summarize it.

The success levels of the participants in textual level comprehension in MCSUM and SUMONLY were found to be statistically significantly different (p=.002) through one-way analysis of variance given in Table 4. The participants performed better in macrostructure formation, that is to say, text level comprehension when they read for the SUMONLY task (M=65.59, SD=21.14). In the MCSUM, the macrostructures they produced of the texts were significantly less successful (M=48.6%, SD=22.46).

Table 3. Descriptive Statistics

	MC	MCSUM	SUMONLY
M	63.35	48.6	65.69
Median	62.5	42.8	71.4
SD	19.71	22.64	21.13
Skewness	-.2	.36	.06
Kurtosis	-.7	-.64	-.93

Table 4. One-way Analysis of Variance

					95% CI	
(I) Method	(J) Method	Mean dif. (I-J)	Std. E.	Sig.	Lower bound	Upper bound
MC	MCSUM	14.75	5.22	.006	4.73	25.122
	SUMONLY	-2.23	5.22	.670	-12.6	8.14
MCSUM	MC	-14.75	5.22	.006	-25.12	-4.37
	SUMONLY	-16.98	5.22	.002	-27.35	-6.6
SUMONLY	MC	2.23	5.22	.670	-8.14	12.6
	MCSUM	16.98	5.22	002	6.6	27.35

The results of the analyses (Tables 3 and 4) indicate that reading a text for the purpose of answering MC questions may not contribute to the formation of the macrostructure of the text as much as reading for summarization purposes does. It is important to note that overall performance difference among the MC and SUMONLY task scores was minimal. Although the scores participants got from the MC test were close to the grades they got from the SUMONLY task (see Table 3), the performance in MC task was not totally transferred to the task following immediately in which formation of the macrostructure of the text was required.

Besides, two-way ANOVA between-subjects analysis (see Table 5) indicates that the difference between the mean scores of the two tasks was merely a result of the test method and that the test results were not affected by the texts used in the study or by text and method interactions; the effect for method was significant (F=6.24, *p*=.003).

Table 5. Two-way ANOVA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7987.41	5	1597.48	3.66	.005
Intercept	336291.53	1	336291.53	771.15	.000
Method	5449.64	2	2724.82	6.24	.003
Text	944.38	1	944.38	2.16	.145
MethodText	1593.39	2	796.69	1.82	.167
Error	39247.85	90	436.08		
Total	383526.81	96			
Corrected Total	47235.27	95			

3.2. Research Question 2

RQ2: How do test takers' reading styles and preferences differ according to multiple choice and oral summary tasks?

The purpose of RQ2 was to identify the types and frequency of reading operations employed by the participants in the study while they completed an MC and a summarization task. By doing so, we aimed at examining the reading operations that were instrumental in the participants' performances during reading for the two tasks and whether the cognitive processes they engaged in while reading contributed to the formation of the macrostructure of the texts. For the MC task completion, eight reading operations (RO) were identified during the verbal protocols. Table 6 lists the reading operations that participants stated they operationalized during reading for the MC task.

Table 6. ROs the Participants Stated they Went Through during the MC Task

		f	%
RO3	I followed a question-to-text sequence, matching the keywords in text and questions	30	93.7%
RO6	I read the whole text from the beginning to the end carefully	12	37.5%
RO5	I read carefully only the selected part(s) of the text that might be relevant to the question	9	28.1%
RO7	During reading, I read a part of the text more than once to understand it	8	25%
RO1	I read the text carefully first, before attempting the task	7	21.8%
RO2	I read the text expeditiously to have a general idea before attempting the task	6	18.7%
RO4	I read expeditiously to find a relevant part that might include the answer	6	18.7%
RO9	I tried to understand how the text was organized, how the ideas and details were connected	1	3.1%

The compilation of reading operations that the participants stated they went through while completing the MC task emphasizes certain characteristics of MC tasks and make it clear how test takers approach these tasks. For the MC task, 93.7% of the participants reported that they "followed a question-to-text sequence, matching the keywords in the text and the questions, (RO3)". Only 21.8% of the participants stated they "read the text carefully before attempting the task (RO1)".

For the SUMONLY task, the reading operations the participants in the study stated they carried out are presented in Table 7. Eight reading operations arose from the participants' statements regarding their reading preferences for the summary task.

Table 7. ROs the Participants Stated they Executed during the SUMONLY Task

		f	%
RO1	I read the text carefully, before attempting the task	31	96.8%
RO6	I read the whole text from the beginning to the end carefully	31	96.8%
RO10	I read to get the main ideas and remember them	12	37.5%
RO7	During reading, I read a part of the text more than once to understand it	11	34%
RO8	I read to make connections between paragraphs or parts	10	31.2%
RO11	I paid further attention to introduction and conclusion paragraphs as they would include the main idea	7	21.8%
RO9	I tried to understand how the text was organized, how the ideas and details were connected	5	15.6%
RO2	I read the text expeditiously to have a general idea before attempting the task	1	3.1%

While reading for the SUMONLY task, 96.8% of the participants stated they "read the text carefully first, before attempting the task, (RO1)" and as a result, they "read the whole text carefully, (RO6)". The second most frequently utilized reading operation was RO10, "I read to get the main ideas and remember them", which was reported by 37.5% of the participants. 21.8% of them reported that they "paid further attention to the introduction and conclusion paragraphs as they would include the main idea of the whole text, (RO11)" when they were dealing with SUMONLY task. 31.2% of the participants asserted that they "read to make connections between paragraphs or parts, (RO8)" and 15.6% of them stated that they "tried to understand how the text was organized and how the ideas and details were connected, (RO9)" to understand the text. It is clear that as expected, these reading processes are genuinely careful reading operations that a test taker can make use of to attain comprehension at text level.

To compare the means of reading operations of MC and SUMONLY tasks, a paired-samples t-test was performed (see Table 8). It indicated that differences were statistically significant between the means of every reading operation executed for MC and SUMONLY tasks except for RO6. However, the effect size (d=0.27) for this analysis was found to be small.

Table 8. Paired Samples T-test – The Comparison of ROs

		M	SD	Std. E.M.	95% CI of the D. Lower	Upper	df	Sig.(2-tailed)
Pair 1	RO1MC-RO1SUM	.75	.44	-.9	.07	-.59	31	.000
Pair 2	RO2MC – RO2SUM	.15	.36	.06	.02	.28	31	.023
Pair 3	RO3MC – RO3SUM	.18	.39	.07	.04	.33	31	.012
Pair 4	RO4MC – RO4SUM	.28	.45	.08	.11	.44	31	.002
Pair 5	RO5MC – RO5SUM	-.59	.49	.08	-.77	-.41	31	.000
Pair 6	RO6MC – RO6SUM	-.09	.64	.11	-.32	.13	31	.414
Pair 7	RO7MC – RO7SUM	-.28	.45	.08	-.44	-.11	31	.002
Pair 8	RO8MC – RO8SUM	-.15	.36	.06	-.28	-.02	31	.023
Pair 9	RO9MC – RO9SUM	-.37	.49	.08	-.55	-.19	31	.000
Pair 10	RO10MC – RO10SUM	-.25	.44	.07	-.4	-.09	31	.003
Pair 11	RO11MC – RO11SUM	.93	.24	.04	.84	1.02	31	.000

4. DISCUSSION and CONCLUSION

This study was set out to investigate whether MC tests of reading comprehension allow test takers to form an integral understanding of the texts they read during test taking process. The aim was to test whether through MC questions it might be possible to make the test takers to process all the information in the text so that they can form a coherent summary in their minds. Two equal forms TOEFL reading tests were used for two different summarization tasks in a counter-balanced way; summarization after taking an MC test and summarization without questions (SUMONLY). The comparison of the results have shown that test takers' summarization was less effective after an MC test; they could remember fewer ideas from the text than they would normally remember if they read the text from the beginning to the end. The analyses conducted showed that the level of comprehension required for answering MC questions was not adequate for a satisfactory summary formation as the test takers could remember only half of the main ideas from the texts (48.6%). Verbal protocols confirmed that reading for MC test is strategic: An overwhelming percentage of participants (93.7%) stated that they "followed a question-to-text sequence" as a problem-solving activity where the problem is the question and the text is only a means to answer it. Therefore, questions are prioritized and reading is guided by the questions and maintained as much as the questions required. Following a question-to-text sequence in test taking means that reading the whole text

is not a requirement, which is supported by only a few participants (37.5%) "reading the whole text from the beginning to the end carefully (RO6)" in this study.

The reading skills that were uniquely operationalized in the MC task, but were not mentioned in the SUMONLY task, are also worth consideration as they help differentiate between the two tasks. RO3 "I followed a question-to-text sequence", RO4 "I read expeditiously to find a relevant part that might include the answer" and RO5 "I read carefully only the selected part(s) of the text that might be relevant to the question" were strategic, expeditious reading operations that predictably did not emerge from how the participants in the study described their reading processes for the SUMONLY task. It is, therefore, fair to say that reading a text for the MC task and the SUMONLY task required the execution of different reading operations and that the MC task directed the participants towards the activation of local level selective reading operations, which served only for the answering of the questions but not necessarily understanding the text as a whole. Therefore, we can conclude that text level understanding might be hindered through task-specific strategic question answering in an MC task whereas in linear, careful reading to summarize a text, as there are no interfering processes, text level understanding is more possible. Thus, we can conclude that task specific MC reading operations do not contribute to a deeper understanding of a text; on the contrary, they may even be hindering it.

That is a critical observation to make in terms of reading activities. When the whole text is not read in a test, as proven by the remaining 62.5% of participants who did not mention reading the text in full, interpretations about a test taker's ability based on a representative sample of reading operations cannot be made. Reading for answering questions, which is inherent in MC design, seems to contribute only to finding the answers to the questions as test takers may devote less effort for reading outside the scope of the questions. If in a reading test, test takers can complete tasks without reading the whole text, the test is more like a puzzle activity where test takers find out which information in a text fits the question.

Attaining comprehension at textual level and formation of macrostructure requires a careful, high-level reading process where readers are able to use their ability to integrate information and draw conclusions (Pressley, 2002). When a careful and a linear reading style is not adopted, and most importantly, when reading activity takes place in a segmented and selective manner, formation of macrostructure is not likely to take place because the processes necessary for it cannot be substantiated during such a reading activity as was confirmed by our MCSUM test results. A local and segmented style of reading, whether it is done in a careful or expeditious manner, leaves test takers with only pieces of information from the text. In this study, none but only one participant indicated that they read to understand organization of the text in MC task.

On the other hand, the responses to the summarization task reflected a careful, linear reading style from the beginning to the end of the text. This is quite similar to text processing when the reader needs to understand all the information in the text and learn from it. As Khalifa and Weir (2009) stated, during summarization, high level processes come into play, directing the reading activity so that it follows a global and holistic manner. Such reading operations were reflected in 96.8% of the participants who stated that they "have read the whole text from the beginning to the end carefully" (RO6). Our findings are also supportive of Taylor (2013), who states that summarization tasks represent "a view of text comprehension as the construction of a mental representation of the whole text and they therefore offer an appropriate format for assessing this" (p.56).

Considering the frequent use of MC tests in the assessment of academic reading ability, the question is whether or not an MC assessment tool is an efficient evaluation technique that can account for the relevant reading skills and sub-skills indicative of academic reading ability. Validity of our interpretations based on test results is closely contingent upon whether the test measures what it intends to measure (Brown, 1996). The task used for assessment purposes,

thus, should have a good representation of what a test taker can do in a real-life context. Reading in a real-life academic context requires readers to read at deeper levels and construct both a text model of comprehension and a situation model of interpretation (Grabe, 2009; Kintsch, 1998). More specifically, an important skill in an academic context is to be able to "read and understand a text in its entirety with the purpose of learning from it" and this skill is associated with high-level reading processes where "reading at the whole-text level" (Khalifa & Weir, 2009) and eventually the formation of the macrostructure of it is required. The formation of the macrostructure of a text necessitates the mastery of global comprehension skills and they are associated with understanding explicit information in the text and extracting main ideas and making connections between them to eventually integrate and synthesize information (Bax, 2013; Weir & Bax, 2012). Thus, tests that fall short of assessing ability to understand a long text and to form efficient macrostructure of it cannot help us to arrive at adequate and accurate interpretation of test takers' readiness for academic life.

MC tasks are practical in terms of their administration and scoring and they are definitely useful in assessing several other reading skills. Note that reading is an umbrella term that covers many sub-skills a learner needs to develop to be able to cope with a written text that they encounter in real life. MC tasks, for instance, can be regarded as effective tools to teach and assess the reading sub-skills that require the mastery of sentence and paragraph level reading comprehension and the ability to search for information. In order to teach or to assess textual level comprehension skill, however, we have seen that asking several MC questions on a text is not a helpful format. Global level deeper comprehension should be emphasized both in teaching reading and in the assessment of it and appropriate techniques should be used for this. Classroom instruction and teaching programs preparing learners for academic life should incorporate practices that cultivate and enhance the required academic reading skills. The focus of classroom practices and materials should be to teach learners how to read texts to perform well in different reading types and also process a text fully to extract complete meanings from it (Weir et al., 2009) and create a text and a situation model (Kintsch, 1998) by reading carefully at the whole text level (Khalifa & Weir, 2009).

As mentioned above, accurate interpretations about a test taker's reading ability cannot be made depending on a test that is not representative of relevant reading skills. Similarly, it is not realistic to expect desired outcomes from teaching programs that include practices that are weak in such coverage. As proven to be instrumental in the assessment of textual level comprehension in the current study, summaries, both oral or written, should be utilized more for teaching purposes. Summaries are good indicators of the formation of macrostructures and the macrostructure of the written material reflects reading comprehension more effectively (Kintsch & E. Kintsch, 2005).

In terms of assessment procedures, we need to make sure that decisions concerning the test format are not made at the expense of construct validity. Tests designed for academic assessment purposes, as well as research purposes, should be assessing test takers' reading ability at several levels as required in real life settings where readers need to tackle with long texts which they have to read and understand from the begining to the end. Test tasks should target such higher level integrative, interpretative reading skills as well as expeditious, selective ones to draw a full picture of a reader's ability. Otherwise, they are risking their validity.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID
Burcu Kayarkaya https://orcid.org/0000-0002-3801-6345
Aylin Ünaldı https://orcid.org/0000-0003-4119-6700

5. REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK; New York, NY, USA.

ALTE (2011). Manual for Language Test Development and Examining, Strasbourg, Council of Europe. Retrieved from https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b

Airasian, P. W. (1994) Classroom assessment (2nd ed.). New York: McGraw-Hill.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. (1985). *Becoming a nation of readers: the report of the Commission on Reading*. Pittsburgh, PA: National Academy of Education.

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Language Testing in Practice. Oxford: OUP.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye tracking. *Language Testing*, *30*(4), 441-465. doi:10.1177/0265532212473244

Bernhardt, E. B. (1983). Three approaches to reading comprehension in intermediate German. *The Modern Language Journal*, *67*(2), 111-115. doi:4781.1983.tb01478.x

Britt, M., Rouet, J.F., Durik, A. M. (2017). Literacy beyond Text Comprehension. New York: Routledge, https://doi.org/10.4324/9781315682860

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Cerdan, R., Vidal-Abarca, E., Martinez, T., Gilabert, R., & Gil, L. (2009). Impact of question-answering tasks on search processes and reading comprehension. *Language and Instruction*, *19*(1), 13-27.

Cohen, A. (1984). On taking language tests. *Language Testing*. *1*(1). 70-81. doi:10.1177/026553228400100106

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. doi:10.1177/0265532207076364

Cutting, L.E., & Scarborough, H.S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277-299. doi:10.1207/s1532799xssr1003_5

Enright, M.K, Grabe, W., Koda, K., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: a working paper*. Princeton, NJ: ETS.

Fuhrman, M. (1996). Developing Good Multiple-Choice Tests and Test Questions. *Journal of Geoscience Education*, *44*(4), 379-384. doi:10.5408/1089-9995-44.4.379

Gernsbacher M. A. (1997). Coherence cues mapping during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships in the production and comprehension of text* (pp. 3-21). Mahwah, NJ: Erlbaum.

Graesser A., Singer M., & Trabasso T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371-395.

Grabe, W. (1991). Current Developments in Second Language Reading Research. *TESOL Quarterly*, *25*(3), 375-406. doi:10.2307/3586977

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

Goodman, K.S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist*, *6*(4), 126-135. doi:10.1080/19388076709556976

Gough, P.B. (1972). One second of reading. *Visible Language*, 6, 290-320.

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology*, *79*(3), 220–227. https://doi.org/10.1037/0022-0663.79.3.220

Haladyna, T.M., & Downing, S.M. (2009). A Taxonomy of Multiple-Choice Item- Writing Rules. *Applied Measurement in Education*, *2*(1), 37-50. doi:10.1207/s15324818ame0201_3

Khalifa, H., & Weir, C. (2009). Examining Reading: Research and Practice in Assessing Second Language Reading. *Studies in Language Testing*, *29*. Cambridge: Cambridge University Press.

Keenan, J.M., Betjemann, R.S., & Olson, R.K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281-300.

Kintsch, W. (1998). *Comprehension: a paradigm for cognition.* Cambridge: Cambridge University Press.

Kintsch, W., & Kintsch, E. (2005). Comprehension. In Paris, S. G. and Stahl, S. A. (eds.) *Children's Reading Comprehension and Assessment*, 71-92. Mahwah, New Jersey: Erlbaum.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*(2), 193-220.

Kurz, T.B. (1999). *A review of scoring algorithms for multiple choice tests.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Lau, P.N.K., Lau, S.H, Hong, K.S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple choice tests. *Educational Technology & Society*, *14*, 99-110.

Lee, J.F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition*, *8*, 201-212.

Lim, H. J. (2014). Exploring the validity evidence of the TOEFL IBT reading test from a cognitive perspective. *Unpublished PhD Thesis*. Michigan State University.

Martinez, M. E., & Katz, I. R. (1995). Cognitive Processing requirements of Constructed Figural Response and Multiple-Choice Items in Architecture Assessment. *Educational Assessment*, *3*(1), 83–98. doi:10.1207/s15326977ea0301_4

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

Myers J. L., & O'Brien E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*, 131-157.

Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing*, *25*(1) 127-143.

Pearson, P. D., Garavaglia, D., Lycke, K., Roberts, E., Danridge, J., & Hamm, D. (1999). The impact of item format on the depth of students' cognitive engagement. Washington, DC: *Technical Report*, American Institute for Research.

Pressley, G.M. (2002). Metacognition and self-regulated comprehension. In Farstrup, A.E., & Samuels, S.J. (Eds.), What research has to say about reading instruction. Newark, DE: International Reading Association.

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

Scouller, K. (1998). The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. *Higher Education*, *35*, 453-472.

Sheehan, K.M. and Ginther, A. (2001). What do passage-based MC verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section. Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147-170.

Smith, M. (2017). Cognitive Validity: Can Multiple-Choice Items Tap Historical Thinking Processes? *American Educational Research Journal*, *54*(6), 1256-1287.

Taylor, L. (2013). Testing reading through summary: investigating summary completion tasks for assessing reading comprehension ability. *Studies in Language Testing*, *39*. Cambridge, England, UCLES/Cambridge University Press.

Unaldi, A. (2004). *Componentiality of the reading construct: Construct validation of the reading subskills of the Boğaziçi University English Proficiency Test.* Unpublished PhD Thesis. Faculty of Education, Boğaziçi University.

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

Watson Todd, R. (2008). The impact of evaluation on Thai ELT. In Ertuna, K., French, A., Faulk, C., Donnelly, D., and Kritprayoch, W. (Eds.), *Proceedings of the 12th English in South East Asia conference: Trends and Directions*, pp.118-127. Bangkok: KMUTT

van Dijk, T.A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction and cognition.* Hillsdale (N.J.): Lawrence Erlbaum Associates.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281-300. https://doi.org/10.1191/0265532205lt309oa

Weir, C, Hawkey, R, Green, T., & Devi, S. (2009). The cognitive processes underlying the academic reading construct as measured by IELTS. IELTS Research Reports, 9, 157–189, British Council, London and IELTS Australia, Canberra.

Weir, C., Bax, S. (2012). Investigating learners' cognitive processes during a computer-based CAE Reading test. Cambridge Research Notes, Cambridge ESOL, 47 (February 2012), 3–14. Retrieved from www.cambridgeesol.org/rs_notes/rs_nts47.pdf

Wolf, D.F. (1991). *The effects of task, language of assessment, and target language experience on foreign language learners' performance on reading comprehension tests.* Dissertation, University of Illinois, ProQuest Dissertations and Theses.