Tabaran Institute of Higher Education
International Journal of Language Testing
In press

ISSN 2476-5880
https://www.ijlt.ir

Behavioral Cognitive Assessment Scrutinized in Language Testing and Vocabulary Size Test

Zari Saeedi[1]*, Hessameddin Ghanbar[2], Mahdi Rezaei[3]

ARTICLE INFO

Article History:
Received: September 2023
Accepted: November 2023

KEYWORDS
Cognitive load
English vocabulary
Assessment
Perceived difficulty
Response-time

ABSTRACT

Despite being a popular topic in language testing, cognitive load has not received enough attention in vocabulary test items. The purpose of the current study was to scrutinize the cognitive load and vocabulary test items' differences, examinees' reaction times, and perceived difficulty. To this end, 150 students were selected using cluster/convenience-sampling, and took the Cambridge Placement Test (CPT) and Vocabulary Size Test (VST; Nation & Beglar, 2007). After uploading the vocabulary-size test's items in PsychoPy software, there was a behavioral stage to measure students' reaction times and correct responses. Out of these 150 high school students, a total of 60 (20 from each proficiency level of elementary/intermediate/advanced groups) were selected. In this quantitative study, all 60 students were interviewed to determine their perceived difficulty of the international VST items and their item's difficulty-index. The data were analyzed quantitatively via simple regression and qualitatively through the examination of the students' perceived difficulty. The results and interview findings revealed a significant connection between cognitive load/reaction time, difficulty estimate, and perceived difficulty at intermediate level. In contrast, at elementary and advanced levels, these variables could not predict the cognitive load. The findings can help to test, course, and syllabus designers by educating them on the significance of cognitive load theory so that they can base their exam designs on its premises and alleviate students' increased cognitive-workload.

1. Introduction

According to a number of scholars (e.g., Gass et al., 2013; Ponce et al., 2020), test developers and psychometricians are concerned about understanding and improving language tests' psychometric qualities. Cognitive processing in test items is crucial for understanding language acquisition and performance, as these processes significantly impact individuals' cognitive processes. In this context, utilizing cognitive load theory (CLT), the influence of item functioning on people's minds has been explored in cognitive processing (Dindar et al., 2014). Sweller's cognitive load concept, developed in the 1980s, analyzes multimedia learning patterns and patterns of cognitive burden (Brünken et al., 2003; Wiebe et al., 2010), instructional materials, learning, and teaching (Sweller et al., 1998, 2019). Cognitive testing refers to the distribution of test-takers' mental activity power during the test (Sweller, 1988). Language test performance is influenced by cognitive load, specifically intrinsic load, which refers to inherent task features attributed to task difficulty (De Jong, 2010; Paas et al., 2003). There has been an emerging tendency toward using item difficulty as a potential objective measure and reliable predictor of mental workload (Ehrich et al., 2021). The fundamental properties of test items are thought to be

[1] Department of English Language and Literature, Allameh Tabataba'i University, Tehran, Iran, saeediz240@yahoo.com
[2] Islamic Azad University, Fereshtegaan International Branch, Iran, hessam.ghanbar@gmail.com
[3] Department of English Language and Literature, Allameh Tabataba'i University, Tehran, Iran, mahdirezaei093940@gmail.com

connected to item difficulty (Ehrich et al., 2021). CLT is a new line of study that attempts to explain why some tasks are more challenging than others (Martin, 2014). As a result, cognitive load studies emphasize the importance of various measures, including subjective (perceived difficulty), objective behavioral (response and reaction times), and physiological (neuroimaging, eye movements, and heart rate). Cognitive abilities are linked to mental effort, while perceived difficulty is correlated to task complexity (van Gog & Paas, 2008).

English is used in various activities outside the classroom, with vocabulary size crucial for determining students' language development and placement in appropriate stages (Masrai, 2022). According to Miralpeix and Munoz (2018), a person's overall language proficiency can be predicted to some extent by the amount of vocabulary they know; this also affects how well they understand what they read (Biemiller, 2005), listening skills (Mathews, 2018), speaking abilities (Miralpeix & Munoz, 2018), grammar, and writing (Alderson, 2005; Miralpeix & Munoz, 2018). Vocabulary ability and vocabulary size have been studied from various perspectives, but cognitive load and psychometric qualities remain unexplored. Cognitive load is a psychological concept that refers to the cognitive resources that an individual uses to learn or perform a task (Minkley et al., 2021). Measures of cognitive load are believed to indicate the working memory resources used or needed during task performance under varied experimental situations. Cognitive load has typically been defined in terms of the perceived difficulty of tasks in assessment contexts, as well as in instructional design contexts (Choi et al., 2014; Krell, 2017; Paas & van Merriënboer, 1994; Skuballa et al., 2019).

In measuring the cognitive load, the first problem of the conflicting pattern in the relationships between behavioral and subjective measures was discovered in some earlier studies (e.g., DeLeeuw & Mayer, 2008). Additionally, there are discrepancies in the findings of the correlation between task complexity and measures of cognitive load (Lee, 2014). An indication of high cognitive load, for instance, can be a quick response time. Response time has been recommended as a way to acquire reliable evidence of an individual's performance that is unaffected by the load of another activity (Kalyuga et al., 2001; Leppink et al., 2013; Paas, 1992). Additionally, objective and subjective cognitive load tests can be used to demonstrate the difference between actual and perceived cognitive load. Furthermore, it appears that cognitive psychologists relied on subjective assessments, such as perceived difficulty (e.g., Prisacari & Danielson, 2017), objective behavioral indicators, such as reaction time (e.g., Antonenko & Niederhauser, 2010; Dindar et al., 2014; Pouw et al., 2016), as well as physiological evaluations like brain imaging (e.g., Dindar et al., 2014; Pouw et al., 2016). To provide more information on the cognitive load of tasks, researchers combined multiple measures (Leppink, 2017). The second problem is that the absence of studies into item functioning and the adoption of cognitive load measurement in the domain of language testing may cause several issues. For instance, disregarding the cognitive load imposed by language test items on the test taker's mind can lead to overload and, hence, poor performance. This can consequently influence their cognitive processes and performance due to their working memory capacity reaching its limit (Goldhammer et al., 2014; Ponce et al., 2020; Sweller et al., 1998). To put it another way, the number of items imposed on the examinees' minds greatly influences whether they pass or fail the examination. The working memory of test-takers may become overloaded if test items do not provide workloads that are compatible with their intended functions and degree of difficulty. This can prevent examinees from answering a question appropriately.

Contrary to the substantial corpus of studies on how important cognitive load is for learning and teaching languages, little attention has been paid to research on test item loads (Ponce et al., 2020). Comparing the number of items on various formats (such as computer vs. paper) or kinds of materials has been the extent of the study of item loads for language tests (such as static vs. animated graphics; Dindar et al., 2014; Prisacari & Danielson, 2017). In general, test items have drawn researchers' attention in fields such as chemistry (Prisacari & Danielson, 2017), mathematics (Gvozdenko & Chambers, 2007), and algebra (Sweller et al., 2011). Particularly, the number of multiple-choice language items has not received enough attention (Ponce et al., 2020). Cognitive psychology and psychometrics have combined as a consequence of the evaluation of item functioning using a variety of cognitive measures together with difficulty estimations. The behavioral measure did assist CLT by providing pertinent information about item difficulty, cognitive processes, and item functioning in response to several criticisms directed against the use of subjective measures, such as being sensitive to under- or over-estimation of individuals (De Jong, 2010; Ponce et al., 2020; van der Linden, 2009). On the other hand,

CLT may support the research on item functioning in psychometrics by integrating a number of concurrent subjective, objective, and behavioral measurements. The strain of the items they develop may then be fully understood by test designers. This knowledge might encourage test creators to proceed cautiously when designing tests, taking into account the potential negative psychological impacts of item malfunction. In other words, they may assess to see if the features observed or perceived by test-takers in terms of difficulty, processing time, and mental effort are reflected in the test items they design. Cognitive load measurement can offer valuable support for this notion. Therefore, a thorough examination of language test items might give more extensive information about the processes behind responding to each item, which can enhance test item analysis and design compared to limiting the performance of test takers to response time or response accuracy. The inconsistency between test takers' perceived difficulty of a test item and their actual performance was explored while taking into consideration the attitudes of the high school students in the study process. The attitudes of test takers got little consideration in language test item research, despite being extremely essential (Prisacari & Danielson, 2017).

Psychometricians have long been interested in the examination of the international vocabulary size test (VST) item functioning, although many of these vocabulary items have not been examined internationally. The present study's purpose was to determine a clearer picture of the load patterns of vocabulary test questions used in the vocabulary size test (VST). The patterns of item loads were illustrated through the application of subjective (i.e., the perceived difficulty self-reported), objective behavioral (i.e., reaction time), the correlation between perceived item difficulties, estimated item difficulties, and the reaction time needed for test items.

The purpose of the current study (as part of a broader research on tests' cognitive-load via carrying out brain-scanning/EEG). The study was carried out in biomedical research at ATU with the ethics committee's consent (IR.ATU.REC.1401.033). The following questions were the focus of the present study:

1. How can behavioral assessment of the cognitive load imposed on EFL learners estimate/predict the difficulty level of the vocabulary size test in different language proficiency levels?
2. How can EFL students' perceived item difficulty in the vocabulary size test predict their cognitive load?

2. Review of Literature

3.2.1. Cognitive Load Theory (CLT)

Numerous different titles for cognitive load include mental workload, mental load, and mental effort. It is a subject of study in which researchers from a diverse range of fields are interested (Ayres et al., 2021; Sweller et al., 2011). The mental strain imposed on a performer while carrying out a task is known as cognitive load. (Sweller et al., 2011; Sweller, 2019; Yin et al., 2008). Subjective measures are common techniques for evaluating cognitive burden (Paas et al., 2003). CLT is a psychological theory that explores the impact of instruction on psychological and behavioral phenomena. It focuses on the "unobservable" behaviors, or cognitive load, experienced by individuals during various tasks, providing a useful explanation for these phenomena (Noroozi & Karami, 2022). The foundation of CLT consists of cognitive load and learning components. Many scholars have investigated cognitive load in a variety of fields, including instructional design and cognitive psychology (Sweller, 2010). The relationships between learning, teaching, and human cognitive architecture could be investigated via the CLT framework (Sweller et al., 1998). According to research, the cognitive load theory explains why some tasks are more difficult than others. It is based on the notion that the cognitive structure is made up of two different sets of memory stores: restricted short-term (working) and limitless long-term (storage) (Martin, 2014). Working memory, a crucial component of cognitive architecture, is responsible for conscious cognitive processing and information processing, making it related to consciousness in humans, as suggested by Paas (1992) and Sweller et al. (1998). However, the capacity of this sort of memory to store and process information is limited (Miller, 1956; Peterson & Peterson, 1959). In light of this limitation, the capacity of the working memory may be greatly exceeded when numerous components of a task are processed simultaneously and connected, which may impede learning (Chandler & Sweller, 1991; Paas, 1992). The demands of the task may directly affect the load imposed on working memory (Sweller et al., 1998).

Limitations on working memory should be considered as a key element in instructional designs so as to prevent excessive cognitive load or overload imposed by a task (Sweller et al., 2011). Cognitive overload or underload can result in negative effects, as Johannsen (1979) suggests, as excessive cognitive load can negatively affect working memory function. Conversely, cognitive underload, for example, from a lack of motivation, might affect how well a task is performed (Young et al., 2014). Regarding the critical role that cognitive load plays in learning, research on conventional instructions showed that the designs needed to be reexamined and changed in order to increase learning by lowering the load on working memory (Schnotz & Kürschner, 2007). *Performance* is also impacted by overload or underload, in addition to learning. Performance is a component of CLT that considers replies given correctly, errors made, and response times (Paas et al., 2003). When cognitive capacity is exceeded by task demands, poor performance results. More precisely, the demands of a task can affect how much strain is placed on test-takers' thoughts and how well they do (Dindar et al., 2014; Gvozdenko & Chambers, 2007). As a result, the test-takers' performances could indicate how much pressure the test items have imposed on them. On the other hand, a multidimensional theory highlights the CLT, which depicts the strain that a task causes on the cognitive structure (Paas & van Merrinboer, 1994).

2.2. Cognitive Studies and Learning

Numerous studies, including ones on multimedia learning, have used cognitive load measurement (DeLeeuw & Mayer, 2008; Dindar et al., 2014), task-based language instruction (Lee, 2018; Révész et al., 2015; Sasayama, 2016), and evaluation (Pouw et al., 2016; Prisacari & Danielson, 2017). DeLeeuw and Mayer (2008) conducted two studies on cognitive strain in multimedia learning. They used subjective and objective assessments, different language complexity, problem-solving settings, and a redundancy program to evaluate the measurement sensitivity. The study found a strong positive link between mental effort and language complexity, with the more mental effort required, the more complicated the sentence. Reaction time and sentence complexity were also positively correlated. However, the second experiment revealed that complex sentences required more mental effort. The study did not prove that reaction time is a reliable indicator of difficulty posed by interacting factors, and there was no clear connection between mental effort and response speed compared to the previous analysis.

Dindar et al. (2014) compared the cognitive burdens of static and graphic accomplishment exams using reaction time, accuracy rate, and mental strain assessment. Results indicated that length of time was a reliable predictor of cognitive burden, with longer response times indicating harder tasks and larger loads. Statistically, no association was found between response time and mental effort, despite previous studies finding a slender relationship between the two variables. The dependability of task complexity with respect to the subjective and objective measures frequently used in different studies in the area of task-based language teaching (TBLT) has focused on cognitive load research (Lee, 2018; Révész et al., 2014; Révész et al., 2015; Sasayama, 2016).

2.3. Test Reaction Time and Task Complexity

Reaction time and task complexity were found to be correlated in reading comprehension and problem-solving tasks by Goldhammer et al. (2014). They found that response time and task difficulty were correlated with higher-order cognitive skills needed for problem-solving tasks. However, task complexity did not correlate with response time in reading tests, suggesting that task difficulty is a complex factor determining reaction time.

In a research study to ascertain the effect of a novel interface on banked cloze tests, Ponce et al. (2020) used eye tracking to quantify response time and accuracy. They found that response time is crucial in determining cognitive load, with longer reaction times causing a higher cognitive load. Gvozdnko and Chambers (2007) assessed examinees' response time on arithmetic test questions, revealing various cognitive processes in terms of reaction time. Aryadoust et al. (2022) analyzed brain activity, eye movements, and listening performance to examine how test methods impacted subjects' cognitive workloads and listening abilities. They discovered that increased brain activity was connected to slower eye movements, which were linked to lower cognitive burden. Pouw et al. (2016) investigated the effects of meaningful vs. non-meaningful activity on competence domains but found no correlation between response speed and perceived cognitive load. In recent research by Noroozi et al. (2023), who

studied the link between perceived item difficulty and reaction time, they discovered a substantial connection between the variables of both grammatical and vocabulary items.

Response time was discovered not to be a reliable predictor of cognitive load. However, other studies (Lee, 2014) asserted that students' refusal to push themselves to complete the tasks as they became difficult indicated a high degree of cognitive load, which they claimed was the cause of their quick reaction times. The relationship between response time and right answer probability is complex, as faster response times increase the probability of providing the right answer. However, the percentage of right responses increases with response time, suggesting that higher item difficulty correlates with longer response times (van der Linden, 2009).

The reason why the cognitive load is important in test development is that recent studies (Aryadoust et al., 2022; Brüggemann et al., 2023; Burton, 2023; Van de Weijer-Bergsma & Van der Ven, 2021; Xiangming et al., 2023) showed that cognitive load has an important role in such language skills as reading comprehension, speaking, and listening. And as a result, this vital parameter should be taken into consideration in developing different tests in language skills.

In another research, for instance, Burton, 2023, examined eye movements during online L2 speaking assessments. The eye movements of the participants were observed and examined between the end of the test question and the start of their response. Additionally, the participants self-reported information on how challenging they perceived the exam questions to be. According to the findings, participants were more inclined to look away from the interlocutor as test questions got harder. However, they did not blink more frequently as the difficulty increased.

Brüggemann et al., 2023 investigated the effects of mode and medium on cognitive burden during reading comprehension tests. This study investigated the cognitive load that fourth-grade children experienced throughout a reading comprehension test in three distinct test formats. This study showed no changes in the cognitive load experienced when reading comprehension exams were administered on paper, on the computer, or in a format that required computer adaptation. At the end of each exam, students had a higher cognitive burden than in the middle part of the exam.

In both individual and group settings, Xiangming et al. (2023) studied long-term reading outcomes and cognitive load. The statistical findings demonstrated that reading on a mobile device produced the highest degree of intrinsic, extraneous, and germane load. Following reading in print, reading on social media provided the second-highest degree of extraneous and germane load. There was no discernible difference between reading in print and on social media. Additionally, the study discovered that reading on mobile devices resulted in a greater degree of additive cognitive load but worse performance on the reading test.

In a different study, Aryadoust et al. (2022) looked at how test administration affected the cognitive load and performance of listeners. During the while-listening performance (WLP) assessments, the test-takers' gaze patterns indicated that they had adopted keyword matching and shallow listening. The neuroimaging and gaze behavioral data showed that the WLP tests put less strain on test-takers' cognitive abilities than the post-listening performance (PLP) tests did. However, compared to the PLP tests, the test-takers scored better on one of two WLP tests, receiving higher test scores.

2.4. Categories of Cognitive Load

Working memory in humans has been shown to be finite, and decades of research have helped to better comprehend it. According to Miller (1956), humans are only capable of holding seven and a half items in short-term memory. Within cognitive load theory, research has mostly concentrated on the issue of limited working memory (Baddeley, 1992; Crapo et al., 2000; Green et al., 2009; Miller & Kintsch, 1994; Van Merrienboer & Sweller, 2005). Prior research on cognitive load theory has identified three types of load that interfere with working memory and reduce cognitive function when learning and solving problems (Anderson et al., 2011). There are three different sorts of cognitive load: germane, extraneous, and intrinsic (Carlson et al., 2003; Paas et al., 2003; Sweller & Chandler, 1994; Sweller et al., 1998, 2019; Van Merrienboer & Sweller, 2005; Young et al., 2016). Task difficulty and intrinsic load are terms that refer to the inherent structure and important components of test items (Paas et al., 2003). The test items' inherent features are connected to the intrinsic load (De Jong, 2010). The task's inherent strain might be referred to as its "intrinsic load" (Sweller et al., 1998). Extraneous load

is the amount of effort that is imposed on working memory as a result of how information is presented (Sweller et al., 1998; Sweller et al., 2011). On the other side, extraneous cognitive load consists of elements that are not essential for the educational purpose or task, although they may have been introduced by the instructional strategy. Nevertheless, they could prevent or inhibit learning (Young et al., 2016). The "Germane load" refers to the effort required to recall information that has previously been learned and stored in long-term memory (Young et al., 2016). Learning was impaired by extraneous cognitive load, but it was facilitated by germane cognitive burden (Paas et al., 2003).

2.5. Measuring Cognitive Load

In the past, the only way to assess cognitive load was to look at the mistake rate. More direct measurements of cognitive load grew in popularity as the idea progressed (Sweller et al., 2011). The lack of a single standardized method necessitates the use of a variety of measures to produce a more accurate picture of cognitive load (Brünken et al., 2010; Leppink, 2017; Skulmowski & Rey, 2017).

Cognitive burden is often assessed using both subjective and objective methods, with behavioral and neurological assessments being two types of objective measurements (Brünken et al., 2010; Sweller et al., 2019). When it comes to quantifying cognitive load in learning, behavioral assessment is quite important (Lee, 2014). The brain mapping process, which rejects behaviorism and promotes experimental psychology, aims to verify ideas or hypotheses in experimental settings. Brain mapping methods use devices to reveal brain operations and anatomy. Functional imaging is used for investigating cognitive processes, while structural imaging is used for anatomical investigations (Daimiwal et al., 2013). Perception, attention, memory, language, decision-making ability, executive functioning, visual and spatial processing, and cognitive ability are among the cognitive processes that have been extensively studied. Some studies use objective, physiological measures as indicators of cognitive load (e.g., various heart rate or pupillometric measures; Solhjoo et al., 2019; Zheng & Cook, 2011. FMRI (functional magnetic resonance 1imaging), MRI (magnetic resonance imaging), and EEG (electroencephalography) are some of the most commonly used non-invasive equipment for educational and linguistic applications. Electroencephalography (EEG), a method of recording brain waves, could be done to directly evaluate cognitive burden (Lee, 2014).

Cognitive burden is now measured by error rate, while the average reaction time is used to assess complexity (Pelánek et al., 2021). The theory's evolution has fostered a wide range of subjective, objective, and physiological measurements of cognitive load. Subjective judgments of both the degree of complexity and the perceived mental effort (Paas, 1992), estimated duration (Baralt, 2013), the Leppink Cognitive Load Scale, and some others may all be used to quantify cognitive load (Andersen & Makransky, 2020).

Time estimation, self-rating, and the secondary task approach were the three techniques applied by Sasayama (2016) to evaluate the cognitive challenges associated with narrating four visual sequences. The tasks varied in difficulty and complexity, with the hardest tasks demanding the longest reaction times. The most difficult task was also considered the most complicated and required the most mental effort. Proficiency had different effects on learner performance and cognitive load measurement. The use of self-reports led to an overestimation of the impact on the complexity of the cognitive task, so caution should be exercised when interpreting subjective measurement.

Lee (2018) used self-reported perceptions of mental strain, anxiety, complexity, and time estimation to investigate whether variations in task complexity may affect cognitive load. The findings showed that the most difficult tasks required the most mental effort and were correlated with response time and task complexity. Accuracy rates are crucial when dealing with challenging tasks (Révész et al., 2015; Révész et al., 2014), but Lee (2018) found that accuracy had less influence than task time. Sasayama (2016) reported the same lack of impact on accuracy, indicating that complicated tasks required longer reaction times. The tactile detection response task and rhythmic tapping approach were studied by Greenberg and Zheng (2022) to understand how secondary tasks affect and obstruct cognitive performance. They highlighted the disturbance generated by these methods and demonstrated how it differed depending on the modality.

The study aimed to explore the correlation between test item difficulty and cognitive strain using a quantitative correlational approach. Data were collected from 60 MA masters and participants. The Rasch model was used to estimate difficulty levels in the Iranian University Entrance Examination

(IUEE) for the vocabulary and grammar sections. Results showed significant connections between reaction times and assessments of difficulty for vocabulary questions, while no significant correlations were found for grammar questions (Noroozi & Karami, 2022). In 2015, Révész et al. examined task difficulty using self-ratings, expert judgment, and response time. They found that task complexity increased mental effort, and harder tasks were considered more complicated. However, response time did not correlate with task difficulty, as reaction times' levels had no statistically significant impact on one another.

Lee (2014) examined cognitive load evaluations using brain imaging, self-ratings, and learning outcomes. Participants assessed the difficulty of a seven-minute documentary and discovered a substantial negative correlation between learning results and difficulty ratings. As the task became more sophisticated and burdened, participants' performance on the comprehension exam became unsatisfactory as they stopped expending mental effort on the difficult task. This interrupted understanding as intrinsic load increased, indicating the need for more effective cognitive load evaluations.

Online objective measurements appear to be more helpful than subjective ones in assessing fluctuations in cognitive load while doing a task (Paas & van Merrinboer, 1993). There is a connection between learners' behavior and learning processes; as a result, several behavioral measures of secondary task technique, length of time on task, and task complexity are all considered to be indicators of cognitive load (Brünken et al., 2010; Sweller et al., 2011; Sweller et al., 2019). Several studies evaluated the difficulty and load of various tasks using the response time of the secondary task technique (Lee, 2018; Sasayama, 2016).

2.6. Vocabulary Size Test

To assess several areas of vocabulary proficiency (Henriksen, 1999; Nation, 2001, 2019; & Read, 2000), so far, a variety of vocabulary tests have been created (e.g., Leech, 1991; Meara & Jones, 1990; Nation & Beglar, 2007; Read, 2000; Schmitt, 2000; Wesche & Paribakht, 1996) since learning vocabulary is a vital element in language learning (Aryadoust, 2012; Effatpanah, 2019; Yu, 2021). Two well-known descriptors of vocabulary knowledge are size (breadth) and depth (depth); therefore, size and depth. Size refers to the quantity of words known, whereas depth refers to the quality of those words (Schmitt, 2000). As a result, one's vocabulary may be assessed for both its breadth and depth. It could be difficult and impractical to assess both types of vocabulary knowledge in classroom settings with time constraints. Moreover, through exposure to language outside of the classroom, students' vocabulary knowledge tends to grow (Sundqvist & Wikström, 2015), both through unintentional and intentional vocabulary learning (Hulstijn, 2012; Laufer, 2017).

Despite the size-depth distinction's widespread adoption (e.g., Read, 2004), size has been considered to be a more accurate measure of L2 vocabulary knowledge, in part because of its simple polling method that assesses several target words at once. As a result of its importance in the form-meaning relationship for vocabulary usage (e.g., Laufer et al. 2004; Meara, 2002; Schmitt, 2010), it has also been thought to be the primary component of vocabulary knowledge. The importance of analyzing learners' vocabulary sizes has generally received more attention from researchers.

There has been a ton of research on assessing second language (L2) vocabulary knowledge over the last three decades; as a result, it is currently one of the greatest fields in applied linguistics (e.g., Carcamo, 2022; Chang, 2020; Jeon, 2021; Lai et al., 2023; Milton, 2009; Nation & Beglar, 2007; Read, 2000; Schmitt et al. 2001; Wesche & Paribakht, 1996).

The term "vocabulary size" is used to describe how much vocabulary a person knows. A native-speaking five-year-old who is about to start school knows about 3000-word families (Biemiller & Slonim, 2001). After a brief period of schooling, this child's receptive vocabulary grows to roughly 5,000 word families by the age of eight (Biemiller, 2005). By the end of high school, a native English speaker aged 17 knows around 13000–14000 word families (Coxhead et al., 2015). Except for experts like surgeons and botanists, native speakers' receptive vocabulary seldom exceeds 20000-word families, although it appears to increase by roughly 1000-word families per year (Biemiller, 2005; Coxhead et al., 2015).

The quantity of words non-native English speakers (NNS) know in a foreign language is significant since it is closely related to what they are able to achieve in that language. An NNS of

English must know at least the most frequent 3000, 5000, and 9000-word families to be able to have a basic conversation (Schmitt & Schmitt, 2012), as well as the most frequent 5000 and 9000-word families to read books and newspapers (Nation, 2006). It has been found that vocabulary size directly affects reading comprehension, significantly affects writing and grammar, and improves listening skills (Biemiller, 2005; Mathews, 2018).

The frequency of the words was used to make pedagogical judgments, such as which words to teach in the school in a clear and concise manner. Nation (2001; 2011) classified word families into four categories: academic, technical, and low-frequency words. They suggested that high-frequency words be explicitly taught in the classroom rather than low-frequency words, which are uncommon, and academic and technical words, which are only required when students want to study in English. In a more recent study, Schmitt (2014) reorganized word families into three categories: high-frequency (most frequent 3000), mid-frequency (between 3000 and 9000), and low-frequency (9000+). They contend that teaching only high-frequency words is insufficient given the requirement for vocabulary knowledge of 8000–9000 word families to read authentic English texts and that we must find ways to address mid-frequency vocabulary in the classroom.

When comparing vocabulary sizes, there is a difference between productive and receptive vocabulary sizes. Various receptive vocabulary size tests exist, including the Eurocenters Vocabulary Size Test (Meara & Jones, 1988), the Vocabulary Size Test (Nation & Beglar, 2007), which includes 14000 and 20000-word family variants and is based on the British National Corpus and the Corpus of Contemporary American English, and The Picture Vocabulary Size Test (Nation & Beglar, 2007). However, there is no agreement on how to measure productive vocabulary size (Nation & Anthony, 2016). The target group for this program is young, preliterate language learners.

3. Methodology

3.1. Participants

The participants in this study were selected based on convenience and cluster sampling. There were 150 high school male students. They took both the vocabulary size test and the Cambridge placement test face-to-face, not virtually, to ensure that they were at the right level. These participants were selected based on their Cambridge Placement Test (*General English*, n.d.) and vocabulary size test scores. They ranged in proficiency in the English language from beginner to advanced. Finally, for the main study, 60 male students were selected based on purposive sampling for the behavioral phase. They were aged between 15 and 18 (M=16.83, SD= 0.50) studying at a high school in Tehran. The reason for selecting just male students was that the researcher was a teacher in high school and he did not have access to female students.

3.2. Instrumentation

3.2.1. Vocabulary Size Test

To evaluate students' knowledge of vocabulary, a vocabulary size test was used (Nation & Beglar, 2007) and the reliability with different sets of items were 0.91 and 0.96. The vocabulary size test gauges a student's comprehension of receptive written words. The exam assesses knowledge of the written word's form, the relationship between form and meaning, and, to a lesser extent, concept knowledge. It is based on estimates of word family frequency taken from the British National Corpus (BNC; Nation, 2006). The exam primarily evaluates vocabulary understanding in isolation despite the fact that the tested term only appears in one non-defining context throughout the whole test. The vocabulary size test is a multiple-choice vocabulary test that employs Read and Chapelle's (2001) technique. It is discrete, selective, and mostly context-independent. Up to the 20th 1000-word level, the test is offered in both monolingual and bilingual formats. Participants in the test must choose one of four possible definitions or translations for each word. Both paper and electronic editions of the test are available.

The vocabulary size test (Nation & Beglar, 2007) was created as a competence test for those learning English as a second language or as a foreign language to measure their overall vocabulary size. Based on a frequency count of word families in the British National Corpus, this exam consists of 140 items, 10 from each of the fourteen 1,000-word levels. There are four items in the multiple-choice test.

Multiple-choice items are highly reliable, administered easily, and scored objectively (Bakytbekovich et al., 2023). An example from the third 1,000-level test question is provided below.

3. jug: He was holding a **jug.**

a. a container for pouring liquids

b. an informal discussion

c. a soft cap

d. a weapon that explodes

The language used in the four alternatives was chosen to be more frequent than the word being assessed in the item's writing. Every test term is given a straightforward, non-defining context. The learner's test score is multiplied by 100 to establish their total vocabulary size since each question on the test represents a family of 100 words (10 items from each 1,000-word frequency level). Therefore, if a student receives a test score of 68 out of 140, their entire vocabulary size is 6,800 words.

Beglar (2010) examined the Rasch-based validation of the vocabulary size test the results showed that both the items and the examinees typically performed as expected by a priori assumptions, that the Rasch model suited the vast majority of the items very well, that the items were quite unidimensional, and that the Rasch model accounted for 85.6% of the variance. Rasch reliability indices >0.96 indicated that different combinations of items provided an accurate measurement for this sample of examinees. The items demonstrated a strong degree of measurement invariance, with disattenuated Pearson correlations for person measures estimated with different sets of items of 0.91 and 0.96. The vocabulary size test offers teachers and researchers a brand-new tool that significantly extends the range of assessment provided by existing tests of written receptive vocabulary size.

3.2.2. Cambridge Placement Test

A quick placement test for English language learners is the online Cambridge Placement Test (*General English*, n.d.), which consists of 25 multiple-choice questions. The test takes 10 to 15 minutes to complete. In this study, the Cambridge placement test was used to determine the individuals' proficiency level for the sake of homogeneity. Participants' competence levels ranged from A1 to C1 based on the standard European framework of reference (CEFR). This online placement exam for grammar and vocabulary was designed to assess candidates' proficiency with English grammar, vocabulary, and phrasing. There were three options for each question, and test-takers could select one of them while providing an answer.

3.2.3. Self-Report Interview

The present study included an interview of self-reports of perceived difficulty, which were measured after the vocabulary test on PsychoPy software. It has also been demonstrated that this self-report is a reliable predictor of cognitive stress (Prisacari & Danielson, 2017). Although task difficulty and mental effort might have a relationship with each other, they measure different constructs: task difficulty corresponds to the task itself, and mental effort relates to a process involving more aspects than being limited to the task itself (van Gog & Paas, 2008). The data were gathered orally.

3.2.4. PsychoPy Software

The vocabulary items and their correct responses, keys, and item numbers were coded and entered into Excel by using PsychoPy Software (PsychoPy-2022.2.2; Peirce et al., 2019). A short training run was conducted before the major test to make sure that the individuals were prepared. The students were told to answer questions as precisely and rapidly as they could by pressing the pre-selected keyboard keys. The students answered the vocabulary-size test items after a brief training session to make sure they were ready for the main exam.

3.3. Data Collection Procedure

In this study, a mixed method was employed by the researchers, and the students were selected through cluster and convenience sampling to select 150 male high school students. The researchers asked students to take the Cambridge Placement Test (CPT) and the Vocabulary Size Test (VST) to

ensure their exact level. To prevent cheating, the researchers administered the test in person. The researchers jotted down their responses on a separate piece of paper so they could not cheat by sharing their answer sheets with their friends. Some of them were excluded from taking the test because they were false beginners. Finally, students were divided into three levels: elementary, intermediate, and advanced.

In the next stage of this study, the students were selected through purposive sampling (N = 60). To establish the fixed time of the multiple-choice question's appearance on the screen, the initial group of participants (N = 15) participated in the pilot study (Elementary = 5, Intermediate = 5, Advanced = 5). The display time of items was proposed to be fixed, which was seen as vital for recording mental processing demands. To record the initial fixation on the stimuli, a fixed presentation time was used (Scharinger et al., 2020). Thus, the participants' further reading or rereading could reveal further cognitive processes. In light of the pilot phase's findings, the current study's fixed timing for the display of language items was established. Fifteen students, whose proficiency levels matched those of the participants in the study's main phase (Elementary = 5, Intermediate = 5, Advanced = 5), were administered the vocabulary tests with the most characters. They were to read the stems and alternatives and answer the questions as quickly as they could. A fixed time was established to ensure reliable load capture and prevent test-takers from guessing, using compensating techniques, or moving to a lower level of attention (Goldhammer et al., 2014; Scharinger et al., 2015). That is, the maximum response time (50 minutes) was then established by the researchers using the average of each group. Further, Jung (2018) argued that time limitations could increase the probability of detecting task complexity effects, despite the side effect of imposing an additional cognitive load on individuals' minds.

The researchers then prepared the data and entered it into the PsychoPy software (Peirce et al., 2019). The vocabulary items were coded and put into Excel along with their proper responses and item numbers. After running the experiment, Excel exports the data to PsychoPy (such as the lexical items used for the stimuli) and imports the millisecond-level reaction times as well as the multiple-choice question answers from PsychoPy. The test items in the vocabulary block were not randomly arranged. A one-hour test session was held for the participants. The process for responding to multiple-choice questions was given to the participants by the researchers. They were told to hit the pre-programmed keys on the keyboard to answer questions as accurately as possible. The vocabulary size test was administered face-to-face to sixty students at school during the summer. They responded to the questions, and their answer files were eventually corrected and entered into an Excel spreadsheet. The researchers calculated the difficulty of 140 vocabulary items for each student and examined whether or not their reaction time could predict the cognitive load.

After taking the VST on PsychoPy software, the researchers asked students to specify their perceived difficulty (Lou & Noels, 2016; Nakamura et al., 2022) for each question orally (N = 60). This session took 20 minutes after the exam, and the students were supposed to state whether the test as a whole was very easy, easy, hard, or very hard. In the next step, simple linear regression was run to see whether or not there was a correlation between response time, perceived difficulty, and cognitive load.

3.4. Data Analysis Procedure

To answer the first research question, the vocabulary size test (VST) by using the PsychoPy software exported every participant's reaction time as an Excel file. Afterward, three simple linear regressions (using the students' mean score) were also conducted to investigate the connection between test difficulty and length of time.

The percentage of test takers who correctly responded to an item was regarded as an indicator of that item's difficulty (Wood, 1960). The difficulty decreased as this percentage increased. This implies that there is an inverse relationship involved: the harder an item is to complete, the lower its index (Wood, 1960). The number of people who properly answered a question was divided by the total number of people who answered it in order to determine how tough the item was. Usually, the letter *p*, which stands for the item's difficulty, shows this proportion (Crocker & Algina, 1986). The formula used to calculate it is as follows:

$$p_i = \frac{A_i}{N_i}$$

where:
pi = the item's difficulty index *i*
Ai = the number of accurate responses to item *i*
Ni = number of correct answers plus number of incorrect answers to item *i*

To answer the second question, that is, examining the prediction of cognitive load from EFL students' perceived item difficulty, another regression analysis was utilized. In this regression analysis, the cognitive load was measured as the reaction time (using the students' mean score), which was considered the independent variable in the regression analysis. Both at the elementary and advanced levels, there was no discernible relationship between perceived difficulty and time. However, at the intermediate level, there was a substantial correlation between time and difficulty.

The researchers decided not to use a 5-point rating scale because of respondents' conservative tendencies, which may result in lower dependability levels (Krosnick et al., 2002). Instead, they used a 4-point rating scale. Answering Likert scale-type questions is culture-bound, meaning that in different cultures, students might answer Likert scale questions differently (Lee et al., 2002; Marefat & Pakzadian, 2017). The correlation between perceived difficulty and reaction time was investigated in such a way that after each student completed the vocabulary size test, the perceived difficulty was evaluated to determine their opinions of the questions and whether they thought the test was very easy, easy, hard, or very hard. If a student stated that most of the questions were very hard, they also thought the test as a whole was very hard (Lou & Noels, 2016; Nakamura et al., 2022). Likewise, if a student stated that most of the questions were very easy, they assumed that they thought the test as a whole was very easy.

4. Results

4.1. Research Question One on Behavioral Evaluation of Cognitive Load

Regarding the first research question, three simple linear regressions were conducted, which investigated to what extent the cognitive load of the vocabulary size test (measured as the time spent on the test, considered the independent variable in the regression analysis) can predict the difficulty level of the vocabulary size test (considered the criterion variable in the regression model) across different levels. It should be stated that the data met the assumption of normality as all skewness values of variables were between -2 and +2 standard errors of their measures (see Ghanbar & Rezvani, 2023; Plonsky & Ghanbar, 2018). Also, all *p* values less than 0.05 are considered statistically significant (see Table 1 for the descriptive statistics of variables in the regression analysis).

Firstly, the cognitive load was not a significant predictor of the difficulty level of the vocabulary test at the elementary level. Additionally, this regression model was not significantly different from zero, with F (1, 18) = 0.58, p =.45, and the adjusted R^2 *at.03*, demonstrating the non-significance of this regression model. This showed that cognitive load was not a significant predictor of the difficulty level of the test (see Table 2 for regression coefficients), and it predicted 3% of the variance of the difficulty level of the vocabulary size test (see Plonsky & Ghanbar, 2018 for more information on R^2 values).

Nonetheless, at the intermediate level, the cognitive load was a significant predictor of the difficulty level of the vocabulary test. Moreover, this regression model was significantly different from zero, with F (1, 18) = 0.42, p =.00, and the adjusted R^2 *at.68*, representing the significance of this regression model. This exhibited that cognitive load was a significant predictor of the difficulty level of the test (see Table 2 for regression coefficients), and it could predict 68% of the variance of the difficulty level of the vocabulary size test at the intermediate level.

Ultimately, it was revealed that cognitive load was not a significant predictor of the difficulty level of the vocabulary test at the advanced level. As well, his regression model was not significantly different from zero, with F (1, 18) = 0.98, p =.*34*, and the adjusted R^2 at.05, demonstrating the non-significance of this regression model. This finding showed that cognitive load was not a significant predictor of the difficulty level of the test at the advanced level (see Table 2 for regression coefficients), and it merely predicted 5% of the variance of the difficulty level of the vocabulary size test.

Table 1

Descriptive Statistics of Independent and Dependent Variables in the Regression Analysis across Proficiency Levels for the Difficultly Level

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	
Elementary							
CL	20	12.41	36.36	23.76	7.71	.19	.51
D	20	.25	.45	0.36	0.05	-.42	.51
Intermediate							
CL	20	15.28	45.22	28.16	8.05	.41	.51
D	20	.42	.67	0.53	0.08	.19	.51
Advanced							
CL	20	9.42	35.28	22.59	6.34	-.31	.51
D	20	.58	.76	0.67	0.06	.09	.51

Note: CL = Cognitive Load, D = Difficulty Level of the Vocabulary Size Test

Table 2

Regression Coefficients of Regression Analyses across Different Proficiency Levels

		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
Elementary	ID	-28.51	37.38	-.17	-.76	.45
Intermediate	ID	84.66	12.92	.83	6.55	.00
Advanced	ID	-23.52	23.71	-.22	-.99	.33

Note: ID= dependent variable in the regression model

4.2. Research Question Two on Students' Perceived Item Difficulty in Vocabulary Size Test

To respond to research question two, another simple linear regression was also conducted to examine the perceived difficulty of the items by the students (the independent variable) can predict the cognitive load (the time spent on the total test, considered as the criterion variable) (see Table 3 for the descriptive statistics of variables in this regression model and skewness measures).

Table 3

Descriptive Statistics of Independent and Dependent Variables in the Regression Analysis across Proficiency Levels for the Cognitive Load

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	
Elementary							
PD	20	1	2	1.95	.22	.72	.51
CL	20	12.41	36.36	23.76	7.71	.19	.51
Intermediate							
PD	20	2	3	2.55	.51	.41	.51
CL	20	15.28	45.22	28.16	8.05	.41	.51
Advanced							
PD	20	3	4	3.15	.36	0.22	51
CL	20	9.42	35.28	22.59	6.34	-.31	.51

Note: PD= Perceived Difficulty, CL = Cognitive Load

In this regression equation, students' perceived item difficulty was not a significant predictor of cognitive load at the elementary level. In fact, this regression model was not significantly different from zero, with F (1, 18) = 0.22, p =.64, and the adjusted R^2 at.04, signifying the non-significance of this regression model. This finding brought to the fore that students' perceived item difficulty was not

a significant predictor of cognitive load (see Table 4 for regression coefficients), and hence it predicted merely 4% of the variance of cognitive load.

In contrast, pertaining to the intermediate level, students' perceived item difficulty was a significant predictor of cognitive load. Likewise, this regression model was significantly different from zero, F (1, 18) = 17.84, p =.*00*, with the adjusted R^2 at.47, showing the significance of this regression model. This demonstrated that students' perceived item difficulty was a significant predictor of cognitive load (see Table 4 for regression coefficients), and it could predict 47% of the variance of cognitive load at this proficiency level.

It was also found that students' perceived item difficulty was not an accurate indicator of cognitive strain at the advanced level. In fact, this regression model was not significantly different from zero; F (1, 18) = 0.22, p =.63, with the adjusted R^2 at.04, indicating the non-significance of this regression model. This finding revealed that students' perceived item difficulty was not an accurate indicator of cognitive load at the advanced level (see Table 4 for regression coefficients), and it purely predicted 4% of the variance of this variable.

Table 4
Regression Coefficients of Regression Analyses across Different Proficiency Levels

		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
Elementary	ID	-3.78	8.07	-.11	-.46	.64
Intermediate	ID	11.13	2.63	.70	4.22	.00
Advanced	ID	1.94	4.05	.12	-.48	.63

Note: ID= dependent variable in the regression model

5. Discussion

5.1. Research Question One on Behavioral Evaluation of Cognitive Load

Three simple linear regression analyses were carried out in order to investigate the relationship between test length and difficulty. At either the elementary or advanced levels, the relationship between time and difficulty was not statistically significant. However, there was a significant relationship between time and difficulty at the intermediate level. Therefore, the difficulty could very well predict how much time someone would spend at the intermediate level.

The results appear to be in line with other research studies that found no correlation between reaction time and task complexity at the elementary and advanced levels (DeLeeuw & Mayer, 2008; Noroozi & Karami, 2022; Pouw et al., 2016; Révész et al., 2014; Révész et al., 2015). The findings suggested that response time, which is unresponsive to task difficulty, could not account for the cognitive load of the vocabulary items at the elementary and advanced levels. The first experiment in DeLeeuw and Mayer's (2008) study examined the correlation between response time and task difficulty. However, these findings are not in line with previous studies that indicated a relationship between task complexity and reaction time (e.g., Aryadoust et al., 2022; DeLeeuw & Mayer, 2008; Dindar et al., 2014; Gvozdenko & Chambers, 2007; Noroozi & Karami, 2022; Ponce et al., 2020; van der Linden, 2009). The correlation between reaction time and difficulty estimations in this study may be due to the task types and character counts, which may differ in different cases. Longer response times may indicate more time spent reading characters rather than reflecting on the answer, indicating cognitive load (Lee, 2018).

Regarding the intermediate level, our findings corroborate earlier studies that discovered a connection between task difficulty and reaction time (e.g., Aryadoust et al., 2022; Dindar et al., 2014; DeLeeuw & Mayer, 2008; Gvozdenko & Chambers, 2007; Noroozi & Karami, 2022; Ponce et al., 2020; van der Linden, 2009). The fact that more time is required to finish a task when it becomes more demanding or difficult was highlighted. Therefore, it appears that reaction time and item complexity both accurately measure cognitive strain. This result is also partially consistent with those of the research conducted by Goldhammer et al. (2014), which emphasized the significance of a question's level of difficulty in providing an accurate response: the tougher the question, the longer the response

time. The fact that Goldhammer et al. (2014) used distinct task types (problem-solving versus reading and literacy) and did not find any consistent patterns of connections may account for the partial agreement. Consequently, it appears that the complexity of the task is influenced by its kind. The results, however, contradicted those of other research (DeLeeuw & Mayer, 2008; Noroozi & Karami, 2022; Pouw et al., 2016; Révész et al., 2014; Révész et al., 2015), in which it was stated that response time was an inappropriate measure of cognitive burden. So, response time is similar to difficulty estimates for vocabulary items, requiring longer responses for higher values. This can reveal evidence of task demands and cognitive processing (Gvozdenko & Chambers, 2007). Examinees' responses and response times indicate difficulty levels for items, allowing for the calculation of difficulty estimates. However, this does not seem to be the case at the elementary and advanced levels.

5.2. Research Question Two on Students' Perceived Item Difficulty in Vocabulary Size Test

To explain the second research question, which focused on the correlation between perceived difficulty and response time, a simple linear regression was also carried out to examine the connection between the length of the test and how tough the students thought the items were. In contrast to the findings of the present research, in a number of other studies (e.g., Pouw et al., 2016; Révész et al. 2016; & Révész et al., 2014), no significant relationship was found between perceived difficulty and time duration at elementary and advanced levels. It might be claimed that the length of the test could not have anticipated the perceived difficulty. While at the intermediate level, there was a significant relationship between perceived difficulty and length of time, which was in line with Lee (2019), Noroozi et al. (2023), and Sasayama (2016). According to them, the easiest and hardest tasks, respectively, needed the least and most time. This means that when a task becomes more difficult and demands greater mental effort from the individual, more reaction time is required to complete the task. The higher the cognitive load, the longer the response time. Response time can indeed provide some evidence as to how deep the processing is or how much cognitive ability is needed to complete the task (Goldhammer et al., 2014). In addition, response time can reveal some information about the level of cognition or the number of cognitive resources needed to complete the task (Goldhammer et al., 2014).

6. Conclusions and Implications

This study's main objective was to provide a thorough explanation of the correlation between the assessed difficulty of vocabulary size test questions, the perceived difficulty, and the length of time. One of the major contributions of this study is to accentuate the significance of cognitive processes and cognitive load perspectives on test development procedures, as Gass et al. (2013) and Ponce et al. (2020) stated.

The researchers used three simple linear regressions in the first research question to investigate the relationship between exam difficulty and length of time with regard to the cognitive load parameter. The correlation between difficulty and time was not significant at the elementary level or at the advanced level. Nevertheless, the correlation between difficulty and length of time was significant at the intermediate level. Consequently, it can be stated that difficulty can predict the length of time at the intermediate level.

The correlation between response time and perceived difficulty was the focus of the second research question. Using simple linear regression, it was also determined whether there was a correlation between the length of the exam and how challenging the students perceived the questions to be. There was no obvious correlation between perceived difficulty and time at the elementary and advanced levels, while at the intermediate level, the correlation between them was significant.

The first research question aimed to examine the relationship between estimated levels of difficulty and reaction times as an indicator of cognitive load for vocabulary items. There was no correlation between estimated difficulty and length of time at elementary and advanced levels. The observed results may be due to the high character counts of vocabulary items. Longer reaction times may indicate more time spent reading characters rather than pondering the answer, reflecting the item's complexity. However, estimates of difficulty are a reliable measure of cognitive load at intermediate levels, and response time is a sound alternative objective index of vocabulary item cognitive load.

The second research question found no relationship between perceived difficulty and length of time as an indicator of cognitive load at elementary and advanced levels. Possible explanations include test takers' potential underestimation of item difficulty or fluctuating judgments of difficulty when tasks or items change. This suggests fundamental differences among vocabulary items (Ary et al., 2019; Lee, 2018). However, at the intermediate level, the correlation between them was significant.

Overall, this study's findings may have significant theoretical ramifications for the CLT. By using item difficulty assessments and cognitive load measures, researchers may better understand item evaluation and the effects of the language items' loads on test-takers' brains. The measures employed in CLT were mainly limited to the widely used subjective and objective measures. In addition, the use of objective measures (response time) and subjective measures (perceived difficulty) has not been prevalent. Hence, the use of item difficulty estimates as a more objective measure, response time, and subjective measures like perceived difficulty can contribute to the outcomes of cognitive load assessments. The discipline of language testing may potentially be affected by the cognitive load and related research findings. By using cognitive load measurements concurrently, cognitive load theory can be an effective way to investigate item functioning in psychometrics. This will make it possible for test designers to properly comprehend the load and purpose of the items they create. This information might help test designers continue cautiously as they plan the exam, and take into account any detrimental psychological effects that an inadequate item might have. In other words, they can determine if the test questions they develop reflect attributes experienced or perceived by test-takers, such as the degree of difficulty and amount of time needed. To this end, cognitive load measures can provide worthwhile evidence. In addition, the danger of over-reliance on the methods used for item design or sole dependence on item difficulty estimates as the measure to determine difficulty can be minimized. Whether the identified language item difficulty is translated into a similar experienced level of difficulty is of concern for item analysis and test development. In this regard, more caution should be taken not to overload the test takers' minds, as this may detrimentally influence their performance.

The loads on the items must be considered while designing a test. Additionally, since tests are developed based on the content, the syllabus and course designers need to give additional consideration to a variety of factors while developing a syllabus. Test designers must also take neurolinguistic issues more into account when developing tests.

Response time data aids CLT by assessing item complexity, cognitive processes, and functioning, addressing concerns about subjective measures and potential under- or overestimation of test items. This approach addresses concerns about subjective measures (De Jong, 2010; Ponce et al., 2020; van der Linden, 2009).

Here, in light of the limitations imposed on the current study, several suggestions for further research are provided. Firstly, the current study was not carried out on a large scale. It would be preferable if more participants were included in possible future studies so that the results could be more generally applicable. The participants in this study were all male, with an age range of 15–18. Another study can be conducted with both genders and with other age groups. The current study has only considered high school students with different levels of proficiency. Note that test takers who are university students can better distinguish the nuances of task difficulty compared with high school students (Ayres, 2006; Sasayama, 2016). Also, the results might vary according to the sampling location and the quality of the schools. For example, government/state schools and private schools might have shown differences in results. Therefore, more exploration can be conducted by future researchers in this field of study.

With the purpose of deepening our understanding of the specific criteria that the test takers used to rate the difficulty of each item, retrospective interviews and think-aloud techniques are strongly recommended. Classification of items into groups of items with the least-to-most complexity can also lead to interesting findings in future investigations. That is, the items can be classified into different groups based on their difficulty. Hence, future studies should address this issue by including vocabulary items with a relatively equal number of characters. In future studies, the researchers can use a large number of participants to get better outcomes.

Declaration of Conflicting Interests

The authors affirm that the current research does not include any conflicts of interest for them.

Funding
For the current study, the authors were not given any financial assistance.

Reference

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. A&C Black.

Andersen, M. S., & Makransky, G. (2020). The validation and further development of a multidimensional cognitive load scale for virtual environments. *Journal of Computer Assisted Learning*. *37*(1), 183–196.

Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., & Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. *Computer Graphics Forum*, *30*(3), 791–800. https://doi.org:10.1111/j.1467-8659.2011.01928.x

Antonenko, P. D., & Niederhauser, D. S. (2010). The influence of leads on cognitive load and learning in a hypertext environment. *Computers in Human Behavior*, *26*(2), 140–150. https://doi.org:10.1016/j.chb.2009.10.014

Aryadoust, V. (2012). How Does " Sentence Structure and Vocabulary " Function as a Scoring Criterion Alongside Other Criteria in Writing. *Language*, *2*(1), 28–58.

Aryadoust, V., Foo, S., & Ng, L. Y. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Language Testing*, *39*(1), 56–89. https://doi.org:10.1177/02655322211026876

Ary, D., Jacobs, L. C., Irvine, S., & Walker, D. (2019). *Introduction to research in education* (10th ed.). Wadsworth Cengage Learning.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*(5), 389–400. https://doi.org:10.1016/j.learninstruc.2006.09.001

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, *12*. https://doi.org:10.3389/fpsyg.2021.702538

Baddeley, A. (1992). Working memory: The interface between memory and cognition. *Journal of Cognitive Neuroscience*, *4*(3), 281–288. https://doi.org:10.1162/jocn.1992.4.3.281

Bakytbekovich, O. N., Mohammed, A., Alghurabi, A. M. K., Alallo, H. M. I., Ali, Y. M., Hassan, A. Y., Demeuova, L., Viktorovna, S. I., Nazym, B., & Afif, A. K. N. S. (2023). Distractor Analysis In Multiple-Choice Items Using the Rasch Model. *International Journal of Language Testing*, *13*, 69–78. https://doi.org/10.22034/IJLT.2023.387942.1236

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition*, *35*(4), 689–725. https://doi.org:10.1017/s0272263113000429

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In *Teaching and learning vocabulary: Bringing research to practice.* (pp. 223–242). Lawrence Erlbaum Associates Publishers.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*(3), 498–520. https://doi.org/10.1037/0022-0663.93.3.498

Brüggemann, T., Ludewig, U., Lorenz, R., & McElvany, N. (2023). Effects of mode and medium in reading comprehension tests on cognitive load. *Computers & Education*, *192*, 104649. https://doi.org/10.1016/j.compedu.2022.104649

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53–61. https://doi.org:10.1207/s15326985ep3801_7

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load. In J. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory (pp. 181-202).* Cambridge University Press. https://doi.org:10.1017/cBO9780511844744.011

Burton, J. D. (2023). Gazing into cognition: Eye behavior in online L2 speaking tests. *Language Assessment Quarterly*, *20*(2), 190-214. https://doi.org/10.1080/15434303.2022.2143680

Carcamo, B. (2022). A Bilingual Version of the Vocabulary Size Test for Spanish Speakers. *International Journal of Language Testing*, *12*(2), 45–58. https://doi.org/10.22034/IJLT.2022.157124

Carlson, R., Chandler, P., & Sweller, J. (2003). Learning and understanding science instructional material. *Journal of Educational Psychology*, *95*(3), 629–640. https://doi.org:10.1037/0022-0663.95.3.629

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293–332. https://doi.org:10.1207/s1532690xci0804_2

Chang, Y. H. (2020). The effect of ambiguity tolerance on learning English with computer-mediated dictionaries. *Computer Assisted Language Learning*, *33*(8), 960–981. https://doi.org/10.1080/09588221.2019.1604550

Choi, H.-H., van Merriënboer, J. J. G., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review*, *26*(2), 225–244. https://doi.org/10.1007/s10648-014-9262-6

Coxhead, A., Nation, P. & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in new zealand secondary schools. *New Zealand Journal of Educational Studies*, *50*(1), 121–135. https://doi.org/10.1007/s40841-015-0002-3

Crapo, A. W., Waisel, L. B., Wallace, W. A., & Willemain, T. R. (2000). Visualization and the process of modeling. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 218–226. https://doi.org:10.1145/347090.347129

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Cengage Learning.

Daimiwal, N., Sundhararajan M., & Shriram, R. (2013). EEG based cognitive workload assessment for maximum efficiency. *IOSR Journal of Electronics and Communication Engineering*, *7*, 34–38

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, *38*(2), 105–134. https://doi.org:10.1007/s11251-009-9110-0

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, *100*(1), 223–234. https://doi.org:10.1037/0022-0663.100.1.223

Dindar, M., Kabakçı Yurdakul, I., & Dönmez, F. I. (2014). Measuring cognitive load in test items: Static graphics versus animated graphics. *Journal of Computer Assisted Learning*, *31*(2), 148–161. https://doi.org:10.1111/jcal.12086

Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, *9*(1), 1–28.

Ehrich, J. F., Howard, S. J., Bokosmaty, S., & Woodcock, S. (2021). An item response modeling approach to cognitive load measurement. *Frontiers in Education*, *6*. https://doi.org:10.3389/feduc.2021.648324

Gass, S. M., Behney, J., & Plonsky, L. (2013). *Second language acquisition: An introductory course* (4th ed.). Routledge.

General English. (n.d.). Www.cambridgeenglish.org. Retrieved September 29, 2023, from https://www.cambridgeenglish.org/test-your-english/general-english

Ghanbar, H., & Rezvani, R. (2023). Structural equation modeling in L2 research: A systematic review. *International Journal of Language Testing*, *13*, 79–108. https://doi.org:10.22034/IJLT.2023.381619.1224

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org:10.1037/a0034716

Greenberg, K., & Zheng, R. (2022). Cognitive load theory and its measurement: A study of secondary tasks in relation to working memory. *Journal of Cognitive Psychology*, 1–19. https://doi.org:10.1080/20445911.2022.2026052

Green, T. M., Ribarsky, W., & Fisher, B. (2009). Building and applying a human cognition model for visual analytics. *Information Visualization*, *8*(1), 1–13. https://doi.org:10.1057/ivs.2008.28

Gvozdenko, E., & Chambers, D. (2007). Beyond test accuracy: Benefits of measuring response time in computerised testing. *Australasian Journal of Educational Technology*, *23*(4), 10 –31. https://doi.org:10.14742/ajet.1251

Hulstijn, J. H. (2012). Incidental Learning in Second Language Acquisition. *The Encyclopedia of Applied Linguistics*, *5*, 2632–2640. https://doi.org/10.1002/9781405198431.wbeal0530

Jeon, J. (2021). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2021.1987272

Johannsen, G. (1979). Workload and workload measurement. *Mental Workload*, *8*, 3–11. https://doi.org:10.1007/978-1-4757-0884-4_1

Jung, J. (2018). Effects of task complexity and working memory capacity on L2 reading comprehension. *System*, *74*, 21–37. https://doi.org:10.1016/j.system.2018.02.005

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology*, *93*(3), 579–588. https://doi.org:10.1037/0022-0663.93.3.579

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186x.2017.1280256

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2001). The impact of "No Opinion" response options on data quality. *Public Opinion Quarterly*, *66*(3), 371–403. https://doi.org:10.1086/341394

Lai, C., Chen, Q., Wang, Y., & Qi, X. (2023). Individual interest, self-regulation, and self-directed language learning with technology beyond the classroom. *British Journal of Educational Technology*, *July*, 1–19. https://doi.org/10.1111/bjet.13366

Laufer, B. (2017). The three "I"s of second language vocabulary learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 343-354). Routledge. https://doi.org:10.1177/1362168816683118

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. https://doi.org/10.1191/0265532204lt277oa

Leech, D. H. (1991). Teaching and learning vocabulary by I.S.P. Nation. New York: Newbury house, 1990. 275 pp. *Issues in Applied Linguistics*, *2*(1). https://doi.org/10.5070/l421005136

Lee, H. (2014). Measuring cognitive load with electroencephalography and self-report: Focus on the effect of English-medium learning for Korean students, Educational Psychology. *An International Journal of Experimental Educational Psychology*, *34*(7), 838-848. https://doi.org:10.1080/01443410.2013.860217

Lee, J. (2018). Task complexity, cognitive load, and L1 speech. *Applied Linguistics*, *40*(3), 506–539. https://doi.org:10.1093/applin/amx054

Lee, J. (2019). Task complexity, cognitive load, and L1 speech. *Applied Linguistics*, *40*(3), 506–539. https://doi.org:10.1093/applin/amx054

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing and Health*, *25*(4), 295–306. https://doi.org:10.1002/nur.10041

Leppink, J. (2017). Cognitive load theory: Practical implications and an important challenge. *Journal of Taibah University Medical Sciences*, *12*(5), 385–391. https://doi.org:10.1016/j.jtumed.2017.05.003

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. https://doi.org:10.3758/s13428-013-0334-1

Lou, N. M., & Noels, K. A. (2016). Changing language mindsets: Implications for goal orientations and responses to failure in and outside the second language classroom. *Contemporary Educational Psychology*, *46*, 22–33. https://doi.org:10.1016/j.cedpsych.2016.03.004

Marefat, F. & Pakzadian, M. (2017). Attitudes towards English as an international language (EIL) in Iran: Development and validation of a new model and questionnaire. *Iranian Journal of Applied Language Studies*, *9*(1), 127–154. https://doi.org:10.22111/ijals.2017.3166

Martin, S. (2014). Measuring cognitive load and cognition: metrics for technology-enhanced learning. *Educational Research and Evaluation*, *20*(7-8), 592–621. https://doi.org:10.1080/13803611.2014.997140

Masrai, A. (2022). The Development and Validation of a Lemma-Based Yes/No Vocabulary Size Test. *SAGE Open*, *12*(1), 215824402210743. https://doi.org/10.1177/21582440221074355

Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System*, *72*, 23–36. https://doi.org:10.1016/j.system.2017.10.005

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, *18*(4), 393–407. https://doi.org/10.1191/0267658302sr211xx

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.) *Applied linguistics in society*, 80–87.

Meara, P., & Jones, G. (1990). Eurocentres vocabulary size test. *TESL Canadian Journal*, *3*(1), 69–79.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org:10.1037/h0043158

Miller, G., & Kintsch, W. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Multilingual Matters.

Minkley, N., Xu, K. M., & Krell, M. (2021). Analyzing relationships between causal and assessment factors of cognitive load: Associations between objective and subjective measures of cognitive load, stress, interest, and self-concept. *Frontiers in Education*, *6*(April), 1–15. https://doi.org/10.3389/feduc.2021.632907

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. https://doi.org:10.1515/iral-2017-0016

Nakamura, Y. T., Gu, Y., Jin, H., Yu, D., Hinshaw, J., & Rehman, R. (2022). Introducing neuroscience methods: An exploratory study on the role of reflection in developing leadership from a HRD perspective. *Human Resource Development International*, *26*(4), 458–470. https://doi.org:10.1080/13678868.2022.2094151

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge university press. https://doi.org/10.1017/cbo9781139524759

Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching*, *44*(4), 529–539. https://doi.org/10.1017/s0261444811000267

Nation, P. (2019). The different aspects of vocabulary knowledge. In *The Routledge handbook of vocabulary studies*, 15–29. https://doi.org/10.4324/9780429291586-2

Nation, P., & Anthony, L. (2016). Measuring vocabulary size. In *Handbook of Research in Second Language Teaching and Learning*, 3, 355–368. https://doi.org/10.4324/9781315716893

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13 https://doi.org:10.1016/J.sbspro.2015.07.546

Noroozi, S., & Karami, H. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Language Testing in Asia*, *12*(1). 1–19. https://doi.org:10.1186/s40468-022-00163-8

Noroozi, S., Karami, H., & Saeedi, Z. (2023). An investigation of the relationship between subjective and objective cognitive load measures of language item difficulty. *Language Horizons*, *7*(1), 7–33. https://doi.org:10.22051/lghor.2022.37418.1551

Prisacari, A. A., & Danielson, J. (2017). Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use. *Computers in Human Behavior*, *77*, 1–10. https://doi.org:10.1016/j.chb.2017.07.044

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org:10.1037/0022-0663.84.4.429

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(4), 737–743. https://doi.org:10.1177/001872089303500412

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351–371. https://doi.org/10.1007/bf02213420

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122–133. https://doi.org:10.1037/0022-0663.86.1.122

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org:10.1207/s15326985ep3801_8

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics*, *24*, 146–161. https://doi.org/10.1017/S0267190504000078

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32. https://doi.org:10.1191/026553201666879851

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org:10.3758/s13428-018-01193-y

Pelánek, R., Effenberger, T., & Čechák, J. (2021). Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*., *32*(1),196–232. https://doi.org:10.1007/s40593-021-00252-4

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198. https://doi.org:10.1037/h0049234

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal*, *102*(4), 713–731. https://doi.org:10.1111/modl.12509

Ponce, H. R., Mayer, R. E., Sitthiworachart, J., & López, M. J. (2020). Effects on response time and accuracy of technology-enhanced cloze tests: an eye-tracking study. *Educational Technology Research and Development*, *68*(5), 2033–2053. https://doi.org:10.1007/s11423-020-09740-1

Pouw, W. T. J. L., Eielts, C., van Gog, T., Zwaan, R. A., & Paas, F. (2016). Does non-meaningful sensori-motor engagement promote learning with animated physical systems? *Mind, Brain, and Education*, *10*(2), 91–104. https://doi.org:10.1111/mbe.12105

Révész, A., Michel, M., & Gilabert, R. (2015). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgements. *Studies in Second Language Acquisition*, *38*(4), 703–737. https://doi.org:10.1017/s0272263115000339

Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning*, *64*(3), 615–650. https://doi.org:10.1111/lang.12061

Sasayama, S. (2016). Is a "complex" task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, *100*(1), 231–254. https://doi.org:10.1111/modl.12313

Scharinger, C., Kammerer, Y., & Gerjets, P. (2015). Pupil dilation and EEG alpha frequency band power reveal load on executive functions for link-selection processes during text reading. *PLOS ONE*, *10*(6), e0130608. https://doi.org/10.1371/journal.pone.0130608

Scharinger, C., Schüler, A., & Gerjets, P. (2020). Using eye-tracking and EEG to study the mental processing demands during learning of text-picture combinations. *International Journal of Psychophysiology*, *158*, 201–214. https://doi.org/10.1016/j.ijpsycho.2020.09.014

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Palgrave MacMillan. https://doi.org/10.1057/978023 0293977

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/s0261444812000018

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, *19*(4), 469–508. https://doi.org:10.1007/s10648-007-9053-4

Skulmowski, A., & Rey, G. D. (2017). Measuring cognitive load in embodied learning settings. *Frontiers in Psychology*, *8*, 11–91. https://doi.org:10.3389/fpsyg.2017.01191

Skuballa, I. T., Xu, K. M., & Jarodzka, H. (2019). The impact of co-actors on cognitive load: When the mere presence of others makes learning more difficult. *Computers in Human Behavior*, *101*, 30–41. https://doi.org/10.1016/j.chb.2019.06.016

Solhjoo, S., Haigney, M. C., McBee, E., van Merrienboer, J. J. G., Schuwirth, L., Artino, A. R., Battista, A., Ratcliffe, T. A., Lee, H. D., & Durning, S. J. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-50280-3

Sundqvist, P., & Wikström, P. (2015). Out-of-school digital gameplay and in-school L2 English vocabulary outcomes. *System*, *51*, 65–76. https://doi.org:10.1016/j.system.2015.04.001

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org:10.1016/0364-0213(88)90023-7

Sweller, J. (2010). Cognitive Load Theory: Recent Theoretical Advances. In J. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive Load Theory* (pp. 29-47). Cambridge: Cambridge University Press. https://doi:10.1017/CBO9780511844744.004

Sweller, J. (2019). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1-16. https://doi.org:10.1007/s11423-019-09701-3

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*(3), 185–233. https://doi.org:10.1207/s1532690xci1203_1

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org:10.1007/s10648-019-09465-5

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org:10.1023/a:1022193728205

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org:10.1111/j.1745-3984.2009.00080.x

Van de Weijer-Bergsma, E., & Van der Ven, S. H. G. (2021). Why and for whom does personalizing math problems enhance performance? Testing the mediation of enjoyment and cognitive load at different ability levels. *Learning and Individual Differences*, *87*. 101982. https://doi.org/10.1016/j.lindif.2021.101982

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*(1), 16–26. https://doi.org:10.1080/00461520701756248

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2),147–177. https://doi:10.1007/S10648- 005-3951-0

Wiebe, E. N., Roberts, E., & Behrend, T. S. (2010). An examination of two mental workload measurement approaches to understanding multimedia learning. *Computers in Human Behavior*, *26*(3), 474–481. https://doi.org:10.1016/j.chb.2009.12.006

Wood, D. A., & Adkins, D. C. (1960). *Test construction: Development and interpretation of achievement tests*. CE Merrill Books.

Xiangming, L., Li, X., & Zhang, J. (2023). Longitudinal reading outcome and cognitive load in individual- and collaboration-based environments. *Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2022.2163666

Yin, B., Chen, F., Ruiz, N., & Ambikairajah, E. (2008). Speech-based cognitive load monitoring system. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2041–2044. https://doi.org:10.1109/icassp.2008.4518041

Young, J. Q., ten Cate, O., O'Sullivan, P. S., & Irby, D. M. (2016). Unpacking the complexity of patient handoffs through the lens of cognitive load theory. *Teaching and Learning in Medicine*, *28*(1), 88–96. https://doi.org:10.1080/10401334.2015.1107491

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: Implications for medical education. *Medical Teacher*, *36*(5), 371–384. https://doi.org:10.3109/0142159X.2014.889290

Yu, X. (2021). Text complexity of reading comprehension passages in the national matriculation english test in China: The development from 1996 to 2020. *International Journal of Language Testing*, *11*(2), 142–167.

Zheng, R., & Cook, A. (2011). Solving complex problems: A convergent approach to cognitive load measurement. *British Journal of Educational Technology*, *43*(2), 233–246. https://doi.org/10.1111/j.1467-8535.2010.01169.x