Education and Science

Original Article

2025 Supplement 1 69-92

Cloze Tests in Measuring Reading Comprehension Levels *

M. Rahman Kalyoncu [1], Muhammet Memiş [2]

Abstract

This study, conducted as correlational research, aims to objectively examine the validity of cloze tests in Turkish, which are commonly used to assess reading comprehension levels, general language proficiency, and the readability of written materials, and to evaluate the procedures for using these tests to measure reading comprehension. The study investigates the consistency between multiple-choice reading comprehension tests, frequently used in national exams for their functionality and objectivity, and cloze tests designed and scored using various methods. During the study, a total of eight measurement tools were administered to a sample group of 90 seventh-grade students. These tools consisted of four multiple-choice reading comprehension tests based on four distinct texts and four cloze tests, each systematically deleting a word at a different position within the text. Two scoring methods were applied to the cloze tests, one considering only the exact words as correct and the other accepting alternative words that preserved the meaning of the sentence. Within this scope, approximately 23.000 test items were presented to the students in the study group, and around 43.000 evaluations were conducted on these items. Data collected during the 2023-2024 academic year were analyzed using Pearson correlation analysis, revealing a significant positive relationship between cloze tests and multiple-choice tests. Tests scored by considering only the exact words demonstrated greater consistency, while the correlation decreased when context-preserving alternatives were accepted as correct. The highest correlation occurred when every sixth word was systematically deleted. Based on the findings, it is recommended that in cloze tests, exact words should be accepted as correct instead of context-preserving words, every 6th word should be systematically deleted, and the tests should be systematically integrated into measurement and evaluation practices.

Keywords

Cloze test
Multiple-choice test
Reading comprehension
Correlation
Turkish

Article Info

Received: 09.30.2024
Accepted: 12.30.2024
Published Online: 03.03.2025

DOI: 10.15390/EB.2025.14079

* A part of this study was presented at the International Symposium on Measurement, Selection and Placement held between 4-6 October 2024 as an oral presentation.

[1] Sinop University, Faculty of Education, Department of Turkish and Social Sciences Education, Türkiye, mrkalyoncu@sinop.edu.tr

[2] Ondokuz Mayıs University, Faculty of Education, Department of Turkish and Social Sciences Education, Türkiye, muhammet.memis@omu.edu.tr

Introduction

Education is an indisputably significant phenomenon in human history. One of its essential components is measurement and evaluation. Systematic educational and instructional activities are designed to achieve specific objectives. Education that fails to meet its predetermined goals is considered ineffective and unsuccessful. At this point, the concepts of measurement and evaluation play a crucial role in assessing the quality and effectiveness of education. Measurement is defined as the process of determining whether an entity possesses certain characteristics or to what extent it does, with the results expressed using symbols (Tekin, 1982). Evaluation, on the other hand, is described as the process of comparing the results obtained from measurement with a specific criterion to reach a decision (Yılmaz, 1998).

One of the key characteristics that must be measured in the educational process is reading comprehension. This is because reading and understanding what is read form the foundation of a significant portion of educational activities (Balcı, 2016; Karatay, 2018). Consequently, individuals who fail to develop adequate reading comprehension skills tend to encounter academic challenges across various stages of their education (Akyol, 2020; Çelenk, 2006; Geske & Ozola, 2008; Uyanık, 2012). In this context, measuring reading comprehension skills is both highly important and necessary. However, the measurement of comprehension, particularly reading comprehension, has been a subject of debate for many years. This is because comprehension is a cognitive process, and the mental operations occurring in the brain cannot be directly observed. Therefore, all measurements related to comprehension are indirect and lack absolute certainty (Akyol, 2020). Under contemporary conditions, reading comprehension levels are measured by evaluating the products derived from a text, the responses provided to questions about the text, or through direct observation of the process itself.

The measurement of reading comprehension levels in educational institutions can be conducted using various tools, such as multiple-choice tests, open-ended questions, matching tests, and true-false tests. Among these, the most commonly used tools are open-ended and multiple-choice questions. Each test type has specific advantages and disadvantages relative to the other. Multiple-choice tests typically consist of a stem and a set of options, requiring students to select the option they believe is correct after reading the question stem. While these tests enable easy and objective scoring, they present certain challenges and limitations, particularly in the context of Turkish language education. In multiple-choice tests designed to measure reading comprehension levels, a paragraph is usually presented, followed by several related questions. However, due to their fixed structure and predefined options, such questions do not allow students to demonstrate creativity, interpret the content, or engage in deeper critical thinking about the material (Özbay, 1997; Temizkan & Sallabaş, 2011; Üstüner & Şengül, 2004). Additionally, since the correct answer is explicitly provided among the options, even students with limited reading skills, comprehension abilities, or familiarity with the text may still succeed in selecting the correct answer (Başaran, 2013; Katz & Lautenschlager, 1994; Özbay, 1997). As a result, multiple-choice tests are often ineffective for skill assessment (Aşılıoğlu, 1993), tend to measure factual knowledge rather than higher-order cognitive skills (Temizkan & Sallabaş, 2011), and fail to comprehensively assess reading comprehension (Ömeroğlu, 2018). Therefore, they are not entirely suitable as tools for a skill-based subject like Turkish. On the other hand, open-ended questions also present certain challenges, with subjectivity being the primary issue. Open-ended exams do not have a single correct answer, and the scores obtained from these questions are largely shaped by the scorer's judgment. As a result, these questions are highly influenced by the scorer's characteristics and opinions, making objective evaluation extremely difficult. Moreover, a language teacher who lacks sufficient knowledge, experience, and expertise may find it challenging to design valid and reliable exams using open-ended questions (Temizkan & Sallabaş, 2011). The same issue applies to multiple-choice tests, as crafting multiple-choice questions is a complex skill that requires significant expertise and experience. The frequent occurrence of errors in exams conducted with multiple-choice questions highlights this reality (Üstüner & Şengül, 2004). Given these disadvantages, the literature emphasizes that no single method is sufficient to comprehensively measure reading skills or can be considered the best approach

(Arıcı, 2018). As discussed, the tools currently in use exhibit clear weaknesses, including a lack of objectivity, the influence of chance, and difficulties in both design and scoring. However, there are alternative tools available that do not carry these weaknesses and can be utilized to measure reading comprehension levels effectively. In the continuation of this study, cloze tests will be introduced as an example of such tools. These tests are believed to address the shortcomings of existing methods and offer features capable of compensating for their deficiencies.

The cloze technique, one of the tools used to measure reading comprehension levels, was introduced in 1953 by Wilson Taylor (1953) in his study titled *Cloze Procedure: A New Tool for Measuring Readability*. This technique is fundamentally based on the Gestalt concept of closure, which refers to the mind's ability to complete incomplete phenomena (Ulusoy, 2009). Cloze tests are used to assess the readability of written material, an individual's reading level on a specific text, vocabulary knowledge in a particular field or subject, language proficiency, general comprehension level, and the overall reading profile of a class or small group (Loewe, 1983; Mariotti & Homan, 2001). In their simplest form, cloze tests involve systematically deleting every "n-th" word in a text (Harmer, 2002). Since their introduction to the literature, they have been the subject of numerous studies and have been used as tools for various purposes (Oller, Bowen, Dien, & Mason, 1972). These include addressing students' reading difficulties (Dağ, 2010), identifying individual differences and perceptions related to a subject (Manis & Dawes, 1961), enhancing language proficiency (Booth, 1998; James, 2004), improving students' reading comprehension skills (Sukarni, 2021; Wahdaniah, Marbun, & Husin, 2013), identifying readers with different linguistic backgrounds (Craker, 1971), and measuring text difficulty in the development of readability formulas (Coleman, 1965; Çetinkaya, 2010). Moreover, in the international literature, some researchers (Carvalho & Souza, 2023) have noted that cloze tests are among the most commonly used tools for measuring reading comprehension levels.

In cloze tests, readers fill in the blanks in a text by reading, interpreting, and making sense of the content. The percentage of correctly completed blanks constitutes the reader's score; as the score decreases, the text is considered more difficult for the reader (Dubay, 2007). However, cloze tests present certain challenges for students. These include cases where students avoid answering the test due to its difficulty, leading to inconsistencies in obtaining valid and reliable results (Karatay, Bolat, & Güngör, 2013; Mariotti & Homan, 2009). Furthermore, even skilled readers generally achieve no more than 65% success on easier texts (Bormuth, 1966; Froese, 1971). In this context, valid scores should not be expected from students encountering cloze tests for the first time, and students should be familiarized with the format before the main application (Ruddell, 2005; Vacca & Vacca, 2005). Similarly, a single cloze test cannot be expected to comprehensively measure students' language proficiency (Brown & Grüter, 2020). On the other hand, the high objectivity of cloze tests, their ease of preparation, simplicity in use and scoring, and most importantly, the ability to use the text itself as the testing tool (Çetinkaya, 2010) make them functional and widely used tools. This functionality has contributed to their widespread application. Nevertheless, it remains challenging to refer to a standardized approach in the preparation, administration, and scoring of cloze tests.

Before delving into the preparation and administration processes of cloze tests, it is important to note that numerous distinct procedures, scoring methods, preparation strategies, perspectives, and findings are associated with this topic. One of the first considerations is the length of the text to be converted into a cloze test. This length is primarily determined by the number of blanks to be included in the test and the regular interval at which words will be deleted. The most widely accepted approach in the international literature recommends using a 300-400-word text, leaving the first and last sentences intact, and systematically deleting every fifth word until 50 blanks are obtained. However, these standards and guidelines are subject to debate. For instance, Nation (2009) argues that the ideal number of blanks in a cloze test should range from 40 to 50, while Shahnazari-Dorcheh, Roshan, and Hesabi (2012) suggest that the number of blanks should be adjusted according to the reader's proficiency level, with 25-30 blanks being optimal for beginners. The procedure for word deletion in cloze tests is also a topic of debate. Cloze tests can be prepared in two formats: fixed-ratio deletion or irregular deletion

(Koda, 2005). In fixed-ratio deletion, words are systematically removed at regular intervals, whereas in irregular deletion, specific word types or contextually significant words are removed. Regarding fixed-ratio deletion, Bormuth (1964) argues that, in a cloze test designed for a native language, deleting a word to the right or left of the "n-th" word makes no significant difference. Oller et al. (1972) assert that irregular deletion can produce valid results but is not practical for widespread use, while Bachman (1982) supports systematic deletion at fixed intervals. On the other hand, some researchers contend that tests prepared using fixed-ratio deletion focus more on surface-level or sentence-level meanings rather than overall comprehension, making them inadequate for fully measuring linguistic competence (Carlisle & Rice, 2004). They argue that such tests often reflect memory performance rather than comprehension ability, suggesting that irregular deletion is a more valid approach (Alderson, 2000; Cain & Oakhill, 2006). However, it should also be acknowledged that each language has unique characteristics, and standards established for one language may not yield reliable results when applied to another (Kalyoncu & Memiş, 2024; Kurudayıoğlu & Karadağ, 2005). In this context, it becomes a scientific necessity to test these perspectives and standards specifically for Turkish, as such studies have not yet been conducted.

Regardless of the ongoing debates in the literature, the general process for creating cloze tests can be outlined as follows: Cloze tests are typically constructed by selecting a 300-400-word text unfamiliar to students, leaving the first and last sentences intact, and systematically deleting words at predetermined intervals until 50 blanks are obtained. Blanks are left in place of the deleted words, and their lengths should remain consistent throughout the text (Stubbs & Tucker, 1974). For Turkish, the blanks should ideally be 12-character spaces (Çetinkaya, 2010). During the blanking process, if the word to be deleted is a proper noun, it is skipped, and the next word is converted into a blank. By following these procedures, an ideal cloze test can be produced. Thus, it can be concluded that even individuals without extensive experience in measurement, evaluation, or educational applications can easily prepare a high-quality cloze test by adhering to these guidelines.

A review of the literature reveals numerous studies (Abraham & Chapelle, 1992; Darnell, 1970; Hinofotis, 1980; Irvine, Atai, & Oller, 1974; Williams, Ari, & Santamaria, 2011) that examine the correlation between cloze tests and standardized reading comprehension tests, consistently identifying high correlations between the two test types. Beyond these strong correlations, another notable strength of cloze tests is their ability to yield valid results across a wide range of languages with distinct characteristics, such as French (Tremblay & Garrison, 2010), Spanish (Vari-Cartier, 1980), Portuguese (Suehiro & Santos, 2015), Indonesian (Sukarni, 2021), Korean (Taylor & Lee, 1954), Thai (Oller et al., 1972), Arabic (Abanami, 1982), and Japanese (Shiba, 1957), thereby demonstrating their universal applicability. Additionally, the ease of creating, administering, and managing cloze tests, combined with their material flexibility and objectivity, makes them highly practical for educational purposes. However, as previously noted, regardless of their validity, it cannot be assumed that a standard or procedure developed for a different language will yield accurate results when applied to Turkish. Therefore, the validity of these standards and procedures must also be specifically tested for Turkish. The primary reason for this necessity lies in the differences between Turkish and many other languages in terms of average sentence length and syntax. For instance, according to R. Flesch (1948), the average sentence length in English is 14-15 words, whereas Ateşman (1997) reports that the average sentence length in Turkish is 9-10 words. This difference may stem from variations in the nature of word order within sentences. Additionally, differences in syntax naturally affect the placement of sentence elements. Consequently, the sequence of words to be deleted in Turkish texts is also likely to differ from that in other languages.

An examination of studies conducted in the national literature on cloze tests reveals a lack of consensus regarding a standardized cloze procedure for Turkish. Furthermore, the consistency of these tests with standardized objective tools has not been sufficiently investigated. Existing studies primarily focus on evaluating the consistency of cloze tests using less objective measurement tools or subjective opinions, determining which order of word deletion yield more consistent results, and debating whether synonyms should be accepted as correct answers. For instance, in his study, Ulusoy (2009) compared students' scores on cloze tests with teachers' evaluations of the students and concluded that tests in which every 5th word was deleted did not yield reliable results, whereas tests in which every 6th word was deleted were more reliable. Conversely, Kaplan and Çiftçi (2021) argued that not accepting synonyms as correct answers in cloze tests creates a disadvantage. They found that the "substitute word" method, in which every 4th word is deleted and context-preserving responses are accepted as correct, produced more consistent results. Similarly, Tunçer and Erden (2015) created two cloze tests in which every 10th word was deleted, with the first letter of the deleted words provided as a clue in one of the tests. In their study examining the correlation between cloze test scores and open-ended questions, they reported that the test with first-letter clues yielded more consistent results. Likewise, Uyanık (2011), who analyzed the correlation between cloze tests and a test containing six open-ended reading comprehension questions, found that deleting the 7th and 8th words in a text resulted in a higher correlation than tests in which the 6th word was deleted.

In conclusion, a review of the national literature reveals that there is no clear procedure specifically designed for cloze tests in Turkish, and the validity of the methods employed has not been tested using objective tools. Based on this justification, this study aims to examine the consistency between cloze tests prepared using different procedures and validated, reliable multiple-choice tests, to determine whether cloze tests yield valid results in measuring reading comprehension levels in Turkish, and to establish a clear procedure for the preparation, administration, and scoring of these tests. In line with this objective, the research seeks to answer the following questions:

1. Is there a significant relationship between the scores of cloze tests, where every 4th word in the same text is deleted and only exact words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?
2. Is there a significant relationship between the scores of cloze tests, where every 4th word in the same text is deleted and contextual words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?
3. Is there a significant relationship between the scores of cloze tests, where every 5th word in the same text is deleted and only exact words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?
4. Is there a significant relationship between the scores of cloze tests, where every 5th word in the same text is deleted and contextual words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?
5. Is there a significant relationship between the scores of cloze tests, where every 6th word in the same text is deleted and only exact words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?
6. Is there a significant relationship between the scores of cloze tests, where every 6th word in the same text is deleted and contextual words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?
7. Is there a significant relationship between the scores of cloze tests, where every 7th word in the same text is deleted and only exact words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?
8. Is there a significant relationship between the scores of cloze tests, where every 7th word in the same text is deleted and contextual words are accepted as correct, and the scores of multiple-choice reading comprehension tests for the students in the study group?

Method

The Model of the Research

This study employed correlational research, a quantitative research method, to investigate whether cloze tests accurately and reliably measure reading comprehension levels in Turkish texts and to evaluate the proposed procedures for the preparation, administration, and scoring of cloze tests. Correlational research aims to identify, describe, and provide insights into cause-and-effect relationships between two or more variables (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz, & Demirel, 2023; Karakaya, 2014; Özdemir & Doğruöz, 2020). To better understand these relationships, correlational studies often examine the associations between different variables collected from the same individuals in natural settings (Mertens, 2015; Tuncer, 2020).

In this context, correlational research was deemed the most appropriate method for the objectives and processes of this study. This approach was selected because the study utilized two different measurement tools for each of four distinct texts, one of which was scored in two different ways, and the correlations between these tools were analyzed. Figure 1 below illustrates the research model designed for this study:

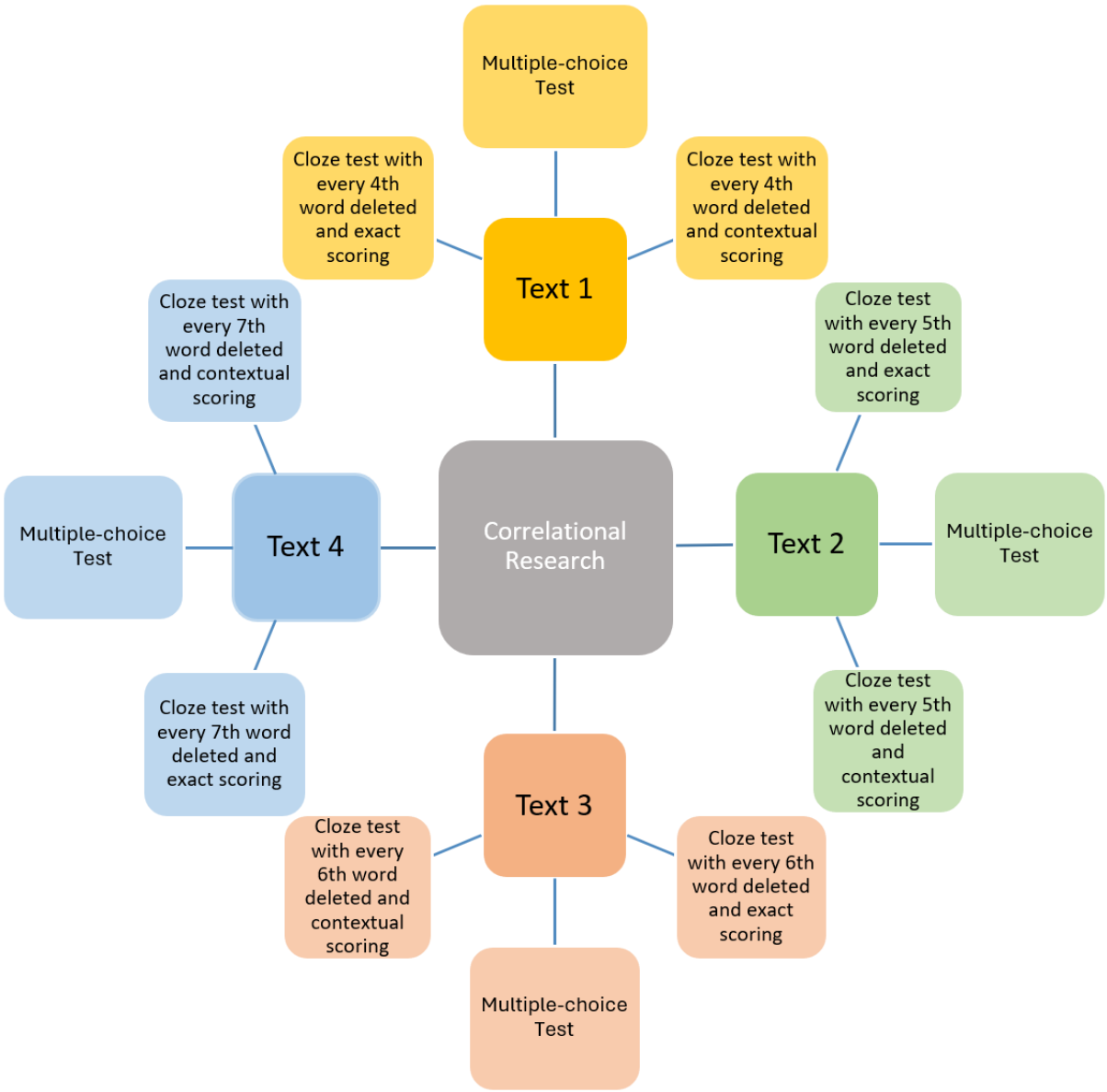

Figure 1. Research Model

When Figure 1 is examined, it becomes evident that the study was conducted using four distinct texts. For each of these texts, two different measurement tools were prepared: one being a multiple-choice reading comprehension test and the other a cloze test. The cloze tests were scored using two distinct methods: one accepting only the exact words as correct and the other accepting context-preserving words as correct. Accordingly, the study utilized eight different measurement tools and involved twelve distinct evaluation processes across the four texts.

Study Group

The study group consisted of 90 seventh-grade students from schools located in the provinces of Samsun and Balıkesir. The students were selected on a voluntary basis from schools that were easily accessible. However, due to the structure of the research, each student in the study group was required to participate in all tests prepared for the same text. Students who did not complete all the tests for a particular text were excluded from the evaluation process. As a result, only the scores of students who participated in all the tests for a single text were considered for the correlation analysis. To ensure reliable correlation comparisons, the study group size was adjusted to match the number of participants in the tests with the lowest participation rate for a given text, which was 60. The students who participated fully but were removed from the study group in order to equalize the groups were removed equally between the highest and lowest students according to their scores. After these adjustments, the final study group consisted of 60 seventh-grade students, and the data obtained from this group were analyzed. Details regarding the characteristics of the study group are presented in Table 1 below:

Table 1. Demographic information of the Study Group

Variable	***f***	**Percentage**
Grade (7)	60	100%
Gender (Male)	22	36.6%
Gender (Female)	38	63.4%
City (Samsun)	36	60%
City (Balıkesir)	24	40%

When Table 1 is examined, it is observed that 22 students (36.6%) in the study group are male, while 38 students (63.4%) are female. Additionally, 36 students (60%) are studying in Samsun, and 24 students (40%) are receiving their education in Balıkesir.

Data Collection Tools

In this study, eight different data collection tools related to four distinct texts were utilized. The primary source for these tools is the *Informative Text Reading Comprehension Scale* developed by Temizkan (2007). This scale was created as part of Temizkan's doctoral dissertation to measure reading comprehension levels. The scale includes four texts titled "*Eğer Gençlik Bilseydi*", "*Okumak*", "*Roman Okumak*", and "*Okul ve Meslek Seçimi*". Each text is accompanied by 14 multiple-choice reading comprehension questions, resulting in a total of 56 questions. The validity and reliability of this scale were tested and confirmed in the aforementioned research. The item discrimination indices (rjx) calculated for the test questions in that study are presented in Table 2 below.

Table 2. Item Discrimination Indices of the Informative Text Reading Comprehension Scale

Item	Eğer Gençlik Bilseydi	Okumak	Roman Okumak	Okul ve Meslek Seçimi
	rjx	rjx	rjx	rjx
1	.33	.83	.33	.33
2	.25	.33	.50	.50
3	.25	.25	.83	.83
4	.50	.25	.83	.83
5	.50	.50	.41	.41
6	.83	.33	.33	.33
7	.83	.41	.83	.83
8	.33	.41	.25	.25
9	.25	.33	.41	.41
10	.33	.25	.33	.33
11	.33	.33	.83	.83
12	.83	.41	.83	.83
13	.41	.50	.33	.33
14	.50	.25	.83	.83

An item's discrimination index (rjx) between 0.20 and 0.30 indicates moderate discrimination, while values above 0.30 signify good discrimination (Güler, 2019). In this context, it was observed that 9 of the 56 items in the tests had moderate discrimination, while the remaining 47 items exhibited good discrimination. In addition to the item discrimination indices calculated by Temizkan (2007), the validity and reliability of the same scale were re-evaluated by Özyılmaz (2010), who found that the KR-20 value was .85. Given that a reliability coefficient of 0.70 or higher is considered sufficient for the reliability of a test (Büyüköztürk, 2006), it can be stated that the scale is reliable. Furthermore, Özyılmaz (2010) consulted expert opinions regarding the level of the test and concluded that it is suitable for seventh-grade students. In this context, it can be concluded that the scale is an appropriate tool for measuring reading comprehension levels and can be administered to seventh-grade students.

In addition to the multiple-choice tests described, cloze tests were also necessary during the research process. These cloze tests were created using texts from the "Informative Text Reading Comprehension Scale," ensuring that the tests corresponded to the same texts. The scale includes four texts: "Eğer Gençlik Bilseydi" (300 words), "Okumak" (313 words), "Okul ve Meslek Seçimi" (368 words), and "Roman Okumak" (370 words).

The text "Eğer Gençlik Bilseydi" was converted into a cloze test by systematically deleting every 4th word, "Okumak" by deleting every 5th word, "Roman Okumak" by deleting every 6th word, and "Okul ve Meslek Seçimi" by deleting every 7th word. During the conversion process, the integrity of the first and last sentences was preserved, and blanks corresponding to proper nouns were skipped. Furthermore, the blanks were standardized to a length of 12 character spaces throughout the texts. A sample excerpt from the cloze tests obtained using the described procedure is presented in Table 3 below.

Table 3. Sample Excerpts from the Cloze Tests Prepared for the Study

Eğer Gençlik Bilseydi	Okumak	Roman Okumak	Okul ve Meslek Seçimi
Fransızların bir sözü vardır: "Gençlik bilseydi, ihtiyarlık yapabilseydi." derler. Ne yazık ki _____ bilmez. Bilmediği için _____ yapabileceği birçok şeyi _____ artık …	Okumak insan için bir zevk, bir eğlence olduğu kadar, hiç kuşkusuz eğitici bir eylemdir de. Bilgilerimizi artırarak, aklımızı işleterek; _____, görüşümüzü genişleten bir eylemdir. _____ insan olayları değerlendirmede, çevresini _____, yaşamın ….	Kişileri roman okumayı sevenlerle roman okumayı sevmeyenler diye ikiye ayırabiliriz. Roman okumayı sevmeyenlerden bir hayır _____ demiyorum, büyük işlere asıl onların _____ söyleseler ona da inanırım. Ama _____ hoşlanmam onlardan. …	Her ders yılı sonunda birçok anneler, babalar, çocuklar, gençler tasalanırlar. Anneler, babalar, öğrenimlerini bir üst derecedeki _____ devam ettirmek isteyen çocuklarını nereye vereceklerini _____ veya seçtikleri okullara göndermek imkânı bulamazlar. _____ fazla …

In cloze tests, the characteristics of the words deleted from the texts are highly significant. This is because variations in the characteristics of these words can influence participants' performance. The characteristics of the words deleted in the cloze tests created during the research process are presented in Table 4 below.

Table 4. Types of Words Deleted in Cloze Tests

Word Classes	Eğer Gençlik Bilseydi	Okumak	Roman Okumak	Okul ve Meslek Seçimi
Noun (*f*)	17 (34%)	21 (42%)	17 (34%)	14 (28%)
Verb (*f*)	11 (22%)	4 (8%)	10 (20%)	4 (8%)
Verbal (*f*)	6 (12%)	10 (20%)	4 (8%)	9 (18%)
Adjective (*f*)	8 (16%)	3 (6%)	6 (12%)	12 (24%)
Adverb (*f*)	3 (6%)	1 (2%)	2 (4%)	4 (8%)
Pronoun (*f*)	-	5 (10%)	7 (14%)	6 (12%)
Conjunction (*f*)	3 (6%)	4 (8%)	3 (6%)	-
Preposition (*f*)	2 (4%)	2 (4%)	1 (2%)	1 (2%)
Total (*f*)	50	50	50	50

When Table 4 is examined, it is evident that nouns constitute the majority of the words deleted in all tests. Similarly, prepositions, conjunctions, and adverbs are the least frequently deleted word types. On the other hand, the frequency of deleted verbs, verbals, pronouns, and adjectives varies across texts. Considering the potential impact of the characteristics of the deleted words on the scores obtained from the tests, it can be stated that variations in word attributes represent a limitation of the study.

Another point related to the cloze tests created in the study is the comprehension levels at which the test items operate. The cloze test items do not measure comprehension at a single level. This is because these tests require the reader to first read the entire test, develop an understanding of the topic, analyze the author's purpose, thought structure, style, and even personality, establish relationships between different parts of the text, interpret the structures within the text, and fill in the blanks using vocabulary knowledge. Subsequently, the reader is expected to evaluate the suitability of the word they provide within the context of the text. Thus, it can be stated that cloze tests incorporate the comprehension levels defined in Barrett's Taxonomy, including literal comprehension, recognition, recall, inferential understanding, and evaluation, as well as the stages of Bloom's Taxonomy, such as remembering, understanding, applying, and analyzing. However, it is not entirely possible to distinctly separate the measurement levels of the items within cloze tests. This is because all items are created in the same format and with the same expectations, and readers often experience similar processes across

many of the items. Nevertheless, due to the nature of fixed-ratio deletion procedures, there are cases where questions can be solved through literal or sentence-level comprehension alone. That said, when the tests are evaluated as a whole, the relatively low occurrence of such questions and the significant proportion of items requiring inferential understanding make this issue negligible. For example, the proportions of items that can be solved at the sentence level or through recognizing word patterns, versus those requiring inferential understanding, in the cloze tests created for this study are presented in Table 5 below.

Table 5. Measurement Levels of Items in Cloze Tests

Comprehension	**Eğer Gençlik Bilseydi**		**Okumak**		**Roman Okumak**		**Okul ve Meslek Seçimi**	
Level	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**
Literal Comprehension	3	6	5	10	7	14	5	10
Inferential Comprehension	47	94	45	90	43	86	45	90

When Table 5 is examined, it is observed that 86% to 94% of the items in the cloze tests measure inferential comprehension levels, requiring an understanding of the overall structure of the text, analysis, establishing relationships between sentences and paragraphs, and making predictions. Upon examining the details of the texts, it is found that the text titled "Eğer Gençlik Bilseydi" contains 3 items (I28, I32, I47), the text titled "Okumak" includes 5 items (I5, I21, I26, I27, I37), the text titled "Roman Okumak" contains 7 items (I1, I6, I8, I17, I25, I35, I45), and the text titled "Okul ve Meslek Seçimi" includes 5 items (I11, I14, I15, I36, I39) that do not require an understanding of the overall meaning of the text but can be answered by understanding the sentence or recognizing the word pattern.

Table 6. KR-20 Reliability Coefficients of Cloze Tests Created Within the Study

Texts	**Exact Scoring**	**Contextual Scoring**
Eğer Gençlik Bilseydi	0.72	0.87
Okumak	0.75	0.89
Roman Okumak	0.73	0.87
Okul ve Meslek Seçimi	0.71	0.84

Table 6 presents the KR-20 values of the cloze tests created during the research process, based on the validated and reliable *"Informative Text Reading Comprehension Scale"*. The KR-20 formula is used when responses to each item in a test are scored as 1 (correct) or 0 (incorrect) (Arastaman & Kısa, 2020; Büyüköztürk et al., 2023). When the reliability coefficients are examined, it is observed that the KR-20 values for all tests, each consisting of 50 items, exceed the threshold of 0.70. Considering that a reliability coefficient of 0.70 or higher is sufficient to establish the reliability of a test (Büyüköztürk, 2006), it can be concluded that the cloze tests developed for this study are reliable measurement tools.

Table 7. Descriptive Analyses of the Tests

Test	N	Items	Mean	Median	Mode	Sd	Variance	Min	Max
Eğer Gençlik Bilseydi Multiple-choice Test	60	14	8.5	9	10	1.7	3.1	3	12
Okumak Multiple-choice Test	60	14	7	7	8	2	2	4	10
Roman Okumak Multiple-choice Test	60	14	9	10	11	2.8	7.8	2	14
Okul ve Meslek Seçimi Multiple-choice Test	60	14	9	10	11	2	6	4	13
Eğer Gençlik Bilseydi Exact Scoring Cloze	60	50	9.6	9	7	4.3	18.9	1	19
Okumak Exact Scoring Cloze	60	50	6	6	1	4	18	0	20
Roman Okumak Exact Scoring Cloze	60	50	5.8	5	5	3.5	12	1	15
Okul ve Meslek Seçimi Exact Scoring Cloze	60	50	6	6	4	3	12	1	15
Eğer Gençlik Bilseydi Contextual Scoring Cloze	60	50	17.1	17	13	8.5	72	2	39
Okumak Contextual Scoring Cloze	60	50	15.4	15	18	8.3	69	2	36
Roman Okumak Contextual Scoring Cloze	60	50	19	19	15	9	86	1	37
Okul ve Meslek Seçimi Contextual Scoring Cloze	60	50	16	16	16	7	54	2	36

Table 7 presents the descriptive statistics for three different categories of measurement tools. An analysis of these statistics reveals that the average difficulty levels of tools with different structures vary. The average difficulty levels for the multiple-choice tests are 0.6 for the text "Eğer Gençlik Bilseydi", 0.5 for "Okumak", 0.64 for "Roman Okumak", and 0.64 for "Okul ve Meslek Seçimi". For the cloze tests where only exact words are accepted as correct, the average difficulty levels are 0.19 for " Eğer Gençlik Bilseydi", 0.12 for " Okumak", 0.11 for " Roman Okumak" and 0.12 for " Okul ve Meslek Seçimi". For the cloze tests where context-preserving words are accepted as correct, the average difficulty levels are 0.34 for " Eğer Gençlik Bilseydi", 0.3 for " Okumak", 0.38 for " Roman Okumak", and 0.32 for " Okul ve Meslek Seçimi". This demonstrates, as supported in the literature, that cloze tests are more challenging compared to conventional and standardized reading comprehension tests. Additionally, all items in all tests were assigned the same point value. For multiple-choice tests consisting of 14 questions, the total number of correct answers was multiplied by 7.14. For cloze tests consisting of 50 items, the total number of correct answers was multiplied by 2. Consequently, all tests were evaluated on a scale of 100 points.

Using the aforementioned data collection tools, four different reading comprehension tests, each containing 14 multiple-choice reading comprehension questions, and four different cloze tests prepared using distinct procedures, were developed. In this context, the data for the study were collected using eight different data collection tools.

Collection and Analysis of Data

The data were collected by the researchers over approximately four months during the 2023-2024 academic year through face-to-face sessions with students. Initially, students practiced cloze test examples that were not used as measurement tools, during two class hours each, to help them become familiar with the application process. Following this, the students were presented with the cloze test versions of the texts, with one class hour allocated for each cloze test. After this phase, no further applications related to this research were conducted with the students for approximately two weeks. Subsequently, the tests from the "Informative Text Reading Comprehension Scale" were administered to the students in sequence. With the completion of this application, the data collection process was finalized.

Once data from both types of tests were obtained, the data analysis process began. In the first step of the analysis, the normality distributions of the scores obtained from the tests were examined. Since two different scoring methods were used for the cloze tests, the scores from eight different measurement tools generated twelve separate results, all of which were subjected to normality analysis. The results of these normality analyses are presented in Table 8 below.

Table 8. Normality Distributions of Scores Obtained from the Tests

Tests	Skewness	Kurtosis
"Eğer Gençlik Bilseydi" Multiple-choice Test	-.646	.539
"Okumak" Multiple-choice Test	-.364	-.551
"Roman Okumak" Multiple-choice Test	-.635	-0.63
"Okul ve Meslek Seçimi" Multiple-choice Test	-.478	-.618
"Eğer Gençlik Bilseydi" Exact Scoring Cloze	.162	-.655
"Okumak" Exact Scoring Cloze	.738	.665
"Roman Okumak" Exact Scoring Cloze	.577	-.345
"Okul ve Meslek Seçimi" Exact Scoring Cloze	.561	-293
"Eğer Gençlik Bilseydi" Contextual Scoring Cloze	.308	-.593
"Okumak" Contextual Scoring Cloze	.082	-.624
"Roman Okumak" Contextual Scoring Cloze	.355	-.218
"Okul ve Meslek Seçimi" Contextual Scoring Cloze	.428	.065

When Table 8 is examined, it is observed that the skewness and kurtosis values for all tests range between -0.7 and 0.8. Considering that skewness values outside the range of -1 to +1 generally indicate a significantly skewed distribution (Hair, Black, Babin, Anderson, & Tatham, 2013) and that kurtosis values within the range of ±1.0 are considered excellent for most purposes (George & Mallery, 2010), it can be concluded that the data obtained through the tests exhibit a normal distribution. Therefore, the correlation analyses conducted in line with the study's objectives were calculated based on normal distributions. The thresholds considered in the evaluation of the calculated correlation coefficients during the research process are presented in Table 9 (Köklü, Büyüköztürk, & Çokluk, 2007; Taşpınar, 2017; Taylor, 1990).

Table 9. Interpretation of Correlation Coefficients

Value	Interpretation
0.00 – 0.19	Very weak relationship
0.20 – 0.39	Weak relationship
0.40 – 0.69	Moderate relationship
0.70 – 0.89	Strong relationship
0.90 – 1.00	Very strong relationship

In line with Table 9, correlation values between 0.20 and 0.39 were interpreted as weak relationships, values between 0.40 and 0.69 as moderate relationships, and values between 0.70 and 0.89 as strong relationships.

Ethics Committee Approval

The data collection procedure and the collected data within the scope of this study were approved by the Social and Human Sciences Ethics Committee of Bartın University under protocol code 2024-SBB-0074, confirming that they do not involve any ethical issues. No personal data were collected from participants, who took part in the study on a voluntary basis, and all ethical principles were adhered to throughout the research process.

Results

The findings obtained during the research process are presented below in alignment with the order of the research questions.

Table 10. Pearson Correlation Analysis Results Between the Multiple-Choice Reading Comprehension Test and the Cloze Test with Every 4th Word Systematically Deleted for the Same Text

	Mean	Sd	N	1	2
1 Multiple-Choice Test	61.3	12.63	60	1	
2 Exact Scoring Cloze	19.2	8.71	60	.538**	1
3 Contextual Scoring Cloze	34.3	17	60	.587**	.850**

**p<0.01

When Table 10 is examined, it is observed that there is a moderate positive relationship between the multiple-choice reading comprehension test and the cloze test in which every 4th word was systematically deleted for the same text. This level of relationship is evident in both scoring methods: one in which only exact words are accepted as correct and the other in which contextual words are accepted as correct. However, an analysis of the correlation coefficients reveals that accepting contextual words as correct yields more consistent results with the multiple-choice tests than the scoring method that accepts only exact words. Additionally, it was found that the cloze test scores obtained using the scoring method that accepts contextual words as correct and the scores obtained using the scoring method that accepts only exact words are highly positively correlated, with a correlation coefficient of .850.

Table 11. Pearson Correlation Analysis Results Between the Multiple-Choice Reading Comprehension Test and the Cloze Test with Every 5th Word Systematically Deleted for the Same Text

	Mean	Sd	N	1	2
1 Multiple-Choice Test	50.11	10.8	60	1	
2 Exact Scoring Cloze	11.97	8.42	60	.383**	1
3 Contextual Scoring Cloze	37.87	18.57	60	.322*	.706**

**p<0.01; *p<0.05

When Table 11 is examined, it is observed that, unlike the moderate significant relationship seen in Table 10, there is a weak significant relationship between the multiple-choice tests and the cloze tests where every 5th word was systematically deleted. This relationship is consistent across both scoring methods used in the cloze tests; however, the scoring procedure that accepts only exact words as correct demonstrates a slightly higher correlation coefficient of .383. When the relationship between the two scoring methods themselves is analyzed, it is found that the methods show a high level of consistency, with a correlation coefficient of .706.

Table 12. Pearson Correlation Analysis Results Between the Multiple-Choice Reading Comprehension Test and the Cloze Test Where Every 6th Word Was Systematically Deleted for the Same Text

		Mean	Sd	N	1	2
1	Multiple-Choice Test	64.76	20.06	60	1	
2	Exact Scoring Cloze	11.6	6.92	60	.639**	1
3	Contextual Scoring Cloze	30.7	16.7	60	.637**	.707**

**p<0.01

Table 12 presents the correlations between the cloze test, in which every 6th word was systematically deleted and scored using different methods, and the multiple-choice test for the same text. When Table 12 is examined, it is found that, similar to the cloze tests where every 4th word was deleted, there is a moderate significant relationship between the cloze tests where every 6th word was deleted and the multiple-choice tests. Additionally, it is noteworthy that this relationship is quite similar across both scoring methods used. Furthermore, there is a high-level correlation of .707 between the two scoring methods themselves. However, the cloze test where every 6th word was deleted demonstrates greater consistency with the multiple-choice tests compared to the cloze tests where every 4th or 5th word was deleted.

Table 13. Pearson Correlation Analysis Results Between the Multiple-Choice Reading Comprehension Test and the Cloze Test with Every 7th Word Systematically Deleted for the Same Text

		Mean	Sd	N	1	2
1	Multiple-Choice Test	65.47	17.25	60	1	
2	Exact Scoring Cloze	12.97	6.89	60	.416**	1
3	Contextual Scoring Cloze	32.93	14.66	60	.405**	.860**

**p<0.01

When Table 13 is examined, it is observed that the cloze tests where every 7th word was systematically deleted also show a moderate significant consistency with the multiple-choice tests. Furthermore, it is noteworthy that the consistency of the results obtained using different scoring methods in this test is similar to that observed with the multiple-choice tests. However, as with the cloze tests where the 5th and 6th words were systematically deleted, it is found that the scoring method accepting only exact words as correct demonstrates a higher correlation coefficient.

Discussion and Conclusion

The conclusions of this study, conducted to determine whether cloze tests effectively measure reading comprehension levels in Turkish and to test the proposed procedures for administering cloze tests, are as follows:

In the study, data collection tools were first prepared, and then the data collection process was initiated. During this process, the same systematic approach was followed for all applications. This approach involved first presenting the students with the cloze test versions of the texts, and after a two-week waiting period, administering tests containing 14 multiple-choice questions related to the same texts. In line with the described application process, the correlations between the cloze test, in which every 4th word was deleted and scored using different methods, and the multiple-choice test for the same text were initially examined. The analysis revealed a moderate, positive, and significant relationship between the two tests, regardless of the scoring procedure of the cloze test. Subsequently, the same procedures were applied to the cloze tests where every 5th word was systematically deleted, and a weak positive relationship was identified between the cloze tests and the multiple-choice reading comprehension test. Continuing with the same approach, a moderate positive relationship was also found between the cloze tests, in which every 6th and 7th words were systematically deleted, and the multiple-choice tests, regardless of the scoring method. Overall, the study concluded that there is a moderate, significant relationship between cloze tests and multiple-choice tests. The scoring method that accepts only exact words as correct yielded results that were relatively more consistent with the multiple-choice tests. The highest consistency with the multiple-choice tests was achieved with the cloze test where every 6th word was systematically deleted and only exact words were accepted as correct. However, it is also noteworthy that in the application with the highest correlations, where every 6th word was systematically deleted, the differences between the results obtained using different scoring methods were minimal.

Within the framework of the processes described, it has been determined that there is a significant relationship between all the cloze tests employed and the multiple-choice reading comprehension tests. While the nature of this relationship varies depending on the tests, it is generally moderately positive. Accordingly, it can be stated that cloze tests are consistent with multiple-choice tests, which are frequently used to measure reading comprehension and are a dominant question type in national exams, and that cloze tests measure reading comprehension at a similar level to multiple-choice tests. In this context, the findings of this study align with previous research indicating that cloze tests can yield valid results for different languages (Klare, 1974; Suehiro & Santos, 2015; Sukarni, 2021; Taylor, 1956; Tremblay & Garrison, 2010), can be used to determine text readability (Bormuth, 1963; Bormuth, 1967; Klare, Simaiko & Stolurow, 1972), and can assess students' reading comprehension levels (Brown, 1982; Brown & Grüter, 2020; Febriyanti, 2017; Huensch, 2013; Şahindokuyucu, 2006; Lu, 2006; Oller, 2006). Similarly, these results are consistent with studies conducted in the national literature in Turkish, which suggest that cloze tests can provide valid results in identifying students' reading levels (Tunçer & Erden, 2015; Ulusoy, 2009; Uyanık, 2012), measuring reading comprehension levels (Akyol, 2020; Çetinkaya, 2010; Hızarcı, 2009), and determining text readability (Çepni, Gökdere, & Küçük, 2002; Keskin & Akıllı, 2013; Köse, 2009). Consequently, it is concluded that cloze tests, with their practical and objective application features and their ability to be directly applied to almost any text without requiring additional tools, can be effectively used to measure reading comprehension levels and determine text readability.

Although cloze tests are considered highly functional tools for measuring reading comprehension levels, certain contentious issues regarding these tests persist in the relevant literature. One of the primary debates centers around which word positions should be deleted during the preparation of cloze tests. Bormuth (1964) argued that in cloze tests developed for native languages, deleting the n-th word and replacing it with either the word to its left or right would yield equivalent results. Similarly, Bachman (1982) stated that testing the specific details of cloze tests is unnecessary and that valid results can be achieved through systematic deletions. On the other hand, Oller et al. (1972) noted that while irregular deletions can also produce valid results, they are less practical for regular use. Furthermore, Potter (1968) suggested that deleting more than 20% of a text makes the test excessively difficult. In addition to these discussions in the international literature, similar debates exist in the national literature concerning Turkish. For example, Kaplan and Çiftçi (2021), who examined the consistency between reading levels determined through "error inventory analysis" (Akyol, 2020) and cloze tests, reported the most consistent results from cloze tests where every 4th word was systematically deleted, among tests where the 4th, 5th, and 6th words were deleted. On the other hand, Uyanık (2011), who analyzed the correlation between cloze tests and a test consisting of six open-ended reading comprehension questions, found that deleting the 7th word systematically yielded better results than deleting the 6th word. Another study by Tunçer and Erden (2015), examining the correlation between cloze tests and a different test with six open-ended comprehension questions, suggested that deleting the 10th word while providing its first letter as a clue was more consistent. In contrast, Keskin and Akıllı (2013), who administered cloze tests where different word positions were deleted from the same text, found no significant differences.

In the present study, it was determined that cloze tests systematically deleting the 6th word yielded the most consistent results for Turkish. This finding aligns with Ulusoy (2009) and Hızarcı (2009), who asserted that deleting the 6th word produces more valid results than deleting the 5th word. However, considering that tests deleting words at different positions were based on different texts, it can be argued that such comparisons may not be entirely reliable, and more detailed research is required to draw definitive conclusions. To enable clear comparisons between procedures, factors other than the varying deletion methods that may affect reading comprehension must be minimized. While significant care was taken in this study, the factors influencing the scores obtained from tests created using different procedures are not limited solely to the procedural differences. The variations in texts and the words deleted complicate this evaluation. Furthermore, it is crucial that future studies on the subject are conducted in an objective and unbiased framework.

Another aspect of the prominent discussions regarding the deletion process in cloze tests concerns the approach adopted in the deletion procedure. As previously explained, the deletion process can be conducted in two different ways: fixed-ratio deletion or rational deletion. Both methods have their respective advantages and disadvantages. For instance, fixed-ratio deletion provides a standardized technique that can be applied instantly and effortlessly to any text. On the other hand, rational deletion allows for a deeper focus on the semantic dimension of the text, enabling the creation of tests that better assess comprehension. Indeed, in languages like Turkish, where flexibility and variations in syntax can significantly impact meaning, rational deletion may be considered more appropriate. However, this study did not include any analyses related to rational deletion. Therefore, no conclusions can be drawn regarding the effectiveness of rational deletion procedures. Within this framework, it remains an open question in the national literature whether rational deletion procedures yield valid results for Turkish.

In addition to the discussions above, the scoring process of cloze tests remains a topic of debate in both national and international literature. These discussions primarily center on whether only the exact words or all words that preserve the meaning of the sentence should be considered correct. There are three different approaches that can be used for scoring cloze tests, and these approaches will be illustrated using an example sentence from the text titled "Roman Okumak":

"Roman ise gerçekten uzaklaşmaz, _____ gerçeği kavratmaya, hep gerçeği anlatmaya _____."

"A novel does not distance itself from reality; _____ aims to convey reality, always _____ to narrate reality."

In the example above, the original word that fills the first blank in the text is *"always (hep)"*. Accepting only this exact word as correct implies that any other word would be considered incorrect. On the other hand, filling the blank with the word *"constantly (daima)"* maintains the overall structure and meaning of the text. This represents a second method in which words that are semantically equivalent to the required word are also accepted as correct. The third method allows for words that preserve the general meaning of the sentence but deviate to some extent from the original intent of the text. An example of this would be filling the first blank with the word *"novel (roman)"*. Although this maintains the sentence's general meaning, it diverges from the overall context and purpose of the text. Similar examples can be given for the second blank. The original word for the second blank is *"strives (çalışır)"*. However, filling this blank with *"endeavors (uğraşır)"* would maintain the intended meaning in a way that aligns with *"strives (çalışır)"*. On the other hand, words such as *"aspires (özenir)"* or *"focuses (yoğunlaşır)"* would still render the sentence meaningful but introduce a slightly different nuance than what the text originally intends.

This study examines two scoring methods: one that accepts only the exact words as correct and another that accepts contextual words as correct. During the research process, it was observed that students who filled in the blanks with the exact words performed better overall on the test. Conversely, students who were unable to use the exact words tended to struggle with filling in the remaining blanks. Additionally, scoring based on contextual words generally exhibited lower correlations with multiple-choice tests. This finding aligns with the views of researchers who advocate for accepting only exact words as correct (Taylor, 1953; Ulusoy, 2009; Uyanık, 2012). On the other hand, there are researchers (Kaplan & Çiftçi, 2021; Oller et al. 1972) who argue that it is necessary to accept words that preserve the context as correct. Despite these differing perspectives, the scores obtained through both scoring methods displayed similar correlations with multiple-choice reading comprehension tests. The necessity for objective scoring of data obtained from a measurement tool (Turgut & Baykul, 2019) supports the reasonableness of the scoring method that accepts only exact words. Cloze tests scored according to this approach have been demonstrated to be highly objective measurement tools. In this study, the highest correlation levels were generally achieved using the scoring method that considers exact words as correct. However, due to Turkish being an agglutinative language, the tense or form in which the words were inflected by students was not accounted for in this scoring method. As is well-known, in Turkish, a proposition can be expressed in multiple forms through inflection. For instance, a sentence indicating that a person named Ahmet will arrive at a location tomorrow can be conveyed in various ways:

"Ahmet yarın buraya gelecek. (Ahmet will come here tomorrow)"

"Ahmet yarın buraya gelecektir. (Ahmet will be coming here tomorrow.)"

"Ahmet yarın buraya geliyor. (Ahmet is coming here tomorrow.)"

"Ahmet yarın buraya gelir. (Ahmet comes here tomorrow.)"

"Ahmet yarın buraya gelebilir. (Ahmet may come here tomorrow.)"

"Ahmet yarın buraya gelmeli. (Ahmet should come here tomorrow.)"

Given these variations, marking different conjugations of the verb *"gelmek (to come)"* as incorrect would not be justifiable. Therefore, the standard for evaluating such cases in Turkish should not be determined based on the international literature that primarily applies to English, a language where morphological variations often result in more significant semantic shifts. In this regard, the scoring methodology applied in this study, which accepts original words as correct, differs from the more rigid approach that only considers exact word matches while rejecting words with different morphological forms.

In conclusion, the explanations and evaluations presented in this study reveal several key findings. There is a significant relationship between cloze tests and multiple-choice reading comprehension tests. Cloze tests are effective tools for measuring reading comprehension levels. A more valid cloze test can be developed by systematically deleting every sixth word, and accepting only exact words as correct ensures more reliable and objective scoring. Based on this, it can be concluded that cloze tests are suitable tools for measuring reading comprehension levels. Furthermore, these tests avoid issues typically associated with open-ended questions, such as subjectivity related to scorer characteristics, difficulties in ensuring objectivity during scoring, or inefficiencies in scoring time. This is because cloze tests are highly objective and easy to score. Similarly, cloze tests do not face the issues seen in multiple-choice questions, such as the element of chance, suppression of students' creativity, restriction of interpretative abilities, and the need for expertise and experience during the creation process. In cloze tests, readers are required to produce answers by interpreting the narratives in the text, using their vocabulary knowledge, and exercising their creativity. As mentioned earlier, cloze tests prepared using fixed-ratio deletion do not require specialized expertise. On the other hand, it is crucial to maintain diversity in measurement and evaluation practices. Otherwise, students continuously assessed with the same type of measurement tools are likely to develop rigid cognitive structures (Üstüner & Şengül, 2004) and may struggle to think beyond these patterns. This poses a significant problem in modern educational systems that value critical thinking, entrepreneurship, and creativity. Therefore, it is essential to diversify the tools used in measurement and evaluation and incorporate tools that can address the weaknesses of those currently in use into the educational process.

Suggestions

- Cloze tests, which are practical and objective tools for measuring reading comprehension levels, should be systematically integrated into educational processes.
- To ensure more valid results, the scoring system that emphasizes objectivity and accepts only exact words as correct should be used during the scoring process of cloze tests. However, considering the structural characteristics of Turkish, which allow expressions to be inflected in various ways, it would be reasonable not to mark differently inflected words as incorrect.

- The additional functions of cloze tests demonstrated for other languages should also be tested for Turkish.
- The ideal number of blanks in cloze tests for Turkish should be examined through further studies.
- Correlations between objective reading comprehension tests and cloze tests where the first letter of the missing word is provided as a clue should be investigated.
- The validity of cloze tests should be explored for different grade levels and age groups.
- The validity of cloze tests created through rational deletion processes should be evaluated specifically for Turkish.
- Different procedures for preparing cloze tests should be tested using texts of similar readability levels and with similar types of deleted words to determine the most suitable procedure for creating cloze tests.
- The validity of cloze tests should be tested with larger study groups who have sufficient experience with the relevant test type.
- The validity of cloze tests should be compared with different types of standardized reading comprehension tests.

Limitations

During the research process, different texts were used in studies aimed at determining which word positions should be systematically deleted. Since many characteristics of the texts can influence the scores obtained from the tests, the differences in the test construction methodology were not the sole factor affecting the correlations in this process. Furthermore, it is known that students' familiarity with the type of test can impact their performance. In this study, the students in the sample group had significantly less exposure to cloze tests compared to multiple-choice tests, which represents another limitation of the research. Another limitation of the study concerns the size of the sample group. To obtain more generalizable results, it would be beneficial to conduct research with larger participant groups.

References

Abanami, A. A. (1982). *Readability analysis of the 11th and 12th grade earth science textbooks used in the public schools in Saudi Arabia* (Doctoral dissertation). Houston University, Houston.

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal*, *76*(4), 468-479.

Akyol, H. (2020). *Türkçe ilk okuma yazma öğretimi*. Ankara: Pegem Akademi.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Arastaman, G., & Kısa, N. (2020). Geçerlik ve güvenirlik. In N. Cemaloğlu (Ed.), *Bilimsel araştırma teknikleri ve etik* (pp. 193-205). Ankara: Pegem Akademi.

Arıcı, A. F. (2018). *Okuma eğitimi*. Ankara: Pegem Akademi.

Aşılıoğlu, B. (1993). *Ortaokullarda Türkçe öğretimi* (Unpublished doctoral dissertation). Ankara University, Ankara.

Ateşman, E. (1997). Türkçede okunabilirliğin ölçülmesi. *Dil Dergisi*, 58, 71-74.

Bachman, L. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*(1), 61-70.

Balcı, A. (2016). *Okuma ve anlama eğitimi*. Ankara: Pegem Akademi.

Başaran, M. (2013). Okuduğunu anlamanın ölçülmesinde paragraftan anlam kurmaya dayalı çoktan seçmeli sorular. *Eğitim Bilimleri Araştırmaları Dergisi*, *3*(2), 107-121.

Booth, D. (1998). *Guiding the reading process*. Portland Maine: Stenhouse Publishers.

Bormuth, J. (1963). Cloze as a measure of readability. *Proceedings of the International Reading Association*, 1, 131-134.

Bormuth, J. R. (1964). Mean word depth as a predictor of comprehension difficulty. *California Journal of Educational Research*, 15, 226-231.

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly*, 1, 79-132.

Bormuth, J. R. (1967). *Cloze readability procedure*. California: University of California.

Brown, J. D. (1982). *Testing EFL reading comprehension in engineering English* (Doctoral dissertation). University of California, California.

Brown, J., & Grüter, T. (2020). The same cloze for all occasions?: Using the Brown (1980) cloze test for measuring proficiency in SLA research. *International Review of Applied Linguistics in Language Teaching*, 60, 1-26.

Büyüköztürk, Ş. (2006). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem Akademi.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2023). *Eğitimde bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi.

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology*, 76, 697-708.

Carlisle, J., & Rice, M. (2004). Assessment of reading comprehension. In A. Stone, E. Silliman, B. Ehren & K. Apel (Ed.), *Handbook of language and literacy* (pp. 521-555). New York, NY: Guilford.

Carvalho, M., & Souza, A. (2023). Reading assessment in Brazil between the years 2014-2020: Instruments and skills. *Educação e Pesquisa*, 49, 1-19.

Coleman, E. B. (1965). *On understanding prose: Some determiners of its complexity. (NSF Final Report GB-2604)*. Washington D.C.: National Science Foundation.

Craker, H. V. (1971). *Clozentropy procedure as an instrument for measuring oral English competencies of first grade children* (Doctoral dissertation). New Mexico University, New Mexico.

Creswell, J. W. (2020). *Eğitim araştırmaları (*H. Ekşi, Ed. & Trans., 5nd ed.). İstanbul: EDAM. (Original work published 2012)

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contribitons of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277-299.

Çelenk, S. (2006). *Etkinlik temelli ilkokuma ve yazma öğretimi*. İstanbul: Morpa Kültür Yayınları.

Çepni, S., Gökdere, M., & Küçük, M. (2002). Adaption of the readability formulas into the Turkish science textbooks. *Energy Education Science and Technology*, *10*(1), 49-58.

Çetinkaya, G. (2010). *Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması* (Unpublished doctoral dissertation). Ankara University, Ankara.

Dağ, N. (2010). Okuma güçlüğünün giderilmesinde 3P metodu ile boşluk tamamlama (cloze) tekniğinin kullanımı üzerine bir çalışma. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi*, *11*(1), 63-74.

Darnell, D. K. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. *Speech Monographs*, 37, 36-46.

Dubay, W. H. (2007). *Smart language: Readers, readability, and the grading of text*. Impact Information.

Febriyanti, P. (2017). The correlation between reading comprehension and students' ability in answering cloze test of the seventh grade students at SMPN I Kalipuro Banyuwangi in the 2014/2015 Academic Year. *Language and Art Journal*, *1*(2), 36-47.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221-233.

Froese, V. (1971). Cloze readability versus the Dale-Chall formula. *International Reading Association*, 1, 19-23.

George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference, 17.0 update*. Boston: Pearson.

Geske, A., & Ozola, A. (2008). Factors influencing reading literacy at the primary school level. *Problem of Education in 21th Century*, 6, 71-77.

Güler, N. (2019). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2013). *Multivariate data analysis*. Harlow: Pearson.

Harmer, J. (2002). *The practice of English language teaching* (3nd ed.). England: Longman.

Hızarcı, S. H. (2009). *İlköğretim 6. sınıf yeni sosyal bilgiler ders kitaplarının okunabilirlik düzeylerinin incelenmesi* (Unpublished master's thesis). Gazi University, Ankara.

Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 121-128). Rowley, MA: Newbury House.

Huensch, A. (2013). *The perception and production of palatal codas by Korean L2 learners of English* (Doctoral dissertation). Illinois University, Illinois.

Irvine, P., Atai P., & Oller J. W. (1974). Cloze, dictation, and the test of English as a foreign language. *Language Learning*, *24*(2), 245-252.

James, W. (2004). *Special education and social development*. New Delhi: Anmol Publications PVT. LTD.

Kalyoncu, R., & Memiş, M. (2024). Türkçe için oluşturulmuş okunabilirlik formüllerinin karşılaştırılması ve tutarlılık sorgusu. *Ana Dili Eğitimi Dergisi*, *12*(2), 417-436.

Kaplan, K., & Çiftçi, M. (2021). Okuma seviyesinin belirlenmesinde ikame kelime uygulaması. *Türk Dili Araştırmaları Yıllığı-BELLETEN*, 72, 209-236.

Karakaya, İ. (2014). Bilimsel araştırma yöntemleri. In A. Tanrıöğen (Ed.), *Bilimsel araştırma yöntemleri* (pp. 57-82). Ankara: Anı Yayıncılık.

Karatay, H. (2018). *Okuma eğitimi kuram ve uygulama*. Ankara: Pegem Akademi.

Karatay, H., Bolat, K. K., & Güngör, H. (2013). Türkçe ders kitaplarındaki metinlerin okunabilirlik ve anlaşılabilirliği. *The Journal Academic Social Science Studies*, *6*(6), 603-623.

Katz, S., & Lautenschlager, G. J. (1994). Answering reading comprehension items without passages on the SAT-I, the ACT, and the GRE. *Educational Assessment*, *2*(4), 295-308.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281-300.

Keskin, H. K., & Akıllı, M. (2013). Fen ve teknoloji ders kitaplarının okunabilirliğinin farklılaştırılmış boşluk doldurma testleri ile ölçülmesi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 27, 47-66.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10, 62-102.

Klare, G. R., Simaiko, H. W., & Stolurow, L. M. (1972). The cloze procedure: A convenient readability test for training materials and translations. *International Review of Applied Psychology*, *21*(2), 77-106.

Kleijn, S., Pander Maat, H., & Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing*, *36*(4), 553-572.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge: Cambridge University Press.

Köklü, N., Büyüköztürk, Ş., & Çokluk, Ö. (2007). *Sosyal bilimler için istatistik*. Ankara: Pegem Akademi

Köse, E. Ö. (2009). Biyoloji 9 ders kitabında hücre ile ilgili metinlerin okunabilirlik düzeyleri. *Journal of Arts and Sciences*, 12, 141-150.

Kurudayıoğlu, M., & Karadağ, Ö. (2005). Kelime hazinesi çalışmaları açısından kelime kavramı üzerine bir değerlendirme. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *25*(2), 293-307.

Loewe, E. E. (1983). *The effect of using cloze data for revising instructional materials* (Doctoral dissertation). Florida University, Florida.

Lu, G. (2006). *Cloze test and reading strategies in English language teaching in China* (Master's thesis). Western Cape University, Cape Town.

Manis, M., & Dawes R. M. (1961). Cloze scores as a function of attitude. *Psychological Reports*, 9, 79-84.

Mariotti, A. S., & Homan, S. P. (2001). *Linking reading assesment to ıntruction: An application worktext for elementary classroom teachers*. New Jersey: Lawrence Erlbaum Associates.

Mariotti, A. S., & Homan, S. P. (2009). *Linking reading assessment to instruction: An application worktext for elementary classroom teachers*. New York: Routledge.

Mertens, D. M. (2015). *Research and evaluation in education and psychology*. Thousand Oaks, CA: Sage.

Nation, I. S. P. (2009). *Teaching ESL/EFL reading and writing*. New York: Routledge.

Oller, J. W. (2006). Close Tests of the second language proficiency and what they measure. *Language Learning, 23*(1), 105-118.

Oller, J. W., Bowen, D. J., Dien, T. T., & Mason, V. W. (1972). Cloze tests in English, Thai, and Vietnamise: Native and non-native performance. *Language Learning, 22*(1), 1-13.

Ömeroğlu, E. (2018). Açık uçlu sınavlarla çoktan seçmeli test sınavlarının karşılaştırılması test sınavlarının yazma becerisine etkisi. *International Journal of Language Academy*, *6*(26), 548-570.

Özbay, M. (1997). Test türü imtihanların Türkçe öğretimindeki yeri. *Bilge*, 11, 13-16.

Özdemir, M., & Doğruöz, E. (2020). Bilimsel araştırma desenleri. In N. Cemaloğlu (Ed.), *Bilimsel araştırma teknikleri ve etik* (pp. 65-98). Ankara: Pegem Akademi.

Özyılmaz, G. (2010). *İlköğretim 7. sınıf öğrencilerine okuduğunu anlama stratejilerinin öğretiminin okuduğunu anlama başarısı üzerine etkisi* (Unpublished master's thesis). Yıldız Teknik University, İstanbul.

Potter, T. C. (1968). *A taxonomy of cloze research, part I: Readability and reading comprehension* (Report No. TR1). California: Southwest Regional Laboratory for Educational Research and Development.

Retorta, M. S. (2001). Multiple-choice and cloze procedures in reading tests: What do they really measure?. *ESPecialist*, *22*(2), 127-154.

Ruddell, M. R. (2005). *Teaching content reading and writing*. New York: John Wiley & Sons.

Shahnazari-Dorcheh, M., Roshan, S., & Hesabi, A. (2012). What is the optimum lenght of a cloze test?. *International Journal of English Linguistics*, *2*(5), 142-153.

Shiba, S. A. (1957). A study of the measurement of readability- application of the cloze procedure to the Japanese language. *Japanese Journal of Psychology*, 28, 67-73.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of ESL proficiency for Arab students. *Modern Language Journal*, *58*(5), 239-241.

Suehiro, A. C. B., & Santos, A. A. A. (2015). Reading comprehension and phonological awareness: Evidence of validity of their measures. *Estudos de Psicologia*, *32*(2), 201-211.

Sukarni, S. (2021). The use of cloze test to test reading comprehension of non-English department students. *Jo-ELT (Journal of English Language Teaching) Fakultas Pendidikan Bahasa & Seni Prodi Pendidikan Bahasa Inggris IKIP*, *8*(1), 74-82.

Şahindokuyucu, A. (2006). *A study of cloze and multiple-choice tests for measuring reading comprehension of preparatory students* (Unpublished master's thesis). Bolu Abant İzzet Baysal University, Bolu.

Taşpınar, M. (2017). *Sosyal bilimlerde SPSS uygulamalı nicel veri analizi*. Ankara: Pegem Akademi.

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35-39.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*(1), 415-433.

Taylor, W. L. (1956). Recent developments in the use of cloze procedure. *Journalism Quarterly*, *33*(1), 42-48.

Taylor, W. L., & Lee, K. W. (1954). KM readers lend hand to science: Cloze method works in written Korean and may serve as a tool for Korean language reform. *Korean Messenger*, 3, 4-5.

Tekin, H. (1982). *Eğitimde ölçme ve değerlendirme*. Ankara: Daily News Ofset Tesisleri.

Temizkan, M. (2007). *İlköğretim ikinci kademe Türkçe derslerinde okuma stratejilerinin okuduğunu anlama üzerindeki etkisi* (Unpublished doctoral dissertation). Gazi University, Ankara.

Temizkan, M., & Sallabaş, M. E. (2011). Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 30, 207-220.

Tremblay, A., & Garrison, M. D. (2010). Cloze tests: A tool for proficiency assessment in research on L2 French. In M. T. Prior, Y. Watanabe, & S. K. Lee (Eds.), *Selected proceedings of the second language resarch forum 2008* (pp. 73-88). MA: Cascadilla Press.

Tuncer, M. (2020). Nicel araştırma desenleri. In B. Oral & A. Çoban (Eds.), *Kuramdan uygulamaya eğitimde bilimsel araştırma yöntemleri* (pp. 205-227). Ankara: Pegem Akademi.

Tunçer, B., & Erden, G. (2015). Boşluk doldurma testlerinin ilkokul 4. sınıf öğrencilerinin okuduğunu anlama düzeylerini belirlemede kullanılabilirliği [Special issue]. *Bartın Üniversitesi Eğitim Fakültesi Dergisi*, 318-324.

Turgut, F., & Baykul, Y. (2019). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi.

Ulusoy, M. (2009). Boşluk tamamlama testinin okuma düzeyini ve okunabilirliği ölçmede kullanılması. *Türk Eğitim Bilimleri Dergisi*, 7(1), 105-126.

Uyanık, G. (2011). *İlköğretim 5. sınıf öğrencilerinin boşluk tamamlama tekniğiyle belirlenen okuma seviyeleri ile okuduğunu anlama düzeylerinin karşılaştırılması* (Unpublished master's thesis). Gazi University, Ankara.

Uyanık, G. (2012). İlköğretim 5. sınıf öğrencilerinin okuma seviyelerinin farklı boşluk tamamlama testleri ile belirlenmesi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *32*(3), 657-672.

Üstüner, A., & Şengül, M. (2004). Çoktan seçmeli test tekniğinin Türkçe öğretimine olumsuz etkileri. *Fırat Üniversitesi Sosyal Bilimler Dergisi*, *14*(2), 197-208.

Vacca, R. T., & Vacca, J. A. L. (2005). *Content area reading: Literacy and learning across the curriculum*. London: Pearson Education.

Vari-Cartier, P. (1980). *The readability and comprehensibility of Spanish prose as determined by the frase graph and the cloze procedure* (Doctoral dissertation). Rutgers University, New Jersey.

Wahdaniah, Marbun, R., & Husin, S. (2013). The use of cloze test in increasing the students' reading comprehension. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa*, *2*(1), 1-12.

Williams, R. S., Ari, O., & Santamaria, C. N. (2011). Measuring college students' reading comprehension ability using cloze tests. *Journal of Research in Reading*, *34*(2), 215-231.

Yılmaz, H. (1998). *Eğitimde ölçme ve değerlendirme*. Konya: Mikro Basım-Yayın-Dağıtım.