ALZAHRA UNIVERSITY
Journal of
Language Horizons
LGHOR

Scientefic Quarterly Journal of Language Horizons, ***Alzahra University***
Volume 7, Issue 1, Spring 2023 (Serial No. 15)

An Investigation of the Relationship Between Subjective and Objective Cognitive Load Measures of Language Item Difficulty

Research Article
pp. 7-33

Shadi Noroozi[1]
Hossein Karami*[2]
Zari Saeedi[3]

Received: 2021/10/01 Accepted: 2022/03/16

Abstract

The current study strived to delve into the response behavior and perceptions of examinees while taking a test in light of cognitive load theory. The empirical data were collected from 60 MA English major graduates and students, with a high level of language proficiency. The participants were required to answer 60 multiple-choice language items (i.e., grammar and vocabulary questions), taken from the high-stakes tests of the MA English majors of the Iranian university entrance examination (IUEE), as fast and as accurately as possible. After completing each test item, they rated their perceptions with regard to the difficulty of test items (Bratfisch et al., 1972) and the amount of mental effort (Paas, 1992). Their response time spent on each language item and their selected options were also stored by the Psychopy software (Peirce et al., 2019). Through running Pearson and Spearman rho correlations, the findings revealed that response time enjoyed a strong positive correlation with mental effort, meaning that both objective and subjective cognitive load measures matched in terms of

* Corresponding Author
[1] MA Holder, English Department, Faculty of Foreign Languages and Literature, University of Tehran, Tehran, Iran. shadi.noroozi@ut.ac.ir
[2] Assistant Professor, English Department, Faculty of Foreign Languages and Literature, University of Tehran, Tehran, Iran. hkarami@ut.ac.ir
[3] Associate Professor, Department of English Language and Literature, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran. saeedi.za@atu.ac.ir

DOI: 10.22051/lghor.2022.37418.1551
DOR: 20.1001.1.2588350.2023.7.1.1.9

sensitivity to cognitive load changes in language test items. Further, the subjective measures of perceived mental effort and perceived level of difficulty revealed to be the sound indicators of cognitive load changes. As predicted, response time also indicated that more difficult language test items imposed a greater amount of load. The implications of the study will be explained.

Keywords: cognitive load, language test-items, multiple-choice questions, response-time, subjective/objective measures

Introduction

Test developers and psychometricians have voiced their concern over the understanding of and improving the psychometric qualities of language tests in recent years. The remarkable effects of high-stakes tests not only on the individuals' academic career (Shohamy et al., 1996) but also on their cognitive architecture or minds (Sweller et al.,1998) have recently received considerable critical attention. The exploration of cognitive processing provoked by test items is becoming fundamental to the examination of language test items because cognitive processes can have undue influences on the individuals' language learning and their performance on tests (Gass et al., 2013; Ponce et al., 2020). In this regard, the effect of tasks (items) on individuals' minds has been investigated with respect to cognitive processing using cognitive load theory (CLT) (Dindar et al., 2015).

Cognitive load theory was presented by Sweller in the 1980s and this theory can account for cognitive load patterns as to multimedia learning (Brünken et al., 2003; Paas et al., 2008; Whelan, 2007; Wiebe et al., 2010), instructional materials, learning, and teaching (Sweller et al., 1998, 2019). Concerning language testing, cognitive load can be expressed as the distribution of cognitive capacity of test takers while taking the test (Sweller, 1988). Hence, cognitive load plays a pivotal role in language test performance. Of the three classifications of cognitive load theory (i.e., intrinsic, extraneous, and germane load), intrinsic load corresponds to inherent tasks (items) characteristics (de Jong, 2010) which can be related to the difficulty of tasks (Paas et al., 2003). Substantially, CLT can be regarded as a new line of research on the explanation for the difficulty of some material in comparison with some other material (Martin, 2014).

Subjective (e.g., perceived mental effort and difficulty level), objective behavioral (e.g., response time and reaction time), and physiological (e.g., electroencephalography, pupil dilation, heart rate, etc.) measures are prevalent in cognitive load studies. Mental effort appears to have a relationship with the processes or the cognitive capacities dedicated to accomplishing a task, whereas perceived level of difficulty seems to be related to the difficulty of a task or item itself (van Gog & Paas, 2008). On the other hand, it has been shown that several measures should be used to paint a complete picture of cognitive load (Leppink, 2017; Skulmowski & Rey, 2017). Thus, considering this issue can provide insights into whether there is a correspondence between cognitive load measures using tasks (items) with varying degrees of difficulty. Consequently, the current study set out to scrutinize the behaviors and perceptions of test takers while taking the multiple choice vocabulary and grammar questions of IUEE in light of CLT.

Literature Review

Cognitive Load Theory (CLT)

This theory centers on the "unobservable" phenomena (i.e., cognitive load) that individuals experience while dealing with different tasks. Central to the entire discipline of cognitive load theory (CLT) are the constructs of *cognitive load* and *learning*. This theory has long been an object of research in a wide range of scientific domains such as cognitive psychology and instructional design (Sweller, 2010). According to Sweller et al. (1998), CLT is a framework that explains the relationships among learning, cognitive architecture, and materials design. It is pivotal to know about the cognitive architecture and the efficiency of instructional designs in CLT.

Cognitive load theory rests on the assumption that cognitive architecture of human beings consists of two sets of memory stores: limited working memory and unlimited storage memory (Martin, 2014). As one of the elements of cognitive architecture, working memory, a place for the occurrence of all conscious cognitive processing, is responsible for information processing (Paas, 1992). Nonetheless, this type of memory is restricted in capacity to hold and process information (Miller, 1956; Peterson & Peterson, 1959). Due to this limitation, simultaneous processing and connecting several elements of a task may considerably surpass working memory capacity. Hence, learning may be hampered (Chandler & Sweller, 1991;

Paas, 1992).

In fact, the imposed load on the working memory is directly affected by task requirements (Sweller et al., 1998). To avoid the unnecessary cognitive load or overload imposed by a task, working memory capacity limitations should be considered in instructional designs (Sweller et al., 2011). Cognitive overload and underload may have adverse effects. As for cognitive overload, Johannsen (1979) reasoned that too much amount of load can negatively affect the functioning of working memory. On the other hand, cognitive underload can influence the performance on a task due to the lack of motivation, for example (Young et al., 2015).

To have effective learning, instructional designs should be reconsidered and modified by reducing working memory load. In other words, it seems necessary that the instructional design be proportional to cognitive architecture (Schnotz & Kürschner, 2007). In addition to learning, performance is also affected by the overload or underload. As an aspect of CLT, *performance* is concerned with the correct responses, error rate, and response time (Paas et al., 2003). Poor performance can be ascribed to task demands surpassing cognitive capacity. Stated more specifically, the demands of a task can influence the amount of load imposed on the minds of the test takers and hence their performance (Dindar et al., 2015; Gvozdenko & Chambers, 2007). Therefore, test takers' performance can indicate the load imposed by the test items.

Intrinsic Load

Due to the interaction between instructional design and cognitive architecture, several forms of loads (i.e., intrinsic, extraneous, and germane) have been identified (Paas et al., 2003; Sweller, 1994; Sweller et al., 2011; Sweller et al., 1998, 2019). Intrinsic load may be defined in terms of the innate nature of a task (Sweller et al., 1998). Central to intrinsic load is *element interactivity* which deals with the simultaneous processing of several elements in the working memory to understand or learn a task (Sweller, 2010). Therefore, the relationship between working memory capacity and element interactivity becomes paramount. Pollock et al. (2002) recognized grammatical syntax as an instance of high interactivity material. In addition, there is a positive relationship among element interactivity, intrinsic load, and working memory capacity (Sweller, 2010).

For de Jong (2010), intrinsic load can refer to the experienced difficulty of a task. In other words, the difficulty of materials can be associated with their intrinsic or inherent nature. According to a definition provided by Sweller (2020), the term *difficulty* may be defined with respect to the nature of the information of a task and the individual's knowledge. Difficulty may have different sources (Sweller et al., 2011). Indeed, the degree to which task elements interact is related to intrinsic load. Regarding a task with low element interactivity, the simultaneous processing of a great number of elements seems to be the cause of difficulty in learning the vocabulary of a language (e.g., learning the translation of "dog" into Persian).

Measuring Cognitive Load

Traditionally, measuring cognitive load was restricted to error rate. As the theory developed, more direct measures of cognitive load burgeoned (Sweller et al., 2011). Due to the lack of any single standardized method (Brünken et al., 2010), the implementation of diverse measures is pivotal to obtain a more precise picture of cognitive load (Leppink, 2017; Skulmowski & Rey, 2017). The validity and reliability of measures are of great concern to assess the load imposed by the instructional designs and experienced by the individuals (Brünken et al., 2010).

Regarding the subjective measures, self-report questionnaires are used as for rating perceived mental effort and level of difficulty. Learners can be asked to reflect on their invested mental effort (Paas, 1992). Brünken et al. (2010) maintained that behavioral parameters were taken as the indicators of cognitive load. Neuro-imaging techniques, time-on-task, the secondary task, and reaction time are different methods identified to measure cognitive load objectively.

Cognitive Load Theory: Review of Empirical Studies

Concerning the implementation of diverse instruments, there is a wave of studies that applied cognitive load measures to the fields of multimedia learning (DeLeeuw & Mayer, 2008; Dindar et al., 2015), task-based language teaching (TBLT) (Lee, 2019; Révész et al., 2016; Sasayama, 2016), and testing (Pouw et al., 2016; Prisacari & Danielson, 2017).

To investigate the underlying trait of cognitive load, DeLeeuw and Mayer (2008) carried out two experiments through subjective and objective measures in multimedia learning. They examined the sensitivity of measures through the use of different levels of sentence complexity, problem solving situations, and redundancy program. As for the sentence complexity, the high complexity sentence contained more interacting elements in comparison with the low complexity sentence. Sentence complexity can potentially cause intrinsic load (Sweller, 1999); thus, the higher the element interactivity of a sentence, the more complex the sentence and hence the greater the intrinsic load. Besides, reaction time, a meaningful indicator of cognitive load, showed that the longer the reaction time, the higher the cognitive load.

In their first experiment, they found a significant positive correlation between mental effort and sentence complexity: the higher the sentence complexity, the greater the amount of invested mental effort. A positive correlation was also found between reaction time and high complexity sentence. Moreover, the study revealed a modest significant correlation between reaction time and mental effort. Similar to the first experiment's results, a higher amount of mental effort was observed in the case of high complexity sentence in their second experiment. In contrast to the first experiment's findings, reaction time did not measure precisely task difficulty with regard to the number of interacting elements. In addition, no significant correlation was found between mental effort and reaction time.

In one study undertaken by Dindar et al. (2015), the difference between the cognitive loads of two different types (static vs. graphic) of achievement tests was investigated by making use of response time, rate of accuracy, subjective measure of mental effort, and the secondary task. The results demonstrated that response time was a reliable index of cognitive load: the longer the response time, the more complex the task, and the greater the load. Previous studies reported a modest correlation between reaction time and mental effort (DeLeeuw & Mayer, 2008; van Gerven et al., 2006); however, no statistically significant correlation was observed between the variables.

As for the context of TBLT, numerous studies have been designed to examine the validity of task complexity with regard to subjective and objective measures widely implemented in cognitive load studies (e.g., Lee, 2019; Révész et al., 2016; Révész et al., 2014; Sasayama, 2016). Sasayama (2016) conducted research to assess the cognitive complexity of the narration of four picture sequences through three measures: time estimation, self-rating, and the secondary task. After each narration, perceived level of difficulty and mental effort in addition to time estimation were measured. The four tasks had varying difficulty, ranging from the simplest to the most complex. The participants were also instructed to respond quickly to the secondary task. The findings revealed that the most complex task required the longest reaction time. Moreover, the most complex task was considered the most difficult and seemed to require the greatest amount of mental effort. Sasayama (2016) highlighted the impact of proficiency not only on the participants' performance on the same task but also on the cognitive load measures. In other words, proficiency seemed to affect the performance of learners and cognitive load measures differentially. Her findings also suggested that the reported effects on the cognitive task complexity were over-estimated through the use of self-reports. Therefore, it was advised that the interpretation of the findings obtained through the subjective measures be made with caution.

In addition, Révész et al. (2016), using the measures of self-ratings, expert judgments, and reaction time, investigated the validity of task complexity. Their results indicated that the more complex the task, the greater the amount of consumed mental effort. The findings also revealed that more complex tasks were identified as more difficult. No statistically significant difference was found between the different primary tasks' reaction time and their different levels of complexity, meaning that reaction time had no relationship with task difficulty.

In another study, Lee (2019) examined whether variations in task complexity could truly lead to alterations in cognitive load by the implementation of the self-ratings of perceived mental effort, stress, difficulty, and time estimation, and the implementation of the objective measure of the secondary task. The results showed that the most complex tasks were perceived to be the most difficult. Furthermore, these tasks enjoyed the highest rate of mental effort. The results indicated a positive relationship between response time and task complexity. In contrast to the earlier studies that highlighted the importance of accuracy rates in the case of more complex tasks (Révész et al., 2016; Révész et al., 2014), Lee (2019) observed significant effects for the amount of time rather than accuracy. Similarly, Sasayama (2016) reported no effects for accuracy. In other words, reaction time was the longest regarding the most complex tasks compared to the ones with the least complexity.

On the other hand, Lee (2014) examined the validity and reliability of cognitive load measures through electroencephalography (EEG), self-ratings, and learning outcomes. After watching a seven-minute documentary video, the participants were directed to report quickly their perceived mental effort and difficulty. The findings revealed a statistically negative correlation between difficulty ratings and learning outcomes. In other words, as the task became more complex and imposed a greater amount of load, it was perceived to be more difficult. The participants appeared to cease expending mental effort on the more difficult task. Hence, their performance on the learning outcome (i.e., comprehension test) was unsatisfactory. To put it simply, when the intrinsic load increased, the comprehension was interrupted and the participants invested less effort in accomplishing the tasks.

As to the realm of testing, Pouw et al. (2016) explored the influence of meaningful versus non-meaningful conditions of physical engagement on different forms of competency (i.e., unreflective, reflective, and motoric) in solving problems. No statistically significant correlations between response time and measures of mental effort and perceived difficulty were reported.

Contrasting with the substantial body of evidence on the role of cognitive load in language teaching and learning, research on item loads has received scant attention (Ponce et al., 2020). In general, test items have attracted the researchers' attention in such fields as chemistry (Prisacari & Danielson, 2017), mathematics (Gvozdenko & Chambers, 2007), and algebra (Sweller et al., 2011); however, examining the load of language test items has been restricted to the comparison of item loads between different modes (e.g., computer vs. paper) or types of materials (e.g., static vs. animated graphics) (Dindar et al., 2015; Prisacari & Danielson, 2017). Specifically, the load of multiple-choice language items has been under-explored (Ponce et al., 2020).

Also, there are inconsistencies in the relationship between mental effort and response time. It has been suggested that the longest response time was recorded for the most mental-effort-consuming task (Sasayama, 2016). On the other hand, a moderate correlation (DeLeeuw & Mayer, 2008) and no statistically significant correlation (DeLeeuw & Mayer, 2008; Dindar et al., 2015; Pouw et al., 2016) between mental effort and response time were reported. Recall that a significant correlation between response time and mental effort was observed in the first experiment of the study conducted by DeLeeuw and Mayer (2008); however, in their second experiment, a non-significant but small correlation was observed. Consequently, there seems to be no consensus among the researchers.

On the other hand, much of the current literature centered on the relationship between task difficulty and mental effort. Many researchers unanimously emphasized a positive relationship between mental effort and task difficulty (Lee, 2019; Révész et al., 2016; Sasayama, 2016). Likewise, DeLeeuw and Mayer (2008) highlighted the significant correlation between mental effort and task complexity. This is in contrast to Lee's (2014) argument.

Hence, due to the scarcity of research on the relationship between difficulty and the cognitive load of language test items, there is a call for the implementation of multiple measures of cognitive load. Specifically, to collect dependable evidence on the individuals' performance, not affected by the load of another task, the use of response time has been suggested. Furthermore, the use of objective and subjective measures can reveal the difference between the actual and the perceived cognitive load.

The paucity of research on the investigation of item functioning as to the measures of cognitive load in the realm of language testing in the Iranian context has incurred several problems because disregarding the cognitive load imposed by language test items on the test taker's minds may lead to imposing a high load on their minds. This can consequently influence their cognitive processes and performance due to the working memory capacity's reaching its limitation (Goldhammer et al., 2014; Ponce et al., 2020; Sweller et al., 1998). Stated another way, success or failure in the test heavily relies on the load of items imposed on the examinees' minds. Test items may not provoke loads commensurate with their designed features and difficulties, which can lead to the overload of test takers' working memory capacity. This can result in the failure in answering an item.

The Present Study

The current study strived to provide a more precise image of load patterns of language test items (i.e., vocabulary and grammar sections) of the MA Iranian university entrance examination (IUEE) of English majors. Given that item loads can influence test takers' performance and can also reveal information about item functioning, the patterns of item loads were portrayed through the simultaneous implementation of subjective (i.e., perceived mental effort and perceived level of difficulty self-reports) and objective behavioral (i.e., response time) measures.

Stated more precisely, the present study investigated the relationship between response time and mental effort to check if response time can reveal cognitive load in line with mental effort. Also, the relationships between the subjective measures of perceived difficulty and mental effort were sought to check whether both measures match and assess difficulty similarly. Further, the relationships between response time and perceived level of difficulty were explored to ascertain whether the experienced difficulty was reflected in the time spent on answering the test items.

The scrutiny of language test items has blossomed in the world; however, the load of language test items of the MA English majors of the Iranian university entrance examination (IUEE) has not been explored in our context. Therefore, there is a need for methods triangulation (Ary et al., 2019) through subjective and behavioral measures to cast light on the actual cognitive processing and the test item functioning.

The present study strived to address the following questions:

1. Is there any statistically significant relationship between the test-takers’ perceived mental effort and the response time for each language test item?
2. Is there any statistically significant relationship between the test takers' perceived mental effort and perceived level of difficulty?
3. Is there any statistically significant relationship between the test takers' perceived level of difficulty and the response time for each language test item?

Method

Participants

Twenty-five male and 35 female MA graduates and students of the University of Tehran, Allameh Tabatabaee University, and Alzahra University, aged between 21 and 39 (*M* = 27.28, *SD* = 4.41) and majoring in teaching English as a foreign language, translation studies, and English literature, attended the study. The participants were selected through convenience sampling (Dornyei, 2007). They were homogenous in terms of language proficiency such that their proficiency level fell into the categories of advanced and very advanced users of the English language after taking the Oxford placement test (OPT) (Dave, 2004). In addition, the piloting phase was necessary to set the fixed timing for presenting each language test item on the screen. To this end, five female and male participants attended the pilot phase. Moreover, regarding the probable influence of practice effect, as the time span between the entrance examination and the time the test was run was approximately two years, it was supposed that this effect would be very unlikely to exist.

Instruments

Multiple-choice vocabulary and grammar items, two self-report ratings, and a proficiency test were the instruments used to collect data. The current study made use of grammar (20 items) and vocabulary (40 items) sections of the MA English majors of the IUEE tests held in 2018 and 2019. The Psychopy software collected the response answer and response time of every multiple-choice question with the precision of milliseconds (Peirce et al., 2019). The present study included two subjective self-reports of perceived mental effort and perceived level of difficulty. As one of the load components, mental effort, developed and validated by Paas (1992), was measured subjectively through the *Mental Effort* rating scale. Examinees can report mental effort on a 7-point symmetrical category scale on a numerical value spanning from *very low* (1) to *very high* (7) mental effort. The reasoning behind the prevalent implementation of mental effort self-report is the simplicity of data gathering and analysis (Paas, 1992; Paas & van Merriënboer, 1993). Further, the reliability of mental effort measure has shown to be acceptable (α = .82).

As for the second rating scale, *Level of Difficulty*, developed and validated by Bratfisch et al. (1972), was rated by a 7-point scale spanning from 1 (*very easy*) to 7 (*very difficult*). Perceived difficulty is concerned with the difficulty of the item itself. This self-report rating scale has also been shown to be a sound indicator of cognitive load (Prisacari & Danielson, 2017). Although task difficulty and mental effort may 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). To check the homogeneity of participants with respect to language proficiency, the grammar section of the OPT was administered. Based on whether their scores fell in the range of 75 to 100, the categories of advanced to very advanced language users, the participants attended the main phase of the study.

Procedure

Participants took the test for about one hour. They were instructed how to deal with the self-ratings, and had received complete explanations on the definitions of mental effort and perceived difficulty level. They were also directed to answer the language test items as fast and as correctly as possible. Upon confirmation of their understanding of the explanations, the participants were then required to answer the multiple choice test items in addition to the self-reports of mental effort and level of difficulty. Indeed, they answered grammar and vocabulary items sequentially presented in two conditions, each including 20 and 40 items, respectively. The order of language test items was randomized by the Psychopy software. A fixation cross (i.e., 500 ms) was also shown after the presentation of each language test item. Having answered each language item, they rated their perceived difficulty and the amount of experienced mental effort.

Results

Before running Pearson and Spearman rho correlations, the normality assumption for response time, perceived mental effort, and task difficulty was evaluated by Shapiro–Wilk's test ($p > 0.05$). Indeed, no violation of this assumption was observed. The data were also checked for outliers with respect to response time, mental effort, and perceived difficulty. Only two cases' response times fell three standard deviations away from the mean and the overall mean replaced their response times. The reason for running Spearman rho correlation was the fact that the data were collected through Likert-type items. Indeed, Likert-type items or ranked data can sometimes be regarded as ordinal in nature (Boone et al., 2014; Pallant, 2016).

The first research question scrutinized the relationship between the test takers' perceived mental effort and response time. To answer the question, Pearson correlation and Spearman rho correlation were obtained. However, due to the similarity of the results, only the results pertinent to Pearson correlation are reported. Note that only correct responses are considered in the analysis in the case of response time because only correct answers seem to reveal cognitive load (Lee, 2019). The results from the correlation between response time and mental effort are summarized in Tables 1 and 2. Also note that the results of each grammar and vocabulary section are presented in two separate sets such that the first set represents the items related to the test held in 2019 and the second set indicates the items associated with the test administered in 2018. This can potentially help cross-validate the results through a second data set.

The results obtained from the correlational analyses are set out in Table 1. As evident, a strong positive correlation between the response time and mental effort of the first set of grammar items was observed (r = .625, p < .001). Additionally, the correlation between response time and mental effort of the second set of grammar items was also significant (r = .631, p < .001).

Table 1

Pearson Product Moment Correlation of Response Time and Mental Effort (Grammar Items)

		1	2	3	4
1. Response time grammar items (first set)	Pearson	__	.625**	.640**	.202
	Sig. (2-tailed)		.000	.000	.121
2. Mental effort grammar items (first set)	Pearson		—	.389**	.566**
	Sig. (2-tailed)			.002	.000
3. Response time grammar items (second set)	Pearson			—	.631**
	Sig. (2-tailed)				.000
4. Mental effort grammar items (second set)	Pearson				__
	Sig. (2-tailed)				

Table 2 provides the correlations between mental effort and response time of 40 vocabulary items. This table is quite revealing in two ways. The results indicate that response time and mental effort of the first set of vocabulary items were positively correlated (*r* = .704, *p* < .001). A strong positive correlation was also found between mental effort and response time of the second set of vocabulary items (*r* = .769, *p* < .001).

Table 2

Pearson Product Moment Correlation of Response Time and Mental Effort (Vocabulary Items)

		1	2	3	4
1. Mental effort vocabulary items (first set)	Pearson	__	.704**	.687**	.515**
	Sig. (2-tailed)		.000	.000	.000
2. Response time vocabulary items (first set)	Pearson		__	.527**	.780**
	Sig. (2-tailed)			.000	.000
3. Mental effort vocabulary items (second set)	Pearson			—	.769**
	Sig. (2-tailed)				.000
4. Response time vocabulary items (second set)	Pearson				—
	Sig. (2-tailed)				

The next research question delved into the relationship between perceived mental effort and level of difficulty of the language test items. To this end, Spearman rho correlation was run as both of the variables were ordinal in nature. Note that the data were analyzed in two ways: considering only correct responses in one case and all responses (i.e., both correct and incorrect answers) in the other case. This was because no evidence was found in the literature for excluding incorrect responses from data analyses. Although results of the correlational analyses of only correct and all responses (i.e., including both incorrect and correct answers) were not too much different, their tables are reported.

As shown in Table 3, a strong positive correlation between mental effort and level of difficulty of the first set of grammar items was detected (r = .891, *p* < .001) considering only correct responses. It can further be seen from the data presented in the table that level of difficulty was significantly correlated with perceived mental effort of the second set of grammar items (r = .907, *p* < .001). As evident in Table 4, a strong positive correlation was found between the aforementioned variables in the case of the first set of grammar items including both correct and incorrect responses (r = .839, *p* < .001). Also, looking at Table 4, it is obvious that a strong positive correlation was observed between the aforesaid variables of the second set of grammar items considering all responses (r = .771, *p* < .001). Hence, it appears that the magnitudes of correlations were reduced in the case of including both correct and incorrect answers.

Table 3

Spearman Rho Correlation of Mental Effort and Difficulty Level (Only Correct Answers to Grammar Items)

		1	2	3	4
1. Mental effort grammar items (first set)	Spearman's rho Sig. (2-tailed)	__	.891** .000	.556** .000	.517** .000
2. Difficulty level grammar items (first set)	Spearman's rho Sig. (2-tailed)		__	.523** .000	.588** .000
3. Mental effort grammar items (second set)	Spearman's rho Sig. (2-tailed)			__	.907** .000
4. Difficulty level grammar items (second set)	Spearman's rho Sig. (2-tailed)				—

Table 4

Spearman Rho Correlation of Mental Effort and Difficulty Level (Correct and Incorrect Answers to Grammar Items)

		1	2	3	4
1. Mental effort grammar items (first set)	Spearman's rho	—	.839**	.881**	.743**
	Sig. (2-tailed)		.000	.000	.000
2. Difficulty level grammar items (first set)	Spearman's rho		—	.718**	.892**
	Sig. (2-tailed)			.000	.000
3. Mental effort grammar items (second set)	Spearman's rho			__	.771**
	Sig. (2-tailed)				.000
4. Difficulty level grammar items (second set)	Spearman's rho				—
	Sig. (2-tailed)				

Table 5 presents the correlations between perceived mental effort and level of difficulty of vocabulary section considering only correct responses. Looking at the table below, a positive correlation was found between the aforementioned variables of the first set of vocabulary items (r = .792, $p < .001$). Besides, perceived mental effort was significantly correlated with perceived difficulty level of the second set of vocabulary items (r = .881, $p < .001$). Furthermore, Table 6 displays an overview of the aforementioned variables' correlations. Looking at Table 6, it is apparent that a strong positive correlation was observed between the variables of the first set of vocabulary items ($r = .689$, $p < .001$). The results, as evident in Table 6, reveal that a significant positive correlation between the variables of the second set of vocabulary items was found ($r = .724$, $p < .001$). Thus, it can also be concluded that the consideration of all responses can reduce the magnitudes of correlations to some extent.

Table 5

Spearman Rho Correlation of Mental Effort and Difficulty Level (Only Correct Answers to Vocabulary Items)

		1	2	3	4
1. Mental effort vocabulary items (first set)	Spearman's rho	__	.792**	.617**	.493**
	Sig. (2-tailed)		.000	.000	.000
2. Difficulty level vocabulary items (first set)	Spearman's rho		__	.542**	.681**
	Sig. (2-tailed)			.000	.000
3. Mental effort vocabulary items (second set)	Spearman's rho			__	.881**
	Sig. (2-tailed)				.000
4. Difficulty level vocabulary items (second set)	Spearman's rho				—
	Sig. (2-tailed)				

Table 6

Spearman Rho Correlation of Mental Effort and Difficulty Level (Correct and Incorrect Answers to Vocabulary Items)

		1	2	3	4
1. Mental effort vocabulary items (first set)	Spearman's rho	__	.689**	.953**	.727**
	Sig. (2-tailed)		.000	.000	.000
2. Difficulty level vocabulary items (first set)	Spearman's rho		—	.624**	.918**
	Sig. (2-tailed)			.000	.000
3. Mental effort vocabulary items (second set)	Spearman's rho			—	.724**
	Sig. (2-tailed)				.000
4. Difficulty level vocabulary items (second set)	Spearman's rho				—
	Sig. (2-tailed)				

The third research question explored the relationship between the test takers' perceived difficulty level and response time. To answer the proposed question, it was necessary to obtain both Pearson and Spearman rho correlations because perceived level of difficulty may also be viewed as an ordinal variable. As the results were akin, only those pertinent to Pearson correlation are reported. Recall that incorrect responses are discarded.

The outcomes of the correlational analyses are presented in Tables 7 and 8. As shown in Table 7, there was a strong positive correlation between perceived level of difficulty and response time of the first set of grammar items (r = .578, p < .001). Besides, looking at Table 7, it is evident that there was a significant positive correlation between the aforesaid variables of the second set of grammar items (r = .602, p < .001).

Table 7

Pearson Product Moment Correlation of Response Time and Difficulty Level (Grammar Items)

		1	2	3	4
1. Response time grammar items (first set)	Pearson	__	.578**	.640**	.236
	Sig. (2-tailed)		.000	.000	.069
2. Difficulty level grammar items (first set)	Pearson		__	.338**	.583* *
	Sig. (2-tailed)			.008	.000
3. Response time grammar items (second set)	Pearson			__	.602* *
	Sig. (2-tailed)				.000
4. Difficulty level grammar items (second set)	Pearson				__
	Sig. (2-tailed)				

Table 8 displays the correlation between perceived level of difficulty and response time for vocabulary items. As this table shows, a positive correlation was found between the mentioned variables of the first set of vocabulary items (r =. 651, p < .001). Also, a strong positive correlation between the variables of the second set of vocabulary items was observed (r =. 785, p < .001).

Table 8

Pearson Product Moment Correlation of Response Time and Difficulty Level (Vocabulary Items)

		1	2	3	4
1. Response time vocabulary items (first set)	Pearson Sig. (2-tailed)	__	.651** .000	.780** .000	.474** .000
2. Difficulty level vocabulary items (first set)	Pearson Sig. (2-tailed)		__	.570** .000	.737** .000
3. Response time vocabulary items (second set)	Pearson Sig. (2-tailed)			__	.785** .000
4. Difficulty level vocabulary items (second set)	Pearson Sig. (2- tailed)				—

Discussion

The first research question sought to examine the relationship between response time and mental effort. By running Pearson and Spearman rho correlations, statistically significant correlations were observed. The outcomes seem to be contrary to those of DeLeeuw and Mayer (2008), Dindar et al. (2015), and Pouw et al. (2016) who found no significant correlation between response time and mental effort. Note that the results obtained in DeLeeuw and Mayer's (2008) research revealed a correlation with a small effect size of .12. Nonetheless, in part of their study, they found a statistically significant correlation between response time and mental effort. Hence, the current study's findings seem to be partly in accord with DeLeeuw and Mayer's (2008) study outcomes. The reason behind the conflicting findings may lie in making use of instruments which were different in nature.

Our findings also mirror those of Sasayama's (2016) study in which the least and most demanding tasks required the least and most amount of time, respectively. Besides, the easiest and the most difficult tasks were perceived as the least and most mental effort consuming, respectively. Our findings also seem to be in agreement with the outcomes of the recent study carried out by Ponce et al. (2020) who regarded response time as an appropriate indicator of cognitive load. They further concluded that the higher the cognitive load, the larger the response time needed to answer the question. Hence, response time and mental effort seem to be the sound indicators of cognitive load. Response time can indeed provide some evidence as to how deep the processing is or how much cognitive resources are required to accomplish a task (Goldhammer et al., 2014).

The second research question investigated the relationship between perceived mental effort and task difficulty through running Spearman rho correlation. Significant correlations with large effect sizes were observed with regard to both grammar and vocabulary items. The outcomes of the current study corroborate the findings of numerous studies (e.g., Lee, 2019; Révész et al., 2016; Sasayama, 2016) in which more difficult tasks were perceived as more demanding and required more mental effort. However, our findings are in disagreement with those of the study conducted by Lee (2014) in which the researcher reported when the task at hand became too demanding (i.e., when perceived to have a high level of difficulty), individuals ceased to invest mental effort in accomplishing the task. That is, when encountered with a very difficult task, they became reluctant to make an attempt to complete the task.

The last question explored the relationship between perceived level of difficulty and response time. Statistically significant correlations were detected between the variables of both grammar and vocabulary items. The outcomes of the current study appear to be consistent with those of Sasayama (2016) and Lee (2019) who reported that the most difficult task required the longest amount of time. Our outcomes are not in line with those of the studies carried out by Pouw et al. (2016), Révész et al. (2016), and Révész et al. (2014) who found no significant correlation between perceived level of difficulty and response time.

Directions for Further Research

This study cast light on the role of cognitive load in exploring the language test items to unravel the cognitive processes underlying test taking. However, this study has some limitations that can open up the line for further research. One limitation of this study concerns the mere focus on vocabulary and grammar items. Hence, the cloze test and reading comprehension sections of the entrance examination were not taken into account due to the execution constraints of the Psychopy software. Another limitation lies in the small number of participants due to the pandemic situation. Several caveats need to be noted about the generalizability of findings concerning the participants. First, the current study has considered only MA students majoring in teaching English as a foreign language, literature, and translation studies with a high level of proficiency. Note that test takers with a high level of proficiency can better distinguish nuances of task difficulty compared to those with a low level of proficiency (Ayres, 2006; Sasayama, 2016). Future studies should include a low proficiency group as well and compare the performance and perceptions of participants with those with a high level of proficiency.

Also, to gain a more profound 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 the least- to the most- complex ones and comparing their differences can also lead to interesting findings in future investigations. The inequivalent number of characters of grammar items can also be considered a confounding variable. Hence, future studies should address this issue by including grammar items of approximately equal number of characters.

Moreover, the present study focused on measuring mental effort merely through self-report rating scales. To develop a deeper understanding of the influence of task difficulty on mental effort and response time as well as perceived difficulty, a variety of nonintrusive physiological measures such as electroencephalography (Antonenko & Niederhauser, 2010) and eye-tracking (Scheiter et al., 2020) can be applied to capture cognitive load while test takers are taking the test. Ultimately, as test takers themselves might not rate their perceived mental effort or perceived level of difficulty based on consistent reasoning, further studies should be carried out to investigate expert judgments as well (Révész et al., 2016).

Conclusion

In conclusion, the results of the present study indicated that the expected relationships and hypotheses were borne out. That is, the outcomes of the statistical analyses provide support for the predictions that objective and subjective measures of cognitive load can reveal that more difficult language items impose a greater amount of cognitive load. Hence, both subjective and objective measures seem to match with respect to the difficulty of language items. Also, subjective measures assess cognitive load in a similar way.

The current study can be of paramount significance from various perspectives. Cognitive load theory can contribute to the exploration of item functioning in psychometrics through the concurrent use of various cognitive load measures. In this way, test designers can have a thorough grasp of load and function of the items they develop. This awareness might urge test designers to proceed with caution in designing test items when considering the possible detrimental effects of item malfunctioning on the test takers' minds. In other words, they can examine whether the test items they design correspond to the characteristics being experienced or perceived by the test takers; for example, when it comes to the difficulty of items, the speed of processing, and the investment of mental effort. To this end, the cognitive load measures can provide worthwhile evidence. In this respect, the examination of language test items of MA English majors of the Iranian university entrance examination (IUEE) can provide insights into the item functioning. Ignoring cognitive load measures can make our understanding of item functioning inadequate.

On the other hand, due to some criticism leveled against the use of subjective measures such as being subject to under- or over-estimation of individuals, the behavioral measures can contribute to CLT through providing helpful information about item difficulty, cognitive processes, and item functioning (de Jong, 2010; Ponce et al., 2020; van der Linden, 2009). All in all, the investigation of test item functioning through the implementation of cognitive load measures can be considered an initial stage in the scrutiny of language test items from the cognitive load perspective.

References

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

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

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Bratfisch, O., Borg, G., & Dornic, S. (1972). *Perceived item-difficulty in three tests of intellectual performance capacity* (Report No. 29). Institute of Applied Psychology.

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

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

Dave, A. (2004). *Oxford placement test*. Oxford University Press.

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., Yurdakul, I. K., & Dönmez, F. I. (2015). 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

Dornyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.

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

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

Gvozdenko, E., & Chambers, D. (2007). Beyond test accuracy: Benefits of measuring response time in computerised testing. *Australasian Journal of Educational Technology*, *23*(4), 542-558. https://doi.org/10.14742/ajet.1251

Johannsen, G. (1979). Workload and workload measurement. In N. Moray (Ed). *Mental workload: Its theory and measurement,* (pp. 3-11). Springer.

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

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

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

Martin, S. (2014). Measuring cognitive load and cognition: Metrics for technology enhanced learning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *20*(8), 592-621. https://doi.org/10.1080/13803611.2014.997140

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

Paas, F. (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., Ayres, P., & Pachman, M. (2008). Assessment of cognitive load in multimedia learning environments: Theory, methods, and applications. In D. Robinson & G. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning* (pp. 11-35). Information Age Publishing.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1)*,* 1-4. https://doi.org/10.1207/S15326985EP3801_1

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurements as a means to advance cognitive load theory. *Educational*

Psychologist, *38* (1), 63-71. https://doi.org/10.1207/S15326985EP3801_8

Paas, F., & van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors, 35*(4), 737-743. https://doi.org/10.1177/001872089303500412

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (6th ed.). George Allen & Unwin.

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

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

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 12*(1), 61-86. https://doi.org/10.1016/S0959-4752(01)00016-0

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

Pouw, W. T., 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,* 91-104. https://doi.org/10.1111/mbe.12105

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

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *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*, 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*, 231-254. https://doi.org/10.1111/modl.12313

Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: Are they biased? *Educational Psychology Review*, *32*(4), 1003-1027.

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

Shohamy, E., Dinitsa-Schmidt, S,. &Ferman, I. (1996). Test impact revisited washback effect over time. *Language Testing*, *13*(3), 298-317. https://doi.org/10.1177/026553229601300305

Skulmowski, A., & Rey, G. D. (2017). Measuring cognitive load in embodied learning settings. *Frontiers in Psychology*, *8*, 1-6. https://doi.org/10.3389/fpsyg.2017.01191

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (1999). *Instructional design in technical areas*. ACER.

Sweller, J. (2010). *Cognitive load theory: Recent theoretical advances.* In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29-47). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.004

Sweller, J. (2020). 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., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296. https://doi.org/10.1023/A:1022193728205

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31,* 261-292.

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

van Gerven, P. W. M., Paas, F., van Merriënboer, J. J. G., & Schmidt, H. G. (2006). Modality and variability as factors in training the elderly. *Applied Cognitive Psychology, 20*(3), 311-320. https://doi.org/10.1002/acp.1247

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

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. *Educational Research Review, 2*(1),1-12. https://doi.org/10.1016/j.edurev.2006.11.001

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

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics, 58*(1), 1-17. https://doi.org/10.1080/00140139.2014.956151