Received: 12 August 2024 | Revised: 29 November 2024 | Accepted: 3 March 2025 | Published online: 24 March 2025

RESEARCH ARTICLE

International Journal of Changes in Education
2025, Vol. 00(00) 1–10
DOI: 10.47852/bonviewIJCE52024077

BON VIEW PUBLISHING

Discovering Levels of Cognitive Demand in Tests Developed by Iranian High School English Teachers Across Grades

Ali Ghezelbash[1], Iman Alizadeh[2,*] and Hoda Divsar[1]

[1]*Department of TEFL, Payame Noor University, Tehran, Iran*

[2]*English Language Department, School of Medicine, Guilan University of Medical Sciences, Rasht, Iran*

Abstract: Despite the significance of the cognitive load of educational procedures, its status quo in teacher-made tests has remained largely uncharted. This study investigated the interplay between cognitive demand levels (CDL) of test items and students' academic levels. Using the content analysis method, English language tests developed by Iranian high school teachers across three grades (10 through 12) were collected, and a scheme was developed to code the CDL of the tests. The coding scheme was developed based on the Revised Bloom's Taxonomy, Webb's Depth of Knowledge, and Hess's Cognitive Rigor Matrix. Three trained coders coded the tests using the piloted and validated scheme. The results showed that the low CDLs of "remembering," "understanding," and "applying" constituted a large proportion (89.49%), while "analyzing," "evaluating," and "creating," as the high CDLs comprised a small number (10.51%) of the test items. Grade 12 and grade 11 tests contained the most frequent high and low CDLs, respectively. In all grades, "understanding" was the most frequent level, and "creating" was the least frequent level. The trend of remembering and creating was normal, understanding and analyzing was descending, and applying and evaluating showed variation across the grades. The study concludes that the tests predominantly required low CDLs and were not developed based on a measure of cognitive demand balancing and/or regulating low and high CDLs across the grades. It recommends that test developers use a cognitive demand guide specifying the underlying nature of tasks required in the tests, the complexity of the expressions and information presented in the tests, and the choice of test item stem format and response format to regulate the cognitive demand of tests for different school grades.

Keywords: cognitive demand, language, tests, high schools

1. Introduction

Language testing involves considerations of measuring language knowledge and designing appropriate procedures for measuring this knowledge [1]. Moreover, language tests involve test takers' cognitive processes, which formulate and structure information for conceptual and procedural understandings necessary for successful completion of assessment tasks [2]. Moreover, the development of students' cognitive domain is one of the primary goals of education systems [3]. The specification of the content of instruction and the determination of assessment components of students' learning in relation to the cognitive domain with an eye to the student's level of cognitive engagement and academic abilities have proved critical [4]. Different test items require varying levels of cognitive processing from test takers. Furthermore, a test for different levels of education should cover different cognitive levels to accommodate the different capabilities of learners. Stein et al. [5] maintain that different tasks cause varying levels of cognitive demand in students, which helps them think creatively and judgmentally. The levels of cognitive demand may be a variable of students' levels of cognitive development, necessitating the development of appropriate tasks to reflect different levels of cognitive demand. Stein et al. [5] define cognitive demand as "the kind and level of thinking required of students to successfully engage with and solve the task" [5]. As students advance through the academic stages, teachers need to adjust and involve them in answering tasks requiring increasing levels of cognitive demand [4]. Therefore, it can be concluded that tests are expected to incorporate items demanding low and high levels of cognitive processing, and the cognitive demand of tests is expected to rise as students move through academic levels. Moreover, as the cognitive processes provoked by test items can play a crucial role in test takers' performance, the analysis of the cognitive load of test items has become of paramount importance [6, 7]. Haladyna and Rodriguez [8] define cognitive demand as "the expected mental complexity involved when a test item is administered to a typical test taker." According to them, the cognitive demand of a test item is influenced by the nature of the underlying task being measured and the test taker's ability which depends on their instructional history. Kan et al. [9] argue that a test taker's type and level of thinking in dealing with a test item determine the cognitive demand of the item, noting that the cognitive demand may change depending on the type of item stem used. Based on the information presented in the

***Corresponding author:** Iman Alizadeh, English Language Department, School of Medicine, Guilan University of Medical Sciences, Rasht, Iran. Email: iman_alizadeh96@gums.ac.ir

item, a test taker may use lower-order thinking skills or higher-order thinking skills to answer the question [10]. Haladyna and Rodriguez [8] maintain that although a test taker's instructional history can influence their cognitive processes and choice of strategies in dealing with a test item, the cognitive demand of a test primarily depends on the underlying nature of the task required in the test. They further note that the cognitive complexity of a test item may change depending on the complexity of the expressions and information presented in the test. Additionally, the choice of an appropriate test item, stem format, and response format can also affect the cognitive demand of the test item. Therefore, it can be concluded that the cognitive demand of a test item has to do with the nature of the task required by the item, which is, in turn, regulated by the information presented in the stem and the information required in the response.

Different methods have been proposed and used to determine the cognitive demand of a test and are discussed under the Cognitive Load Theory [11]. Among many other dimensions, the theory is also concerned with the cognitive load a task exerts on learners. Cognitive psychologists have sought recourse to physiological measures such as functional magnetic resonance imaging (fMRI) and electroencephalography, subjective measures of perceived level of difficulty or cognitive load [7], a combination of subjective cognitive load measures and Rash Model [12], as well as objective behavioral measures like secondary task technique and response time [13]; moreover, a number of frameworks rooted in educational objectives and cognitive processes like that of Hess's Cognitive Rigor Matrix (CRM) [14], Webb's four-level Depth of Knowledge(DoK) framework [15], and Revised Bloom's taxonomy [16]. Scholars in the field have recommended that content experts in the field review the items with regard to a cognitive taxonomy to discover the cognitive complexity of the cognitive demand of tests [17]. A scrutiny of the literature in this regard attests that the CRM [14], DoK [15], and revised Bloom's taxonomy [16] are used to determine and predict the cognitive demand of test items. The frameworks treat cognitive demand as an inherent characteristic of the test item and are considered a standard approach to studying the cognitive demand associated with a test item [18]. Bloom [3] identifies six major categories for cognitive domain from the lowest to the highest: knowledge, comprehension, application, analysis, synthesis, and evaluation. Anderson and Krathwohl [16] revised the original Bloom's taxonomy and presented a hierarchy of six thinking skill levels, with remembering, understanding, and applying falling under lower-order thinking skills and analyzing, evaluating, and creating falling under higher-order thinking skills. Bloom's Taxonomy is crucial for measuring higher- and lower-level cognitive demand in assessments. On the other hand, DoK is another instrumental perspective of cognitive complexity, which considers the content assessed in a test item and the expected depth of content understanding on the side of the learner [14]. Therefore, the researchers decided to use the elements of the CRM, RBT, and DoK frameworks to address the purposes of the study.

Previous studies have investigated the cognitive complexity of questions [19–24], the relationship between test complexity and testing method with test difficulty [12, 25–27], the effect of behavioral factors on tests [28–30], and many other relevant issues [31–40]. The present study targeted the cognitive demands of tests developed by high school teachers across school grades; therefore, studies in this area are reviewed. Moreover, as cognitive processes frameworks and taxonomies like DoK and RBT play a crucial role in measuring higher- and lower level cognitive skills in assessments [41], studies using such frameworks are covered. Despite the importance attached to the assessment of learners' cognitive skills and the role of cognitive complexity of test items on learners' test performance [42], there is a lack of evidence on how language teachers formulate the cognitive load of test items while constructing tests in the Iranian context. Moreover, the need for a better understanding of candidates' cognitive processing in language tests has been emphasized in the literature [43]. Discovering the cognitive demand of English tests can help identify how comprehensively and deeply English teachers target students' English cognitive abilities in their tests. Moreover, it will reveal whether the teachers balance high and low cognitive demands in their tests across different school grades. In the Iranian education system, English is one of the foreign languages offered to students. English language testing in Iran has a long tradition of discrete point tests that focus on language knowledge rather than the ability to perform in real situations [44]. Language performance evaluation in Iranian schools is conducted both locally and nationally. Locally, students are assessed annually through mid-term and final-term examinations. However, students also sit for a nationwide examination at the end of grade 12. According to Jahangard [45], listening and speaking skills receive little to no attention in Iranian schools and textbooks, and students are often evaluated on these skills during the term in the class and not in the final exam. Final exams focus on reading, writing, grammar, and vocabulary. While testing the cognitive level of students is considered one of the highest objectives of any assessment system [42], the literature review on language tests in the Iranian context shows that little attention is given to other aspects of language knowledge in the English language tests developed by teachers. Although cognitive demand levels have become important aspects of assessment, obtaining reliable test items on cognitive complexity is challenging. Informed with the need for the adjustment of the cognitive demand levels of tests with the academic stages of the students [4], and the fact that tests for different levels of education should cover different cognitive levels to accommodate different capabilities of learners, the present study attempted to map the levels of cognitive demand of the English language tests developed by Iranian high school English teachers, non-native to the English language, across three grades.

2. Literature Review

Testing the cognitive level of students is among the highest objectives of assessment systems [42]. Studies in the area of testing and assessment have approached the issue of cognitive load or complexity variously. Golshan and Rezaee [46], for example, analyze and codify the cognitive levels of two nationwide English final exams based on Bloom's Taxonomy. The results of the study showed that low cognitive level questions (knowledge, comprehension, and application) were prominent. Similarly, Köksal and Ulum [41] analyze exam questions in specific Turkish universities, reporting that exam papers did not settle on higher cognitive skills. Chandio et al. [47] also analyze the incorporation of Bloom's Taxonomy in assessment, learning, and teaching. The study revealed that the questions focused on the lower degrees of learning like remembering, understanding, and applying, which promote memorization, and there was less focus on the higher domains of learning such as analyzing, evaluation, and creating. Setiyana and Muna [48] investigate the application of Bloom's Taxonomy in test items at a university in Indonesia. They collected 471 test items and used a rubric on verb usage of Bloom's Taxonomy to identify certain cognitive skills in items. The results of the study revealed that most of the test items were from the "remember" level, and the "understand" level was the second highest, followed by the "apply" level, and the "analyze" level. The "evaluate" and "create" levels were ignored in the test items. Fitri et al. [49] analyzed the levels of questions used in reading texts based on Bloom's Revised Taxonomy,

reporting that only three levels of Bloom's Taxonomy were found in the test, and almost half of them were in the form of the first level (remembering). Al-Massarweh [50] studies the extent to which the final exam questions for students from sixth to eleventh in schools of the Ministry of Education in Jordan matched the levels of DoK. It was discovered that the tests were at the level of remembering, and hardly tapped extended thinking. It was reported that, unlike essay writing questions, objective exams achieved the depth of knowledge in all its dimensions. It was also reported that the dimensions of depth of knowledge (remembering, concepts, and skills, strategic thinking, and extended thinking) differed in terms of the subject type. It was also reported that there were differences between the tests developed for grades sixth, seventh, and eighth on the one hand and the eleventh grade on the other hand.

3. Research Methodology

3.1. Research design

The present study required the identification and codification of the components of RBT, DoK, and CRM frameworks (remembering, understanding, applying, analyzing, evaluating, and creating) in a set of sixty high school English exam questions. To achieve the purposes of the study, the quantitative content analysis method was used. The method, which involves the systematic coding and quantification of content, is used to determine the presence of certain themes or concepts within given written, visual, or oral content data [51]. The study also employed a descriptive quantitative research method to examine the frequencies of the components in the tests across the three grades. This study was extracted from part of a Master's degree thesis approved at Payame-Noor University of Guilan Province, Rasht on 17 – 4 – 2021, under Student No. 980007716. All ethical considerations were observed, and the required informed consent was obtained from the participants in the present study.

3.2. Participants

The participants were three male English teachers aged between 24 and 42 years. The English teachers were the coders, and a convenient sampling method was used to select them. They had at least five years of experience in teaching English and developing tests. They used a valid coding scheme developed based on RBT, DoK and CRM frameworks. The coders separately analyzed the tests using the coding scheme.

3.2.1. Instruments

A coding scheme, which was based on RBT, DoK, and CRM frameworks, was developed and validated in the study. The frameworks provide the standard for exploring the cognitive demand of test items [18]; studies conducted in the areas of learners' cognitive processing and language testing have used such frameworks as they enable teachers to design well-balanced tests with different cognitive skills and guide them in the order in which they should teach students to reach higher levels of thinking. The coding scheme was developed in a number of steps: (1) the scrutiny of the RBT, DoK, and CRM frameworks and related concepts in the literature, (2) the extraction of the verbs, statements, and questions denoting each of the components of the RBT, DoK, and CRM frameworks, (3) formulation of the components of the RBT, DoK, and CRM frameworks into a coding scheme, (4) the piloting of the coding scheme on a sample of tests with the participation of 3 coders, (5) the collection of coders' comments on the weaknesses of the scheme, (6) the revision and modification of the scheme based on the coders' comments, and (7) the coders' approval of the coding scheme. The scheme contained the six components of the RBT across the four levels of DoK as shown in CRM. For feasibility and administrative purposes, all manifestations of each of the cognitive levels of the RBT across the DoK framework and CRM matrix were named under the given RBT level. For example, all manifestations of understanding (as a cognitive level of RBT) in the DoK levels 1 (Recall and Reproduction), 2 (Skills and Concepts), 3 (Strategic Thinking or Reasoning), and 4 (Extended Thinking) were labeled as solely understanding. Next, the cognitive levels of remembering, understanding, and applying (DoK1 and 2 levels) were assigned as lower-cognitive demand levels; and analyzing, evaluating, and creating were targeted (DoK3 and DoK 4 levels) as higher-cognitive demand levels [16]. To help and secure the accuracy and appropriateness of the association between a test item and the cognitive levels, verbs, statements, and questions denoting the meaning and concept of each of the cognitive levels were appended to the coding scheme.

3.2.2. Procedure

To collect the data, the researchers visited six high schools to collect exam questions. Twelve school teachers and six principals were explained about the purpose of the study, and permission was obtained from them to analyze the exam questions. They were told that only the test items on the exam questions would be studied and analyzed and all other information on the questions would remain confidential. The exam questions were obtained and then grouped in terms of grades. Then, three coders were recruited and trained by experienced experts teaching at the university level on how to analyze the exam questions. The training included a module on the components of the RBT, DoK, and CRM frameworks so that the coders master relevant concepts and components. In this phase, they were trained on the theoretical concepts underlying each framework and were familiarized with the verbs, statements, and questions denoting each of the levels of the RBT, DoK, and CRM frameworks. The coders were also trained on how to use the coding scheme to analyze the test items. Next, they were asked to use the scheme to code sample exam questions to remove any ambiguities and potential problems in coding the questions. The coders performed the rating independently and then a focus group of the coders and trainers was formed, needed consultation and advice were made, and the coding phase started subsequently. In the coding phase, the coders were provided with a parcel of questions which were grouped in terms of grades. They studied the test items on each of the exam questions and matched the items with cognitive levels on the scheme using sample verbs, questions, and statements. They reviewed all the items in each of the exam questions. They also indicated the test items, and the grade in which the test items had been used, the language skill(s) the test items targeted, and the total frequency of the test items associated with each of the cognitive levels. Each of the coders performed the coding independently; the codings were subsequently cross-checked to secure the reliability of the codings. In case of any inconsistencies, the coders were asked to convene to reach a consensus.

4. Results

The materials used for this study were questions and final exam papers developed by high school teachers. A pool of questions extracted from sixty exams from different high schools was created. Most of the tests contained four language components of vocabulary, grammar, writing, and reading. The total number of tests

reviewed over the eight years and the language skills incorporated in the tests are given in Table 1.

Table 1
The language skills reviewed

Grades Skills	Vocabulary	Grammar	Writing	Reading
Grade 12	304	195	147	215
Grade 11	250	183	124	169
Grade 10	287	232	137	173
Total	841	610	408	557

The two language skills of listening and speaking were not reviewed as they were not incorporated in all exams. Samples of the teacher-developed tests falling under each of the cognitive levels are given below.

1) Remembering:
Example 1: Match the words with their definitions.

1. The time after now. a. to protect someone or something from danger.
2. a person b. future
3. healthy c. human
4. defend d. strong and well

Example 2: Match the pictures with the following sentences.



. . . .

1) A book in which you record your thoughts or feelings or what has happened every day.
2) Try to avoid foods that contain a lot of fat.
3) We have to take care of elderly people.
4) He couldn't figure out what his mother was talking about.

Example 3: A. Match the words (1–4) with the definitions (a–e). There is one extra definition.

1. craftsman (. . . .) a. all people
2. humankind (. . . .) b. without taking any notice of
3. rarely (. . . .) c. one who makes beautiful things by hand
4. despite (. . . .) d. starting a short time ago
e. not very often

2) Understanding:
Example 1: A. Fill in the blanks using given words (a–e). There is one extra word.

a. prevent b. absolutely c. diet d. tilework e. ability

1. You're . . . . right – we can't all fit in one car.
2. The physical or mental power or skill to do something is known as . . . .
3. We hope to . . . . anything unpleasant from happening.
4. There are lots of . . . . in Sheikh Lotfollah Mosque.

Example 2: Unscramble the words to make a complete sentence.

1. any – solve – have – suggestions – do – the – you – problem – to?
. . . .
2. shopping – somebody – car – while – she – took – was – her.
. . . .

Example 3: Choose the best answer.

1. It could be cold and wet so . . . . some sensible clothes.
a. satisfy b. pack c. weave d. appreciate
2. Praying decreases stress and gives people a calm and . . . . life.
a. unsafe b. irregular c. balanced d. unhealthy

3) Applying:
Example 1: Fill in the blanks with correct form of the verbs.

1. She promised . . . . (not object) to his bad habit of . . . . (smoke) in the car.
2. Just avoid . . . . (make) unnecessary mistakes.

Example 2: Fill in the blanks with your own words.

1. When my brother saw his score in the final exam, ... into tears.
2. The problem was really difficult and I couldn't . . . . out its answer.

Example 3: Combine two sentences with (so, and, or, but).

1. I was feeling very tired. I want to bed very early.
2. It was an easy exam. I can't pass it.

4) Analyzing:
Example 1: One odd out. Circle the different words in each group.

1. a) lower b) develop c) improve d) boost
2. a) elementary b) advanced c) technical d) intermediate

Example 2: The pronoun "they" in thesecond paragraph refers to. . . . .
a. traditions b. customs c. immigrants d. cultures
Example 3: Read the following sentences and find the subjects, verbs, objects

1. The French built the church in the 13th century.
2. I picked up the wrong bag by mistake.
3. The baby laughed very loudly.

5) Evaluating:
Example 1: There is one error in each of the following sentences. Find and correct them.

1. look! It is a beautiful wooden small house.
2. All human are going to work together to have a beautiful planet.

Example 2: Which one is the main idea of the text?

a. Paying attention to physical health improves lifestyle.
b. To enjoy a better lifestyle, we need to develop healthy habits.
c. The most important thing to enjoy a good life is having emotional health.
d. The best factor to keep people healthy is having a light breakfast followed a light lunch.

Example 3: We can conclude from the passage that atoms move most slowly . . . .

a) in a frozen object
b) in a small object
c) in a still object
d) in a melting object.

6) Creating:
Example 1: Complete the sentences.

1. I like learning Chinese, but .....
2. You must study well, and .....

Example 2: Generate one question and then answer it.
Question: .....
Answer: .....
Example 3: Write a topic sentence for following paragraph.

Most of them have good taste. They are really delicious. Examples of fruits are apples, oranges, and bananas. Fruits have many vitamins and minerals, including vitamin C, vitamin E, and potassium. For these reasons, it is a good idea to eat lots of fruits.

To analyze the data, the frequency and percentages of the codes within and across the three grades were computed. The study investigated the cognitive demand of all tests developed by the school teachers. The results of the analyses are summarized in Table 2.

Table 2
Frequency and percentages of the cognitive demand levels

Cognitive levels: RBT DoK CRM	Frequency	Percentage
Remembering	651.00	26.95
Understanding	1099.00	45.49
Applying	412.00	17.05
Analyzing	146.00	6.04
Evaluating	74.00	3.06
Creating	34.00	1.41
Total	2416.00	100.00

As shown, the frequencies of the low cognitive demand levels of "remembering," "understanding," and "applying" constituted a large proportion (89.49%) of the questions. In contrast, the frequencies of "analyzing," "evaluating," and "creating," as the high cognitive demand levels constituted a small proportion (10.51%) of the whole package. Moreover, "understanding" was the most frequent cognitive level and creating was the least frequent one.

The study also investigated whether there were any significant differences in the cognitive demand levels of the tests across the grades. The results are given in Table 3 below.

As shown in the table, in all grades, the most frequent cognitive level was "understanding" and "creating" was the least frequent one. Remembering took the logical course and increased across grades from 25.81% in grade 10 to 26.45% in grade 11 and 28.46 in grade 12. Understanding, on the other hand, took a descending order across the grades. Applying just fluctuated as it increased in grade 10, dropped in grade 11, and increased in grade 12 to roughly match up with the frequency in grade 10. Analyzing had its peak in grade 10 and dropped in grades 11 and 12 where it roughly leveled. Evaluating, underwent fluctuation as it declined from 2.90 in grade 10 to 1.79 in grade 11; then, it reached its peak in grade 12 (2.56). Creating had a steady logical rise from 0.48 % in grade 10 to 1.10 in grade 11 and 2.56 in grade 12.

The results indicated that the tests mostly contained low cognitive levels. The proportions of the understanding, analyzing, and evaluating in the tests developed for grade 10 were more than those of grade 11. However, creating, applying, and remembering were more frequent in the tests designed for grade 11. The frequency of "remembering," evaluating," and "creating" in the tests developed for grade 12 was higher than the tests in grades 10 and 11.

The distribution of the low-cognitive demand levels (remembering, understanding, and applying) and the high-cognitive demand levels (analyzing, evaluating, and creating) was calculated across the three grades. The results are summarized in Table 4.

Table 3
The cognitive demand levels across the grades

Cognitive levels	Grade 10		Grade 11		Grade 12	
	Frequency	Percentage	Frequency	Percentage	frequency	percentage
Remembering	214.00	25.81	192.00	26.45	245.00	28.46
Understanding	394.00	47.53	336.00	46.28	369.00	42.86
Applying	137.00	16.53	135.00	18.60	140.00	16.26
Analyzing	56.00	6.76	42.00	5.79	48.00	5.57
Evaluating	24.00	2.90	13.00	1.79	37.00	4.30
Creating	4.00	.48	8.00	1.10	22.00	2.56
Total	829	100	726	100	861	100

Table 4
High- and low-demanding cognitive levels across grades

Grades	Low Demand		High Demand		
	Frequency	Percentage	Frequency	Percentage	Total
Ten	745.00	89.86	84.00	10.13	829
Eleven	663.00	91.32	63.00	8.67	726
Twelve	754.00	87.57	107.00	12.42	861
Total	2162	89.48	254	10.51	2416

The results showed that out of 2416 codes, more than two-thirds manifested low cognitive demand levels (89.48%), and only a small proportion represented high cognitive levels (10.51%).

To determine if there was a balanced representation of low and high demand levels in the tests across the three grades and eight academic years, a Chi-square test was performed. The results are presented in Table 5.

Table 5
Chi-square test for low and high cognitive demand levels across grades and years

	Across grades
Chi-Square	1521.709
Df	5
Asymp. Sig.	.000

The results revealed a statistically significant difference among the observed frequencies of low and high cognitive demand tests across the three grades (X2 (5, n = 2416) = 1521.709, $p < 0.05$). The results indicated that the tests did not represent a balanced pattern of low and high cognitive demand levels across the three grades.

5. Discussion

This study explored the cognitive demand levels of the English tests developed by high school English teachers across three grades. The findings showed that low-demanding cognitive levels with the dominance of "understanding" were more frequent in the questions. The results also indicated that the tests did not engage the students' higher cognitive levels adequately. It was, however, discovered that the tests developed for grade 12, which is the final grade in the high school, included high cognitive levels (analyze, evaluate, and create) compared to the English tests for the two other grades. The results of this study were similar to a study by Setiyana and Muna [48], where most of the test items were at the remembering (45%) and understanding (42%) levels. Similarly, in this study, remembering and understanding were the predominant levels, with the remembering level representing 26.95% and the understanding level representing 45.49% of the tests. Likewise, Nakkam and Khamoja [43], who evaluate the reading comprehension questions in Moroccan ELT textbooks, reported that a few questions in the textbooks were at high cognitive levels with analyzing, evaluating, and creating levels representing only 6.04%, 3.06%, and 1.41% of the questions, respectively. Similarly, in the present study, high cognitive levels were almost absent in the items that assessed the students' reading; the test items which assessed the students' reading comprehension were at the remembering level. In the same vein, Nakkam and Khamoja [52] analyze reading tests in the "Bahasa Inggris" textbook, reporting that the remembering level represented 26.95%, the understanding level represented 45.49%, and applying represented 17.05% in the lower-order levels of thinking, while the analyzing level represented 6.04%, and the evaluating and creating levels represented only 3.06% and 1.41%, respectively. The reason for this variation might be the difference between the materials in the two studies. The present study analyzed exam papers with four language skills, while Arvianto et al. [53] analyzed textbook reading exercises. Questions on different language skills such as vocabulary or grammar may focus on different cognitive levels. The findings of this study also differed from those by Assaly and Smadi [54], which evaluates the cognitive levels of questions in a master class textbook. They found that 40% of the questions focused on high cognitive levels, while 52% were at the comprehension level, 3.7% were at the remembering level, and 6% of the questions were at the applying level. The discrepancy may be due to the English language proficiency level of the master's class textbook, where the author emphasizes higher-order thinking skills.

The findings in the present study also showed a statistically significant difference between the representation of low- and high-demanding cognitive levels in exam questions developed by English teachers. The results showed that the understanding level had the highest frequency among lower-order levels, with 1099 items in all English tests. Remembering was present in 651 items while applying had the lowest frequency (412) among lower-order skills. The findings in the present study are in line with the results of the study by Golshan and Rezaee [46], who analyzed and codified the cognitive levels of two nationwide English final exams in the Iranian context based on Bloom's Taxonomy, showing that low cognitive level questions (knowledge, comprehension, and application) were prominent. Similarly, likewise, in the Turkish context, Köksal and Ulum [41] analyzed exam questions in certain Turkish universities, reporting that exam papers did not settle on higher cognitive skills.

The findings of the present study also showed that the frequencies of higher-order levels of analyzing, evaluating, and creating were 146, 74, and 34, respectively. The chi-square test across the three grades revealed that the proportion of lower-order thinking skills was statistically different from that of higher-order thinking skills. The results confirmed the findings of Chandio et al. [47], which indicate that questions developed by teachers in Pakistan focused on lower-order skills of remembering, understanding, and applying, promoting memorization, with less attention to higher-order domains of learning. The preference for lower-order questions may stem from their minimal preparation requirements in writing and evaluating, or from traditional methodologies that emphasize lower cognitive levels such as information recall. The cause of this similarity could be a result of the fact that in the educational system of both countries and final exam papers developed by teachers, the major emphasis is on high school textbook exercises. Teachers may be influenced by traditional systems of teaching and testing and focus on recalling or memorizing information [55, 56]. Likewise, Tangsakul et al. [57] used Bloom's revised taxonomy to analyze reading comprehension questions in Team Up in English 1–3 and Grade 9 English O-NET Tests. The results showed that the levels of reading comprehension questions were at lower levels, with a significant difference between the representation of higher-order and lower-order thinking levels of Bloom's revised taxonomy in the questions. The majority of the questions in their study were at the remembering level (71.15%). In the same vein, Razmjoo and Kazempourfard [58] investigated the representation of cognitive levels in Interchange course books and found a significant difference between the books in their inclusion of different levels. Similar to the present study, lower-order thinking skills were the most dominant learning levels in these books, with less focus on the higher domains, and the content of the books revolved around lower-order thinking skills. Coursebook developers and test designers should strive to develop exercises and activities that include higher-order thinking skills. The results of this study are also different from those of Qasrawi and BeniAbdelrahman's [59] study, which analyzes the frequency of lower and higher thinking processing in tests and

exercises included in Unlock English textbooks. The results showed that Unlock textbooks enhance both higher- and lower-order thinking levels. However, in the present study, there was a statistically significant difference between the representation of higher-order and lower-order thinking levels of Bloom's revised taxonomy in final exam questions. The differences between the two studies suggest that Iranian test developers and teachers may not have been adequately trained in designing higher-order questions to develop test-takers' higher cognitive skills.

6. Conclusion

This study aimed to map CDL of English language tests developed by high school non-native English teachers across three grades (10 through 12). The results showed that the low-cognitive demand components (remembering, understanding, and applying) were more frequent than the high-cognitive demand components (analyzing, evaluating, and creating). Few questions were at the high-cognitive demand level, particularly the "creating" component, which was rarely used in the English tests. The "understanding" component was dominant in the tests. The results also showed that the level of cognitive demand of the tests differed across school grades with grade 12 containing the highest number of tests requiring high-cognitive demand components. The study also discovered that most vocabulary items in high school final exams were measured at the remembering level. The majority of the reading comprehension questions focused on remembering, while most grammar tests focused on understanding. As for the writing skill, the understanding level had the highest proportion. The findings run contrary to the need for the students to be able not only to remember or understand what they have learned but also to analyze, evaluate, and synthesize the facts to solve real-life problems. The study concludes that the tests predominantly required low-cognitive demanding components and were not developed based on a measure of cognitive demand involving both lower-order and higher-order thinking processes. Therefore, a revision in test design may be necessary to balance and regulate the cognitive demand of the tests. Teachers and test designers may still be under the influence of traditional testing systems placing emphasis on lower-order learning objectives such as recalling or memorizing information. Teachers seem to need training on the cognitive demand of tests and how it differs from linguistic difficulty; they also need training on techniques for the incorporation of different levels of cognitive demand in tests to tap students' higher levels of thinking. Teachers and test developers are recommended to use a cognitive demand guide specifying the underlying nature of tasks required in the tests, the complexity of the expressions and information presented in the test, and the choice of test item stem format and response format for different school grades to regulate the cognitive demand of test items. While the complexity of item stems and response options in a test item can affect its difficulty [8], the cognitive demand of a test item is not necessarily associated with its difficulty. For example, two vocabulary test items may have the same cognitive demand, requiring either lower or higher-order thinking skills, but their difficulty level for a test taker may be different. Therefore, when developing tests in terms of levels of cognitive complexity, the test developer should consider what students have learned from the lessons and then choose a topic and make questions based on each level. To put it another way, the cognitive complexity of a language test item is not solely determined by how difficult the linguistic knowledge required in the item is; rather, it involves the cognitive processes that the item requires. As such, any given language test item, whether requiring difficult or simple linguistic knowledge, can be developed in a way that involves either a simple or complex cognitive load. It is up to the test developer to calibrate the cognitive complexity of a language test item. For instance, a test item involving difficult linguistic knowledge may involve simple cognitive processes, while a test item involving simple linguistic knowledge may involve complex cognitive processes. Cognitive demand levels of tests and students' academic levels need to be adjusted and tests for different education levels should cover different cognitive levels.

The findings have practical implications for the development of tests that measure not only students' knowledge of the English language but also their capacity to deal with levels of cognitive complexity incorporated in the tests. The study also has practical implications for teachers and authorities who make policies for testing English language skills in education systems:

1) They can use the results of this research to adapt the existing testing system based on the cognitive and linguistic requirements of students.
2) Education policymakers can formulate standards requiring considerations for cognitive demand and load in the development of school tests as well as high-stake national tests.

Based on the findings and results of this study, some suggestions for further research can be made. The following suggestions deserve further research:

1) A different study can be conducted to investigate the representation of Bloom's revised taxonomy in the tests developed by junior high school teachers.
2) Further studies may investigate whether Bloom's revised taxonomy is reflected in the English high school textbooks.
3) Studies must be conducted to find out whether test designers and teachers from both government and private schools are familiar with the Bloom's revised taxonomy.
4) Finally, a similar study can investigate the representation of Bloom's revised taxonomy in the tests developed by other high school teachers.

Acknowledgement

The authors are grateful to all those who contributed to the completion of the study.

Ethical Statement

This study was conducted as part of a Master's degree thesis and was reviewed and approved by the relevant institutional authority at Payame-Noor University, Guilan Province, Rasht. All ethical considerations were observed, and the required informed consent was obtained from the participants in the present study.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

The data that support this work are available upon reasonable request from the corresponding author.

Author Contribution Statement

Ali Ghezelbash: Validation, Formal analysis, investigation, Data curation, Writing – original draft, Project administration. **Iman Alizadeh:** Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision. **Hoda Divsar:** Methodology, Validation, Investigation, Resources, Data curation, Project administration.

References

[1] Shohamy, E. (1996). Language testing: Matching assessment procedures with language knowledge. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 143–160). Springer. https://doi.org/10.1007/978-94-011-0657-3_6

[2] Jang, E. E. (2017). Cognitive aspects of language assessment. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment. Encyclopedia of language and education*. Springer. https://doi.org/10.1007/978-3-319-02261-1_11

[3] Bloom, B. S. (1956). *The taxonomy of educational objectives, the classification of education goals, handbook I: Cognitive domain*. USA: David McKay Company.

[4] Wakhata, R., Mutarutinya, V., & Balimuttajjo, S. (2023). Exploring the impact of Stein et al.'s levels of cognitive demand in supporting students' mathematics heuristic problem-solving abilities. *Frontiers in Education*, *8*, 949988. https://doi.org/10.3389/feduc.2023.949988

[5] Stein, M. K., Smith, M. S., Henningsen, M., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development*. USA: Teacher College.

[6] Ehrich, J. F., Howard, S. J., Bokosmaty, S., & Woodcock, S. (2021). An item response modeling approach to cognitive load measurement. *Frontiers in Education*, *6*, 648324. https://doi.org/10.3389/feduc.2021.648324

[7] 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

[8] Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. USA: Routledge.

[9] Kan, A., Bulut, O., & Cormier, D. C. (2018). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment*, *24*(1), 13–32. https://doi.org/10.1080/10627197.2018.1545569

[10] Berk, R. A. (1996). *A consumer's guide to multiple-choice item formats that measure complex cognitive outcomes*. Retrieved from: http://images.pearsonassessments.com/images/NES_Publications/1996_12Berk_368_1.pdf

[11] 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

[12] 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), 13. https://doi.org/10.1186/s40468-022-00163-8

[13] 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*, 2033–2053. https://doi.org/10.1007/s11423-020-09740-1

[14] Hess, K. K. (2010). *Applying Webb's depth-of-knowledge (DOK) levels in science*. Retrieved from: https://www.veronaschools.org/cms/lib/NJ01001379/Centricity/Domain/17/DOKscience_KH11.pdf

[15] Webb, N. L. (2002). *Depth-of-knowledge levels for four content areas*. Retrieved from: http://ossucurr.pbworks.com/w/file/fetch/49691156/norm%20web%20dok%20by%20subject%20area.pdf

[16] Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. USA: Addison Wesley Longman, Inc.

[17] Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, *34*(5), 348–364. https://doi.org/10.1177/0146621609349801

[18] Das, S., Das Mandal, S. K., & Basu, A. (2020). Identification of cognitive learning complexity of assessment questions using multi-class text classification. *Contemporary Educational Technology*, *12*(2), ep275. https://doi.org/10.30935/cedtech/8341

[19] Hayes, H., Demeter, M., Morris, J. G., & Trajkovski, G. (2021). Transparency of cognitive complexity in performance assessments: A validity study. *The Journal of Competency-Based Education*, *6*(2), e1244. https://doi.org/10.1002/cbe2.1244

[20] Moghaddam, A. M., Azizmalayeri, F., & Bayat, A. (2022). Impact of cognitive complexity of tasks on EFL learners' individual and collaborative written performance. *Journal of Language and Translation*, *4*(12), 131–142.

[21] İlhan, M., Boztunc Öztürk, N., & Şahin, M. G. (2020). The effect of the item's type and cognitive level on its difficulty index: The sample of TIMSS 2015. *Participatory Educational Research*, *7*(2), 47–59. https://doi.org/10.17275/per.20.19.7.2

[22] Yang, Q., Chang, S., Hwang, G., & Zou, D. (2020). Balancing cognitive complexity and gaming level: Effects of a cognitive complexity-based competition game on EFL students' English vocabulary learning performance, anxiety and behaviors. *Computers & Education*, *148*, 103808.

[23] Muhayimana, T., Kwizera, L., & Nyirahabimana, M. R. (2022). Using Bloom's taxonomy to evaluate the cognitive levels of Primary Leaving English Exam questions in Rwandan schools. *Curriculum Perspectives*, *42*(1), 51–63. https://doi.org/10.1007/s41297-021-00156-2

[24] Baghaei, S., Bagheri, M. S., & Yamini, M. (2021). Learning objectives of IELTS listening and reading tests: Focusing on revised Bloom's Taxonomy. *Research in English Language Pedagogy*, *9*(1), 182–199. https://doi.org/10.30486/relp.2021.1916940.1244

[25] Kumaar, J. A. M., & Srinivas, R. (2022). Do linguistic proficiency and task cognitive complexity affect vocabulary retention of ESL learners? *Journal of Positive School Psychology*, *6*(4), 2725–2737.

[26] Karami, M., & Karami, Z. (2023). Increasing the level of cognitive complexity of learners in reading and writing through the model of teaching based on depth of knowledge. *Advances in Cognitive Sciences*, *25*(2), 144–157.

[27] Krell, M., Khan, S., & van Driel, J. (2021). Analyzing cognitive demands of a scientific reasoning test using the linear logistic test model. *Education Sciences*, *11*(9), 472. https://doi.org/10.3390/educsci11090472

[28] Azamnouri, N., Pishghadam, R., & Naji Meidani, E. (2020). The role of emotioncy in cognitive load and sentence com-

prehension of language learners. *Issues in Language Teaching*, *9*(1), 29–55. https://doi.org/10.22054/ilt.2020.51543.485

[29] Demir, B. (2022). Effect of anxiety on cognitive load in foreign language oral tests. *Baskent University Journal of Education*, *9*(1), 102–111.

[30] Woodard, S. R., Chan, L., & Conway III, L. G. (2021). In search of the cognitively complex person: Is there a meaningful trait component of cognitive complexity? *Personality and Social Psychology Review*, *25*(2), 95–129. https://doi.org/10.1177/1088868320972299

[31] Prasad, G. (2021). Evaluating student performance based on Bloom's taxonomy levels. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1797/1/012063

[32] Jumaniyozova, F. T. (2021). The importance of five main principles of language assessment in designing the language tests in Uzbek universities. *Current Research Journal of Philological Sciences*, *2*(6), 1–7. https://doi.org/10.37547/philological-crjps-02-06-01

[33] Arslan, R. Ş., & Üçok-Atasoy, M. (2020). An investigation into EFL teachers' assessment of young learners of English: Does practice match the policy? *International Online Journal of Education and Teaching*, *7*(2), 468–484.

[34] Moslehi, S., & Razmjoo, S. A. (2021). On the representation of Bloom's revised Taxonomy in TOEFL iBT and IELTS academic. *Journal of English Language Teaching and Learning*, *13*(28), 173–200. https://doi.org/10.22034/ELT.2021.46190.2391

[35] Mortazavizadeh, G., Golshan, M., & Rezai, M. J. (2022). Metacognitive strategies awareness and use and reading comprehension of Iranian EFL university students: A structural equation modeling approach. *Iranian Evolutionary Educational Psychology Journal*, *4*(3), 516–535. https://doi.org/10.52547/ieepj.4.3.516

[36] Baghaei, S., Bagheri, M. S., & Yamini, M. (2020). Analysis of IELTS and TOEFL reading and listening tests in terms of Revised Bloom's Taxonomy. *Cogent Education*, *7*(1), 1720939. https://doi.org/10.1080/2331186X.2020.1720939

[37] Rahmawati, Y., Afrizal, A., Dwi Astari, D., Mardiah, A., Budi Utami, D., & Muhab, S. (2021). The integration of dilemmas stories with STEM-project-based learning: Analyzing students' thinking skills using Hess' cognitive rigor matrix. *Journal of Technology and Science Education*, *11*(2), 419–439. https://doi.org/10.3926/jotse.1292

[38] Haataja, E. S., Tolvanen, A., Vilppu, H., Kallio, M., Peltonen, J., & Metsäpelto, R. L. (2023). Measuring higher-order cognitive skills with multiple choice questions – Potentials and pitfalls of Finnish teacher education entrance. *Teaching and Teacher Education*, *122*, 103943. https://doi.org/10.1016/j.tate.2022.103943

[39] Marpaung, M. B., & Sihombing, H. B. M. (2021). The complexity of reading comprehension processes: A review from reader, text, and cognitive process perspectives. *Jurnal Littera: Fakultas Sastra Darma Agung*, *1*(1), 53–64.

[40] 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

[41] Köksal, D., & Ulum, Ö. G. (2018). Language assessment through Bloom's Taxonomy. *Journal of Language and Linguistic Studies*, *14*(2), 76–88.

[42] Sivaraman, S. I., & Krishna, D. (2015). Blooms taxonomy – Application in exam papers assessment. *Chemical Engineering (VITU)*, *12*(12), 32.

[43] Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. B. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge University Press.

[44] Razavipour, K., & Rezagah, K. (2018). Language assessment in the new English curriculum in Iran: Managerial, institutional, and professional barriers. *Language Testing in Asia*, *8*, 1–18. https://doi.org/10.1186/s40468-018-0061-8

[45] Jahangard, A. (2007). Evaluation of EFL materials taught at Iranian public high schools. *The Asian EFL Journal*, *9*(2), 130–150.

[46] Golshan, M., & Rezaee, M. (2016). Investigating the cognitive levels of English final exams based on Bloom's Taxonomy. *International Journal of Educational Investigations*, *3*(4), 57–68.

[47] Chandio, M. T., Pandhiani, S. M., & Iqbal, R. (2017). Bloom's Taxonomy: Improving assessment and teaching-learning process. *Journal of Education and Educational Development*, *3*(2), 203–221.

[48] Setiyana, R., & Muna, H. (2019). Evaluating Bloom's Taxonomy in test items. In *Proceedings of the 2nd English Education International Conference*, *2*, 79–83.

[49] Fitri, Y. K., Syarif, H., & Anwar, D. (2019). The levels of questions used on reading test with regards to Revised Bloom's Taxonomy. *Jurnal Ta'dib*, *22*(1), 27–32.

[50] Al-Massarweh, S. (2023). The extent of knowledge of achieving the final exam questions for the sixth to eleventh grades to the levels depth of knowledge "DOK" Webb in Jordan. *Modern Applied Science*, *17*(2), 1–70.

[51] Huxley, K. (2020). *Content analysis, quantitative*. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *SAGE research methods foundations*. SAGE Publications. https://doi.org/10.4135/9781526421036880564

[52] Nakkam, J., & Khamoja, A. (2020). Evaluation of reading comprehension questions in Moroccan ELT textbooks. *International Journal for Innovation Education and Research*, *8*(6), 315–324. https://doi.org/10.31686/ijier.vol8.iss6.2420

[53] Arvianto, Z., Faridi, A., & Linggar Bharati, D. (2017). The compatibility of reading exercises with Bloom Revised Taxonomy and 2013 curriculum (A case of English textbook entitled Bahasa Inggris for grade XI published by Department of National Education 2014). *English Education Journal*, *6*(2), 42–51.

[54] Assaly, I. R., & Smadi, O. M. (2015). Using Bloom's Taxonomy to evaluate the cognitive levels of master class textbook's questions. *English Language Teaching*, *8*(5), 100–110.

[55] Divsar, H. (2020). Bloom's Revised Taxonomy across the learning objectives in English translation curricula. *Iranian Journal of English for Academic Purposes*, *8*(4), 136–152.

[56] Hassan, S. (2023). Analysis based on Bloom's Taxonomy: Pakistan's Federal English curriculum and examination content for matric. *Journal of Positive School Psychology*, *7*(1), 1324–1338.

[57] Tangsakul, P., Kijpoonphol, W., Linh, N. D., & Kimura, L. N. (2017). Using Bloom's Revised Taxonomy to analyze reading comprehension questions in team up in English 1–3 and grade 9 English o-net tests. *International Journal of Research*,

5(7), 31–41. https://doi.org/10.29121/granthaalayah.v5.i7.2017.2106

[58] Razmjoo, S. A., & Kazempourfard, E. (2012). On the Representation of Bloom's Revised Taxonomy in interchange coursebooks. *The Journal of Teaching Language Skills*, *4*(1), 171–204.

[59] Qasrawi, R., & BeniAbdelrahman, A. (2020). The higher and lower-order thinking skills (HOTS and LOTS) in unlock English textbooks (1st and 2nd Editions) based on Bloom's Taxonomy: An analysis study. *International Online Journal of Education and Teaching*, *7*(3), 744–758.

How to Cite: Ghezelbash, A., Alizadeh, I., & Divsar, H. (2025). Discovering Levels of Cognitive Demand in Tests Developed by Iranian High School English Teachers Across Grades. *International Journal of Changes in Education*. https://doi.org/10.47852/bonviewIJCE52024077