

Design and Development of a Diagnostic Reading Comprehension Test

Rabia BÖREKÇİ¹ & Aysun YAVUZ²

¹PhD, Ministry of National Education, TURKEY
rabia_borekci@hotmail.com
<https://orcid.org/0000-0001-5678-7365>

²Prof. Dr., Çanakkale Onsekiz Mart University, Çanakkale, TURKEY
yavuzaysun@comu.edu.tr
<https://orcid.org/0000-0001-6838-8695>

Abstract

This study reports the development and validation of a diagnostic reading comprehension test designed for Turkish 8th-grade EFL learners at the CEFR A2 level. The test was constructed following established test-development procedures, including construct definition, text selection, expert review, piloting, and statistical analysis. Five reading texts and 52 items targeting literal, reorganization, and inferential comprehension were developed, along with a mixed scoring system combining dichotomous scoring and weighted marking for sequencing items. Readability analyses (Flesch–Kincaid levels 4.3–6.3) confirmed that the texts were appropriate for the target proficiency level. Content and construct validity were examined through expert review and student reflections, while internal consistency reliability was assessed using KR-20, yielding strong coefficients (0.91 in the pilot; 0.81 in the revised version). Qualitative data from expert reviewers indicated strong alignment with CEFR A2 descriptors and MoNE curriculum outcomes, though some items displayed vocabulary-heavy or grammar-oriented characteristics that may introduce construct-irrelevant variance. Student reflections highlighted the overall clarity and accessibility of the test but echoed concerns about vocabulary load in certain sections. Overall, the findings suggest that the diagnostic test is psychometrically sound and pedagogically useful for identifying learners' reading comprehension strengths and weaknesses. Recommendations for refinement and future large-scale validation studies are also discussed.

Keywords: Reading comprehension, EFL learners, test development, validity, reliability

Tanısal Okuma Anlama Testinin Tasarımı ve Geliştirilmesi

Özet

Bu çalışma; 8. sınıf seviyesindeki, CEFR A2 düzeyinde yabancı dil öğrenimi gören Türk öğrenciler için tasarlanmış bir tanısal okuma anlama testinin geliştirilmesi ve geçerlilik çalışmalarını rapor etmektedir. Test, yapı tanımı, metin seçimi, uzman değerlendirilmesi, pilot uygulama ve istatistiksel analiz gibi yerleşik test geliştirme adımları izlenerek oluşturulmuştur. Beş okuma metni ve literal, yeniden düzenleme ve çıkarımsal anlama becerilerini hedefleyen 52 madde geliştirilmiş; sıralama maddeleri için ikili puanlama ve ağırlıklı puanlamayı birleştiren karma bir değerlendirme sistemi kullanılmıştır. Okunabilirlik analizleri (Flesch–Kincaid 4,3–6,3) metinlerin hedef düzeye uygun olduğunu göstermiştir. İçerik ve yapı geçerliliği uzman değerlendirmeleri ve öğrenci yansımalarıyla incelenmiş; iç tutarlılık güvenirliği KR-20 ile test edilmiş ve güçlü katsayılar elde edilmiştir (pilot: 0,91; revize: 0,81). Uzman geribildirimleri CEFR A2 tanımlayıcıları ve MEB müfredatıyla yüksek uyum göstermiş; ancak bazı kelime ağırlıklı veya gramer odaklı maddelerin yapı ile ilgisiz varyansa yol açabileceği belirtilmiştir. Öğrenci yansımaları testin genel olarak açık ve erişilebilir olduğunu, ancak bazı bölümlerde kelime yükünün zorluğu artırdığını göstermiştir. Bulgular, geliştirilen testin psikometrik açıdan sağlam ve öğrenenlerin okuduğunu anlama becerilerini tanılama açısından pedagojik olarak kullanışlı olduğunu ortaya koymaktadır. Çalışma, ayrıca iyileştirme önerileri ve gelecekte yapılacak geniş ölçekli geçerlilik araştırmalarına ilişkin öneriler sunmaktadır.

Anahtar Kelimeler: Okuduğunu anlama, yabancı dil olarak İngilizce öğrenenler, test geliştirme, geçerlik, güvenirlik

1. Introduction

In today's information-rich society, the ability to identify and prioritize relevant information from vast amounts of content is essential. This skill underpins the construction of meaning from text, which is critical for an effective language learning process (Sulfa et al., 2023). Thus, reading comprehension is a vital life skill that enables individuals to understand and interpret written texts. It serves as a foundation for academic success by directly influencing students' capacity to engage with and process information (Kendeou et al., 2009). Accordingly, assessing reading comprehension is a key concern in educational settings, as it encompasses multiple levels of understanding.

As the complexity of reading comprehension extends beyond simply decoding information, a structured model is needed to clarify the skills involved in reading comprehension. Barrett's Taxonomy (1976) responds to this need by detailing five levels of comprehension, each contributing to a fuller understanding of how learners construct meaning from texts. At the literal level, readers recall and interpret explicitly stated information, which serves as a foundational skill for higher-level comprehension. (Wahyuni, 2021). The reorganization level requires synthesizing ideas across the text by comparing, contrasting, or summarizing, thereby deepening engagement (Kardoğan & Geçgel, 2024). Inferential comprehension involves integrating prior knowledge with textual cues to interpret implicit meanings, draw conclusions, or identify themes, and is closely associated with overall EFL proficiency (Kor et al., 2014). At the evaluation level, readers critically assess validity, reliability, and bias, fostering critical literacy and reflective engagement (Hamdollahi et al., 2024). Finally, the appreciation level emphasizes aesthetic and emotional responses, enabling personal connections and deeper enjoyment of the text (Simşek & Direkçi, 2023). Given its comprehensive scope, Barrett's Taxonomy provided the theoretical foundation for this study, enabling reading comprehension to be systematically defined and operationalized into testable sub-skills suitable for A2-level learners.

Barrett's Taxonomy of Reading Comprehension, although introduced in 1976, remains a coherent and widely applied framework for analyzing the cognitive processes involved in reading, particularly in EFL contexts. Its continued relevance is demonstrated by its frequent use in recent research (Junaidi et al., 2024; Koparan, 2025) and its conceptual alignment with contemporary assessment frameworks such as the CEFR, PIRLS, and PISA, all of which distinguish between surface-level comprehension and deeper interpretive skills. The hierarchical structure of the taxonomy clearly articulates how comprehension develops from basic information retrieval to more complex interpretive judgments, making it a dependable model for defining measurable reading skills.

In this study, Barrett's Taxonomy was adopted because it aligns closely with the reading behaviors expected of CEFR A2 learners and with the Ministry of National Education (MoNE, 2018) curriculum outcomes. These frameworks emphasize foundational skills such as locating explicit information, organizing textual details, identifying main ideas, and making simple inferences—subskills operationalized within Barrett's lower three levels. While more recent models tend to foreground advanced critical literacy skills, Barrett offers a practical and developmentally appropriate structure for early-stage EFL readers by capturing both essential literal skills and emerging inferential abilities. This balance makes the taxonomy particularly effective for constructing a diagnostic tool intended to pinpoint strengths and weaknesses in A2-level reading comprehension. Beyond theoretical alignment, it was also necessary to ensure that the construct accurately reflected national and international proficiency expectations. Therefore, the development of the diagnostic reading comprehension test was guided by the CEFR A2 reading descriptors and the MoNE (2018) 8th-grade learning outcomes, as these frameworks define the specific comprehension processes expected at this proficiency level in Türkiye. Both frameworks emphasize the ability to identify explicitly stated information, understand main ideas and supporting details, synthesize simple textual information, and make basic inferences—skills closely corresponding to the lower levels of Barrett's Taxonomy.

Considering these descriptors ensures that the assessment reflects curriculum expectations, targets developmentally appropriate skills for 8th-grade EFL learners and addresses the need for level-appropriate diagnostic tools identified in the literature.

Having established both the theoretical and curricular foundations for defining the construct, it is essential to consider how such constructs can be measured in practice. Effectively assessing the multiple dimensions of comprehension is complex but essential, as assessments serve several purposes: evaluating reading proficiency, supporting classroom learning, identifying key ideas, understanding context, and measuring curricular effectiveness (Florit & Cain, 2011). Two primary approaches dominate these assessments: standardized tests and customized assessments. Standardized tests offer consistent and objective measures that facilitate comparisons across learner populations and support data-driven decision making (Collins & Lindström, 2021). However, they often focus narrowly on specific skills, neglect higher-order thinking and creativity, and may disadvantage learners from diverse linguistic and socioeconomic backgrounds because of cultural biases (Alonzo et al., 2009). In contrast, customized reading comprehension assessments can be tailored to learners' proficiency levels, interests, and cultural contexts, enabling more equitable and accurate measurement (Georgiou & Das, 2012; Sönmez & Çetinkaya, 2022). However, developing these assessments requires significant time, resources, and expertise, and their lack of standardization can limit their generalizability and comparability (Fernandes et al., 2018; Kim 2014).

Many studies assessing learners' reading comprehension have relied on standardized instruments such as the Woodcock–Johnson III Test of Achievement (Al-Janaideh et al., 2022), TOEFL and its earlier versions (Ghaith & El-Sanyaura, 2019; Ghorbani Shemshadsara et al., 2019; Khaghaninejad, 2020; Nergis, 2013), the TOEFL Junior Practice Test (Yang et al., 2021), Cambridge First Certificate in English (Bahmani & Farvadin, 2017), IELTS (Endley, 2016), TOEIC (Maghsoudi, 2022), G-TELP (Jeon, 2011), Longman Preparation Test (Gorjian, 2013), and the CET-4 (Li et al., 2022). Reading-specific assessments, such as the Gates–MacGinitie Reading Tests (Cromley, 2006; Relyea et al., 2020), the E-LQ English Reading Comprehension Assessment (Fitzgerald et al., 2015), and the Critical Reading Inventory (Xu & Durgunoğlu, 2020), have also been widely used. While these standardized instruments provide reliable benchmarks and support comparability across diverse learner groups, they often lack sensitivity to the linguistic, cultural, and curricular realities of specific populations within those groups. Consequently, they tend to offer limited diagnostic insights into sub-skills—such as inferencing, identifying main ideas, or understanding vocabulary in context—that are essential for meaningful instructional support.

Because reading comprehension performance is strongly influenced by passage length, linguistic complexity, task type, and topic familiarity, researchers have increasingly emphasized the value of tailored reading assessments that align with learners' characteristics and educational contexts (Colenbrander et al., 2016). Such assessments can reveal fine-grained comprehension difficulties, particularly among lower-achieving students (Steensel et al., 2012), whereas the use of diverse task formats has been shown to yield more reliable and comprehensive profiles of learners' skills (Cardoso-Martins et al., 2015). Tailor-made reading assessments therefore offer important diagnostic advantages by highlighting the cognitive and linguistic processes that shape comprehension and support the design of targeted interventions (Biancarosa et al., 2025). Simultaneously, appropriate readability levels and contextual relevance contribute to greater validity, motivation, and fairness (Badrasawi et al., 2017).

In response to these needs, several studies have developed or adapted reading comprehension instruments, although their design procedures vary considerably. For example, Liu (2021) and Villanueva Aguilera (2014) incorporated multiple rounds of piloting and rigorous statistical analyses, whereas Köse and Güneş (2021) combined expert review with classroom piloting. In contrast, the limited methodological transparency in Sen's (2009) test construction process reduces confidence in replicability and validity. Validation procedures also differ across studies: while many employ internal consistency indices such as

Cronbach's alpha or the Kuder–Richardson Formula (KR) 20/21 (Darabi Bazvand, 2019; Liu, 2021; Pinninti, 2024; Sen, 2009; Zhang & Lin, 2021), some extend this through test–retest reliability (Wu et al., 2023) or expert review for content validity (Darabi Bazvand, 2019; Köse & Güneş, 2019; Pinninti, 2024). Only a few studies have applied more advanced statistical validation techniques, such as construct validation and test-form equivalence analyses, as reported by Villanueva Aguilera (2014). Overall, the inconsistency in reporting design procedures and validation evidence highlights a persistent gap in methodological rigor in this field.

Variations in item formats also contribute to uneven assessment quality. Studies have used multiple-choice, short-answer, true/false, cloze, and free recall tasks, each offering distinct advantages and limitations. Multiple-choice items increase scoring reliability and efficiency (Pinninti, 2024), whereas short-answer formats reduce guessing and enhance the validity of the assessment (McNeil, 2011). Free recall, as employed by Erten and Razi (2009), can yield deeper insights into comprehension but poses scoring challenges for researchers. Differences in scoring procedures, such as inconsistent rubrics or insufficient attention to inferential item difficulty, further complicate comparisons across studies. Cultural and linguistic appropriateness also varies widely: Erten and Razi (2009) used culturally nativized texts to improve ecological validity, while Darabi Bazvand (2019) ensured readability-adjusted passages for Iranian learners. By contrast, Zhang and Lin's (2021) use of passages from high-stakes exams such as TEM-8 and GRE risked misalignment with learners' backgrounds, potentially threatening validity. Similarly, the range of targeted sub-skills is inconsistent across studies. For instance, Kusiak (2001) assessed both local and global comprehension, whereas Li et al. (2022) employed free recall, sentence completion, and multiple-choice questions to capture a wider construct domain. However, several studies (Darabi Bazvand, 2019; Sen, 2009) lacked a clear rationale for selecting particular comprehension skills, raising concerns regarding theoretical alignment. The test administration procedures also differ substantially. While some studies used whole-class administration (Zhang, 2012), others opted for individualized formats (Wu et al. 2023). These environmental variations may have influenced the learners' performance or test fairness. In contrast, Köse and Güneş (2021) refined their administration protocol during piloting to enhance consistency, suggesting that implementation quality can significantly affect the assessment outcomes.

Taken together, the literature demonstrates that customized reading comprehension assessments produce more contextually responsive and diagnostically meaningful insights than standardized instruments; however, they require careful construct definition, systematic item design, thoughtful text selection, and rigorous validation. Despite the availability of standardized reading tools, a notable gap remains in diagnostic instruments specifically tailored to CEFR A2-level Turkish 8th-grade EFL learners. Existing assessments in Türkiye tend to emphasize achievement rather than diagnose subskill-level comprehension difficulties, limiting their usefulness for formative instructional planning and intervention. To address this gap, the present study developed a diagnostic reading comprehension test (see Appendix) designed to identify specific strengths and weaknesses in the reading skills of Turkish 8th-grade EFL learners. The test was constructed through a systematic process that incorporated construct definition, CEFR and MoNE (2018) alignment, item specification, readability analysis, piloting, and expert reviews. This study was guided by two research questions:

1. What are the psychometric properties of the developed diagnostic reading comprehension test?
2. How do students and expert reviewers perceive the validity, difficulty, and overall appropriateness of the A2-level diagnostic reading test?

2. Methodology

In this descriptive study, the methodology focused on developing a diagnostic reading comprehension test that identifies learners' specific comprehension abilities and difficulties, offering detailed insights that can

inform targeted instructional support. Barrett's Taxonomy of Reading Comprehension was selected as the conceptual foundation for task design due to its clear hierarchical structure and its long-standing use in defining and assessing reading comprehension processes. In line with the descriptive purpose of the study, only the first three levels (literal comprehension, reorganization, and inferential comprehension) were operationalized, as these levels align most closely with CEFR A2 reading descriptors and the MoNE (2018) 8th-grade curriculum outcomes. Both frameworks emphasize fundamental skills, such as locating explicit information, identifying main ideas, understanding basic sequences, and making simple inferences. Higher-level processes, such as evaluation and appreciation, were excluded because they fall beyond the cognitive and linguistic expectations of A2 learners and are not emphasized in the national curriculum.

2.1. Participants

The study involved eighth-grade Turkish EFL learners from a public lower secondary school, with different participant groups contributing to each stage of the instrument development process. The initial micro-pilot included three 8th-grade students whose feedback informed early revisions related to item clarity and task instructions. The subsequent pilot study was conducted with 28 students (15 females, 13 males), aged 13–14 years, who were enrolled in the 8th grade during the 2022–2023 academic year. The main study sample consisted of 30 students, evenly split by gender and within the same age range, who were 8th graders during the 2023–2024 academic year. All student participants had been learning English since the second grade and had no learning disabilities. Eighth graders were purposefully selected because this level represents a pedagogically significant stage in Türkiye: it is the final year of lower secondary education and culminates in the high-stakes LGS (Liseye Geçiş Sınavı) administered by the MoNE. Given the central role of reading comprehension in academic readiness for the transition to high school, developing a diagnostic tool for this group is both timely and relevant.

In addition to the student participants, two expert English language teachers contributed to establishing the content and construct validity of the test. These experts were selected because of their extensive experience working with lower-secondary EFL learners and their familiarity with the MoNE curriculum. One expert was a male teacher with a BA degree and 28 years of English teaching experience at the lower secondary level, and the other was a female teacher with an MA degree and 20 years of experience. Both experts reviewed the test items and provided structured feedback regarding their alignment with the construct, clarity of instructions, and overall appropriateness for A2-level learners. Informed consent was obtained from all student participants and their parents, and the study adhered to the ethical guidelines for educational research.

2.2. Setting

The study was conducted at a state lower-secondary school in northwest Türkiye, which was purposefully selected for its Intensive English Program and manageable student population. A distinctive feature of the school is the 5th-grade Intensive English Program, where students receive 11 hours of instruction per week from multiple teachers, integrating both core content and language skills. In the following grades, the intensity decreased to three hours per week in 6th grade and four hours in 7th and 8th grades. This instructional structure, combined with the school's curricular consistency and relatively small size, provided a practical and well-organized environment for developing, piloting, and administering diagnostic tests.

2.3. Test Development Procedure

2.3.1. Construct Definition

The construct for this study was defined as A2-level reading comprehension, operationalized through key sub-skills including literal comprehension, reorganization of information, inferential comprehension, main idea identification, and contextual understanding. These sub-skills reflect the reading behaviors expected of A2 learners and provide the foundation for developing a diagnostic tool capable of identifying specific areas of strength and difficulty.

To ensure content representativeness and curriculum alignment, the construct and corresponding test tasks were systematically cross-checked against MoNE (2018) English curriculum outcomes for 8th grade and the CEFR A2 reading descriptors. These descriptors specify essential reading behaviors such as locating explicit information, recognizing key ideas, understanding sequence, and making basic inferences and therefore offer clear criteria for determining which sub-skills a diagnostic test at this level must validly measure. Aligning the test with both MoNE expectations and CEFR benchmarks ensured that the assessment reflected national curricular demands while maintaining coherence with internationally recognized proficiency standards, thereby strengthening its construct validity. As mentioned above, Barrett's Taxonomy of Reading Comprehension was selected as the theoretical framework guiding the operationalization of the construct. The taxonomy's hierarchical structure, which differentiates literal recall, reorganization, and inferential reasoning, aligns closely with the cognitive demands of A2-level reading. It also supports diagnostic assessment by enabling the identification of specific comprehension processes in which learners may experience difficulty. Prior research has demonstrated that Barrett's levels offer a systematic means of evaluating comprehension and contribute to construct validity in studies involving adolescent EFL learners (Junaidi et al., 2024; Marlinton et al., 2023).

Additionally, readability considerations were incorporated to ensure that the texts used in the assessment were appropriate for eighth-grade A2 learners. The readability analysis focused on features such as sentence length, lexical load, and syntactic complexity. This process helped prevent construct-irrelevant difficulty and supported the selection of passages that were accessible yet still sufficiently challenging for the target proficiency level. Such linguistic features play a decisive role in determining text comprehensibility in L2 contexts (Crossley & McNamara, 2016). Ensuring alignment between CEFR descriptors and textual difficulty is essential, as CEFR level descriptions outline the communicative and processing demands expected of learners at different proficiency stages (Council of Europe, 2001). Integrating CEFR specifications with readability metrics therefore helped ensure that the chosen texts matched the cognitive and linguistic requirements of A2 learners while still enabling meaningful diagnostic differentiation (Alderson, 2000; Nation, 2006).

2.3.2. Task Design and Text Selection

In line with Alderson's framework, the test tasks were designed to assess both explicit and implicit comprehension, deliberately excluding script-based questions. To capture a broad range of comprehension processes, multiple item formats were incorporated, including chart completion, title selection, true/false statements, gap-filling, main idea identification, multiple-choice items, ordering tasks, and open-ended responses. Texts were selected based on their thematic relevance, potential to engage learners, and appropriateness for the target's proficiency level. Table 1 summarizes how each task aligns with the MoNE (2018) reading objectives, CEFR A2 descriptors, and the corresponding levels of Barrett's Taxonomy, thereby illustrating the construct representation underlying the diagnostic assessment.

Table 1.*Alignment of MoNE (2018) Reading Objectives, CEFR A2 Descriptors, Barrett's Levels, and Test Tasks*

MoNE Reading Objective (2018)	CEFR A2 Reading Descriptor	Text / Task Type	Barrett's Level(s)	Targeted Subskills	Expected Outcomes
Locate specific information in everyday materials	Can find and understand specific, predictable information in simple everyday material	<i>Reunion Invitation</i> – Chart Completion	Literal, Inferential	Extract explicit details; identify purpose (accept/refuse)	Correctly classify responses; retrieve factual details; infer implied intentions
Identify main idea and supporting details	Can identify the main point in short, clear, simple texts	<i>Cooking Tips</i> – Title Selection; True/False	Literal, Reorganization	Identify main idea; verify factual statements; link details to theme	Select correct title; judge truth of statements; recognize thematic coherence
Interpret relationships (sequence, cause–effect)	Can understand the sequence of events in straightforward narratives	<i>Jennifer Lawrence</i> – Ordering; WH-Questions; <i>Cooking Tips</i> – True/False	Literal, Reorganization	Sequence events; recognize cause–effect relations	Arrange events logically; answer WH-questions accurately
Infer meaning from familiar context	Can deduce the meaning of a word from context in familiar topics	<i>Internet Use</i> – MCQs; <i>Household Chores</i> – Gap Filling	Literal, Inferential	Interpret vocabulary from context; infer implied ideas	Select correct vocabulary; infer contextual meanings and implications
Understand sequence of events in narratives/biographies	Can understand chronological order in simple descriptive texts	<i>Jennifer Lawrence</i> – Ordering	Literal, Reorganization	Identify chronological structure; understand biographical progression	Correct sequence of events; demonstrate comprehension of narrative flow
Recognize text type and purpose	Can understand the purpose of short, simple texts (letters, brochures, stories)	All text types: Invitation, Cooking Tips, Internet Use, Biography, Chores	Literal, Reorganization	Identify functional purpose and structure of text types	Distinguish functional, informative, and narrative texts
Develop vocabulary knowledge through context	Can understand words/phrases on familiar topics	<i>Household Chores</i> – Gap Filling; <i>Cooking Tips</i> – Vocabulary in T/F	Literal, Inferential	Use contextual clues for vocabulary and collocations	Apply contextual meaning accurately; select appropriate lexical items
Combine literal and inferential understanding	Can understand short, simple texts on familiar matters	<i>Reunion Invitation</i> , <i>Cooking Tips</i> , <i>Internet Use</i>	Literal, Inferential	Recall explicit details; infer missing or implied meaning	Demonstrate integrated literal and inferential comprehension

As shown in Table 1, this systematic alignment ensured that each test component measured construct-relevant reading behaviors. During test development, passages and tasks were selected and adapted based on these frameworks, enabling each item to target clearly defined subskills such as locating explicit information, identifying main ideas, interpreting sequence, and inferring meaning from context. Readability was evaluated using the Flesch Reading Ease (FRE) and Flesch–Kincaid Grade Level (FKGL) indices, two standardized metrics widely used to determine the accessibility of written materials (Badarudeen & Sabharwal, 2010; Friedman & Hoffman-Goetz, 2006). Both indices quantify text difficulty based on sentence length and syllable count: the FRE provides a score ranging from 0 to 100, with higher values indicating easier readability, whereas the FKGL converts the same linguistic features into an estimated U.S. school grade level required for comprehension.

Rather than creating entirely new materials, test activities were compiled from diverse existing sources. An extensive review of MoNE coursebooks, national and international publications, graded readers, English

language learning websites, news outlets, and scientific resources was conducted to identify suitable items. Books from Turkish publishers and online story platforms commonly used by teachers were examined to maximize contextual relevance and familiarity. Items were selected based on their alignment with A2-level proficiency descriptors, relevance to curriculum learning objectives, and capacity to assess explicit and implicit comprehension skills. This systematic selection process, supported by expert review and attention to learner interests, strengthened the content validity and reliability of the assessment.

Text selection followed a structured, context-sensitive procedure. Reading passages were chosen according to CEFR A2 difficulty criteria, lexical load, syntactic complexity, and topic familiarity appropriate for 8th-grade learners. Familiarity was ensured by drawing on themes embedded in the 8th-grade MoNE curriculum. Most texts were adapted from widely used coursebooks in public school settings and were further refined to meet the diagnostic aims of the test. One passage was retrieved from an open-access language learning website to address the specific sub-skills required in the test specifications. No passages were generated using AI. To maintain authenticity and contextual relevance, only sources aligned with national curriculum expectations and commonly encountered by students were used in this study. All passages were carefully adapted by adjusting their complexity, content, and structure so that they aligned with the targeted comprehension processes. All necessary permissions for text use and adaptation were obtained prior to data collection.

2.4. Piloting and Validation

The study was conducted at a public lower secondary school in northwest Türkiye. All participants in the trial, pilot, and main studies were drawn from this single institution to ensure contextual consistency, although no student participated in more than one phase. As previously noted, the eighth grade represents a high-stakes examination year in Türkiye because of the national LGS. Therefore, developing a diagnostic tool for this level is pedagogically meaningful as it enables teachers to identify reading comprehension gaps before students' transition to upper secondary education. A systematic two-stage piloting process was implemented. The initial micro-pilot was conducted with three eighth-grade students who were not involved in the main study. Using guided reflections and researcher field notes, feedback was gathered on item clarity, instruction comprehensibility, text difficulty, and overall test length. This stage allowed for early refinements of wording, difficulty levels, and formatting, in line with recommendations for iterative test development (Alderson, 2000).

To ensure content validity and curriculum alignment, the draft test was reviewed by two expert English language teachers with extensive lower-secondary EFL teaching experience. The experts reviewed each text and item based on key criteria, including alignment with CEFR A2 reading descriptors, MoNE (2018) 8th-grade curriculum outcomes, text difficulty, instruction clarity, task format suitability, and potential cultural bias. They confirmed that the majority of items were age-appropriate and construct-relevant but highlighted concerns regarding several vocabulary-dependent items and noted a minor cultural bias in the biography text. Their feedback led to adjustments, such as rewording ambiguous instructions, refining text-dependent items, and revising tasks that relied excessively on isolated, lexical knowledge.

The revised instrument was piloted with 28 eighth-grade students from the same school. This second pilot provided additional evidence regarding the test's alignment with curriculum objectives and the clarity of its instructions and task formats under near-final conditions. Analysis of pilot responses indicated that several items continued to impose an unnecessarily high vocabulary load, placing disproportionate emphasis on lexical recall rather than on the intended comprehension processes. These observations informed the final set of revisions, such as simplifying lexical demands and strengthening text dependence, prior to administering the fully revised diagnostic test to a separate sample of 30 eighth-grade students who participated in the main study.

2.5. Scoring and Reliability

Reading comprehension can be assessed through various task types, as demonstrated in the literature. Accordingly, the diagnostic test incorporated multiple formats: chart completion, WH-questions, true/false items, sequencing tasks, title selection, main-idea matching, multiple-choice questions, and gap-filling, resulting in a total of 52 items. Most items were scored dichotomously, awarding 1 point for a correct response and 0 points for an incorrect response, in line with standard discrete-point scoring procedures. Sequencing tasks were scored using the Weighted Marking Protocol (Razi, 2005), which assigns partial credit to partially correct orderings, thus capturing more nuanced aspects of comprehension. This mixed scoring approach enabled the test to measure both discrete and process-oriented comprehension skills while preserving the scoring consistency. Internal consistency reliability was calculated using the Kuder–Richardson Formula 20 (KR-20), which is appropriate for assessments with dichotomously scored items.

2.6. Data Collection

Quantitative data were gathered through the administration of a diagnostic reading comprehension test to the main study sample. The test provided item-level response data that were used to evaluate the psychometric properties of the instrument. Qualitative data were collected from two sources: (a) written guided reflections completed by students immediately after test administration, which captured their perceptions of task clarity, and difficulty, and (b) structured expert review obtained during the validation phase, which addressed issues of content alignment, construct representation, and potential sources of bias. Together, these datasets offer complementary insights into both performance-based and perception-based dimensions of test validity.

2.7. Data Analysis

Quantitative data were analyzed using KR-20 to determine the internal consistency reliability of the diagnostic test, which is appropriate for dichotomously scored data. Descriptive statistics were also used to evaluate item performance and overall test behavior. Qualitative data from student reflections and expert reviews were thematically analyzed using MAXQDA. The analysis involved initial open coding, followed by the development and refinement of categories that captured recurring patterns related to perceived difficulty, clarity, vocabulary load, and construct relevance. This combined analytic approach supports a richer interpretation of the test's validity and usability.

3. Findings

3.1. Psychometric Properties of the Diagnostic Reading Comprehension Test (RQ1)

The construct of the study—A2-level reading comprehension encompassing the subskills of literal understanding, reorganization, inferential reasoning, main idea identification, and contextual interpretation—was broadly supported by the findings.

Table 2.
Scores of readability

Readability Analysis		Reading Test					
		Text 1	Text 2	Text 3	Text 4	Text 5	Total
Counts	Words	238	261	405	196	120	1220
	Characters	1001	1266	1919	893	551	5640

Words	Paragraphs	12	9	6	10	3	40
	Sentences	28	26	37	13	11	115
	Sentences per paragraph	3.1	3.7	6.1	3.2	5.5	4.1
	Words per sentence	7.5	8.9	10.9	12.0	10.4	9.7
	Characters per word	4.0	4.4	4.5	4.1	4.5	4.3
Readability	Passive sentences	0.0%	0.0%	0.0%	7.6%	0.0%	0.08%
	Flesch reading ease	77.5	79.2	70.1	76.0	68.2	74.3
	Flesch-Kincaid grade level	4.3	4.4	6.1	5.6	6.3	5.3

The results of the readability analysis showed that four of the five texts (Texts 1, 2, 4, and 5) fell within the 4th–6th grade readability range (Flesch Reading Ease = 68–79), indicating high accessibility for eighth-grade A2 learners. Only one passage, Internet Use, was slightly more demanding (Flesch–Kincaid Grade Level = 6.1), which aligns with student reflections describing this text as comparatively more challenging. The analysis also revealed that passive sentence use was minimal across all passages, contributing to the structural clarity and reducing the syntactic load. Sentence length varied across texts, with longer sentences in Texts 3 and 4, suggesting a moderate increase in processing demands for these passages. All readability statistics were generated directly from the test passages using an automated tool. The table reports raw text features such as number of words, sentences, and characters, calculated averages such as words per sentence and characters per word, and standardized readability indices (Flesch Reading Ease and Flesch–Kincaid Grade Level), both of which are based on sentence length and syllable density. Taken together, these indicators confirm that the selected texts were appropriate for CEFR A2 learners and that the overall linguistic complexity of the test aligned well with expectations for the target population. The scoring system also contributes to the reliability and fairness of the assessment. Most items were scored dichotomously (1 = correct and 0 = incorrect). Internal consistency reliability was very high for the pilot version (KR-20 = 0.91) and remained strong for the revised main study version (KR-20 = 0.81), exceeding the commonly accepted minimum threshold for adequate reliability (DeVellis, 2017).

3.2. Student and Expert Perceptions of Test Validity, Difficulty, and Appropriateness (RQ2)

Expert reviewers confirmed that the test appropriately represented the national curriculum outcomes and aligned with the CEFR A2 reading descriptors. In terms of Barrett’s Taxonomy, literal and reorganization items were strongly represented, inferential items were moderately addressed, and evaluation and appreciation items were largely absent, reflecting both the strengths and inherent limitations of the construct coverage. The initial micro-pilot with three students and the subsequent pilot with 28 students further supported the clarity of the instructions and overall curricular alignment, but both stages also highlighted several vocabulary-dependent items that risked shifting the focus away from reading comprehension. Revisions made in response to this feedback improved item clarity, strengthened text dependence, and helped balance task difficulty across sections.

Expert review supported the content validity of the test. The reviewers emphasized consistency with CEFR descriptors and MoNE outcomes, noting that most items successfully assessed A2-level behaviors such as locating explicit information, identifying main ideas, and making basic inferences. As one expert stated, “the first three texts clearly target A2 descriptors and follow the thematic structure students are familiar with in the national curriculum.” Readability and difficulty levels were also judged appropriate, although the biography text (Text 4) was described as “noticeably more challenging due to longer sentences and unfamiliar proper nouns.” Experts affirmed the clarity of instructions (“Students will understand what to do in each section because the instructions resemble those in MoNE coursebooks”) but cautioned that certain gap-filling items measured “lexical recall more than reading ability,” which may weaken construct

validity. Cultural relevance was generally achieved, though one reviewer noted that the biography text “may introduce mild cultural bias” due to references unfamiliar to some learners. Despite these concerns, both reviewers agreed that the test has strong potential for repeated use following minor vocabulary-related revisions.

Student reflections also offered valuable insights into the perceptions of validity, difficulty, task design, and engagement. In terms of overall text difficulty, learners generally found the initial passages accessible, describing them as “clear and easy to understand,” whereas the final section was viewed as noticeably more demanding. One student explained that “the last page was entirely based on vocabulary; frankly, I found the last page challenging,” highlighting a perceived inconsistency in the difficulty across sections. Students responded positively to instructional clarity, noting that familiar formats, such as multiple-choice and WH-questions, supported their understanding. At the same time, some expressed interest in more authentic and engaging task types, with one suggesting that “it can be more useful if there are some grammar questions... and maybe better if no options, like suitable title questions.”

A recurring theme was the interplay between reading comprehension and vocabulary load. One learner commented, “I think that if I know so much words that will be easier... So the test is easy,” illustrating how lexical knowledge strongly shapes perceptions of task difficulty. Another remarked that “the first parts... measured my reading, but the last page felt more like a vocabulary exam,” suggesting that some items may have introduced construct irrelevant demands. Students generally found the themes familiar and supportive of their comprehension, although engagement varied. While several learners reported that the test was clear and manageable, others noted that vocabulary-heavy items reduced their motivation and expressed a preference for more interactive tasks, such as title selection or open-ended questions.

Overall, the findings from expert and student feedback converge to show that the diagnostic reading comprehension test demonstrates strong reliability, high face validity, and alignment with CEFR A2 descriptors and MoNE curriculum outcomes. Simultaneously, both groups highlighted areas requiring refinement, notably reducing vocabulary dependence, minimizing grammar-oriented items, and incorporating additional text-dependent tasks to strengthen construct validity and better assess higher-order comprehension skills. Taken together, these findings indicate that the test is psychometrically sound and pedagogically meaningful, while also providing a clear direction for future refinements.

4. Discussion and Conclusion

This study aimed to develop and validate a diagnostic reading comprehension test aligned with the CEFR A2 descriptors and the Turkish national curriculum for lower-secondary learners. The discussion interprets the findings in relation to the instrument’s reliability, validity, diagnostic sensitivity, and instructional usefulness, with particular attention to its function as a diagnostic tool, rather than a summative achievement measure. The findings suggest that the test yields stable and interpretable outcomes suitable for diagnostic purposes. The combined use of dichotomous scoring and partial credit for sequencing tasks appears to enhance diagnostic sensitivity by capturing partial comprehension and emerging control of reading processes that would otherwise remain obscured in strictly binary scoring systems. From a measurement perspective, this supports the interpretation that the instrument can differentiate between learners who have trouble with specific comprehension processes and those who demonstrate developing competence across targeted sub-skills.

Taken together, expert review, piloting, and readability control support the interpretation that the test adequately represents the construct of A2-level reading comprehension. Alignment with CEFR descriptors and MoNE curriculum outcomes (2018) supports the interpretation that the instrument measures instructionally relevant and developmentally appropriate reading behavior. Attention to text difficulty and

linguistic complexity appears to have reduced construct-irrelevant difficulty, allowing learner performance to reflect comprehension processes rather than peripheral linguistic barriers. This interpretation is consistent with prior research highlighting the importance of accessible input in strengthening the interpretability of diagnostic reading assessments (McNeil, 2011; Shahnazari, 2023).

The assessment also shows conceptual coherence with Barrett's Taxonomy by foregrounding literal comprehension, reorganization, and inferential processing—sub-skills that constitute the core of reading comprehension at lower proficiency levels and provide a foundation for later higher-order development. The inclusion of varied task formats such as multiple-choice, true/false, chart completion, sequencing, and open-ended items appears to strengthen the diagnostic function of the instrument by eliciting different manifestations of learner's comprehension. In this respect, the assessment aligns with recommendations advocating the integration of objective and constructed response tasks to generate a more differentiated profile of learners' reading abilities (Kusiak 2001; Li et al. 2022).

Interpretation of student and expert feedback provides further insight into the diagnostic quality of the test. Although the overall structure, instructions, and task formats were generally perceived as clear and familiar, feedback from both groups suggests that certain sections placed comparatively greater demands on vocabulary knowledge. From a construct validity perspective, this pattern points to the possibility that some items may have shifted attention away from comprehension processes toward lexical recall, thereby narrowing the test's diagnostic focus. Expert observations regarding culturally specific references in one biographical passage similarly highlight the role of background knowledge and topic familiarity in shaping performance. Rather than weakening the overall validity of the instrument, such feedback usefully identifies areas for refinement that can enhance construct representation and improve diagnostic precision in future iterations (Darabi Bazvand, 2019; Erten & Razi, 2009).

From an instructional perspective, the findings suggest that diagnostically oriented reading assessments can support teachers in moving beyond global proficiency scores toward a more fine-grained understanding of learners' reading strengths and weaknesses. Patterns of performance across different types appear to indicate whether learners experience greater difficulty with locating explicit information, sequencing ideas, or drawing inferences, thereby enabling more targeted and responsive instructional planning. This diagnostic function is particularly valuable in lower-secondary EFL classrooms and in high-stakes examination contexts such as the LGS, where instructional time is constrained and the efficient targeting of reading subskills is essential. Beyond the development of a single assessment instrument, this study contributes a curriculum-aligned and CEFR-referenced diagnostic test development model that can be replicated in similar EFL contexts. The primary contribution lies in the systematic and transparent process through which the test was designed and validated, rather than in the test product itself. By integrating CEFR descriptors, national curriculum outcomes, Barrett's Taxonomy of Reading Comprehension, expert review, piloting procedures, and readability control, the proposed model foregrounds principled decision-making and construct alignment at each stage of test development. In doing so, this study extends existing work on diagnostic assessment by demonstrating how pedagogical relevance and diagnostic sensitivity can be intentionally embedded in test design.

Importantly, the proposed framework is transferable beyond the CEFR A2 level and the specific context of Turkish lower secondary EFL education. With appropriate recalibration of text difficulty, task formats, and targeted comprehension processes, the model can be adapted to other proficiency levels, such as A1 or B1, and to diverse instructional settings. At lower levels, the framework may prioritize literal comprehension and basic sequencing, whereas at higher levels, it may emphasize inferencing, summarization, and integration of information across texts. In addition to summative diagnostic use, the framework lends itself to formative assessment and classroom-based action research, enabling teachers to identify reading needs, implement targeted interventions, and evaluate the instructional impact over time. Overall, the study

illustrates how diagnostically oriented, curriculum-aligned reading assessments can be systematically developed to support both instructional decision-making and research-informed practice in lower-secondary EFL contexts.

This study has several limitations. The relatively small sample sizes used during the micro-pilot and pilot phases limit the generalizability of item-level interpretations. The absence of a systematic inter-rater reliability analysis for open-ended items also constrains the interpretation of subjective scoring. Despite careful attention to cultural relevance, some tasks may have been influenced by learners' background knowledge and vocabulary size. In addition, higher-level comprehension processes, such as evaluation and appreciation, were underrepresented, an expected constraint at the A2 level but nevertheless an area for future expansion.

Future research may build on this work by administering the test to larger and more diverse samples, systematically examining inter-rater reliability for constructed response items, and further exploring the role of cultural schemata in shaping reading assessment performances. Adapting the proposed test development framework to other CEFR levels and educational contexts may also support longitudinal investigations into reading development and the effectiveness of targeted instructional interventions.

Acknowledgments

This research paper originated from a part of a Ph.D. thesis named “The Effects of Strategy Training on EFL Learners' Reading Comprehension.” at Onsekiz Mart University, School of Graduate Studies, English Language Teaching.

Note on Ethical Issues

The study has been approved by the Ethics Committee of Çanakkale Onsekiz Mart University, Graduate School of Education (E-84026528-050.01.04-2200276280, 18.11.2022) and the Turkish Ministry of National Education (E-99191664-605.01-67036022, 28.12.2022)

Conflict of Interest

The author(s) have declared no conflict of interest in this study.

Funding

The author/authors did not receive any funding for this article from any institution.

References

- Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732935>
- Al-Janaideh, R., Hipfner-Boucher, K., Cleave, P., & Chen, X. (2022). Contributions of code-based and oral language skills to Arabic and English reading comprehension in Arabic–English bilinguals in the elementary school years. *International Journal of Bilingual Education and Bilingualism*, 25(7), 2495–2510. <https://doi.org/10.1080/13670050.2020.1862775>
- Alonzo, J., Basaraba, D., Tindal, G., & Carriveau, R. S. (2009). They read, but how well do they understand? An empirical look at the nuances of measuring reading comprehension. *Assessment for Effective Intervention*, 35(1), 34–44. <https://doi.org/10.1177/1534508408330082>

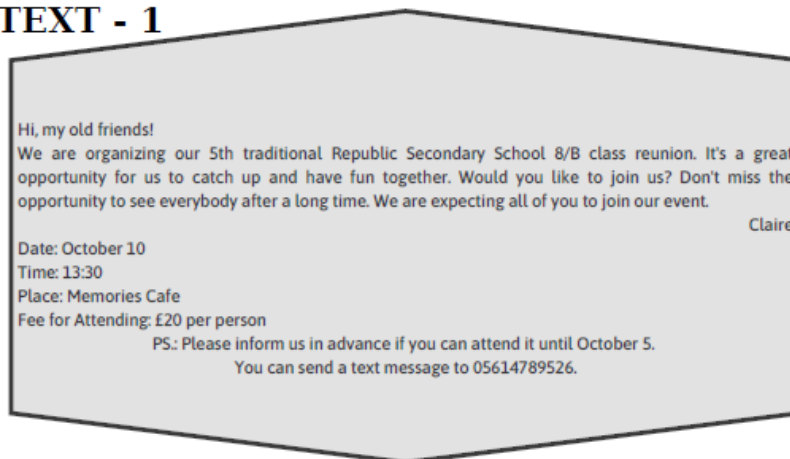
- Badrasawi, K. J. I., Kassim, N. L. A., & Daud, N. M. (2017). The effects of test characteristics on the hierarchical order of reading skills. *Malaysian Journal of Learning and Instruction*, 14(1), 63–82. <https://doi.org/10.32890/mjli2017.14.1.3>
- Bahmani, R., & Farvardin, M. T. (2017). Effects of different text difficulty levels on EFL learners' foreign language reading anxiety and reading comprehension. *Reading in a Foreign Language*, 29(2), 185–202. <https://doi.org/10.64152/10125/66912>
- Biancarosa, G., Kennedy, P. C., Carlson, S. E., Seipel, B., & Davison, M. L. (2025). Diagnostic and instructionally relevant measurement of reading comprehension. *Intervention in School and Clinic*. Advance online publication. <https://doi.org/10.1177/10534512251327734>
- Cardoso-Martins, C., Gonçalves, D. T., Magalhães, C. G. D., & Silva, J. R. D. (2015). Word reading and spelling ability in school-age children and adolescents with autism spectrum disorders: Evidence from Brazilian Portuguese. *Psychology & Neuroscience*, 8(4), 479–487. <https://doi.org/10.1037/pne0000029>
- Colenbrander, D., Nickels, L., & Kohnen, S. (2016). Similar but different: Differences in comprehension diagnosis on the Neale Analysis of Reading Ability and the York Assessment of Reading for Comprehension. *Journal of Research in Reading*, 40(4), 403–419. <https://doi.org/10.1111/1467-9817.12075>
- Collins, A. A., & Lindström, E. R. (2021). Making sense of reading comprehension assessments: Guidance for evaluating student performance. *Intervention in School and Clinic*, 57(1), 23–31. <https://doi.org/10.1177/10534512211001854>
- Cromley, J. G., & Azevedo, R. (2006). Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning*, 1(3), 229–247. <https://doi.org/10.1007/s11409-006-9002-5>
- Crossley, S. A., & McNamara, D. S. (2016). Text-based recall and extra-textual generations resulting from simplified and authentic texts. *Reading in a Foreign Language*, 28(1), 1–19. <https://doi.org/10.64152/10125/66713>
- Darabi Bazvand, A. (2019). L1 domain-specific knowledge as predictor of reading comprehension in L2 domain-specific texts: The case of ELT student teachers. *Cogent Education*, 6(1), 1631019. <https://doi.org/10.1080/2331186X.2019.1631019>
- DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). Sage Publications.
- Endley, M. J. (2016). Proficiency as a variable in Gulf EFL students' employment of reading strategies. *Reading in a Foreign Language*, 28(2), 183–223. <https://doi.org/10.64152/10125/66900>
- Erten, İ. H., & Razi, S. (2009). The effects of cultural familiarity on reading comprehension. *Reading in a Foreign Language*, 21(1), 60–77. <https://doi.org/10.64152/10125/66632>
- Fernandes, S., Querido, L., Verhaeghe, A., & Araújo, L. (2018). What is the relationship between reading prosody and reading comprehension in European Portuguese? *Journal of Research in Reading*, 41(S1), S37–S57. <https://doi.org/10.1111/1467-9817.12248>
- Fitzgerald, J., Stenner, A. J., Sanford-Moore, E. E., Koons, H., Bowen, K., & Kim, K. H. (2015). The relationship of Korean students' age and years of English-as-a-foreign-language exposure with English-reading ability. *Reading Psychology*, 36(2), 173–202. <https://doi.org/10.1080/02702711.2013.843063>
- Florit, E., & Cain, K. (2011). Developmental changes in the relationships between listening comprehension and reading comprehension. *Journal of Educational Psychology*, 103(3), 778–791. <https://doi.org/10.1037/a0024848>
- Georgiou, G. K., & Das, J. P. (2012). Reading comprehension in university students: Relevance of PASS theory. *Journal of Research in Reading*, 37(S1), S1–S16. <https://doi.org/10.1111/j.1467-9817.2012.01542.x>
- Ghaith, G., & El-Sanyoura, H. (2019). Reading comprehension: The mediating role of metacognitive strategies. *Reading in a Foreign Language*, 31(1), 19–43.
- Ghorbani Shemshadsara, Z., Ahour, T., & Hadidi Tamjid, N. (2019). Raising text structure awareness as a strategy for improving EFL undergraduate students' reading comprehension. *Cogent Education*, 6(1), 1644704. <https://doi.org/10.1080/2331186X.2019.1644704>
- Gorjian, B. (2013). The effect of passage content on multiple-choice reading comprehension tests. *Procedia – Social and Behavioral Sciences*, 84, 160–164. <https://doi.org/10.1016/j.sbspro.2013.06.510>
- Hamdollahi, M. H., Mohamadi, R., Sadeghi, A., Ahadi, H., Poormohammadi, F., & Bahrainian, B. (2024). Investigating the relationship between the scores of the Persian reading comprehension assessment

- and reading and dyslexia test. *Function and Disability Journal*, 6(1), Article e84. <https://doi.org/10.32598/fdj.6.84.4>
- Jeon, E. H. (2011). Contribution of morphological awareness to second-language reading comprehension. *The Modern Language Journal*, 95(2), 217–235. <https://doi.org/10.1111/j.1540-4781.2011.01179.x>
- Kardoğan, M., & Geçgel, H. (2024). An analysis of questions based on reading texts in Turkish textbooks prepared for foreigners according to Barrett's taxonomy. *Base for Electronic Educational Sciences*, 5(2), 153–169. <https://doi.org/10.29329/bedu.2024.1064.9>
- Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school. *Journal of Educational Psychology*, 101(4), 765–778. <https://doi.org/10.1037/a0015956>
- Khaghaninejad, M. S. (2020). Are reading comprehension ability and strategies transferable from L1 to L2? *Southern African Linguistics and Applied Language Studies*, 38(4), 293–306. <https://doi.org/10.2989/16073614.2020.1854796>
- Kim, Y. S. (2014). Language and cognitive predictors of text comprehension. *Child Development*, 86(1), 128–144. <https://doi.org/10.1111/cdev.12293>
- Kor, C. P., Low, H. M., & Lee, L. W. (2014). Relationship between oral reading fluency and reading comprehension. *GEMA Online Journal of Language Studies*, 14(3), 19–32. <https://doi.org/10.17576/gema-2014-1403-02>
- Köse, N., & Güneş, F. (2021). Undergraduate students' use of metacognitive strategies while reading. *Journal of Education and Learning*, 10(2), 99–108. <https://doi.org/10.5539/jel.v10n2p99>
- Kusiak, M. (2001). The effect of metacognitive strategy training on reading comprehension. *EUROSLA Yearbook*, 1(1), 255–274. <https://doi.org/10.1075/eurosla.1.19kus>
- Li, Y., Brantmeier, C., Gao, Y., & Hogrebe, M. (2022). Comparing reading strategy measures. *Reading in a Foreign Language*, 34(2), 271–305.
- Liu, H. (2021). Does questioning strategy facilitate L2 reading comprehension? *Journal of Research in Reading*, 44(2), 339–359. <https://doi.org/10.1111/1467-9817.12339>
- Maghsoudi, M. (2022). Contributions of motivation to read in L2. *Reading & Writing Quarterly*, 38(3). <https://doi.org/10.1080/10573569.2021.1931591>
- Marlinton, M., Syahri, I., & Mayasari, S. (2023). The efficacy of task-based learning. *Elsya: Journal of English Language Studies*, 5(2), 250–269. <https://doi.org/10.31849/elsya.v5i2.14082>
- McNeil, L. (2011). Investigating the contributions of background knowledge and strategy use. *Reading and Writing*, 24(8), 883–902. <https://doi.org/10.1007/s11145-010-9230-6>
- Nation, I. S. P. (2006). How large is a vocabulary needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82. <https://doi.org/10.3138/cmlr.63.1.59>
- Nergis, A. (2013). Exploring factors that affect reading comprehension. *Journal of English for Academic Purposes*, 12(1), 1–9. <https://doi.org/10.1016/j.jeap.2012.09.001>
- Pinninti, L. R. (2024). The impact of peer-collaborative strategic reading. *TESL-EJ*, 27(3). <https://doi.org/10.55593/ej.27108a3>
- Razi, S. (2005). A fresh look at ordering tasks: Weighted marking protocol. *Reading Matrix*, 5(1), 1–15.
- Relyea, J. E., Zhang, J., Liu, Y., & Lopez Wui, M. G. (2020). Contribution of home language and literacy environment. *Reading Research Quarterly*, 55(3), 473–492. <https://doi.org/10.1002/rrq.288>
- Sen, H. Ş. (2009). The relationship between metacognitive strategy use and reading comprehension. *Procedia-Social and Behavioral Sciences*, 1(1), 2301–2305. <https://doi.org/10.1016/j.sbspro.2009.01.404>
- Shahnazari, M. (2023). The role of working memory in L2 reading comprehension. *Learning and Motivation*, 82, 101875. <https://doi.org/10.1016/j.lmot.2023.101875>
- Simşek, B., & Direkçi, B. (2023). The effects of augmented reality storybooks. *British Journal of Educational Technology*, 54(3), 754–772. <https://doi.org/10.1111/bjet.13293>
- Sönmez, M., & Çetinkaya, F. C. (2022). The effect of formative assessment on reading comprehension. *International Journal of Assessment Tools in Education*, 9(SI), 88–108. <https://doi.org/10.21449/ijate.1104868>
- Steensel, R. van, Oostdam, R., & van Gelderen, A. (2012). Assessing reading comprehension in adolescent low achievers. *Language Testing*, 30(1), 3–21. <https://doi.org/10.1177/0265532212440950>
- Sulfa, S., Ernawati, E., & Fatmawati, F. (2023). Investigating literal and inferential comprehension achievement. *Technium Social Sciences Journal*, 39(1), 127–133. <https://doi.org/10.47577/tssj.v39i1.8057>

- Villanueva Aguilera, A. B. (2014). Strategy Intervention to Enhance Reading Comprehension Of 15-Year-Old Students in Mexico. Doctoral Dissertation, University of York, Education, York, UK.
- Wahyuni, K. B. (2021). The levels of questions in the textbook “Stop Bullying Now.” *Jurnal Pendidikan Bahasa Inggris Undiksha*, 9(2), 191. <https://doi.org/10.23887/jpbi.v9i2.34393>
- Xu, H., & Durgunoğlu, A. Y. (2020). Motivational factors underlying different levels of reading comprehension. *TESOL Journal*, 11(1), e00448. <https://doi.org/10.1002/tesj.448>
- Yang, Y. H., Chu, H. C., & Tseng, W. T. (2021). Text difficulty in extensive reading. *Reading in a Foreign Language*, 33(1), 1–19. <https://doi.org/10.64152/10125/67394>
- Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The modern language journal*, 96(4), 558-575. <https://doi.org/10.1111/j.1540-4781.2012.01398.x>
- Zhang, H., & Lin, J. (2021). Morphological knowledge in L2 reading comprehension. *Educational Psychology*, 41(5), 563–581. <https://doi.org/10.1080/01443410.2020.1865519>

Appendix

TEXT - 1



This is an invitation to a reunion. Read the replies to the invitation and complete the guest list.

A reunion party sounds great, but I must visit my grandparents with my family then. See you later.
Johanna

A reunion party? That sounds fun. See you at the café.
Janet

I don't want to miss the opportunity to see everybody after a long time. Hope to see you all.
Candy

Awesome! Of course, I will. We can talk about our memories and have fun together. Take care until we meet up there.
Mike

Thank you for the nice invitation. I'd love to come to the event. What time does it finish? I can attend after 14:30.
Kate

Thank you for organizing the event, Claire. I'd love to see all my old friends, but I must finish my project. I'm really busy.
Donna

Thanks for inviting me. It will be so nice to meet up again. I will be there.
Sam

ACCEPTED

REFUSED

TEXT - 2

Read the text.



•Have good kitchen tools. Using good kitchen tools will make cooking easy and fun. For example, you can't stand dicing tomatoes with a knife that is not sharp. If your kitchen tools aren't good, you don't want to be in your kitchen

•Wash your ingredients first. Clean ingredients make the dishes tasty and healthy. There are many different viruses or microbes on the products. So, never forget to wash them.

•Read the full recipe. You can cook some dishes very well, you should know all the steps very well to cook better. Always read and reread the recipe. If you forget even one of the steps, your dish won't be tasty

•Taste your dish. You should taste your dishes before serving them or adding salt and spices to them. All chefs do that.

•Low and slow. You should cook your dishes slowly at low temperatures. Don't try to cook your dishes in a short time.

•Use the Internet. You can find many news and tasty recipes on the internet. You can also watch videos and learn how to cook easily.

•Organize your kitchen with the FIFO System. FIFO is "First In-First Out. It means using old products before using new ones. For example, you have some cheese in your fridge, and buy some cheese from the supermarket, too. The FIFO system tells you to use the old cheese first. All supermarkets work with the FIFO system. New yogurt packages are always at the bottom of the fridges in the supermarkets and old yogurt packages are always at the top.

What is the suitable title for this text?

- A. How to Become a Famous Chef
- B. How to make a delicious meal
- C. Tips for a Good Kitchen
- D. Home-made meals

Write True or False. If it is False, write the correct sentence.

1. We don't need good kitchen tools to cook easily.

2. Dirty ingredients give a good flavor to the dishes.

3. Following each step of a recipe is important.

4. Chefs taste their dishes after serving them.

5. Cooking fast is not a good way of cooking.

6. The internet can help us to cook.

TEXT - 3

Read the texts.

•People spend a lot of time in front of their screens, so they have many problems with their bodies. First, too much screen time is very bad for the eyes. These eye problems cause headaches. Also, people spend too much time on the Internet, and they don't do sports. According to the latest research, too much screen time is one of the biggest reasons for obesity. The Internet causes stress, too. People feel worried, and they become ill more often.

Main idea:

•The Internet helps people in many different ways. People can easily search for information on the Internet and learn new things. Also, communication on the Internet is very easy. We don't have to pay any extra money to talk and have video chats with our friends and family. The Internet helps us to save time, too. We don't waste our time paying the bills, doing the shopping or making arrangements when we use the Internet.

Main idea:

•Nearly all of the students love using the Internet. They do many different things online: They play games, use social media and watch videos. However, many students forget to do their homework because they are always online. They don't even want to go to school. They can't concentrate on their lessons. They aren't interested in their lessons, and they get very low grades. Teachers are worried about their students.

Main idea:

•People send each other friend requests on social media every day, but is this real friendship? Billions of people use social networking sites every day, and they only keep in touch with each other via these websites. They don't meet once. They chat online with each other, or they have video chats. This isn't real communication. Today's teens aren't good at face-to-face interaction, and they don't have many friends. Today's teens also have difficulties expressing themselves, and they are usually alone.

Main idea:

•The Internet is an important part of our lives. We use it all the time. It makes our lives easier. The Internet helps us to do research, keep in touch with each other, have fun, and share ideas. However, we should be very careful when we use it. The Internet can be very dangerous, or it can affect our lives negatively. We can have health problems, or we can meet bad people on the Internet. If we use the Internet carefully, it is very useful. If we don't use it carefully, it can be harmful.

Main idea:

Match the main ideas with the texts. One is extra.

- a. The Internet is the enemy of education.
- b. The Internet is bad for social life.
- c. The Internet has both its advantages and disadvantages.
- d. The Internet is very useful.
- e. The Internet is bad for health.
- f. The internet is time-consuming


Choose the correct options.

1. What health problems do people have when they spend too much time in front of the screen?
a. dry skin b. backache c. sore throat d. headache
2. For what purposes do students use the internet?
a. do shopping b. play games c. pay bills d. make appointments
3. Why do people use social networking sites?
a. having video chats b. watching videos
c. finding our way in traffic d. Searching for information
4. What is the text about?
a. safety problems of the Internet b. addiction to the Internet
c. positive and negative sides of using the internet. d. education opportunities of the internet

TEXT - 4

CELEBRITY PROFILE

We all know **Katniss Everdeen** in *The Hunger Games*, but who is the actress who plays her, **Jennifer Lawrence**?



Fast Facts

Name: **Jennifer Shrader Lawrence**
 Place of birth: **Kentucky, USA**
 Date of birth: **August 15, 1990**
 Profession: **TV and film actress**

Jen's family

Mother: **Karen Lawrence**
 Father: **Gary Lawrence**
 Brothers: **Ben and Blaine Lawrence**

Did you know?
 Jennifer has never had acting classes.

When she was a child Jennifer liked sports and she played hockey and basketball for an all-boys team. She also worked as a model. At the age of 14 she knew she wanted to be an actress, so she went to New York City to look for work. She appeared in advertisements for MTV and the fashion company H&M and got work as an actress on TV. Her family moved to Los Angeles so that Jennifer could work on TV and in films. In 2010 she acted in the film *Winter's Bone* and she was nominated for many awards including an Oscar. In 2012 she starred in the film *The Hunger Games* as Katniss Everdeen. When she isn't working, Jen likes painting, surfing and playing the guitar.

Write a number (1-8) and put the sentences into correct order 8 pts

- She moved to Los Angeles.
- She moved to New York.
- She appeared on TV for the first time.
- At school, she played basketball for a boys' team.
- She was nominated for an Oscar.
- Jennifer was born in Kentucky.
- She played Katniss Everdeen in *The Hunger Games*.
- She did modeling.

Answer the questions

1. What sports did Jennifer play at school?

2. What films has Jennifer Lawrence starred in?

3. What cities has she lived in?

4. What jobs has she done?

5. What does Jennifer do in her free time?

6. What are her two brothers called?

Read the text and circle the correct options for each blank.

Hello! My name is Jessica. I _____ (1) in London with my two _____ (2), Mark and Brian and my parents in my _____ (3), everybody does different _____ (4). For example, I'm in _____ (5) of doing the laundry _____ (6) the ironing. My mother is _____ (7) for preparing the meals and _____ (8) the floor. Doing the grocery _____ (9) and setting the table are my father's _____ (10). It's Mark's duty to _____ (11) care of the dog. He also _____ (12) the windows. Brian always takes out the _____ (13) and loads the dishwasher. We always _____ (14) responsibilities as a family _____ (15) we respect each other.

- | | | | |
|-------------------|--------------|-------------|--------------|
| 1. a) was | b) live | c) go | d) study |
| 2. a) brothers | b) friends | c) cousins | d) relatives |
| 3. a) relatives | b) siblings | c) family | d) neighbors |
| 4. a) activities | b) hobbies | c) routines | d) chores |
| 5. a) responsible | b) task | c) charge | d) fond |
| 6. a) but | b) because | c) so | d) and |
| 7. a) responsible | b) duty | c) change | d) happy |
| 8. a) washing | b) vacuuming | c) doing | d) setting |
| 9. a) shopping | b) doing | c) planning | d) taking |
| 10. a) jobs | b) interests | c) tasks | d) plans |
| 11. a) have | b) do | c) take | d) make |
| 12. a) has | b) cleans | c) vacuums | d) washes |
| 13. a) furniture | b) leaves | c) floor | d) garbage |
| 14. a) have | b) share | c) set | d) respect |
| 15. a) because | b) but | c) However | d) so |