Iraqi Graduate Students' Analysis of the Graduate Record Examination (GRE) Verbal Reasoning Content: A Think-Aloud Study

Iman Ibrahim Khudhair
University of Mosul, Iraq

David Boote
University of Central Florida, United States

This study describes difficulties Iraqi graduate students encounter while answering Verbal Reasoning Measure (VRM) questions of the Graduate Record Examination (GRE). Participants thought aloud while answering questions from three VRM sub-sections: Reading Comprehension (RC), Text Completion (TC), and Sentence Equivalence (SE). Data were collected via audio recordings, from seven Iraqi graduate students, and transcribed orthographically, coded, and analyzed. The analysis yielded several themes: Difficulties related to the structure of the question and/or the meaning of its answers, the content is pertinent to USA culture and history, SE is the most difficult part followed by TC, and RC is time consuming. These findings are consistent with previous research results that the structure of the GRE VRM questions and/or the answers are difficult to understand, SE items were much harder than TC and RC items due to their unfamiliar content and vague meanings, and RC is time consuming for international graduate students. Pedagogical implications were identified and recommendations for ETS and students were suggested.

Keywords: think aloud, GRE, verbal reasoning section, difficulties, graduate students.

Assessing students' performance has long been the core aim of teaching and learning domains. For English language, this assessment process is conducted via quantitative or qualitative methods based on the procedure used to collect data and the type of the research design. Quantitatively, various versions of survey, online voting, and longitudinal research were and are the used methods. Qualitatively, interviews, observations, personal journals, and audio and video recordings are the prominent methods for assessing language knowledge. Graduate Record Examination (GRE) is one of the quantitative methods used by Educational Testing Services (ETS) to assess a wide range cognitive abilities of graduate students to choose the best students to continue their next graduate education. Most US universities depend on the (GRE) score, interviews, undergraduate grades, and recommendation letters as vital factors to nominate master or doctoral students (Stack & Kelley, 2002). A particular GRE total score, with sometimes specified score for each section, is required from both domestic and international students to start their graduate academic courses. However, international graduate students seem to face more difficulties in this test than their domestic counterparts due to various reasons. Although ETS attempts to ensure the impartiality and fairness of the GRE test items via assessment specialists and statistical procedures like Differential Item Functioning (DIF) (Raquel, 2019), international graduate students are still complaining about the problematic parts of this test. Therefore, it is necessary to use methods that are based on students' experience and perception to obtain a clearer

Correspondence concerning this article should be addressed to Iman Ibrahim Khudhair, **Ph.D. ,** University of Mosul, College of Arts, Iraq .Email: iman.ibrahim@uomosul.edu.iq

picture about the cognitive process that occurs while taking the test under time pressure. Hence, this study aimed at: first identifying the difficulties international Iraqi graduate students face during taking the VRM using an innovative qualitative method known as think-aloud approach, and second to see if the test's time-limitation impact students' performance. This method was used in this study because it is an efficient educational strategy and a vital tool in cognitive assessment as it relies on immediate articulated thoughts of graduate international students (Charters, 2003).

Literature Review

In this section, the think-aloud approach, the reasons behind choosing it to analyze data, and the structure of the GRE Verbal Reasoning measure are presented.

What is Think- aloud Approach?

A Think-aloud method is a research approach that involves vocalizing participants' internal thinking as they are engaged in completing a given educational task or a problem. A review of prior research has indicated that this method has solid theoretical grounds when it started in the 1960s originating in the cognitive psychology field (Boren, & Ramey, 2000). However, it developed through years and stabilized in 1984 when Ericsson and Simon published their book *Protocol Analysis: Verbal Reports as Data* in which they set a model that values heedful consecutive verbal data produced by participants while solving a task. Today, think- aloud approach is considered as an effective instructional strategy and a critical tool in cognitive assessment as it is based on articulated thoughts (Charters, 2003). Also, it has sundry advantages, such as: providing an unmodified performance response, real situational assessment about the given task, and flexibility of cognitions according to the situation (Davison, Vogel, & Coffman, 1997). In addition, some scholars argue that the efficient use of this approach can positively leverage students' achievement (Ness, 2016).

Think aloud approach has two types of interviews or analyses: protocol and verbal (Leighton, 2009). *Protocol analysis* (interview) is used to measure and identify a problem-solving process where participants are required to provide a verbal record of their instant thoughts as they solve a task. The focus on simultaneous solving and thinking loudly is vital in protocol analysis to recognize immediately the working memory contents or short-term memory of the student (Ericsson, 2006). *Verbal analysis*, on the other hand, is used to measure and identify student's knowledge including attitudes and beliefs. Verbalizing thoughts simultaneously while solving a task is not needed in verbal interviews as students can articulate what they are thinking after or during solving the problem. This is due to the reliance of this kind of interview on the information obtained from long-term memory instead of short-term memory (Chi,1997). As explained above, the difference between both types is based on the memory location used during the think- aloud interview. Protocol interviews utilize short -term memory, while verbal interviews exploit long-term memory (Leighton, 2009).

Since the main purpose of this paper is to identify various difficulties graduate students encounter in the VRM section of the GRE test, protocol analysis was used. In the data collection and analysis, we assumed that answering every question was a problem-solving task and the goal of the analysis is to understand the cognitive processes participants used while answering the question, correctly or incorrectly. By collecting the immediate verbal record of their thoughts simultaneously as they solve each VRM questions, we rely only on their short-term memory and avoid the problems associated with replying on their long-term memory.

Verbal Reasoning Measure

GRE is an aptitude test that is required from both native and non-native students to begin their graduate degrees. It is mainly used in the United Sates to measure students' cognitive skills

to select the best students who are expected to excel in a graduate program (Bridgeman, Cline, & Hessinger, 2004). It has three sections, each designed to assess certain cognitive traits: Analytical Writing, Verbal Reasoning, and Quantitative Reasoning. For the purposes of this study, only the VRM is considered.

VRM items are designed to measure graduate students' ability to evaluate and analyze written texts and deduce information from them. This section also aims to analyze connections among sentence constituents and identify relationships among concepts and words (ETS, 2022). VRM has three parts: Reading Comprehension, Text Completion, and Sentence Equivalence.

Reading Comprehension (RC) items measure various cognitive abilities that are critical to read, analyze and understand the types of texts they may encounter in their graduate programs. Such abilities include recognizing between major and minor points, comprehending the meaning of words and sentences, deducing conclusions from the given text, specifying the assumptions and options of the author. This section contains about ten passages, and each question relies on the passage that may include one paragraph or several paragraphs. Passages are selected by assessment specialists from materials that are taught and studied in United States' schools in various academic fields. Items may include any of the above listed topics, from identifying an individual word's meaning to evaluating sentences to provide evidence that may weaken or strengthen points mentioned in the text. Most of the items are multiple-choice in form, and students are supposed to choose one correct option; other item types require selecting two correct answers; while others require choosing a whole sentence from the text (ETS, 2022).

Text Completion (TC) items take the form of short passages that have omitted key words. Graduate students are required to use the available text to infer and select short phrases or words to complete the given sentence(s). The most common TC item structure consists of one to five sentences, with space(s) to be filled and three choices for answering each blank. TC items will sometimes include one blank where five choices of words or phrases are provided (ETS, 2022).

Sentence Equivalence (SE) items focus on the synonymous meaning of two answer choices. The structure of these questions is a single sentence with one blank space. Test-takers are required to select two answers, from six choices, to form two coherent and complete sentences that have almost the same meaning when using the two selected choices. Both TC and SE rely mostly on the accumulative knowledge of English language vocabulary.

Think-aloud Approach and VRM

The use of the think-aloud approach in educational environments was impelled by cognitive psychological researchers to find new instruments and methods to collect qualitative data that are as close as possible to the real-world. Thus, scholars like Simon and Ericsson (1984) originated the innovative method of the think-aloud approach that relies entirely on students' articulated thoughts to attain data that are as real, constant, undisturbed, and undirected as possible. Although their model was developed in the field of cognitive psychology to study certain kinds of problem-solving behaviors, its use has expanded widely to collect cognitive process data in many other fields, including educational measurement (Snow & Lohman, 1989), reading comprehension (Pressley & Afflerbach, 2012), and different engineering fields (Sanderson, 1990). Some of the aforementioned fields have similar research purposes to those of cognitive psychology in that they want to deduce how cognitive processes are happening in specific domains (Boren, & Ramey, 2000). It is also used to ensure the validity of psychological and educational tests and ensure that the test or its items are not inadvertently assessing something other than what it claims to be assessing (AERA/APA/NCME, 2014). The use of protocol analysis in the field of educational measurement helps to ensure that participants' thinking processes are aligned with what the test

claims to be assessing (Snow & Lohman, 1989). This integration has since become the core topic for research in the last four decades not only by scholars (e.g.(Li, Kannry, Kushniruk, Chrimes, McGinn, Edonyabo, & Mann, 2012) but also by governmental organizations and assessments boards in the United States like the National Research Council (Grubbs, Strimel, & Kim, 2018)

We selected think-aloud approach as a data collection method for this study for two reasons: previous evidence of this approach's effectiveness in assessing students' cognitive process in educational and assessing environments (Leighton, 2009), and it helps in highlighting individual variances in response among the sample. From the two interview methods the think-aloud approach has, protocol analysis was chosen for this study not the verbal analysis as the researchers needed to know through the data produced by the participants the difficulties they face while solving the VRM items, which requires students to depend on the immediate cognitive analysis they do after reading a question and while solving the question. This immediate cognitive analysis is based on the information available in the short-term memory that the protocol analysis relies on. In addition, the data collected by protocol analysis is more reliable and valid as they are unaltered or modified (Kuusela, & Paul, 2000).

Protocol analysis has been used to study the four communication skills: reading, writing, listening, and speaking (Ness, 2016). However, no research has been conducted to investigate graduate students' thinking processes while they are answering items like those used in the GRE VRM section. Most published research focuses on the predictivity of the GRE test scores on graduate students' performance in their graduate programs, using quantitative methods (Petersen, Erenrich, Levine, Vigoreaux, & Gile, 2018). None of these studies has addressed Iraqi doctorate students' thinking processes while taking the GRE test which is the main goal of this study.

The main aim of this investigation is to understand the difficulties that Iraqi graduate students encounter while answering GRE VRM items. The results of this study may be interesting to test designers who wish to ensure that test items are as relevant as possible to the academic fields of the students. The results will also interest those who help international students prepare for the GRE VRM. To accomplish this goal, the following questions guided the design of the study:

1. What are the difficulties that Iraqi graduate students face while solving GRE VRM questions?

2. How does the time-limitation of the VRM impact students' performance in the test?

Method

Before starting data collection for this study, in person informed consent was obtained from the participants who agreed to take part in the study voluntarily and to audio record their voice during the interviews. The purpose and importance of the study were explained in the consent and the researchers assured that participants that their personal information will be kept completely anonymous.

Sample

The sample consisted of seven Iraqi graduate students at two large research-focused universities in the Southeastern United States, four males and three females. Both universities require all applicants to submit recent GRE scores to be considered for admissions. Those students were selected due to their nationality (Iraqi), Ph.D. candidates, and have taken the GRE test at least once before or during their PhD study. Participants were in various stages of their Ph.D.

programs, including Computer Engineering, Teaching English to Speakers of Other Languages (TESOL), Analytical Chemistry, and Biology.

Table 1

Demographics of the Participants

Participant #	Program	Gender	Year	University	# of times	Place	Required Score
1	Comp. Eng.	male	Fourth	UCF	2	Iraq and USA	309
2	Comp. Eng.	male	Third	UCF	2	USA	309
3	Comp. Eng.	male	Third	UCF	3	USA	309
4	Comp. Eng.	male	Second	UCF	2	USA	309
5	Analyt. Chem.	female	Fifth	UCF	2	USA	310
6	TESOL	female	Second	UCF	3	USA	300
7	Biology	female	Fifth	UF	2	Iraq and USA	312

Table 1 above illustrates the demographics of the participants, the number of times they have taken the test, and the total GRE score required by each program according to both universities' websites.

Methodologically, recent research has questioned the validity of protocol analysis data when participants are not fluent in the language of data collection (Ericsson & Simon, 1984). This is simply because the added cognitive effort of translating one's thought reduces the cognitive resources available for task completion. To avoid these problems, the interviewer made two adjustments to standard think aloud methods. First, because all participants shared the same dialect of Arabic as their first language, participants were told they could code switch during the tasks, thinking aloud in the language in which they were thinking. It also allowed ease of communication with the first author if the researcher or the participants needed clarification. Second, some participants were silent when they first read the questions, especially the longer reading comprehension passages. When they did this, the interviewer allowed them an initial period of silently reading so that they could focus on comprehension before reminding them to "keep talking." This change led to the loss of some data but made the study design viable.

Data Collection

Methodological research on protocol analysis has shown that think aloud is not an appropriate method of data collection for certain tasks and populations, including very young children, problem tasks that are either too easy or too difficult for the participants, people with difficulties processing verbal information, and people who are not adequately fluent in the language of data collection (Ericsson & Simon, 1993). However, we used it as we assumed that VRM items were neither too easy nor too difficult. Anyone who had too much difficulty processing verbal information would not have been able to complete our training tasks.

In this study, the authors used VRM questions from the paper-based GRE test taken from *The Official Guide to the GRE Revised General Test,* (2nd edition) that was published by Educational Testing Services (ETS) in (2012). A full version of the VRM, from pages 309-318, with its assigned allocated time was applied to the sample of this study to simulate the situation of a real GRE test to obtain as close results from the participants as possible to those of the real test. The VRM used in this study included 25 questions that varied in length and difficulty and 35 minutes to answer those questions. The number of questions for each section in the VRM is almost fixed in each try with a largest portion for the RC (12-13 questions) followed by TC (7-8 questions), and SE (4-5 questions). In our study, RC questions were 12, TC were 8, and SE were 5.

The data collected for this study included seven audio-recorded protocol interviews answering GRE VRM questions. Each interview was conducted individually. Data were collected through a period of three months in 2021-2022 year. The test was administered in English, but participants could code switch to reflect their thinking. At the beginning of each interview, participants were required to complete a series of "training" tasks to ensure that they understood what it meant to think aloud during the test, especially the need to continually verbalize their thinking and to avoid efforts to explain or justify their thinking to the interviewer. When the participants demonstrated that they could think aloud effectively, they began answering the GRE test items. The researchers did not explain the test items or how they should be answered, presuming that they had recent familiarity with the test and question types.

Although efforts to remind participants to not explain or justify their thinking, several participants did so spontaneously and persisted in doing so. This behavior was most evident when participants struggled to comprehend the questions and/or answer choices. The interviewer allowed participants to explain their strategies for answering the question or their frustrations with the questions, assuming that these explanations were at least in part an authentic reflection of their thinking even if they were not a verbatim record of their cognitive processes.

Data Analysis

The interviews are divided into four sub-sections according to the fields of the participants: Computer Engineering (4 participants), Analytical Chemistry (1 participant), TESOL (1 participant), and Biology (1 participant). However, the same VRM version was used, and the same procedures were followed with all participants.

Computer Engineering Participants

Four Iraqi graduate males participated in the study who were pursuing their Ph.D. in various years of the program. The first participant was in his second academic year in the Ph.D. program The interview took 47 minutes and 55 seconds. When he started with RC, the participant did not want to understand the meaning of the question but wanted to understand the overall meaning of the passage and tried then to know the correct answer. Since this strategy failed, he went back again to the meaning of the sentence to get clues from the context. The participant said that "RC was time consuming as it took me about 15 minutes to understand the first paragraph of one of the passages because it talked about America's history, which was unfamiliar to me". Although he used deletion and replacing strategies in the TC and SE sections, he commented that "they have difficult options with similar meaning, or the question was sometimes difficult to understand."

The second participant was in his third year of the academic program. The interview went for 45 minutes and 45 seconds. When the participant started with the RC questions, he read the passages silently for himself for 2-5 minutes and then started solving the questions. However, he kept going back and forth between the text and questions saying: "I am not very familiar with most of the words or vocabularies mentioned in the text". Another issue according to this participant is the unfamiliar culturally subject of some passages that addressed for example US history, space, biology, etc.

As for the TC questions, the participant used the substitution strategy where he used each of the given options in the blank to see if it makes sense for him. The participant mentioned that he faced more difficulties in answering the questions that have two-blanks than those with one blank. SE part was according to the participant "a nightmare" as he stated that "I understand the meaning of the question but the problem when I read the given choices, I become shocked

since I do not have any idea about most of them, therefore; I selected the answer relying on my sense. This part is definitely the hardest one among all three."

The third participant is in the same academic year of the previous participant, third year. The interview went for 30 minutes and 34 seconds. When he started working on the RC questions, he analyzed the question and its options simultaneously to see which answer fits more. He used elimination strategy to get closer to the correct answer. However, he stated that "this part (RC) of VRM is time consuming as I re-read some of the texts multiple times to understand them." As for the TC and SE questions, the participant chose the answers randomly as he thought "it is a waste of time to try to answer something I do not understand."

The fourth interviewee was in his last year in the program. The interview took 38 minutes and 57 seconds. He used the elimination strategy with RC questions. Then, the participant used logic in interpreting the given text to reach to the correct answer. Although it is a very good strategy, it is time consuming. He added saying that "this section was somehow difficult due to its limited time and type of passages that were non-academic in nature". As for SE and TC items, the participant mentioned that their "questions were much more difficult than those of the RC," which explains his random choices of the answers.

The Analytical Chemistry Participant

The fifth participant was from Analytical Chemistry department, College of Sciences at UCF. The participant is a female in the final year of the Ph.D. program. The interview took 38 minutes and 53 seconds. She read every item of the VRM loudly while solving them one by one. The researcher code switched a little bit to explain the procedures but did not help the participant with anything else about the test. Due to her specialization in Chemistry and sciences in general, she mentioned that "the passages of RC were understandable, but some of their questions were difficult, especially the inferring items." The most difficult part of the VRM for her was SE since according to her "it requires a massive vocabulary knowledge about every topic, which unfortunately I do not have." As for TC questions, the participant mentioned that "the structure of the sentences was difficult and also the choices were really hard to understand". Therefore, she said that she had two choices: either to understand the sentence and be able to answer it which requires more time, or not getting the meaning of the sentence and thus depend on luck when picking the answer.

The TESOL Participant

The sixth interviewee showed higher linguistic skills and more sophisticated thinking than previous participants in terms of English vocabularies due to her background knowledge in English language. The participant was a female graduate student in the Ph.D. Teaching English to Speakers of Other Languages (TESOL) track, in the second year. The think aloud procedure went for 40 minutes and 6 seconds. The interesting thing is that she worked on the choices first before the question itself, a technique she reported that she learned from GRE preparation books. Yet, the participant admitted that "it was not helpful with the given questions of the RC."

In addition, she had similar opinion to that of other participants about the difficulty of either the question or the options and sometimes both saying that "sometimes I know the meaning of all choices, but the question is difficult to understand, and sometimes it is vice versa." In these situations, the participant used the structure of complex sentences to reach the correct option(s). For example, although she said that the questions of SE "were the most difficult ones," she was able to identify the key words in the sentence depending on the structure of the sentence.

The participant added that the one-blank items in the TC section were easier than the two blank ones. Agreeing with the other participants, she stated that "some SE items were very difficult as they require previous knowledge about the US history. Also, both the structure and

options were not understandable." According to her, SE questions was harder than those of TC because she" (has) to figure out the two choices that are similar in meaning." Therefore, she looked at the choices first and after locating every pair, she tried to see if they would fit the meaning of the sentence.

The Biology Participant

The last participant was a female graduate student in a Ph.D. Biology program, in the fifth year. The think aloud procedure went for 44 minutes and 6 seconds. When the participant started with RC questions, she stated that "she needs more time for this section due to the unclear content of some passages. However, I was able to answer the questions related to sciences as it is part of my specialization." According to her, "TC questions were difficult though I relied on the structure and context to analyze and answer them." The participant shared the same viewpoint expressed by other participants about SE questions when she stated: "the questions of the SE were the most difficult among all as they require rich background knowledge in vocabularies and their use in various contexts." After the interview, the participant said that she felt pressured due to the time limitation required to complete solving VRM section, which impacted her answers and her total grade in the test.

Table 2 shows the main strategies used by each participant to solve VRM questions. Most of the participants used elimination, context clues, deletion, sentence structure, and matching the options with the question strategies to solve the RC items, while they used substitution, key words, logic to solve SE and TC items.

Table 2

Participants' Strategies for Solving VRM's Questions

Participant #	Program	Gender	Strategies
1	Comp. Eng.	Male	-Elimination of direct information/ logic (sense) (RC). - Random answer selection for (TC) and (SE)
2	Comp. Eng.	Male	-Analyzing the question and answers simultaneously (RC) - NO STRATEGIES WERE USED FOR (TC and SE).
3	Comp. Eng.	Male	-Reading silently (RC). -Context Clues (RC) -Substitution (TC) -Reading the Q more than once (TC) - Logic (sense) (SE)
4	Comp. Eng.	Male	-Structure Analysis (RC). - Substitution (TC and SE) - Deletion (TC and SE)
5	Analyt. Chem.	Female	-Reading loudly (RC/TC/SE) -Code switching (RC) -Context clues (RC) -Trying to get the meaning of the sentence (TC) -Answer randomly (TC)
6	TESOL	Female	-working on the answers first (RC/TC/SE) -code switching (RC) -Sentence structure analysis (TC) -Key words (TC) -Identifying the synonymous pair first and fit them in the question. (SE)
7	Biology	Female	*-Background knowledge (RC)* -Sentence structure analysis (TC) -Context clues (TC) -Random answers (SE)

Results and Discussion

The think aloud audio recordings were transcribed carefully, then differences and similarities among the main statements of the participants were distinguished. After that, those sentences were unified and coded to have the following four themes:

1. *Difficulty of either the structure of the question or the meaning of its options.*

All the participants agreed that most of VRM questions are difficult either in terms of the structure of the question or the meaning of its options and sometimes both. Because analyzing the structure and/ or selections is time consuming, participants preferred to select the answers randomly and move to the next question. These difficulties were reported in the SE and TC sections by Computer Engineering and Analytical Chemistry participants. Although participants used various strategies to solve questions, they faced problems in the two blank questions of SE more than those including one blank. The difficulties faced in SE and TC sections might be attributed either to complexity of the structure of the question or their unfamiliarity with the meaning of the options in addition to the limited time needed to analyze and solve such questions. Therefore, they had two choices: either to understand the question and maybe answering it which requires more time or could not understand the meaning of the question and depend on their sense in selecting choice(s).

2. *Culture or history related topics.*

Some VRM questions need previous knowledge about US history and/or culture. Since most Iraqi students are non-native speakers who do not have such knowledge, they struggle in answering those questions and therefore become obliged to select any answer(s) they think related. This difficulty was reported by most interviewees in this study which means that it is not related to the knowledge level a participant has, it is rather an essential one also for many international students who take this test. Inserting cultural and historical items in the GRE test are obvious examples of content bias against international students who should be tested for their analytical skills not cultural and historical knowledge of USA. Therefore, the validation of the items could be improved by ETS through taking into consideration international students' direct opinions about the difficulty level of some items before considering them in a real test or having a panel of non-native experts who can point out issues in the test items.

3. *Sentence Equivalence is the most difficult section followed by Sentence Completion.*

All participants agreed on the fact that SE questions were the hardest ones in the VRM followed by TC and RC questions. Even though the participants used different techniques they learned from GRE preparation books, they reported that SE questions were so frustratingly difficult as they require knowing the precise meaning of two synonymous words that are new. Also, this might be attributed to the double- task issue they must do on SE questions of understanding the meaning of the question and locating the synonymous pair that would fit the meaning of the sentence. In addition, having more than one synonymous pair per item adds more difficulty for the students who needed more time to cluster the synonymous pairs and see which pair fits more to the given question. This issue needs a deeper investigation in a collaboration with ETS itself by getting a GRE test blueprint and evaluating its items by English language experts to understand better if those difficulties result from the students' lack of knowledge or the unfairness of the questions to non- native students.

4. *Reading Comprehension is time consuming.*

Most of the participants agreed that although RC is the least hard section, its questions require more time to answer due to either the content of the RC questions that is drawn from (non) academic books and journals from various domains or the cultural and historical topic of the question that is unfamiliar for international students. The students used various strategies to help them complete this section successfully, but this did not work all the time. In other words,

analyzing structure and trying to understand the meaning of the sentence and choices are time consuming. This difficulty was doubled for the sample of this study since they are international students and need longer time to understand the questions than their domestic counterparts. For these reasons, Iraqi graduate students who took the test reported that they wasted almost half of the allotted time to only read and understand the questions instead of answering them.

Implications and Suggestions

The most important implication of this study's findings is related to GRE test designers who should consider using either analysis of the think-aloud approach; protocol or verbal to develop test items since this approach allows them to identify the kind of psychological variables needed to be measured, especially those related to knowledge structure or problem-solving process. The results of this study are also important for test takers as they can also utilize the think aloud approach, to first monitor their performance and comprehension, and second to prepare for the test in a better way. Thus, when students think aloud, they can assist themselves and others since connections between cognition, oral language, reading, and writing, are made while acting as a paradigm for their student audience (Khudhair, 2020). Finally, psychometricians should consider using protocol analysis to develop or validate a cognitive model of performance tasks, especially in large-scale standardized tests like the GRE.

The findings of this study are also important for teachers who can use the think-aloud protocol and explicit meta-cognitive instruction to train the attention and working memory of students (test-takers) since preparing to answer large-scale tests, like the GRE, relies mainly on the working or short memory and the attention to fine details in the questions (Makmee, 2022; Nasim et al., 2022)

Limit and Scope of the Study

This study has some limitations. First, since the study was implemented only to explore difficulties that Iraqi doctoral students face while solving VRM questions, it mainly focused on this section of the GRE. However, another study can be conducted to the other sections of the GRE test with a different sample. Moreover, the present study is limited to two research questions, more questions can be added. The sample size was also limited to seven students only which makes the results ungeneralizable.

Conclusion

Difficulties in GRE's VRM were investigated in this study via recording and analyzing Iraqi Graduate students' direct opinions using think aloud approach. It has been found that questions which require merely vocabulary knowledge, like those of SE and TC, were much more difficult than the questions that depend on the context, like RC's questions. Also, it has been concluded that international students face more difficulties in solving VRM's questions than native students due to their unfamiliarity with the meaning of many vocabularies and the nature of the content of many questions that are related to the culture or history of USA. As far as RC questions are concerned, the results indicated that although this section was the easiest in the VRM, it was very time consuming because of the topics of the given texts that were sometimes non-academic and /or heritage based. Future research can tap into other important issues related to the large-scale test, such as comparing difficulties faced by graduate students from various cultural backgrounds in TEOFL or IELTS test, examining the effect of linguistic proficiency on performance, or exploring the efficiency of certain test preparation strategies.

References

American Educational Research Association, APA, NCME. (2014). Retrieved in Oct., 2022 from:https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Bridgeman, B., Cline, F., & Hessinger, J. (2004) "Effect of Extra Time on Verbal and Quantitative GRE Scores", *Applied Measurement in Education*,17:1,p25-37.

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. IEEE transactions on professional communication,43(3),261-278.

Charters, E. (2003). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. Brock Education Journal,12(2).

Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. The Journal of the Learning Sciences,6,271-315.

Ericsson, K. A., & Simon, H. A. (1993). Protocol Analysis. Cambridge, MA: The MIT Press.

Educational Testing Services. (2022). Overview of the Verbal Reasoning Measure Retrieved in February2022,from:https://www.ets.org/gre/revised_general/prepare/verbal_reasoning/#:~:text=The%20Verbal%20Reasoning%20measure%20of,relationships%20among%20words%20and%20concepts

Educational Testing Services. (2022). Reading Comprehension Questions. Retrieved in February2022,from:https://www.ets.org/gre/revised_general/prepare/verbal_reasoning/reading_comprehension

Educational Testing Services. (2022). Text Completion Questions. Retrieved in February 2022, from:https://www.ets.org/gre/revised_general/prepare/verbal_reasoning/text_completion/

Educational Testing Services. (2022). Sentence Equivalence Questions. Retrieved in February 2022,fromhttps://www.ets.org/gre/revised_general/prepare/verbal_reasoning/sentence_equivalen

Educational Testing Services.(2012). The Official Guide to the GRE Revised General Test, 2nd edition. New York: McGraw-Hill.

Davison, G. C., Vogel, R. S., & Coffman, S. G. (1997). Think-aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm. Journal of consulting and clinical psychology,65(6),950.

Ericsson, K.A. (2006). Protocol analysis and expert thought: concurrent verbalizations of thinking during experts' performance on representative tasks (pp.223-241). In K.A. Ericsson, N. Charness, P.J. Feltovich, R.R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge, UK: Cambridge University Press.

Grubbs, M. E., Strimel, G. J., & Kim, E. (2018). Examining design cognition coding schemes for P-12 engineering/technology education. *International Journal of Technology and Design Education*, *28*, 899-920.

Khudhair, Iman Ibrahim. (2020). "An Investigation of Factors Predicting Academic Writing Difficulties Among First-Year Doctoral Students." A Ph.D. dissertation, UCF.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. American journal of psychology,113(3),387-404.

Leighton, J. P. (2009). Two types of think aloud interviews for educational measurement. NCME.

Li, A. C., Kannry, J. L., Kushniruk, A., Chrimes, D., McGinn, T. G., Edonyabo, D., & Mann, D. M. (2012). Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. *International journal of medical informatics*, *81*(11), 761-772.

Makmee P. (2022). Increasing Attention and Working Memory in Elementary Students Using Mindfulness Training Programs. FWU Journal of Social Sciences,Vol.16,No.3, 107-119.

Nasim, S. M. et al. (2022). Identifying Metacognitive Listening Comprehension Strategies of Saudi ESP Students. FWU Journal of Social Sciences,16(2).

Ness, M. (2016). Learning from K – 5 teachers who think aloud. Journal of Research in Childhood Education. 30,(3),282–292.

Petersen, S. L., Erenrich, E. S., Levine, D. L., Vigoreaux, J., & Gile, K. (2018). Multi-institutional study of GRE scores as predictors of STEM PhD degree completion: GRE gets a low mark. *PLoS One*, *13*(10), e0206570.

Pressley, M., & Afflerbach, P. (2012). *Verbal protocols of reading: The nature of constructively responsive reading*. Routledge.

Raquel, M. (2019). The Rasch measurement approach to differential item functioning (DIF) analysis in language assessment research. In *Quantitative Data Analysis for Language Assessment Vol. I* (pp. 103-131). Routledge.

Sanderson, P. M. (1990). Verbal protocol analysis in three experimental domains using SHAPA. In Proceedings of the Human Factors Society Annual Meeting (Vol. 34, No. 17, pp. 1280-1284). Sage CA: Los Angeles, CA: SAGE Publications.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp.263-331).

Stack, S. & Kelley, T. (2002) The graduate record examination as a predictor of graduate student performance: The case of criminal justice", *Journal of Criminal Justice Education*, 13:2,335-349.

Wang, W. (2013) Testing the Validity of GRE Scores on Predicting Graduate Performance for Engineering Students, *Open Access Theses and Dissertations from the College of Education and Human Sciences.* Paper 192.