E-ISSN: 2458-8342, P-ISSN: 1301-3718
Ankara University Journal of Faculty of Educational Sciences
2026, 59(1), 221-271
DOI: 10.30964/auebfd.1693237
Research Article

A Comparative Analysis of Multiple Choice and Cloze Tests in Terms of Reading Comprehension and Reader Proficiency Levels

Yusuf Aydın [1], Ezgi Kaya Filik[2]

1 Department of Turkish and Social Studies Education, Faculty of Education, Akdeniz University, Antalya, Türkiye
2 Ministry of National Education, İzmir, Türkiye

ABSTRACT

The aim of this study is to compare student performance obtained from a multiple choice test and a cloze test in terms of the cognitive processes and operational characteristics underlying these test formats. A total of 102 middle school students enrolled in the 7th and 8th grades participated in the study. The research was conducted using a correlational design, and the data were analyzed through quantitative statistical methods. The results revealed a strong positive correlation between multiple-choice test scores and cloze test scores and showed that cloze test performance significantly predicted multiple-choice test performance. These findings indicate that the two test formats measure overlapping cognitive processes and may be used as complementary assessment tools. In addition, the study examined the predictive power of total scores obtained from multiple-choice items corresponding to different cognitive levels of Bloom's Taxonomy on cloze test performance. The results showed that items at the analysis level had a higher predictive power for cloze test performance compared to items at other cognitive levels. While items at the remembering and understanding levels also significantly predicted cloze test performance, items at the application level did not demonstrate a significant effect. This pattern suggests that cloze tests are particularly associated with mid-level cognitive processes, such as text comprehension and analysis.with regard to the classification of reading proficiency levels, a limited level of agreement was found between the two test formats. More than half of the participants were classified into different reading proficiency levels depending on their performance on the two tests.

Keywords: Multiple-choice test, cloze test, reading proficiency level, reading proficiency classification, revised Bloom's Taxonomy

Corresponding Author: Yusuf Aydın
Department of Turkish and Social Studies Education, Faculty of Education, Akdeniz University, Antalya, Türkiye
E-mail: ysf.aydn66@gmail.com **ORCID ID:** 0000-0003-0898-9020 **ROR ID:** https://ror.org/01m59r132
Received Date: 05.07.2025 **Accepted Date:** 02.28.2026 **Publication Date:** 04.15.2026
Citation: Aydın, Y., & Kaya-Filik, E. (2026). A Comparative analysis of multiple choice and cloze tests in terms of reading comprehension and reader proficiency levels. *Ankara University Journal of Faculty of Educational Sciences*, 59(1), 221-271. https://doi.org/10.30964/auebfd.1693237
All ethical declarations related to this article are provided on the final page of the manuscript (Page: 271).

Reading comprehension is the process of constructing meaning from written texts and processing information. This complex process relies on the coordination of multiple cognitive operations (Kendeou, McMaster, & Christ, 2016; Tighe & Schatschneider, 2016). Reading skills are essential for individuals' academic achievement as well as their social functioning. Through reading, students gain access to information and acquire new concepts (Pretorius, 2002). Beyond academic contexts, reading ability also plays a significant role in the development of social and emotional competencies. Research has demonstrated a positive correlation between socioemotional skills and reading comprehension (Yu, Yu, & Tong, 2023). In addition, reading has been shown to enhance mental imagery abilities (Pino & Mazza, 2016). In particular, engagement with fictional texts provides a learning context that supports social and emotional learning (Kozak & Recchia, 2019). Therefore, the development of reading comprehension skills is crucial for individuals' success in both academic and social domains. Given the multidimensional and functional nature of reading comprehension, its assessment must be conducted in a valid and reliable manner. Accordingly, how reading comprehension is measured and the extent to which the observed performance reflects underlying cognitive processes constitute a central issue in the field of educational measurement and assessment.

Assessing reading comprehension is important for determining individuals' information processing abilities and their level of meaning construction from texts. This assessment is conducted through various tests designed to measure reading skills. The main test types used in the assessment of reading comprehension include multiple choice tests, open ended questions, yes/no or true/false questions, cloze tests, short answer questions, sequencing questions, interpretive and evaluative questions, and summarization tasks (Ülper, 2010, pp. 123–136). These diverse testing techniques reflect the multidimensional nature of reading comprehension and are designed to assess different aspects of this skill.

Reading comprehension involves not only recalling information from a text but also higher order cognitive processes such as meaning construction, analysis, synthesis, and evaluation. Accordingly, the assessment techniques used to evaluate this skill have been diversified to encompass these cognitive processes. Nevertheless, among the assessment methods mentioned above, multiple choice questions remain the most widely used format (Ozuru et al., 2007). In Turkey, multiple choice items are extensively employed at almost all levels of education, not only to measure students' reading comprehension performance but also to assess academic achievement across various subject areas. At the international level, this item format has long been included in large-scale assessments such as PISA, TOEFL, IELTS, SAT, and GRE.

The preference for multiple choice tests can be attributed to several factors, including their capacity to assess higher order thinking skills, their cost-effectiveness, and their high reliability (Little & Bjork, 2015). In addition, well-constructed multiple choice tests demonstrate high content validity and are not affected by rater bias during

the evaluation process. Because students are not required to produce extended written responses, writing skills are not involved in the assessment process, allowing reading comprehension to be evaluated more directly (Ülper, 2010, pp. 123–124). Owing to these characteristics, multiple choice tests occupy an important place in educational systems and measurement and assessment practices.

Despite their advantages, multiple choice tests also have several limitations. One major concern is the possibility of selecting the correct answer by chance (Ülper, 2010, p. 124; Cooper & Foy, 1967). In addition, the development of valid and reliable multiple choice tests is a complex process and typically requires pilot testing and item analysis (Ülper, 2010, p. 124; Fuhrman, 1996; Gierl et al., 2017). Furthermore, some researchers have raised concerns about whether multiple choice tests directly assess reading ability. For example, Rupp et al. (2006) demonstrated that participants often perceive responding to multiple choice questions as a problem-solving task rather than a reading comprehension task. This finding suggests that students may adopt strategies that differ from those they typically use while reading a text. Similarly, Daneman and Hannon (2001) showed that some multiple choice questions can be answered through reasoning alone, without reading the accompanying text. Keenan and Betjemann (2006) also reported that certain items in multiple choice tests could be answered without requiring students to engage with the text itself. Taken together, these findings point to the limitations of multiple choice tests in measuring reading ability and underscore the importance of considering alternative assessment methods.

One of the important test types used to assess reading comprehension is the cloze test. In these tests, specific words or phrases are removed from a text, and students are expected to complete the blanks with appropriate words. Numerous studies have demonstrated that cloze tests can be used as valid and reliable instruments for measuring language skills in both first and second language contexts (Aitken, 1977; Sumita et al., 2005; Ulusoy, 2009; Zhang et al., 2023). In addition, several studies have shown that cloze tests exhibit high correlations with other language proficiency measures and provide reliable assessments of language proficiency (Gooskens & van Heuven, 2017; Gaillard & Tremblay, 2016; Luchkina et al., 2021). Taken together, these findings indicate that cloze tests constitute reliable tools for the assessment of language skills.

Like multiple choice tests, cloze tests also exhibit several limitations. Many researchers argue that cloze tests are insufficient for measuring reading comprehension beyond the sentence level (Kleijn, Pander Maat, & Sanders, 2019; Gellert & Elbro, 2012). In addition, these tests are often reported to assess primarily lower-level language skills (Alderson, 1980). Wismann et al. (2018) demonstrated that cloze tests do not support students' ability to transfer acquired knowledge to deductive reasoning tasks, indicating limitations in their capacity to assess higher-order cognitive skills. Furthermore, the study by Baghaei and Ravand (2019) showed that cloze items account for variance beyond general reading comprehension ability. In other words, performance on cloze items cannot be explained solely by reading

comprehension skills. This finding suggests that cloze tests may measure abilities that differ from the targeted reading skill. Consequently, serious concerns have been raised regarding the validity of cloze tests and the scope of the skills they purport to measure.

The studies summarized above indicate that multiple choice and cloze tests possess different strengths and limitations in assessing reading comprehension. This situation raises important questions regarding how scores obtained from these tests should be interpreted and to what extent they reflect the targeted reading skill. Accordingly, the validity of the assessment tools used to evaluate reading comprehension performance emerges as a central issue that warrants careful consideration.

Multiple choice tests, which are widely used in the assessment of reading comprehension, can provide objective and reliable measurements when they are well constructed. However, they also entail notable limitations, particularly with respect to representing higher-order thinking skills and the possibility that correct answers may be selected through guessing (Puthiaparampil & Rahman, 2020).

Cloze tests, by contrast, require the simultaneous use of syntactic structure, lexical knowledge, and contextual cues, thereby allowing for a more integrated assessment of multiple linguistic processes associated with reading comprehension (Kleijn, Maat, & Sanders, 2019). High correlations observed between cloze tests and standard reading and language proficiency measures indicate that these tests can offer valid assessments, particularly in distinguishing lower and mid level reading skills (Fotos, 1991; Kleijn, Maat, & Sanders, 2019; Bråten, Haverkamp, & Anmarkrud, 2024). Nevertheless, the heavy reliance of traditional item formats on sentence level linguistic processing may limit the adequate representation of discourse level coherence building and higher-order comprehension processes.

Determining Reading Levels Using Multiple-Choice and Cloze Tests

Identifying readers' proficiency levels using various reading comprehension tests is essential for determining the sources of difficulties in reading comprehension and for defining different reader subtypes Auphan et al., 2019). In the relevant literature, research comparing multiple choice tests and cloze tests in terms of their effectiveness in measuring reading levels dates back to the 1960s. Bormuth (1967), for instance, compared cloze tests and multiple choice tests with respect to their capacity to assess reading comprehension levels. The findings of this study indicate that scores obtained from cloze tests can be aligned with scores from multiple choice tests. For example, a 38% success rate on a cloze test corresponds to a 75% success rate on a multiple choice test. Similarly, a student who achieves a 50% score on a cloze test is expected to obtain approximately 90% on a multiple choice test. In his study, Bormuth categorized scores from both cloze and multiple choice tests into three levels; however, he did not assign specific labels to these levels.

In a similar line of research, Rankin and Culhane (1969) also reported that cloze test scores could be interpreted as equivalent to the percentage scores obtained from multiple choice tests, thereby supporting Bormuth's findings. In their study, the researchers classified readers into two levels: independent and instructional. To reach the independent reading level, students were required to achieve a 90% success rate on multiple choice tests. At this level, students are able to read and comprehend the material on their own. To attain the same independent reading level on a cloze test, students were required to answer 61% of the cloze items correctly, which indicates the ability to comprehend the text independently. For the instructional reading level, a success rate of 75% on multiple choice tests was required. Students at this level are able to comprehend the material with teacher guidance or within a supportive instructional context. To meet the instructional reading level on the cloze test, a success rate of 41% was considered sufficient, indicating that instructional support is necessary for adequate comprehension of the text.

In his study, Çetinkaya (2010, p. 35) synthesized previous research and compared reading proficiency levels based on cloze tests and multiple choice tests. These levels were classified into three categories: frustrated, instructional, and independent. Readers at the frustrated level correctly completed less than 35% of the cloze items and answered fewer than 50% of the multiple choice questions correctly. Instructional level readers correctly completed 35–50% of the cloze items and answered 50–70% of the multiple choice questions correctly. Independent readers, by contrast, achieved more than 50% accuracy on cloze items and exceeded a 70% correct response rate on multiple choice tests. This classification provides a framework for determining reading proficiency levels using both test formats.

The alignment between these reading levels has been examined in only a limited number of empirical studies. In his study, Wait (1987) reported a significant positive relationship between participants' scores on a cloze test and their scores on a separate reading comprehension assessment. Similarly, Kızılaslan Tunçer and Erden (2015) classified participants into three group-independent, instructional, and frustrated-based on their scores on cloze tests and multiple-choice tests, and found a moderate, positive correlation between the two test types in determining reading proficiency levels.

Investigating the relationship between multiple choice tests and cloze tests, which are widely used to measure reading comprehension, is important for improving measurement and assessment practices in education. This study examines the alignment between these two test types in a comprehensive manner. Previous research has generally compared multiple-choice and cloze tests in terms of reading comprehension outcomes and the relationships between the scores obtained from different item formats. These studies have focused on the extent to which scores derived from different item types converge or diverge, as well as on whether observed performance differences are associated with test format. Unlike these approaches, the present study does not limit the comparison to overall test scores. Instead, multiple

choice items are classified according to the cognitive levels targeted by Bloom's Revised Taxonomy, and the relationships between the subscores obtained at these cognitive levels and cloze test performance are examined. In this way, the study explores which cognitive processes are more strongly aligned with each test type in the assessment of reading comprehension, particularly within the framework of Bloom's cognitive levels, and provides a more detailed analysis of the cognitive foundations underlying performance on cloze tests.

Reading proficiency levels were determined using the classification developed by Çetinkaya (2010), which is based on findings from previous research. By examining the relationship between students' scores on multiple choice and cloze tests, the study aims to contribute important evidence to the development of assessment tools used to measure reading comprehension.

The aim of this study is to examine the relationship between the measurement outcomes of multiple-choice tests and cloze tests used to assess reading comprehension, and to identify the patterns these two test types exhibit in relation to the cognitive processes underlying reading comprehension, particularly within the framework of Bloom's cognitive levels. Accordingly, the following research questions are addressed:

1. Do cloze test scores significantly predict multiple choice test scores?

2. What is the relationship between students' performance on multiple choice items targeting different cognitive levels of Bloom's Revised Taxonomy and their scores on the cloze test?

3. To what extent do multiple choice and cloze tests show alignment in determining participants' reading proficiency levels?

Method

The Method section describes the research design, participants, data collection tools, and procedures used in the study. It outlines the development and psychometric properties of the multiple choice and cloze reading comprehension tests, explains how data were collected and scored, and details the statistical analyses conducted to examine the relationship between test scores and reading proficiency.

Research Model

This study, which compares student performance on multiple choice and cloze tests of reading comprehension, adopts a correlational research design. In correlational research, the relationship between two or more variables is examined without any manipulation of those variables (Büyüköztürk et al., 2013, p. 184). The primary focus of the study is to examine the association between participants' scores obtained from a multiple-choice test and a cloze test designed to assess reading comprehension performance.

To examine the relationship between scores obtained from the multiple-choice test (MCT) and the cloze test (CT) in greater detail, regression analysis was employed. In this analysis, variables were handled within an explanatory statistical modeling framework without assuming a causal predictive relationship. In the first stage, MCT scores were treated as the explained variable, while CT scores were included in the model as the explanatory variable. In the second stage, CT scores were treated as the explained variable, and scores obtained from multiple choice items targeting the remembering, understanding, applying, and analyzing levels of Bloom's Revised Taxonomy were entered into the model as explanatory variables. This approach aims not to determine the direction of causality between the tests, but rather to reveal how different item types and cognitive levels are associated with student performance.

In this study, the items in the MCT were classified according to the cognitive levels of Bloom's Revised Taxonomy: remembering, understanding, applying, and analyzing. Subscores for each level (e.g., the total score obtained from five items at the remembering level) were calculated separately. These subscores were included in the regression analysis to examine their associations with CT scores. In this way, the study investigates which cognitive levels of Bloom's Revised Taxonomy show stronger alignment with performance on the cloze test. No cognitive level classification was applied to the cloze test itself, as it was constructed through systematic word deletion from a single text, and individual blanks cannot be meaningfully assigned to specific cognitive levels. Therefore, the cloze test was treated as a holistic measure of performance. This analysis aims to provide explanatory insights into the types of cognitive processes with which cloze test performance is more strongly associated.

Study Group

The study was conducted in two phases. In the first phase, the data collection instruments were developed and administered. A total of 216 middle school students participated in this phase of instrument development. These participants were enrolled in a public middle school located in the city center of Antalya and characterized by a middle-level socioeconomic background. Of the students, 115 were in the 7th grade and 101 were in the 8th grade. The sample consisted of 52% female students (n = 112) and 48% male students (n = 104).

In the second phase of the study, the tests—whose validity and reliability had been established—were administered in a different public middle school in Antalya with a similar socioeconomic profile. In this phase, 102 students participated, including 60 7th-grade students and 42 8th-grade students. Of these participants, 55% were female (n = 56) and 45% were male (n = 46).

Participants were selected using convenience sampling based on accessibility and voluntary participation. The sample therefore consisted of the accessible target student population during the data collection process.

Data Collection Tools and Process

To measure participants' reading comprehension performance, the researchers developed two instruments: a Multiple-Choice Reading Comprehension Test (MCT) and a Cloze Test (CT). The development process of these tests is described below:

Psychometric Properties of the Multiple choice and Cloze Reading Comprehension Tests

MCT was developed based on the text titled "İyi Uykular, Tatlı Rüyalar" (Gençer, 2014) and initially consisted of 24 multiple choice items, each with four options. Item development was guided by Bloom's Revised Taxonomy. At this stage, six items targeted the remembering level, nine the understanding level, seven the analyzing level, and two the applying level. During the test development process, it was acknowledged that the psychometric properties of reading comprehension tests are influenced not only by item characteristics but also by features of the source text. Accordingly, the reading text and the items developed based on it were reviewed by two experts in curriculum and instruction and one expert in Turkish language education. The experts evaluated the items in terms of linguistic appropriateness and alignment with the intended measurement objectives, as well as the suitability, readability, and informational/event density of the text for the target age group and their potential impact on test difficulty. Based on expert feedback, the number of items was reduced to 20, and necessary revisions were made to strengthen text–item alignment.

A key focus of the study was to ensure that the multiple choice items targeted different cognitive levels of Bloom's Revised Taxonomy. For this reason, each item was independently examined by the experts with respect to the cognitive level it was intended to measure. Experts were asked to evaluate which cognitive level each item was intended to measure, and based on these evaluations, the distribution of the items across cognitive levels was determined. Items for which consensus could not be reached regarding their targeted cognitive level were either revised or removed from the test. In this way, expert-judgment–based validity evidence was obtained for the representation of the intended cognitive levels.

Following test administration, item statistics and the distribution of items across cognitive levels were jointly considered to evaluate the measurement instrument from both statistical and theoretical perspectives. After revisions, the test consisted of six remembering, seven understanding, five analyzing, and two applying items. During implementation, one remembering level item was identified as flawed due to an incorrectly constructed answer option and was therefore excluded from the validity and reliability analyses. The error was a technical implementation mistake, and the necessary correction was made in a way that did not affect the overall findings of the study; the analyses were subsequently conducted based on the remaining items. Consequently, the final version of the MCT consisted of 19 items, with five items at

the remembering level. Table 1 presents the difficulty and discrimination indices for each of the 19 items in the MCT.

Item difficulty indices were calculated based on Classical Test Theory as the proportion of students who answered each item correctly (p-value). Item discrimination indices were computed by taking the difference between the percentages of correct responses in the upper 27% and lower 27% groups for each item.

Table 1

Item Analysis of the Reading Comprehension Test

Item	Difficulty Index	Discrimination Index
Item 1	0.87	0.35
Item 2	0.44	0.6
Item 3	0.74	0.47
Item 4	0.72	0.36
Item 5	0.73	0.57
Item 6	0.87	0.36
Item 7	0.69	0.54
Item 8	0.81	0.4
Item 9	0.48	0.43
Item 10	0.62	0.51
Item 11	0.70	0.67
Item 12	0.51	0.71
Item 13	0.32	0.4
Item 14	0.46	0.42
Item 15	0.65	0.61
Item 16	0.68	0.64
Item 17	0.49	0.67
Item 18	0.51	0.5
Item 19	0.79	0.46

The difficulty indices range between 0.32 and 0.87, and the discrimination indices range between 0.35 and 0.71, indicating that most items have good discriminative power. These ranges indicate that the test includes a variety of difficulty levels and demonstrates satisfactory discriminative power. The average difficulty index of the test is 0.64, and the average discrimination index is 0.51. In the MCT, participants receive 1 point for each correct answer, with no penalty for incorrect responses. The score range for the MCT is 0 to 19.

Validity Study

Exploratory Factor Analysis (EFA) was conducted using a tetrachoric correlation matrix to assess the unidimensionality of the MCT (Embretson & Reise, 2000). In addition, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was found to be above .80, indicating that the sample was sufficient for exploratory

factor analysis (EFA). Moreover, Bartlett's Test of Sphericity was significant ($p < .001$), suggesting that the correlations among the variables were adequate for conducting factor analysis. These findings support the suitability of the dataset for factor analysis and the validity of the results obtained from the analysis.The analysis results were examined based on eigenvalues, factor loadings, and parallel analysis. A clearly dominant first factor was considered evidence of unidimensionality (Crocker & Algina, 1986). Items were evaluated based on whether they loaded sufficiently on a single factor (≥.30) (Çokluk et al., 2012). In parallel analysis, eigenvalues from randomly generated datasets were compared to actual data to assess unidimensionality (Horn, 1965).

Factor loadings for the 19-item MCT are presented in Table 2, and those for the 44-item CT are presented in Table 3.

Table 1

Factor Loadings of MCT Item Pool

Item Number	Factor loading	Item Number	Factor loading
M1	.75	M11	.79
M2	.54	M12	.69
M3	.64	M13	.39
M4	.44	M14	.34
M5	.75	M15	.67
M6	.75	M16	.72
M7	.63	M17	.65
M8	.66	M18	.41
M9	.31	M19	.65
M10	.48		

Table 3

Factor Loadings of the CT Item Pool

Item Number	Factor loading	Item Number	Factor loading
M1	.58	M12	.36
M2	.66	M13	.37
M3	.40	M14	.45
M4	.38	M15	.42
M5	.33	M16	.36
M6	.32	M17	.43
M7	.46	M18	.47

(continued)

Table 3 (continued)

M8	.30	M19	.54
M9	.43	M20	.38
M10	.41	M21	.43
M11	.37	M22	.53
M23	.33	M34	.32
M24	.44	M35	.50
M25	.38	M36	.50
M26	.63	M37	.39
M27	.54	M38	.38
M28	.49	M39	.45
M29	.51	M40	.54
M30	.40	M41	.51
M31	.34	M42	.52
M32	.42	M43	.59
M33	.36	M44	.57

The EFA results show that both tests largely have a unidimensional structure. The first component of the MCT accounts for 37.03% of the total variance, while the CT explains 21.01%. The eigenvalues for the first factor in both tests were significantly higher than those of other factors, indicating that a major portion of items load onto a common dimension. The factor loadings ranged from .31 to .79 in the MCT and from .30 to .66 in the CT, suggesting that most items meaningfully loaded onto a single underlying construct representing reading comprehension.

Parallel analysis graphs for both tests are presented in Figure 1 (MCT) and Figure 2 (CT), supporting the unidimensional structure of each test.

Figure 1

Parallel Analysis Graph-Multiple Choice Test

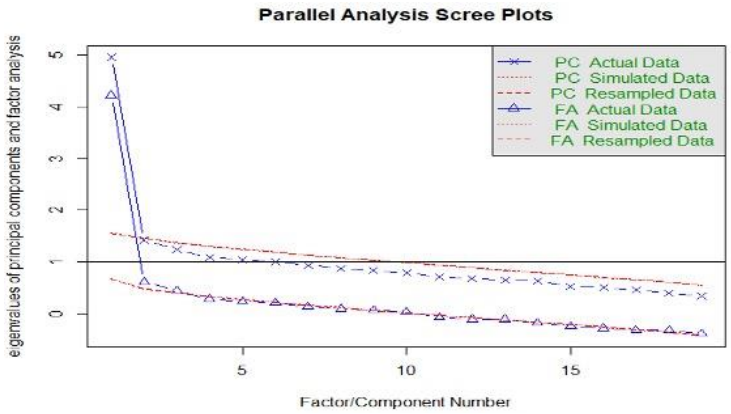
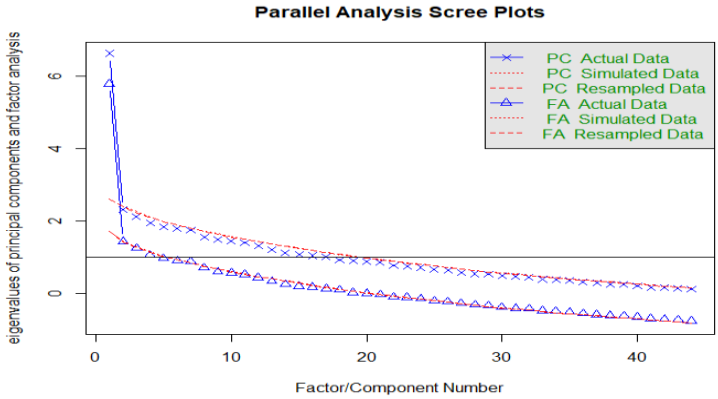


Figure 2

Parallel Analysis Graph-Cloze Test



The analyses conducted on the factor structures of the two tests reveal that both exhibit a unidimensional structure. Results from the parallel analysis indicate that the first factor in both the MCT and CT is clearly distinguishable from the subsequent factors, supporting the assumption of unidimensionality.

When examining the model fit indices, the following values were obtained for the MCT: $\chi^2(152) = 196.22$, $p = .009$, RMSEA = .037 (90% CI [.019, .051]), SRMR = .060, TLI = .972, and CFI = .975. For the CT, the results were $\chi^2(902) = 920.14$, $p = .329$, RMSEA = .013 (90% CI [.00, .031]), SRMR = .088, TLI = .987, and CFI = .976.

These results indicate that both tests demonstrate good model fit and support a unidimensional structure (Hu & Bentler, 1999). However, the nonsignificant χ^2 value for the CT ($p = .329$) suggests that the cloze test provides a better overall fit to the data.

Cloze Test

CT was developed based on the text titled "İyi Uykular, Tatlı Rüyalar." The text consists of a total of 446 words, including the title. It is explanatory in nature and was selected with consideration of the linguistic and cognitive characteristics of middle school students. Its plot grounded in everyday life, clear causal relationships, and semantic coherence indicate that the text provides suitable content for assessing general reading comprehension. Most of the words used in the text are high-frequency items commonly encountered in Turkish language textbooks published by the Ministry of National Education and are expected to be part of students' receptive vocabulary. This design choice aims to ensure that the test primarily reflects processes of meaning construction from context and the use of textual coherence rather than measuring vocabulary knowledge per se.

The CT was constructed in accordance with standard practices of the cloze procedure. No words were deleted from the first sentence of the text, beginning with the second sentence, every sixth word was systematically removed. Punctuation marks, proper nouns, and numbers were excluded from the deletion process. As a result, a total of 69 blanks were created. For each blank, participants were required to produce the word that best fit the textual context. In scoring, only responses that exactly matched the target word were considered correct; spelling errors, synonyms, or semantically related but non-identical words were scored as incorrect. Each correct response was awarded one point, while incorrect or omitted responses received zero points. Accordingly, total scores on the CT range from 0 to 69.

The suitability of the text for use in a cloze test was evaluated by three faculty members specializing in Turkish language education. The experts agreed that the text was appropriate for assessment through the cloze technique in terms of vocabulary level, semantic coherence, and the availability of contextual cues. Nevertheless, it cannot be claimed that cloze tests constructed through systematic word deletion directly or equally measure all cognitive dimensions of reading comprehension. Such tests primarily reflect performance related to processes such as syntactic processing, lexical knowledge, and the use of contextual cues. For this reason, in the present study, the CT was not treated as a comprehensive measure representing the full multidimensional nature of reading comprehension, but rather as an indicator reflecting a dominant general reading comprehension factor.

Reliability Values

The reliability analyses of both tests indicate a high level of internal consistency. For the MCT, the KR-20 coefficient was .83, and McDonald's ω was .85. For the CT, KR-20 was .86, and McDonald's ω was .88. These values demonstrate that both the

MCT and CT possess strong internal consistency and reliably assess the intended construct (Cronbach, 1951; McDonald, 1999).

Ethical Committee Approval

This study was conducted with the approval of the Social and Humanities Sciences Scientific Research and Publication Ethics Committee of Akdeniz University, dated 10/06/2022 and numbered 10.06.2022-379259.

Data Analysis

During the data collection process, participants were first administered the CT. One week later, the MCT was administered to the same group of participants. Both administrations were conducted in a single session and within one class period. Responses to the CT were scored by two independent raters using a pre-prepared answer key, and inter-rater agreement was calculated as 100%. Scores for the MCT were obtained using the test's automated scoring system.

Prior to the analyses, the dataset was examined to determine its suitability for statistical analysis. No missing data were identified. Potential outliers in the total scores obtained from the CT and MCT were examined using boxplots. The boxplot for the CT indicated that Participant 78 had an extreme value; however, upon inspection of this participant's responses, the high score was judged not to be attributable to random responding and was therefore retained in the dataset.

To address the first research question—Do cloze test scores significantly predict multiple-choice test scores?—the distributions of total CT and MCT scores were examined. The total score of the CT and the total score of the MCT were considered as the dependent and independent variables. Prior to the analysis, the distributional properties of both variables were examined. Normality was assessed using the Shapiro–Wilk test and graphical inspections, which indicated that neither variable followed a normal distribution. Accordingly, non-parametric correlation techniques were employed to examine the association between the two tests.

Within the scope of the second research question ("What is the relationship between students' performance on multiple-choice items targeting different cognitive levels of Bloom's Taxonomy and the scores they obtained from the cloze test?"), the total CT score was treated as the dependent variable, while the independent variables consisted of the total scores derived from item groups in the MCT corresponding to different cognitive levels (e.g., knowledge, comprehension, application, etc.). Accordingly, the MCT was divided into sub-scores based on cognitive levels, and the distributional properties of each sub-score were examined separately. Since the assumption of normality was not met, non-parametric statistical techniques were also employed at this stage.

Following the correlational analyses, regression analyses were conducted to examine whether CT scores were statistically explained by MCT cognitive-level scores. In these models, the CT total score was specified as the explained variable,

and the total scores corresponding to the different cognitive levels of the MCT were specified as explanatory variables. Because the study aimed to evaluate the joint contribution of all cognitive-level scores rather than their isolated effects, the enter method was used to include all explanatory variables simultaneously.

Before conducting the regression analyses, key assumptions were examined. Linearity between the explained and explanatory variables was assessed using scatterplots. The normality of residuals was evaluated using Normal Q–Q plots and the Shapiro–Wilk test, and homoscedasticity was checked by inspecting the distribution of residuals against predicted values. To assess multicollinearity among explanatory variables, Variance Inflation Factor (VIF) and tolerance values were calculated; all VIF values were below 10 and all tolerance values exceeded .10, indicating that multicollinearity was not a concern. These results suggest that the assumptions required for regression analysis were adequately met.

To address the third research question—How do multiple-choice and cloze tests align in terms of participants' reading proficiency classifications?—each participant's total scores on the MCT and CT were converted into three reading proficiency levels (struggling, instructional, and independent) based on the threshold values proposed by Çetinkaya (2010). In this study, reading proficiency level refers to the classification of participants' competence based on their reading comprehension performance.

According to these thresholds, participants scoring below 50% on the MCT were classified as struggling, those scoring between 50% and 70% as instructional, and those scoring above 70% as independent. Similarly, participants scoring below 35% on the CT were classified as struggling, those scoring between 35% and 50% as instructional, and those scoring above 50% as independent. Each participant was assigned a reading proficiency level separately for each test type, and the consistency between the two classifications was examined.

Based on these classifications, a 3×3 contingency matrix was constructed, and frequency and percentage values were calculated for each level combination. Finally, Cohen's Kappa coefficient was computed to determine the degree of agreement between reading proficiency classifications derived from the MCT and the CT. In this analysis, the independent variable was the test type (MCT–CT), and the dependent variable was the reader level classification.

Results

Before proceeding with the analyses aimed at answering the research questions, descriptive statistics were conducted using the scores participants obtained from the CCT and the CDT, and the results are presented in Table 4.

Table 4

Descriptive Statistics for the Scores Obtained from the Tests

	MCT	CT
N	102	102
Mean	10,83	19,79
Sdt. Deviation	3,874	9,81
Minimum	1	2
Maximum	18	46

According to Table 4, a total of 102 students completed both tests. The mean score obtained from the multiple-choice test (MCT) was 10.83, while the mean score obtained from the cloze test (CT) was 19.79. The standard deviation was 3.87 for the MCT and 9.81 for the CT. The minimum and maximum scores ranged from 1 to 18 for the MCT and from 2 to 46 for the CT, respectively.

In order to address the first and second research questions together, this section examines the relationship between cloze test (CT) scores and multiple-choice test (MCT) scores, as well as the relationships between these test scores and different cognitive levels of Bloom's Taxonomy. Accordingly, descriptive statistics for the total MCT and CT scores were first calculated, and the relationships between overall test scores and sub-scores corresponding to Bloom's cognitive process levels were examined using correlation analysis. Subsequently, regression analyses were conducted to determine the extent to which CT performance explains MCT performance. Prior to the analyses, the distributional properties of the variables were examined using the Shapiro–Wilk test and graphical inspections. Although deviations from normality were observed, regression analyses were nonetheless applied in order to explore the predictive relationships between the variables. The results of these analyses are presented in Tables 5 and 6.

Table 5

Means, Standard Deviations, and Correlations Among Test Scores

Test	X̄±S.S	1	2	3	4	5
(1) MCT	10.83±3.87	1.00				
(2) CT	19.79±9.81	.69*	1.00			
(3) Remembering	3.54±1.31	.77*	.55*	1.00		
(4) Understanding	4.00±1.67	.85*	.56*	.47*	1.00	
(5) Applying	.71±.69	.59*	.35*	.39*	.39*	1.00
(6) Analyzing	2.58±1.36	.80*	.58*	.48*	.57*	.26*

*p<.01

Descriptive statistics for the reading comprehension tests and the correlations between them are presented in Table 5. The mean score for the MCT was 10.83 (SD = 3.87), while the mean score for the CT was 19.79 (SD = 9.81). Correlation analysis

revealed a strong positive relationship between MCT and CT scores, r(MCT, CT) = .69, p < .01.

When examining the relationship between test scores and the cognitive levels of Bloom's Taxonomy—specifically remembering, understanding, applying, and analyzing—the remembering-level scores were found to correlate significantly and positively with both the MCT (r = .77, p < .01) and the CT (r = .55, p < .01). The understanding-level scores also demonstrated a strong correlation with MCT performance (r = .85, p < .01), while the correlation with CT performance was moderate (r = .56, p < .01).

Application-level scores showed a moderate correlation with MCT scores (r = .59, p < .01) and a low correlation with CT scores (r = .35, p < .01). In contrast, analysis-level scores were strongly correlated with MCT scores (r = .80, p < .01) and moderately correlated with CT scores (r = .58, p < .01).

Table 6

Results of the Regression Analysis on the Prediction of Reading Comprehension Test Scores

Dependent	Predictor	B (e)	β	t	p	sd	F	p	R	R2
MCT	Constant	6.34 (.55)		11.52	<.001	1 100	90,02	<.001	.69	.47
	CT	.35 (.04)	.69	9.49	<.001					
CT	Constant	-2.11 (1.76)		1.19	.236	3 98	29.55	<.001	.69	.48
	Analyzing	1.68 (.52)	.30	3.23	.002					
	Remembering	1.63 (.50)	.28	3.23	.002					
	Understanding	1.19 (.42)	.26	2.81	.006					

The regression analysis was conducted in two stages: the first involved a simple linear regression model predicting MCT scores from CT scores, while the second stage involved a multiple regression model predicting CT scores from the cognitive levels of Bloom's Taxonomy.

In the first model, CT scores were found to be a significant predictor of MCT scores (B = .35 (.04), β = .69, p < .001). The regression model was statistically significant, F(1, 100) = 90.02, p < .001, and CT scores explained 47% of the total variance (R^2 = .47). This result indicates that MCT performance is largely associated with CT performance and that CT provides a strong indicator of reading comprehension ability.

In the second model, CT scores were significantly predicted by the analysis, remembering, and understanding levels of Bloom's Taxonomy. Specifically, the

analysis level (B = 1.68 (.52), β = .30, p = .002), remembering level (B = 1.63 (.50), β = .28, p = .002), and understanding level (B = 1.19 (.42), β = .26, p = .006) all made significant contributions to the model. The overall regression model was significant, $F(3, 98) = 29.55$, $p < .001$, and explained 48% of the variance in CT scores ($R^2 = .48$). These findings demonstrate that CT performance is meaningfully associated with Bloom's cognitive levels, particularly the analysis, remembering, and understanding levels.

Each student was assigned a separate reading level for each test type based on these score intervals, and the consistency of the two tests in determining reading proficiency was then compared accordingly. Table 7 presents the comparison of participants' reading levels.

Table 7

Comparison of MCT and CT in Terms of Reading Proficiency Levels

MCT	CT	f	%
Independent	Independent	14	13.7
Instructional	Instructional	11	10.8
Frustrated	Frustrated	27	26.5
Independent	Instructional	19	18.6
Independent	Frustrated	17	16.7
Instructional	Independent	4	3.9
Instructional	Frustrated	7	6.9
Frustrated	Instructional	3	2.9
Total		102	100

As shown in Table 7, 52 participants were classified at the same reader level based on their scores on the Multiple-Choice Test (MCT) and the Cloze Test (CT). These participants were placed in the frustrated, instructional, or independent reader levels in both tests. In contrast, 50 participants were classified at different reader levels across the two test types. Specifically, 19 participants were classified as independent according to the MCT but instructional according to the CT, while 17 participants were classified as independent based on the MCT but frustrated based on the CT. In addition, 7 participants were classified as instructional according to the MCT but frustrated according to the CT, and 3 participants were classified as frustrated according to the MCT but instructional according to the CT.

To evaluate the agreement between MCT and CT in classifying students' reading proficiency levels, a Kappa analysis was conducted. Although the Kappa coefficient

was statistically significant, the level of agreement between the two tests was relatively low ($\kappa = 0.231$, $p = .000$), suggesting that the two test formats do not consistently place students in the same reading proficiency categories.

Discussion, Conclusion and Suggestions

This study examined the relationship between scores obtained from multiple-choice tests (MCT) and cloze tests (CT), which are widely used to assess reading comprehension, and evaluated the extent to which these two test types align in classifying reading proficiency levels. The findings of the study can be summarized as follows.

A strong and positive correlation was found between MCT and CT scores ($r = 0.783$), indicating a high degree of association between the two test formats in measuring reading comprehension performance.

The results of the regression analysis showed that CT performance statistically explained MCT performance. Specifically, 47% of the total variance in MCT scores was explained by CT scores. The regression model indicated that a one-unit increase in CT score was associated with a 0.35-unit increase in MCT score.

Analysis of the cognitive-level sub-scores within the MCT revealed that items targeting the remembering, understanding, and analyzing levels of Bloom's Revised Taxonomy significantly explained CT performance ($R^2 = 0.48$). Among these, analyzing-level items emerged as the strongest explanatory variable ($\beta = 0.30$), suggesting that analytical processing plays a critical role in performance on the cloze test. Remembering ($\beta = 0.28$) and understanding ($\beta = 0.26$) levels also showed significant associations with CT performance, whereas application-level items did not contribute significantly. These findings indicate that cloze test performance is more closely associated with mid-level cognitive processes.

Despite the strong associations observed at the score level, inconsistencies were identified between the reading proficiency classifications derived from the two test types. The results of the Kappa analysis ($\kappa = 0.189$) indicated a low level of agreement between the MCT and CT in classifying participants' reading proficiency levels. Notably, 55% of the participants were assigned to different reading proficiency levels depending on the test type.

This study revealed a strong positive relationship between scores obtained from multiple-choice tests (MCT) and cloze tests (CT). In addition, CT performance was found to significantly explain MCT performance. These findings are consistent with previous research reporting substantial associations between cloze and multiple-choice measures of reading comprehension (Baldauf Jr. & Propst Jr., 1979; Bensoussan, 1984; Bormuth, 1967; Rankin & Culhane, 1969; Sattarpour & Ajideh, 2014). The strong association between MCT and CT scores can be attributed to the fact that both test types are sensitive to shared cognitive processes involved in reading comprehension. Both formats require readers to extract meaning from text, use

contextual cues, and demonstrate linguistic awareness (Andreassen & Bråten, 2010; Das et al., 2019; Jonz, 1990). From this perspective, MCTs and CTs appear to provide broadly consistent indicators of reading comprehension performance. However, it should also be noted that the scores obtained from the two test formats do not exhibit perfect alignment, which is an expected outcome.

Although MCTs and CTs target overlapping cognitive processes, the specific cognitive operations required by their formats differ. Multiple-choice tests emphasize recognition, discrimination among alternatives, and test-taking strategies, whereas cloze tests place greater demands on contextual integration, lexical retrieval, and production-based processes. In addition, cloze tests may be more sensitive to variables such as vocabulary knowledge and morphological awareness, which are related to but not identical with reading comprehension. Therefore, observed discrepancies between individual scores across the two test types should be interpreted not as reflecting differences in the underlying construct, but rather as consequences of format-specific cognitive and psychometric characteristics.

These findings indicate that both test types are sensitive to certain common cognitive processes associated with reading comprehension. However, considering the structure of the cloze test, which is based on systematic word deletion, the results do not suggest that this test alone encompasses all cognitive dimensions of reading comprehension. The statistical relationship observed with items at the analysis level suggests that the cloze test is indirectly related to higher-order cognitive processes rather than directly measuring them. The findings provide explanatory insights into the cognitive processes associated with reading comprehension as reflected in cloze tests. Nevertheless, more comprehensive research is required to reach a generalizable conclusion regarding whether cloze tests can replace multiple-choice tests or function as complementary assessment tools.

The finding that understanding-level items significantly explained CT performance indicates that comprehension-based cognitive processes play a central role in cloze test performance. Cloze tasks require readers to infer missing words by using contextual cues (Kleijn et al., 2019), while understanding-level items assess students' ability to construct a coherent representation of the text and identify relationships among its components (Anderson et al., 2001, p. 31; Anikin & Sychev, 2020; Verenna et al., 2018). Accordingly, the predictive relationship between understanding-level performance and CT scores is theoretically expected, as both rely on similar interpretive skills.

The significant contribution of analysis- and remembering-level items further suggests that CT performance is not solely dependent on contextual guessing, but is also related to processes involving the analysis and retrieval of textual information. Analysis-level items require learners to decompose information and examine relationships among elements (Anderson et al., 2001, p. 31; Jensen et al., 2014), a skill that may support the accurate completion of missing words in cloze tests. Similarly, the influence of remembering-level items implies that the ability to retain and retrieve

specific information from the text contributes to CT performance. These findings indicate that CTs are aligned with both lower-level linguistic processes and information-access skills.

In contrast, the limited contribution of application-level items to CT performance can be explained by the nature of these items. Application-level questions aim to assess the ability to use learned knowledge in novel or real-life situations (Anderson et al., 2001, p. 31; Monrad et al., 2021), whereas cloze tests primarily focus on understanding the immediate textual context and selecting a word that fits that context. Consequently, application-level performance appears to be less closely aligned with the cognitive demands of CTs.

Overall, the findings suggest that cloze tests are effective in capturing mid-level cognitive processes such as remembering, understanding, and analyzing. However, this should not be interpreted as evidence that cloze tests comprehensively assess higher-order cognitive comprehension. Measuring advanced cognitive processes would require test formats specifically designed to elicit complex reasoning. Nevertheless, the present results differ from studies suggesting that cloze tests assess only low-level skills, indicating instead that CTs can serve as effective tools for evaluating both lower- and mid-level cognitive processes associated with reading comprehension.

With respect to the classification of readers based on comprehension performance, the study found a low level of agreement between MCT- and CT-based reading level classifications. This finding appears to contrast with the strong correlation observed at the score level. The primary reason for this discrepancy lies in the cut-off scores used to assign readers to categorical levels. Small differences in raw scores may lead to different categorical classifications, thereby reducing agreement between test formats. As a result, relying solely on test scores for reader classification may lead to unstable or inconsistent categorizations.

Regarding the classification of students based on their reading comprehension achievement, the level of agreement between the MCT and CT was found to be quite low. This result appears to contradict the findings concerning the relationship between the scores obtained from the two tests. The primary reason for this discrepancy lies in the cut-off scores used to classify students into reader categories. For example, in this study, a student who scored 8 points on the MCT was classified as a frustration-level reader, whereas a student who scored 9 points was classified at the instructional level. Similarly, a reader who obtained 23 points on the CT was placed at the frustration level, while a reader with 24 points was classified at the instructional level. In such cases, small differences in scores become decisive in determining students' reading levels. For this reason, the two test types demonstrated low agreement in identifying reader categories. It can therefore be

argued that relying solely on test scores to classify readers into different levels may not be appropriate.

This study employed a single cloze test constructed using a systematic word deletion procedure. However, previous research has shown that deletion frequency can influence both the cognitive difficulty of cloze tests and the extent to which scores reflect textual properties (Alderson, 1979). Similarly, the type of deleted words has been shown to affect test difficulty (Henk, 1981). Studies by Ulusoy (2009) and Bilki (2011) further demonstrate that different deletion strategies and text types can yield different assessment outcomes, even when based on the same text. Future research employing cloze tests with systematically varied deletion frequencies and word types (e.g., nouns, verbs, connectives) may provide deeper insights into test design and the measurement of reading comprehension.

In conclusion, this study demonstrates that multiple-choice and cloze tests developed from the same text exhibit largely similar patterns in measuring reading comprehension performance. While a strong and significant relationship was observed between total scores, the agreement between reading level classifications was limited. These findings suggest that cloze tests, when used alongside multiple-choice tests, may provide complementary information about reading comprehension. At the same time, differences in scores and classifications should not be attributed solely to item format, but rather to the broader cognitive and psychometric characteristics inherent in different measurement approaches.

E-ISSN: 2458-8342, P-ISSN: 1301-3718
Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi
2026, 59(1), 221-271
DOI: 10.30964/auebfd.1693237
Araştırma Makalesi

Çoktan Seçmeli ve Boşluk Doldurma Testlerinin Okuduğunu Anlama ve Okur Düzeyleri Açısından Karşılaştırılması

Yusuf Aydın [1], Ezgi Kaya Filik [2]

1 Türkçe ve Sosyal Bilimler Eğitimi Anabilim Dalı, Eğitim Fakültesi, Akdeniz Üniversitesi, Antalya, Türkiye
2 Milli Eğitim Bakanlığı, İzmir, Türkiye

ÖZ

Bu çalışmanın amacı, bir çoktan seçmeli test ve boşluk doldurma testinden elde edilen öğrenci performanslarını, bu testlerin dayandığı bilişsel süreçler ve işleyiş özellikleri açısından karşılaştırmaktır. Araştırmaya yedinci ve sekizinci sınıfta öğrenim gören toplam 102 ortaokul öğrencisi katılmıştır. Araştırma korelasyonel bir desenle yürütülmüş ve veriler nicel istatistiksel analizler kullanılarak çözümlenmiştir. Sonuçlar çoktan seçmeli test puanı ile boşluk doldurma testi puanı arasında pozitif yönde yüksek bir ilişki olduğunu ve boşluk doldurma testi puanlarının çoktan seçmeli test puanlarını anlamlı düzeyde yordadığını ortaya koymuştur. Bu bulgular, her iki test türünün ortak bilişsel süreçleri ölçtüğünü ve birbirlerini tamamlayıcı ölçme araçları olarak kullanılabileceğini göstermektedir. Ayrıca araştırmada, Yenilenmiş Bloom Taksonomisi'nin farklı bilişsel düzeylerine karşılık gelen çoktan seçmeli test maddelerinden elde edilen toplam puanların, boşluk doldurma testi başarısını yordama gücü incelenmiştir. Çözümleme düzeyindeki soruların, boşluk doldurma testindeki başarıyı yordamada diğer bilişsel düzeylere kıyasla daha yüksek bir yordama gücüne sahip olduğu görülmüştür. Hatırlama ve anlama düzeyindeki sorular da boşluk doldurma test başarısını anlamlı bir şekilde yordarken uygulama düzeyindeki soruların anlamlı bir etkisi görülmemiştir. Bu durum, boşluk doldurma testlerinin özellikle metni çözümleme ve anlama gibi orta düzey bilişsel süreçlerle ilişkili olduğunu göstermektedir. Okur düzeylerinin belirlenmesi konusunda iki test türü arasında sınırlı bir uyum tespit edilmiştir. Katılımcıların yarısından fazlası, iki test türündeki performasnlarına göre farklı okur düzeylerinde yer almaktadır.

Anahtar sözcükler: Çoktan seçmeli test, boşluk doldurma testi, okur düzeyi, okur düzeyi sınıflaması, Yenilenmiş Bloom'un Taksonomisi

Sorumlu yazar: Yusuf Aydın
Türkçe ve Sosyal Bilimler Eğitimi Anabilim Dalı, Eğitim Fakültesi, Akdeniz Üniversitesi, Antalya, Türkiye
E-posta: ysf.aydn66@gmail.com **ORCID ID:** 0000-0003-0898-9020 **ROD ID:** https://ror.org/01m59r132
Geliş tarihi: 07.05.2025 **Kabul tarihi:** 28.02.2026 **Yayın Tarihi:** 15.04.2026
Atıf Bilgisi: Aydın, Y., & Kaya-Filik, E. (2026). Çoktan seçmeli ve boşluk doldurma testlerinin okuduğunu anlama ve okur düzeyleri açısından karşılaştırılması *Ankara University Journal of Faculty of Educational Sciences*, 59(1), 221-271. https://doi.org/10.30964/auebfd.1693237
Makaleye ilişkin tüm etik beyanlar, çalışmanın son sayfasında yer almaktadır (Sayfa: 271).

Okuduğunu anlama, yazılı metinlerden anlam çıkarma ve bilgiyi işleme sürecidir. Bu karmaşık süreç çok sayıda bilişsel işleme dayanır (Kendeou, McMaster ve Christ, 2016; Tighe ve Schatschneider, 2016). Okuma becerisi bireylerin akademik başarıları ve sosyal yaşamları için oldukça önemlidir. Okuma, öğrencilerin bilgiye erişimini sağlar ve yeni kavramlar öğrenmelerine yardımcı olur (Pretorius, 2002). Okuma becerisi akademik gelişimin yanı sıra sosyal ve duygusal yeteneklerin gelişiminde de önemli bir rol oynar. Araştırmalar sosyal ve duygusal yetenekler ile okuduğunu anlama arasında pozitif bir korelasyon olduğunu göstermektedir (Yu, Yu ve Tong, 2023). Ayrıca okumanın zihinde canlandırma becerisini geliştirdiği ortaya koyulmuştur (Pino ve Mazza, 2016). Özellikle kurgu metin okuma, sosyal ve duygusal öğrenmeyi destekleyen bir öğrenme bağlamı sunar. (Kozak ve Recchia, 2019). Bu nedenle okuduğunu anlama becerisinin geliştirilmesi, bireylerin hem akademik hem de sosyal alanlarda başarılı olmaları için önem taşımaktadır. Bu denli çok boyutlu ve işlevsel bir beceri olan okuduğunu anlamayı ölçmenin geçerli ve güvenilir biçimde yapılması gerekmektedir. Nitekim okuduğunu anlama becerisinin nasıl ölçüldüğü, ölçülen performansın hangi bilişsel süreçleri ne ölçüde yansıttığı ölçme ve değerlendirme alanı açısından temel bir tartışma konusudur.

Okuduğunu anlama becerisinin değerlendirilmesi, bireylerin bilgi işleme yeteneklerini ve metinlerden anlam çıkarma düzeylerini belirlemek açısından önemlidir. Bu değerlendirme okuma becerisini ölçen çeşitli testler aracılığıyla gerçekleştirilir. Okuma becerisini ölçme ve değerlendirmede kullanılan başlıca test türleri arasında çoktan seçmeli testler, açık uçlu sorular, evet-hayır/doğru-yanlış türü sorular, boşluk doldurma testleri, kısa yanıtlı sorular, sıralama soruları, yorum ve değerlendirme soruları ile özetleme soruları yer almaktadır (Ülper, 2010, s. 123–136). Bu çeşitli test teknikleri, okuduğunu anlama becerisinin çok yönlü doğasını yansıtmaktadır ve bu becerinin farklı boyutlarını ölçmek için tasarlanmıştır.

Okuduğunu anlama, yalnızca metindeki bilgiyi hatırlama değil, aynı zamanda anlamlandırma, çözümleme, sentezleme ve değerlendirme gibi üst düzey bilişsel süreçleri de içerir. Dolayısıyla bu becerinin değerlendirilmesi için kullanılan test teknikleri de bu bilişsel süreçleri kapsayacak şekilde çeşitlendirilmiştir. Bununla birlikte yukarıda bahsedilen değerlendirme teknikleri arasında çoktan seçmeli sorular en yaygın olarak kullanılmaktadır (Ozuru ve diğerleri, 2007). Türkiye'de çoktan seçmeli sorular, öğretimin neredeyse her düzeyinde, öğrencilerin yalnızca okuduğunu anlama başarılarını değil, aynı zamanda diğer tüm alanlardaki akademik başarılarını ölçmek için de yaygın bir biçimde kullanılmaktadır. Uluslararası ölçekte de bu tür sorular, uzun yıllardır PISA, TOEFL, IELTS, SAT ve GRE gibi sınavlarda yer almaktadır.

Çoktan seçmeli testlerin tercih edilmesinin başlıca nedenleri arasında, üst düzey düşünme becerilerini ölçebilme yetenekleri, düşük maliyetleri ve yüksek güvenilirlikleri yer almaktadır (Little ve Bjork, 2015). Ayrıca iyi hazırlanmış çoktan seçmeli testler yüksek kapsam geçerliliğine sahiptir ve değerlendirme sürecinde değerlendiricinin önyargılarından etkilenmezler. Bu testlerin yanıtlanması için

öğrencilerin bir metin oluşturmaları gerekmediğinden yazma becerisi ölçme sürecine dahil olmamakta ve böylece sadece okuduğunu anlama becerisi etkin bir şekilde değerlendirilebilmektedir (Ülper, 2010, s. 123–124). Bu özellikleriyle çoktan seçmeli testler, eğitim sistemlerinde ve ölçme ve değerlendirme uygulamalarında önemli bir yere sahiptir.

Çoktan seçmeli testlerin çeşitli avantajları bulunmakla birlikte, bu testlerin bazı dezavantajları da mevcuttur. Çoktan seçmeli testlerde soruların doğru yanıtlarının rastlantısal olarak seçilme olasılığı bulunmaktadır (Ülper, 2010, s. 124; Cooper ve Foy, 1967). Ayrıca bu tür testlerin geliştirilmesi daha karmaşıktır. Geçerli ve güvenilir testler oluşturmak için ön uygulamaların yapılması gerekmektedir (Ülper, 2010, s. 124; Fuhrman, 1996; Gierl ve diğerleri, 2017). Bunun yanı sıra bazı araştırmacılar çoktan seçmeli testlerin doğrudan okuma becerisini ölçüp ölçmediği konusunda kuşkulara sahiptir. Örneğin Rupp ve arkadaşları (2006) katılımcıların çoktan seçmeli sorulara yanıt vermeyi bir anlama görevi yerine problem çözme görevi olarak algıladıklarını ortaya koymuşlardır. Bu durum, öğrencilerin metin okurken ve soru yanıtlarken farklı stratejiler kullandıklarını göstermektedir. Daneman ve Hannon (2001) ise çoktan seçmeli testlerin metni okumadan yalnızca akıl yürütme yoluyla yanıtlanabildiğini ortaya koymuştur. Keenan ve Betjemann (2006) da çoktan seçmeli testlerdeki bazı soruların öğrenciler tarafından metni okumaya gerek kalmadan yanıtlanabildiğini ifade etmektedir. Bu bulgular, çoktan seçmeli testlerin okuma becerisini ölçmedeki sınırlılıklarını göstermekte ve alternatif değerlendirme yöntemlerinin önemli olduğuna işaret etmektedir.

Okuduğunu anlama becerisini ölçmede kullanılan önemli test türlerinden biri de boşluk doldurma testleridir. Bu testler metinden belirli kelime veya ifadelerin çıkarıldığı ve öğrencilerin bu boşlukları uygun kelimelerle tamamlamalarının beklendiği ölçme araçlarıdır. Boşluk doldurma testlerinin hem ana dili hem de ikinci dil öğretiminde dil becerilerinin ölçülmesinde geçerli ve güvenilir araçlar olarak kullanılabileceği çeşitli araştırmalarla ortaya koyulmuştur (Aitken, 1977; Sumita ve diğerleri, 2005; Ulusoy, 2009; Zhang ve diğerleri, 2023). Ayrıca çeşitli çalışmalar boşluk doldurma testlerinin diğer dil yeterlilik testleriyle yüksek korelasyon gösterdiğini ve dil yeterliliğini güvenilir bir şekilde ölçtüğünü ortaya koymaktadır (Gooskens ve van Heuven, 2017; Gaillard ve Tremblay, 2016; Luchkina ve diğerleri, 2021). Bu bulgular, boşluk doldurma testlerinin dil becerilerinin değerlendirilmesinde kullanılabilecek güvenilir araçlar olduğunu göstermektedir.

Çoktan seçmeli testlerde olduğu gibi boşluk doldurma testlerinin çeşitli zayıf yönleri bulunmaktadır. Birçok araştırmacı, boşluk doldurma testlerinin cümle düzeyindeki anlama becerisinin ötesini ölçmede yetersiz kaldığını belirtmektedir (Kleijn, Pander Maat ve Sanders, 2019; Gellert ve Elbro, 2012). Ayrıca bu testlerin daha çok alt düzeydeki dil becerilerini ölçtüğü de vurgulanmaktadır (Alderson, 1980). Wismann (2018) ve arkadaşları, boşluk doldurma testlerinin öğrencilerin öğrendikleri bilgileri tümdengelimsel akıl yürütme görevlerine transfer etme yeteneğini desteklemediğini ortaya koymuştur. Bu bulgu, boşluk doldurma testlerinin üst düzey

bilişsel becerilerin ölçülmesinde yetersiz kaldığını göstermektedir. Baghaei ve Ravand'ın (2019) araştırması ise boşluk doldurma sorularının genel okuduğunu anlama becerisinin ötesinde bir varyans oluşturduğunu göstermiştir. Yani bu soruların sonuçları yalnızca okuduğunu anlama yeteneğiyle açıklanamamaktadır. Bu durum, boşluk doldurma sorularının hedeflenen okuma becerisinden farklı bir beceriyi ölçtüğünü ortaya koymaktadır. Dolayısıyla boşluk doldurma testlerinin geçerliliği ve ölçtükleri becerilerin kapsamı konusunda ciddi sorgulamalar bulunmaktadır.

Yukarıda özetlenen çalışmalar, çoktan seçmeli ve boşluk doldurma testlerinin okuduğunu anlama becerisini ölçmede çeşitli güçlü ve sınırlı yönlere sahip olduğunu göstermektedir. Bu durum, söz konusu testlerden elde edilen puanların nasıl yorumlanması gerektiği ve bu puanların hedeflenen okuma becerisini ne ölçüde yansıttığı sorusunu gündeme getirmektedir. Dolayısıyla okuduğunu anlama başarısının değerlendirilmesinde kullanılan ölçme araçlarının geçerliliği, üzerinde durulması gereken temel bir kavram olarak ortaya çıkmaktadır.

Okuduğunu anlama başarısının ölçülmesinde yaygın olarak kullanılan çoktan seçmeli testler, iyi yapılandırıldıklarında nesnel ve güvenilir ölçümler sunabilmektedir. Ancak özellikle üst düzey düşünme becerilerini yansıtma ve doğru yanıtın tahmin yoluyla seçilebilme olasılığı bakımından önemli sınırlılıklar taşımaktadır (Puthiaparampil ve Rahman, 2020).

Boşluk doldurma testleri, sözdizimsel yapı, sözcük bilgisi ve bağlamsal ipuçlarından eşzamanlı olarak yararlanmayı gerektirdiği için okuduğunu anlama becerisiyle ilişkili birden fazla dilsel süreci bütüncül biçimde ölçmeye olanak tanımaktadır (Kleijn, Maat ve Sanders, 2019). Standart okuma ve dil yeterliği testleriyle elde edilen yüksek korelasyonlar, bu testlerin özellikle alt ve orta düzey okuma becerilerini ayırt etmede geçerli ölçümler sunabildiğini göstermektedir (Fotos, 1991; Kleijn, Maat ve Sanders, 2019; Bråten, Haverkamp ve Anmarkrud, 2024). Bununla birlikte geleneksel soru formatlarının büyük ölçüde cümle düzeyinde dilsel işlemeye dayanması, metinler arası bütünlük kurma ve üst düzey anlama süreçlerinin yeterince temsil edilememesine yol açabilmektedir.

Çoktan Seçmeli Testler ve Boşluk Doldurma Testleri ile Okur Düzeylerinin Belirlenmesi

Çeşitli okuma anlama testleri kullanarak okurların düzeylerini belirlemek okuduğunu anlama becerisinde karşılaşılan problemlerin kaynağını tespit etmek ve farklı okur alt tiplerini tanımlamak için gereklidir (Auphan ve diğerleri, 2019). İlgili alanyazında çoktan seçmeli testler ile boşluk doldurma testlerinin okur düzeylerini ölçme açısından karşılaştırıldığı araştırmalar 1960'lı yıllara kadar uzanmaktadır. Bormuth (1967) boşluk doldurma ve çoktan seçmeli testlerin okuduğunu anlama seviyelerini ölçme açısından karşılaştırmıştır. Bu araştırmanın sonuçları boşluk doldurma testi puanlarının çoktan seçmeli test puanları ile eşleştirilebileceğini ortaya koymaktadır. Örneğin boşluk doldurma testindeki %38'lik başarı oranı çoktan seçmeli testte %75'e karşılık gelmektedir. Aynı şekilde, boşluk doldurma testinde

%50 başarı gösteren bir öğrencinin çoktan seçmeli testte %90 başarı göstermesi beklenmektedir. Bormuth, araştırmasında boşluk doldurma testleri ile çoktan seçmeli testlerinden alınan puanları 3 düzeye ayırmış ancak bu düzeyler için herhangi bir ad belirlememiştir.

Benzer bir araştırma yürüten Rankin ve Culhane (1969) de boşluk doldurma testi puanlarının çoktan seçmeli testlerdeki yüzdelik başarı oranlarına göre eşdeğer puanlar olarak değerlendirilebileceğini ifade etmiştir. Böylece araştırmacılar Bormut'un araştırmasını desteklemiştir. Ayrıca bu araştırmalarında okurları bağımsız ve eğitsel olmak üzere iki sınıfa ayırmışlardır. Bağımsız okuma seviyesine ulaşmak için çoktan seçmeli testlerde %90 başarı oranı gereklidir. Bu seviyede bir öğrenci, materyali kendi başına okuyup anlayabilir. Aynı bağımsız okuma seviyesini boşluk doldurma testinde yakalayabilmek için öğrencinin boşluk doldurma testindeki maddelerin %61'ine doğru yanıt vermesi gerekmektedir. Bu, öğrencinin metni bağımsız olarak anlama kapasitesine işaret eder. Eğitsel okuma seviyesi için %75 başarı oranı gerekmektedir. Bu seviyedeki bir öğrenci, öğretmen rehberliğinde ya da destekleyici bir öğrenme ortamında materyali anlayabilir. Eğitsel okuma seviyesini karşılamak için boşluk doldurma testine %41 başarı oranı yeterli görülmüştür. Bu düzeyde okuduğunu anlamak için öğretim desteği gerekmektedir.

Çetinkaya'nın (2010, s. 35) konuyla ilgili yapılan araştırmaları bütünleştirerek boşluk doldurma testleri ve çoktan seçmeli testler okur düzeyleri bakımından karşılaştırmıştır. Bu düzeyler engelli, eğitsel ve bağımsız olmak üzere üç grupta sınıflandırılmıştır. Engelli düzeydeki okurlar, boşluk doldurma testlerinde boşlukların %35'inden azını doğru tamamlarken çoktan seçmeli testlerde soruların %50'sinden azına doğru yanıt vermektedir. Eğitsel düzeydeki okurlar, boşluk doldurma testlerinde boşlukların %35-50'sini doğru yanıtlarken çoktan seçmeli testlerde soruların %50-70'ini doğru cevaplayabilmektedir. Bağımsız okur düzeyindekiler ise boşluk doldurma testlerinde %50'den fazla başarı sağlarken çoktan seçmeli testlerde %70'ten fazla doğru yanıt oranına ulaşmaktadır. Bu sınıflandırma, her iki test türünü kullanarak okur düzeyleri belirlemeye yönelik bir çerçeve sunmaktadır.

Bu düzeyler arasındaki uyum az sayıdaki uygulamalı araştırmayla sınanmıştır. Wait (1987) araştırmasında katılımcıların boşluk doldurma testinden aldıkları puanlar ve farklı okuma anlama başarı testinden aldıkları puanlar arasında olumlu yönde anlamlı bir ilişki olduğunu ifade etmektedir. Kızılaslan Tunçer ve Erden (2015) araştırmalarında katılımcıların boşluk doldurma testinden ve çoktan seçmeli testlerden aldıkları puanlar doğrultusunda onları bağımsız, eğitsel ve engelli olmak üzere üç gruba ayırmış ve her iki test türünün okur düzeyini belirlemede orta düzeyde ve pozitif bir korelasyon gösterdiğini belirlemiştir.

Okuduğunu anlama becerisini ölçmek için yaygın olarak kullanılan çoktan seçmeli ve boşluk doldurma testleri arasındaki ilişkinin incelenmesi, eğitimde ölçme ve değerlendirme süreçlerinin iyileştirilmesi açısından önemlidir. Bu çalışma, iki test türü arasındaki uyumu ele almaktadır. Önceki araştırmalar, çoktan seçmeli ve boşluk doldurma testlerini genellikle okuduğunu anlama başarısını ölçme sonuçları ve bu

sonuçların birbiriyle ilişkisi üzerinden karşılaştırmıştır. Bu araştırmalarda farklı madde türleriyle elde edilen puanların ne ölçüde örtüştüğü ya da ayrıştı incelemiştir. Bunun yanı sıra öğrencilere farklı test formatları uygulanarak gözlenen performans farklılıklarının test türüyle ilişkisi de ele alınmıştır. Bu araştırma ise söz konusu yaklaşımlardan farklı olarak testleri yalnızca toplam puan düzeyinde karşılaştırmakla yetinmemekte, çoktan seçmeli test maddelerini Yenilenmiş Bloom Taksonomisi'nin farklı bilişsel düzeylerinde ölçme yapmayı hedefleyen maddeler olarak sınıflandırmakta ve bu düzeylere ait alt puanların boşluk doldurma testi başarısıyla ilişkisini de incelemektedir. Böylece çalışmada, iki test türünün okuduğunu anlama becerisini ölçerken hangi bilişsel süreçlerle daha fazla örtüştüğü, özellikle Bloom'un bilişsel düzeyleri bağlamında ele alınmakta ve boşluk doldurma testinin ölçtüğü performansın bilişsel temellerine ilişkin daha ayrıntılı bir çözümleme sunulmaktadır.

Okur düzeylerini belirlemek için Çetinkaya'nın (2010) önceki araştırma bulgularına dayalı olarak oluşturduğu sınıflandırma kullanılmıştır. Araştırma, öğrencilerin çoktan seçmeli testler ile boşluk doldurma testlerinden aldıkları puanlar arasındaki ilişkiyi ve test sonuçlarının sınıf düzeyine göre değişimini inceleyerek okuduğunu anlama becerisini ölçen araçlarının geliştirilmesine yönelik bulgular sunmayı amaçlamaktadır.

Bu araştırmanın amacı, okuduğunu anlama becerisini ölçmeye yönelik olarak kullanılan çoktan seçmeli test ve boşluk doldurma testinin ölçme sonuçları arasındaki ilişkiyi incelemektir. Bu iki test türünün okuduğunu anlama ile ilişkili bilişsel süreçlerle, özellikle Bloom'un bilişsel düzeyleri bağlamında, nasıl bir örüntü sergilediğini ortaya koymaktır. Bu amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır:

1. Boşluk doldurma testi puanları, çoktan seçmeli test puanlarını anlamlı şekilde yordamakta mıdır?

2. Öğrencilerin Bloom Taksonomisi'nin farklı bilişsel düzeylerinde ölçme yapmayı hedefleyen çoktan seçmeli maddelerdeki performansları ile boşluk doldurma testinden elde ettikleri puanlar arasındaki ilişki nasıldır?

3. Çoktan seçmeli test ile boşluk doldurma testi katılımcıların okur düzeyi ile ilgili nasıl bir uyum göstermektedir?

Yöntem

Yöntem bölümü, araştırma desenini, katılımcıları, veri toplama araçlarını ve çalışmada kullanılan uygulama süreçlerini açıklamaktadır. Bu bölümde, çoktan seçmeli ve boşluk doldurma okuduğunu anlama testlerinin geliştirilme süreci ve psikometrik özellikleri ele alınmakta, verilerin nasıl toplandığı ve puanlandığı açıklanmakta, test puanları ile okur yeterliği arasındaki ilişkiyi incelemek için yapılan istatistiksel analizlere yer verilmektedir.

Araştırma Modeli

Çoktan seçmeli ve boşluk doldurma türünde okuduğunu anlama testlerindeki öğrenci performanslarını karşılaştıran bu araştırma korelasyonel bir araştırmadır. Korelasyonel araştırmalarda "iki ya da daha çok değişken arasındaki ilişki herhangi bir şekilde bu değişkenlere müdahale edilmeden incelenir" (Büyüköztürk ve diğerleri, 2013, s. 184). Bu araştırmanın odağında katılımcıların okuduğunu anlama başarılarını ölçen çoktan seçmeli testten ve boşluk doldurma testinden aldıkları puanlar arasındaki korelasyonun incelenmesi yer almaktadır.

Bu çalışmada çoktan seçmeli test (ÇST) ve boşluk doldurma testinden (BDT) elde edilen puanlar arasındaki ilişkiyi daha ayrıntılı biçimde incelemek amacıyla regresyon analizi kullanılmıştır. Bu analizde yer alan değişkenler, nedensel bir yordama ilişkisi varsayılmaksızın istatistiksel açıdan açıklayıcı bir modelleme çerçevesinde ele alınmıştır. İlk aşamada ÇST puanları modelde açıklanan değişken olarak ele alınmış, BDT puanları ise açıklayıcı değişken olarak modele dâhil edilmiştir. İkinci aşamada ise BDT puanları açıklanan değişken olarak ele alınmış, Yenilenmiş Bloom Taksonomisi'nin hatırlama, anlama, uygulama ve çözümleme düzeylerinde yapılandırılmış çoktan seçmeli maddelerden elde edilen puanlar modele açıklayıcı değişkenler olarak dâhil edilmiştir. Bu yaklaşım, testler arasındaki ilişkinin yönünü değil, farklı madde türlerinin ve bilişsel düzeylerin öğrenci performansıyla nasıl ilişkilendiğini ortaya koymayı amaçlamaktadır.

Bu çalışmada ÇST'deki maddeler Yenilenmiş Bloom Taksonomisi'ne göre hatırlama, anlama, uygulama ve çözümleme düzeylerine göre sınıflandırılmıştır. Her düzey için elde edilen alt puanlar (örneğin hatırlama düzeyine ait 5 maddenin toplam puanı) ayrı ayrı hesaplanmıştır. Bu alt puanlar BDT puanlarını yordamak amacıyla regresyon analizine dahil edilmiştir. Böylece BDT'nin Bloom'un hangi bilişsel düzeylerine daha fazla karşılık geldiği incelenmiştir. BDT için ise bilişsel düzey sınıflaması yapılmamıştır. Çünkü BDT tek bir metinden sistematik sözcük silme yoluyla oluşturulmuştur ve her bir boşluk ayrı ayrı bilişsel düzeye atanabilir nitelikte değildir. Bu nedenle BDT, bütüncül bir performans ölçüsü olarak ele alınmıştır. Bu analiz BDT'nin özellikle hangi tür bilişsel işlemlerle daha fazla örtüştüğüne dair açıklayıcı bilgiler sunmayı amaçlamaktadır.

Çalışma Grubu

Araştırma iki aşamadan oluşmaktadır. İlk aşamada veri toplama araçları geliştirilmiş ve uygulanmıştır. Veri toplama araçlarının geliştirilmesi aşamasında araştırmaya 216 ortaokul öğrencisi katılmıştır. Bu aşamada araştırmaya Antalya il merkezinde yer alan orta düzey sosyoekonomik yapıya sahip bir devlet ortaokulunda öğrenim gören toplam 216 öğrenci katılmıştır. Bu öğrencilerden 115'ü 7. sınıf, 101'i ise 8. sınıf öğrencisidir. Katılımcıların %52'si kız (n=112), %48'i erkek (n=104) öğrencilerden oluşmaktadır.

Araştırmanın ikinci aşamasında ise geçerlik ve güvenirlik analizleri yapılmış olan testler, yine Antalya ilinde benzer sosyoekonomik düzeye sahip farklı bir devlet

ortaokulunda uygulanmıştır. Bu aşamada araştırmaya katılan 102 öğrencinin 60'ı 7. sınıf, 42'si 8. sınıf seviyesindedir. Katılımcıların %55'i kız (n=56), %45'i erkek (n=46) öğrencilerden oluşmaktadır.

Katılımcılar, erişilebilirlik ve gönüllülük esasına dayalı olarak uygun örnekleme yöntemiyle seçilmiştir. Örneklem veri toplama sürecinde erişilebilen hedef öğrenci kitlesidir.

Veri Toplama Araçları

Katılımcıların okuduğunu anlama başarısını ölçmek için araştırmacılar tarafından Çoktan Seçmeli Okuduğunu Anlama Testi (ÇST) ve Boşluk Doldurma Testi (BDT) geliştirilmiştir. Aşağıda bu testlerin geliştirilme sürecine ilişkin açıklamalar yapılmıştır:

Çoktan* Seçmeli *ve Boşluk Doldurma Okuma Anlama Testlerine Ait Psikometrik Özellikler

ÇST "İyi Uykular, Tatlı Rüyalar" (Gençer, 2014) adlı metne dayalı 19 çoktan seçmeli sorudan oluşmaktadır. Testte yer alan her soru için 4 seçenek yer almaktadır. Testi geliştirmek için aynı metne dayalı olarak 24 soru hazırlanmıştır. Soruların hazırlanmasında Yenilenmiş Bloom Taksonomisi dikkate alınmıştır. Sorulardan 6'sı hatırlama, 9'u anlama, 7'si çözümleme ve 2'si uygulama düzeyindedir. Okuduğunu anlama testlerinin geliştirilmesi sürecinde, testin psikometrik özelliklerinin yalnızca maddelerden değil, aynı zamanda kullanılan metnin özelliklerinden de etkilendiği göz önünde bulundurulmuştur. Bu doğrultuda hazırlanan okuma metni ve buna bağlı olarak geliştirilen maddeler, iki eğitim programları alan uzmanı ve bir Türkçe eğitimi alan uzmanının görüşüne sunulmuştur. Uzmanlar, maddelerin dilsel uygunluğu ve ölçme amacına hizmet etme düzeyinin yanı sıra, kullanılan metnin hedef yaş grubuna uygunluğu, okunabilirliği ve metnin bilgi/olay yoğunluğunun testin güçlüğü üzerindeki olası etkileri açısından da değerlendirme yapmıştır. Uzman görüşleri doğrultusunda soru sayısı 20'ye düşürülmüş, metin ve madde uyumunu güçlendirecek biçimde gerekli düzenlemeler yapılmıştır.

Çalışmanın temel odak noktalarından biri, çoktan seçmeli maddelerin Yenilenmiş Bloom Taksonomisi'nin farklı bilişsel düzeylerinde ölçme yapmayı hedeflemesidir. Bu nedenle her bir çoktan seçmeli madde, uzmanlar tarafından hedeflediği bilişsel düzey açısından ayrı ayrı incelenmiştir. Uzmanlardan, maddelerin hangi bilişsel düzeyi ölçmeyi amaçladığına ilişkin değerlendirme yapmaları istenmiş, bu değerlendirmeler doğrultusunda maddelerin bilişsel düzeylere dağılımı belirlenmiştir. Uzman görüşleri arasında görüş birliği sağlanamayan maddeler yeniden düzenlenmiş veya ölçme aracından çıkarılmıştır. Böylece testte yer alan maddelerin hedeflenen bilişsel düzeyleri temsil etmesine yönelik uzman yargısına dayalı geçerlik kanıtları elde edilmiştir.

Ayrıca testin uygulama sonrasında elde edilen madde istatistikleri ile maddelerin bilişsel düzeylere göre dağılımı birlikte ele alınarak ölçme aracının hem istatistiksel

hem de kuramsal açıdan değerlendirilmesi amaçlanmıştır. Bu biçimiyle testteki soruların 6'sı hatırlama, 7'si anlama, 5'i çözümleme ve 2'si uygulama düzeyindedir. Hatırlama düzeyinde olan 10. sorunun uygulama sırasında hatalı olduğu fark edilmiş ve soru testten çıkarılmıştır. Uygulanma sürecinde yapılan kontrol sırasında söz konusu sorunun seçeneklerinden birinin hatalı oluşturulduğu fark edilmiştir. Bu nedenle ilgili soru geçerlilik ve güvenilirlik analizlerinde değerlendirmeye alınmamıştır. Hata teknik bir uygulama hatasıdır ve çalışmanın genel bulgularını etkilemeyecek şekilde gerekli düzenleme yapılmış, analizler geri kalan maddeler üzerinden sürdürülmüştür. Bu nedenle testteki toplam madde sayısı 19'a ve hatırlama düzeyinde yer alan madde sayısı 5'e inmiştir. Tablo 1'de 19 sorudan oluşan ÇST'nin her sorusu için güçlük ve ayırt edicilik değerleri sunulmaktadır.

Madde güçlük indeksleri, her bir sorunun doğru yanıtlanma oranını (p) gösteren klasik test kuramına dayalı olarak hesaplanmıştır. Madde ayırt edicilik indeksleri ise 27% üst ve alt grupların her maddeye verdikleri doğru yanıt yüzdeleri arasındaki fark alınarak belirlenmiştir.

Tablo 1

ÇST Madde Analizleri

Soru	Madde Güçlük İndeksi	Madde Ayırt Edicilik İndeksi
S 1	0.87	0.35
S 2	0.44	0.60
S 3	0.74	0.47
S 4	0.72	0.36
S 5	0.73	0.57
S 6	0.87	0.36
S 7	0.69	0.54
S 8	0.81	0.40
S 9	0.48	0.43
S 10	0.62	0.51
S 11	0.70	0.67
S 12	0.51	0.71
S 13	0.32	0.40
S 14	0.46	0.42
S 15	0.65	0.61
S 16	0.68	0.64
S 17	0.49	0.67
S 18	0.51	0.50
S 19	0.79	0.46

Tablo 1'deki güçlük indeksi değerleri 0.32 ile 0.87 arasında değişmektedir. Ayırt edicilik indeksi ise 0.35 ile 0.71 arasında değişmektedir ve soruların çoğu iyi düzeyde ayırt ediciliğe sahiptir. Bu aralıklar, testin hem çeşitli zorluk seviyelerini içerdiğini hem de başarılı bir şekilde ayırt edici olduğunu göstermektedir. Testin ortalama güçlüğü 0.64, ortalama ayırt ediciliği ise 0.51'dir. ÇST'de katılımcılar, her doğru yanıt

için 1 puan almakta olup yanlış yanıtlanan sorular için herhangi bir puan verilmemektedir. ÇST'den alınabilecek puanların aralığı 0 ile 19 arasında değişmektedir.

Geçerlilik Çalışması

Açımlayıcı Faktör Analizi (AFA), ÇST'nin tek boyutluluğunu değerlendirmek amacıyla tetrakorik korelasyon matrisi kullanılarak gerçekleştirilmiştir (Embretson ve Reise, 2000). Ayrıca Kaiser-Meyer-Olkin örneklem yeterlilik katsayısı .80'in üzerinde bulunmuş ve örneklemin AFA için yeterli olduğu anlaşılmıştır. Ayrıca Bartlett Küresellik Testi anlamlı çıkmış ($p < .001$) ve değişkenler arasında faktör analizi yapılmasına olanak sağlayacak düzeyde ilişki olduğu belirlenmiştir. Bu bulgular, veri setinin faktör analizine uygun olduğunu ve analizden elde edilen sonuçların geçerliliğini desteklemektedir. Analiz sonuçları, faktör öz değerleri, faktör yükleri ve paralel analiz kriterlerine göre incelenmiştir. İlk faktörün diğer faktörlere kıyasla belirgin şekilde büyük olması, tek boyutluluğun göstergesi olarak değerlendirilmiştir (Crocker & Algina, 1986). Maddelerin tek faktör altında yeterli düzeyde faktör yüküne (≥.30) sahip olup olmadığı kontrol edilmiştir (Çokluk ve diğerleri, 2012). Paralel analiz ile rastgele oluşturulan veri setlerinden elde edilen öz değerler, gerçek veriden elde edilen öz değerlerle karşılaştırılarak tek boyutluluğun sağlanıp sağlanmadığı yorumlanmıştır (Horn, 1965).

19 maddeden oluşan ÇST'ye ait faktör yükleri Tablo 2'de, 44 maddeden oluşan BDT'ye ait faktör yükleri ise Tablo 3'te sunulmuştur.

Tablo 2

ÇST Madde Havuzuna Ait Faktör Yükleri

Madde No	Faktör yükü	Madde No	Faktör yükü
M1	.75	M6	.75
M2	.54	M7	.63
M3	.64	M8	.66
M4	.44	M9	.31
M5	.75	M10	.48
M11	.79	M16	.72
M12	.69	M17	.65
M13	.39	M18	.41
M14	.34	M19	.65
M15	.67		

Tablo 3

BDT Madde Havuzuna Ait Faktör Yükleri

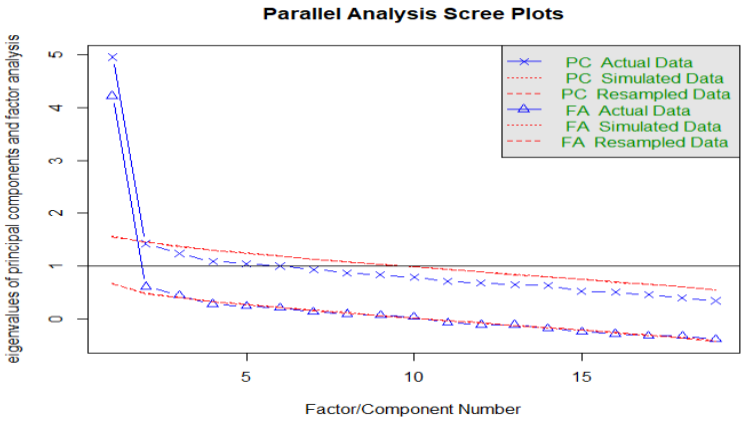
Madde No	Faktör yükü	Madde No	Faktör yükü
M1	.58	M12	.36
M2	.66	M13	.37
M3	.40	M14	.45
M4	.38	M15	.42
M5	.33	M16	.36
M6	.32	M17	.43
M7	.46	M18	.47
M8	.30	M19	.54
M9	.43	M20	.38
M10	.41	M21	.43
M11	.37	M22	.53
M23	.33	M34	.32
M24	.44	M35	.50
M25	.38	M36	.50
M26	.63	M37	.39
M27	.54	M38	.38
M28	.49	M39	.45
M29	.51	M40	.54
M30	.40	M41	.51
M31	.34	M42	.52
M32	.42	M43	.59
M33	.36	M44	.57

Her iki test için yapılan AFA sonuçları, testlerin büyük ölçüde tek faktörlü bir yapıya sahip olduğunu göstermektedir. ÇST'de ilk bileşen toplam varyansın %37.03'ünü açıklarken BDT'de %21.01'ini açıklamaktadır. Her iki testte de ilk faktörün öz değeri diğer faktörlerden belirgin şekilde yüksek olup test maddelerinin önemli bir kısmı aynı temel boyuta yüklenmektedir. Faktör yükleri açısından değerlendirildiğinde, ÇST'de faktör yükleri .31 ile .79 arasında değişirken BDT .30 ile .66 arasında değişmektedir. Bu durum, her iki testte de maddelerin büyük ölçüde ortak bir faktöre anlamlı şekilde yüklendiğini ve bu faktörün testlerin ana yapısını temsil ettiğini göstermektedir. Sonuç olarak her iki test de genel olarak tek faktörlü bir yapıya sahip olup okuduğunu anlama becerisini ölçme konusunda tutarlı bir yapı sunduğu söylenebilir.

AFA sonuçlarına ilişkin paralel analiz grafiklerine göre belirlenen faktör yapıları Şekil 1 ve Şekil 2'de sunulmuştur.

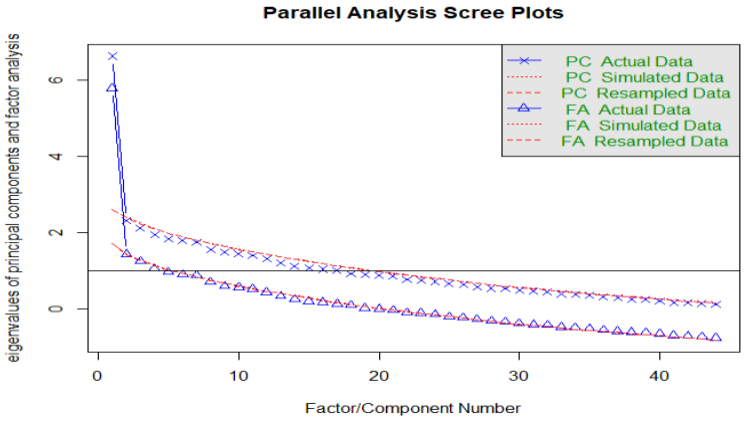
Şekil 1

Paralel Analiz Grafiği-ÇST



Şekil 2

Paralel Analiz Grafiği-BDT



İki testin faktör yapısına ilişkin yapılan analizler, her iki testin de tek boyutlu bir yapı gösterdiğini ortaya koymaktadır. Paralel analiz sonuçları, ÇST ve BDT'de ilk faktörün diğer faktörlerden belirgin şekilde ayrıldığını ve bu durumun tek boyutluluk varsayımını desteklediğini göstermektedir.

Model uyum indeksleri incelendiğinde, ÇST için $X^2(152) = 196.22$, p=.009, RMSEA = .037 (%90 GA [.019, .051]), SRMSR = .060, TLI = .972 ve CFI = .975 olarak hesaplanmıştır. BDT'de ise $X^2(902) = 920.14$, p = .329, RMSEA = .013 (%90 GA [.00, .031]), SRMSR = .088, TLI = .987 ve CFI = .976 değerleri elde edilmiştir.

Bu sonuçlar, her iki testin de iyi düzeyde model uyumu sağladığını ve tek faktörlü bir yapıya sahip olduğunu göstermektedir (Hu ve Bentler, 1999). Ancak BDT'de X^2 değerinin anlamlı olmaması ($p = .329$), bu testin veriyle daha iyi örtüştüğünü göstermektedir.

Boşluk Doldurma Testi

BDT, "İyi Uykular, Tatlı Rüyalar" adlı metne dayalı olarak geliştirilmiştir. Metin, başlık dâhil olmak üzere toplam 446 sözcükten oluşmaktadır. Metin öyküleyici türdedir ve ortaokul düzeyindeki öğrencilerin dilsel ve bilişsel özellikleri dikkate alınarak seçilmiştir. Günlük yaşam bağlamına dayalı olay örgüsü, açık nedensel ilişkiler ve anlamsal bütünlük içeren yapısı, metnin genel okuduğunu anlama becerisini ölçmeye elverişli bir içerik sunduğunu göstermektedir. Metinde yer alan sözcüklerin büyük bölümü, Millî Eğitim Bakanlığı Türkçe ders kitaplarında sıklıkla karşılaşılan ve öğrencilerin alıcı sözcük dağarcığında yer alması beklenen yüksek frekanslı sözcüklerden oluşmaktadır. Bu durum, testin ağırlıklı olarak sözcük bilgisini değil, bağlamdan anlam çıkarma ve metinsel bütünlüğü kullanma süreçlerini yansıtmasını amaçlamaktadır.

BDT, boşluk doldurma tekniğinin standart uygulamalarına uygun biçimde yapılandırılmıştır. Metnin ilk cümlesinde herhangi bir sözcük silinmemiş, ikinci cümleden başlanarak her altıncı sözcük sistematik olarak çıkarılmıştır. Sözcük silme işleminde noktalama işaretleri, özel adlar ve sayılar kapsam dışında bırakılmıştır. Bu işlem sonucunda testte toplam 69 boşluk oluşturulmuştur. Her bir boşluk için katılımcılardan metnin bağlamına uygun sözcüğü üretmeleri beklenmiştir. Yanıtların değerlendirilmesinde yalnızca hedef sözcükle birebir örtüşen yanıtlar doğru kabul edilmiş; yazım hataları, yakın anlamlı sözcükler ya da anlamsal olarak yakın ancak hedef sözcükten farklı yanıtlar yanlış olarak değerlendirilmiştir. Her doğru yanıt için 1 puan verilmiş, yanlış veya boş bırakılan yanıtlar için puan verilmemiştir. Bu doğrultuda BDT'den alınabilecek toplam puanlar 0 ile 69 arasında değişmektedir.

Metnin boşluk doldurma testine uygunluğu konusunda Türkçe eğitimi alanında uzman üç öğretim üyesinin görüşüne başvurulmuş, uzmanlar, metnin sözcük düzeyi, anlamsal tutarlılığı ve bağlamsal ipuçları bakımından boşluk doldurma tekniğiyle ölçmeye uygun olduğu yönünde görüş bildirmiştir. Bununla birlikte sözcük silme yoluyla yapılandırılan boşluk doldurma testlerinin okuduğunu anlama becerisinin tüm bilişsel boyutlarını doğrudan ve eşit düzeyde ölçtüğü iddia edilemez. Bu tür testler, özellikle sözdizimsel yapı, sözcük bilgisi ve bağlamsal ipuçlarından yararlanma gibi süreçlerle ilişkili performansı yansıtmaktadır. Bu nedenle BDT, bu çalışmada okuduğunu anlama becerisinin çok boyutlu doğasını bütünüyle temsil eden bir ölçme aracı olarak değil, baskın bir genel okuduğunu anlama faktörünü yansıtan bir gösterge olarak ele alınmıştır.

Güvenirlilik Değerleri

Her iki testin güvenilirlik analizi sonuçları, testlerin yüksek düzeyde iç tutarlılığa sahip olduğunu göstermektedir. ÇST için Kuder-Richardson 20 (KR-20) katsayısı .83,

McDonald's Omega (ω) katsayısı ise .85 olarak hesaplanmıştır. BDT'de ise KR-20 .86, McDonald's Omega (ω) ise .88 olarak bulunmuştur. Bu değerler, her iki testin de güçlü iç tutarlılığa sahip olduğunu ve ölçtükleri yapıyı güvenilir bir şekilde değerlendirdiğini göstermektedir (Cronbach, 1951; McDonald, 1999).

Etik Kurul Kararı

Bu araştırma, Akdeniz Üniversitesi Sosyal ve Beşeri Bilimler Bilimsel Araştırma ve Yayın Etiği kurulunun 10/06/2022 tarihli 10.06.2022-379259 sayılı kararı ile alınan izinle yürütülmüştür.

Verilerin Toplanması ve Analizi

Veri toplama sürecinde katılımcılara öncelikle BDT uygulanmış, bir hafta sonra aynı katılımcı grubuna ÇST verilmiştir. Her iki uygulama da tek oturumda ve bir ders saatinde gerçekleştirilmiştir. Katılımcıların BDT'ye verdikleri yanıtlar iki bağımsız değerlendirici tarafından önceden hazırlanan yanıt anahtarı doğrultusunda puanlanmış, değerlendiriciler arası uyum %100 olarak hesaplanmıştır. ÇST puanları ise testin otomatik puanlama sistemi kullanılarak elde edilmiştir.

Analizlere geçilmeden önce veri seti istatistiksel çözümlemeye uygunluk açısından incelenmiştir. Veri setinde kayıp değer bulunmadığı belirlenmiştir. Katılımcıların BDT ve ÇST'den elde ettikleri toplam puanlara ilişkin uç değerler kutu grafikleri aracılığıyla incelenmiştir. BDT toplam puanına ilişkin kutu grafiğinde 78 numaralı katılımcının uç bir değer aldığı görülmüş, ancak ilgili katılımcının yanıtları incelendiğinde elde edilen yüksek puanın rastgele yanıtlamayla açıklanamayacağı değerlendirilmiş ve bu değer veri setinden çıkarılmamıştır.

Birinci araştırma sorusunu (Boşluk doldurma testi puanları, çoktan seçmeli test puanlarını anlamlı şekilde yordamakta mıdır?) yanıtlamak amacıyla bağımlı ve bağımsız değişken olarak BDT toplam puanı ve ÇST toplam puanı ele alınmıştır. Analiz öncesinde her iki değişkenin dağılım özellikleri incelenmiştir. BDT ve ÇST toplam puanlarının normalliği Shapiro–Wilk testi ve grafiksel incelemelerle değerlendirilmiş, her iki değişkenin de normal dağılım göstermediği belirlenmiştir. Bu nedenle testler arası ilişkiyi incelemek için parametrik olmayan korelasyon analizi tercih edilmiştir.

İkinci araştırma sorusu (Öğrencilerin Bloom Taksonomisi'nin farklı bilişsel düzeylerinde ölçme yapmayı hedefleyen çoktan seçmeli maddelerdeki performansları ile boşluk doldurma testinden elde ettikleri puanlar arasındaki ilişki nasıldır?) kapsamında bağımlı değişken olarak BDT toplam puanı, bağımsız değişkenler olarak ise ÇST'de yer alan farklı bilişsel düzeylere (örneğin bilgi, kavrama, uygulama vb.) ait madde gruplarından elde edilen toplam puanlar ele alınmıştır. Bu doğrultuda ÇST, bilişsel düzeylerine göre alt puanlara ayrılmış ve her bir alt puan için dağılım özellikleri ayrı ayrı incelenmiştir. Normallik varsayımının sağlanmaması nedeniyle bu aşamada da parametrik olmayan istatistiksel teknikler kullanılmıştır.

İlişkisel analizlerin ardından, BDT puanlarının ÇST bilişsel düzey puanları tarafından yordanıp yordanmadığını incelemek amacıyla regresyon analizleri gerçekleştirilmiştir. Bu analizlerde bağımlı değişken BDT toplam puanı, bağımsız değişkenler ise ÇST'nin farklı bilişsel düzeylerine ait toplam puanlar olarak tanımlanmıştır. Regresyon analizlerinde tüm bağımsız değişkenlerin modele aynı anda dâhil edilmesi araştırmanın amacına uygun görüldüğünden "enter" yöntemi kullanılmıştır.

Regresyon analizine geçilmeden önce temel varsayımlar kontrol edilmiştir. Bağımlı ve bağımsız değişkenler arasındaki doğrusal ilişki scatter plot'lar aracılığıyla incelenmiştir. Artıkların normal dağılım gösterip göstermediği normal Q-Q grafiği ve Shapiro–Wilk testi ile değerlendirilmiş, varyansların homojenliği artıklar ile yordanan değerlerin dağılımı incelenerek kontrol edilmiştir. Ayrıca bağımsız değişkenler arasında çoklu doğrusal bağlantı olup olmadığını belirlemek amacıyla VIF (Variance Inflation Factor) ve tolerans değerleri hesaplanmış; VIF değerlerinin 10'un altında, tolerans değerlerinin ise.10'un üzerinde olduğu görülmüştür. Bu bulgular, regresyon analizleri için gerekli varsayımların karşılandığını göstermektedir.

Üçüncü araştırma sorusunu (Çoktan seçmeli test ile boşluk doldurma testi, katılımcıların okur düzeyi sınıflandırmaları açısından nasıl bir uyum göstermektedir?) yanıtlamak amacıyla her bir katılımcının ÇST ve BDT'den elde ettiği toplam puanlar, Çetinkaya (2010) tarafından önerilen eşik değerler doğrultusunda üç okur düzeyine (zorlanan, eğitsel, bağımsız) dönüştürülmüştür. Her iki test için yapılan bu sınıflama sonucunda, aynı katılımcının ÇST'ye ve BDT'ye göre hangi okur düzeyinde yer aldığı belirlenmiştir. Bu çalışmada "okur düzeyi", katılımcıların okuduğunu anlama performanslarına dayalı olarak sınıflandırıldıkları yeterlik düzeyi olarak ele alınmıştır. Üçüncü araştırma sorusu kapsamında, katılımcıların ÇST ve BDT'den elde ettikleri toplam puanlar, Çetinkaya (2010) tarafından önerilen eşik değerler doğrultusunda zorlanan, eğitsel ve bağımsız olmak üzere üç okur düzeyine dönüştürülmüştür. Bu sınıflandırma, test türlerine göre okur düzeyi belirlemenin tutarlılığını incelemek amacıyla gerçekleştirilmiştir.

ÇST ile BDT'nin katılımcıların okur düzeylerini belirleme konusundaki tutarlılığını incelemek amacıyla her bir katılımcının bu iki testten aldığı puanlara göre okur düzeyi sınıflandırması yapılmıştır. Bu sınıflandırmada, Çetinkaya'nın (2010) çalışmasına dayalı olarak "engelli", "eğitsel " ve "bağımsız" olmak üzere üç düzey belirlenmiştir: Katılımcılar, ÇST puanları %50'nin altında ise "engelli", %50–70 arasında ise "eğitsel", %70'in üzerindeyse "bağımsız" düzeyde sınıflandırılmıştır. Aynı şekilde, BDT puanı %35'in altında olanlar "engelli", %35–50 arasında olanlar "eğitsel", %50'nin üzerindekiler ise "bağımsız" düzeyde değerlendirilmiştir. Her bir öğrenci, her bir test türü için ayrı ayrı bu puan aralıklarına göre bir düzeye atanmış ve bu doğrultuda testlerin okur düzeyini belirlemede ne ölçüde tutarlı oldukları karşılaştırılmıştır.

Elde edilen sınıflandırmalar temel alınarak 3×3 düzey matrisi oluşturulmuş, her bir düzey kombinasyonu için frekans ve yüzde değerleri hesaplanmıştır. Son aşamada,

ÇST ve BDT'ye dayalı okur düzeyi sınıflandırmalarının birbiriyle ne derece uyumlu olduğunu belirlemek amacıyla Cohen's Kappa katsayısı hesaplanmıştır. Bu analizde bağımsız değişken test türü (ÇST–BDT), bağımlı değişken ise okur düzeyi sınıflandırması olarak ele alınmıştır.

Bulgular

Araştırma sorularına yanıt bulmak amacıyla yapılan çözümlemelere geçilmeden önce katılımcıların ÇST ve BDT'den aldığı puanlar kullanılarak betimsel istatistikler yapılmış ve sonuçlar Tablo 4'te sunulmuştur.

Tablo 4

Testlerden Alınan Puanlara İlişkin Betimsel İstatistikler

	ÇST	**BDT**
N	102	102
Ortalama	10,83	19,79
Sdt. Sapma	3,874	9,81
Minimum	1	2
Maksimum	18	46

Tablo 4'e göre her iki testi toplam 102 öğrenci yanıtlamıştır. Katılımcıların ÇST'den aldığı ortalama puan 10,83, BDT'den aldığı puan 19,79'dur. ÇST'ye ait standart sapma değeri 3,874, BDT'ye ait standart sapma değeri ise 9,81'dir. ÇST'den alınan en düşük puan 1, en yüksek puan 18, BDT'den alınan en düşük puan 2, en yüksek puan ise 46'dır.

Birinci ve ikinci araştırma sorularını birlikte ele almak amacıyla, bu bölümde BDT ile ÇST puanları arasındaki ilişki ile bu testlerin Bloom Taksonomisi'nin farklı bilişsel düzeyleriyle olan ilişkileri incelenmiştir. Bu doğrultuda, öncelikle ÇST ve BDT toplam puanlarına ilişkin betimsel istatistikler ile testler ve Bloom'un bilişsel süreç basamaklarına göre oluşturulan alt puanlar arasındaki ilişkiler korelasyon analizi ile değerlendirilmiştir. Ardından BDT başarısının ÇST başarısını ne ölçüde yordadığını ortaya koymak amacıyla regresyon analizleri gerçekleştirilmiştir. Analiz öncesinde değişkenlerin dağılım özellikleri Shapiro–Wilk testi ve grafikler aracılığı ile incelenmiş; normallik varsayımından sapmalar görülmekle birlikte, yordama ilişkisini ortaya koymak üzere regresyon analizleri uygulanmıştır. Elde edilen bulgular Tablo 5 ve Tablo 6'da sunulmuştur.

Tablo 5

Okuduğunu Anlama Test Puanlarına Ait Betimsel Değer ve İlişkiler

Test	X̄±S.S	1	2	3	4	5
(1) ÇST	10.83±3.87	1.00				
(2) BDT	19.79±9.81	.69*	1.00			
(3) Hatırlama	3.54±1.31	.77*	.55*	1.00		
(4) Anlama	4.00±1.67	.85*	.56*	.47*	1.00	
(5) Uygulama	.71±.69	.59*	.35*	.39*	.39*	1.00
(6) Çözümleme	2.58±1.36	.80*	.58*	.48*	.57*	.26*

*p<.01

Okuduğunu anlama testlerine ilişkin betimsel istatistikler ve testler arasındaki korelasyonlar Tablo 5'te sunulmuştur. ÇST ortalama puanı 10.83 (SS = 3.87), BDT ortalama puanı ise 19.79 (SS = 9.81) olarak hesaplanmıştır. Korelasyon analizine göre, ÇST ile BDT arasında pozitif ve güçlü bir ilişki bulunmaktadır, $r_{ÇST,\ BDT}$= .69, p < .01.

Bloom'un bilişsel süreç basamaklarına göre ÇST'de yer alan hatırlama, anlama, uygulama ve çözümleme düzeyindeki soruların test puanlarıyla olan ilişkileri incelendiğinde, hatırlama puanları hem ÇST (r = .77, p < .01) hem de BDT (r = .55, p < .01) ile anlamlı ve pozitif bir ilişki göstermektedir. Anlama düzeyindeki soruların puanları da ÇST ile yüksek bir korelasyon (r = .85, p < .01) gösterirken, BDT ile olan ilişkisi daha düşük düzeyde kalmıştır (r = .56, p < .01).

Uygulama düzeyindeki soruların toplam test puanları ile ilişkisi incelendiğinde, ÇST ile orta düzeyde (r = .59, p < .01), BDT ile düşük düzeyde (r = .35, p < .01) bir ilişki tespit edilmiştir. Çözümleme düzeyindeki soruların test puanları ile olan ilişkisi ise ÇST ile yüksek (r = .80, p < .01), BDT ile orta düzeyde (r = .58, p < .01) anlamlı bir korelasyon göstermektedir.

Tablo 6

Okuduğunu Anlama Test Puanlarını Yordamaya İlişkin Regresyon Analizi Sonuçları

Yordanan	Yordayıcı	B (e)	β	t	p	sd	F	p	R	R2
ÇST	Sabit	6.34 (.55)		11.52	<.001	1 100	90,02	<.001	.69	.47
	BD	.35 (.04)	.69	9.49	<.001					
BDT	Sabit	-2.11 (1.76)		1.19	.236	3 98	29.55	<.001	.69	.48
	Çözümleme	1.68 (.52)	.30	3.23	.002					
	Hatırlama	1.63 (.50)	.28	3.23	.002					
	Anlama	1.19 (.42)	.26	2.81	.006					

Regresyon analizi iki aşamada gerçekleştirilmiştir. İlk aşamada, ÇST toplam puanlarının BDT toplam puanlarını yordama gücü basit doğrusal regresyon modeli ile; ikinci aşamada ise BDT puanlarının Bloom Taksonomisi'nin farklı bilişsel düzeylerine ait ÇST puanları tarafından yordanması, tüm bağımsız değişkenlerin modele aynı anda dâhil edildiği enter yöntemi kullanılarak incelenmiştir.

İlk modelde BDT puanları, ÇST puanlarının anlamlı bir yordayıcısı olarak bulunmuştur ($B = .35(.04)$, $\beta = .69$, $p < .001$). Regresyon modeli anlamlıdır, $F(1,100) = 90.02$, $p < .001$ ve toplam varyansın %47'si ($R^2 = .47$) BDT puanları ile açıklanmaktadır. Bu sonuç, ÇST puanlarının büyük ölçüde BDT puanlarıyla ilişkili olduğunu ve BDT'nin okuduğunu anlama becerisini ölçmede güçlü bir gösterge sunduğunu göstermektedir.

İkinci modelde, BDT puanlarının çözümleme, hatırlama ve anlama basamakları ile anlamlı bir şekilde yordanabildiği görülmüştür. Çözümleme düzeyi, ($B = 1.68(.52)$, $\beta = 0.30$, $p = .002$), hatırlama düzeyi, ($B = 1.63(.50)$, $\beta = 0.28$, $p = .002$) ve anlama düzeyi ($B = 1.19(.42)$, $\beta = 0.26$, $p = .006$). BDT puanlarını anlamlı şekilde yordamaktadır. Regresyon modeli genel olarak anlamlıdır, $F(3, 98)= 29.55$, $p < .001$ ve BDT puanlarındaki varyansın %48'ini açıklamaktadır ($R^2 = .48$). Bu sonuçlar, BDT puanlarının Bloom Taksonomisi'nin bilişsel düzeyleri ile anlamlı ilişkiler gösterdiğini ve özellikle çözümleme, hatırlama ve anlama basamaklarının test performansını anlamlı şekilde yordadığını ortaya koymaktadır.

ÇST ile BDT'nin katılımcıların okur yetkinliği ile ilgili nasıl bir uyum gösterdiğini ortaya koymak için ÇST puanları ve BDT puanları dikkate alınarak her bir katılımcının okur düzeyi belirlenmiş ve testlerin türüne göre karşılaştırılmıştır. Tablo 7'de katılımcıların okur düzeylerine yönelik yapılan karşılaştırma sunulmaktadır.

Tablo 7

ÇST ve BDT'nin Okur Düzeyleri Açısından Karşılaştırılması

ÇST	BDT	f	%
Bağımsız	Bağımsız	14	13.7
Eğitsel	Eğitsel	11	10.8
Engelli	Engelli	27	26.5
Bağımsız	Eğitsel	19	18.6
Bağımsız	Engelli	17	16.7
Eğitsel	Bağımsız	4	3.9
Eğitsel	Engelli	7	6.9
Engelli	Eğitsel	3	2.9
Toplam		102	100

Tablo 7 incelendiğinde, ÇST ve BDT testlerinden elde edilen puanlara göre aynı okur düzeyinde sınıflandırılan 52 katılımcı bulunduğu görülmektedir. Bu katılımcılar

her iki testte de engelli, eğitsel ya da bağımsız okur düzeyinde yer almaktadır. Buna karşılık, 50 katılımcı iki testte farklı okur düzeylerinde sınıflandırılmıştır. 19 katılımcının ÇST puanlarına göre bağımsız, BDT puanlarına göre ise eğitsel düzeyde yer aldığı, 17 katılımcının ÇST'ye göre bağımsız, BDT'ye göre engelli düzeyde sınıflandırıldığı görülmektedir. Ayrıca 7 katılımcının ÇST puanına göre eğitsel, BDT puanına göre engelli, 3 katılımcının ise ÇST'ye göre engelli, BDT'ye göre eğitsel düzeyde yer aldığı belirlenmiştir.

ÇST ve BDT öğrencilerin okur düzeylerini belirleme konusundaki uyumunu değerlendirmek için Kappa analizi yapılmıştır. Kappa analizi sonuçları istatistiksel olarak anlamlı olsa da her iki test türünün öğrencileri aynı okur düzeylerine yerleştirme konusunda yeterince tutarlı olmadığını söylenebilir (κ=0.231, p=0.000).

Tartışma, Sonuç ve Öneriler

Bu çalışmada, okuduğunu anlama becerisini ölçmede yaygın olarak kullanılan çoktan seçmeli testler ve boşluk doldurma testlerinin puanları arasındaki ilişki incelenmiş, her iki test türünün okur düzeyleriyle uyumu değerlendirilmiştir. Araştırma bulguları, aşağıdaki sonuçları ortaya koymuştur.

ÇST ve BDT puanları arasında pozitif yönde yüksek bir ilişki tespit edilmiştir (r=0,783). Bu, her iki test türünün okuduğunu anlama becerisini ölçme açısından uyumlu olduğunu göstermektedir.

Regresyon analizi sonuçlarına göre BDT başarısı ÇST başarısını anlamlı bir şekilde yordamaktadır. ÇST başarısındaki toplam varyansın %47'si BDT başarısı tarafından açıklanmaktadır. Regresyon modeli, BDT puanındaki her bir birimlik artışın, ÇST puanında 0.35 birimlik bir artışa yol açtığını göstermiştir.

ÇST'deki soru türlerinden hatırlama, anlama ve çözümleme düzeyindeki soruların BDT başarısını anlamlı bir şekilde yordadığı belirlenmiştir (R^2 = 0.48). En güçlü yordayıcı değişken, çözümleme düzeyindeki sorular olmuştur (β = 0.30). Bu sonuç, çözümleme becerisinin BDT'deki başarı için kritik bir rol oynadığını göstermektedir. Hatırlama (β = 0.28) ve anlama (β = 0.26) düzeyindeki sorular da BDT başarısını anlamlı bir şekilde yordarken uygulama düzeyi anlamlı bir yordayıcı olarak bulunmamıştır. Bu bulgular, BDT'nin özellikle orta düzey bilişsel becerilerle ilişkili olduğunu ortaya koymaktadır.

ÇST ve BDT puanlarına göre katılımcıların okur düzeyleri arasında uyumsuzluklar olduğu tespit edilmiştir. Kappa analizi sonuçları (κ=0.189), iki test türünün okur düzeylerini belirlemede düşük bir tutarlılık gösterdiğini ortaya koymuştur. Katılımcıların %55'i iki test türlerine göre farklı okur düzeylerinde yer almıştır.

Çalışma, ÇST ve BDT puanları arasında pozitif yönde yüksek bir ilişki olduğunu göstermiştir. Ayrıca BDT başarısı ÇST başarısını anlamlı bir şekilde yordamaktadır. Bu sonuç, önceki araştırmaların (Baldauf Jr. ve Propst Jr., 1979; Bensoussan, 1984; Bormuth, 1967; Rankin ve Culhane, 1969; Sattarpour ve Ajideh, 2014) bulgularını

desteklemektedir. ÇST ve BDT arasındaki yüksek ilişkinin temel nedeni, her iki testin de okuduğunu anlama sürecindeki ortak bilişsel becerileri ölçmesidir. Her iki test, metinden anlam çıkarma, bağlam ipuçlarını kullanma ve dilsel farkındalık gibi temel becerilere dayanır (Andreassen ve Bråten, 2010; Das ve diğerleri, 2019; Jonz, 1990). Bu nedenle her iki test türü okuma becerilerinin değerlendirilmesinde tutarlı ve uyumlu görünmektedir. Bununla birlikte her iki test türünü puanlarının mükemmel bir uyum göstermediği de ifade edilmelidir ve bu beklenen bir sonuçtur. ÇST ve BDT benzer bilişsel süreçleri hedeflese de test formatlarının gerektirdiği bilişsel işlemler farklılaşmaktadır. Çoktan seçmeli testlerde tanıma, seçenekler arasında ayırt etme ve test stratejileri ön plana çıkmaktadır. Boşluk doldurma testlerinde bağlam bütünleştirme, sözcüksel geri çağırım ve üretim temelli süreçler daha belirleyici olmaktadır. Ayrıca boşluk doldurma testlerinin sözcük bilgisi ve biçimbilimsel farkındalık gibi okuduğunu anlama dışı değişkenlere daha duyarlı olması, bireysel puanların testler arasında tam olarak örtüşmemesine yol açabilmektedir. Bu nedenle gözlenen puan farklılıkları, ölçülen yapının değil, ölçme yaklaşımına özgü bilişsel ve psikometrik özelliklerin bir yansıması olarak değerlendirilmelidir.

Bu sonuçlar, her iki test türünün okuduğunu anlama ile ilişkili bazı ortak bilişsel süreçlere duyarlı olduğunu göstermektedir. Ancak boşluk doldurma testinin sistematik sözcük silmeye dayalı yapısı dikkate alındığında, elde edilen bulgular bu testin okuduğunu anlama becerisinin tüm bilişsel boyutlarını tek başına kapsadığını göstermemektedir. Çözümleme düzeyindeki sorularla kurulan istatistiksel ilişki, boşluk doldurma testinin üst düzey bilişsel süreçleri doğrudan ölçtüğünü göstermekten çok bu süreçlerle dolaylı olarak ilişkili olduğunu düşündürmektedir. Bulgular, boşluk doldurma testlerinin okuduğunu anlama ile ilişkili bilişsel süreçlere dair açıklayıcı ipuçları sunmaktadır. Bununla birlikte bu testlerin çoktan seçmeli testlerin yerine geçebilecek ya da onları tamamlayıcı ölçme araçları olarak kullanılabileceğine ilişkin genellenebilir bir sonuca ulaşmak için daha kapsamlı araştırmaların yapılması gerekmektedir.

Anlama düzeyindeki soruların BDT başarısını yordaması, metni anlamaya dayalı bilişsel süreçlerin boşluk doldurma testlerinde belirleyici bir rol oynadığını göstermektedir. BDT, bağlam ipuçlarını kullanarak eksik kelimenin tamamlanmasını gerektirirken (Kleijn ve diğerleri, 2019), anlama soruları öğrencilerin metni bütüncül olarak kavrama ve parçalar arasındaki ilişkileri görme becerilerini ölçer (Anderson ve diğerleri, 2001, s. 31; Anikin ve Sychev, 2020; Verenna ve diğerleri, 2018). Bu bağlamda, anlama düzeyindeki soruların BDT başarısını yordaması beklenen bir durumdur. Çünkü anlama sorularını yanıtlamada kullanılan beceriler, BDT'de eksik bırakılan kelimenin doğru şekilde tamamlanması için de gereklidir.

Çözümleme ve hatırlama düzeyindeki soruların da BDT başarısını anlamlı şekilde yordaması, BDT'nin yalnızca bağlam kullanımına değil, aynı zamanda bilgi yapısının analizi ve hatırlanmasına dayalı süreçlerle de ilişkilendirilebileceğini göstermektedir. Çözümleme soruları, öğrencilerin bilgiyi parçalarına ayırarak incelemelerini ve parçalar arasındaki ilişkileri anlamalarını öğretmeyi amaçlar

(Anderson ve diğerleri, 2001, s. 31; Jensen ve diğerleri, 2014). Bu beceri, BDT'de eksik bırakılan kelimenin tahmini için de önemli bir rol oynayabilir. Hatırlama sorularının etkisi ise öğrencilerin metindeki belirli bilgileri akılda tutma becerilerinin BDT sonuçlarını etkilediğini düşündürmektedir. Bu durum, BDT'nin hem alt düzey dil becerileri hem de bilgi erişimiyle uyumlu olduğunu ortaya koymaktadır.

Uygulama düzeyindeki soruların BDT başarısındaki etkisinin daha sınırlı olması, bu soruların yapısıyla açıklanabilir. Uygulama soruları, öğrencilerin öğrendikleri bilgileri gerçek dünyada veya yeni durumlarda kullanma becerilerini ölçer (Anderson ve diğerleri, 2001, s. 31; Monrad ve diğerleri, 2021). Ancak BDT daha çok mevcut bağlamı anlamaya ve eksik kelimenin bu bağlama uygun şekilde tahmin edilmesine dayanır. Bu nedenle uygulama düzeyindeki soruların BDT ile daha az uyumlu olduğu söylenebilir.

Boşluk doldurma testlerinin anlama, hatırlama ve çözümleme gibi orta düzey bilişsel süreçleri ölçmede etkili olduğu söylenebilir. Ancak bu durum testin üst düzey bilişsel anlamayı değerlendirdiği anlamına gelmemektedir. Üst düzey bilişsel süreçleri ölçebilmek için test formatlarının yeniden tasarlanması ve daha karmaşık düşünme becerilerini hedefleyen bir yapı oluşturulması gerekmektedir. Bununla birlikte, boşluk doldurma testlerinin yalnızca alt düzey bilişsel becerileri ölçtüğünü öne süren araştırmaların bulgularından farklı sonuçlara ulaşıldığı söylenebilir. Bu araştırmanın sonuçları, boşluk doldurma testlerinin alt ve orta düzey bilişsel becerileri değerlendirmede etkili bir ölçme aracı olarak kullanılabileceğini göstermektedir.

Öğrencileri okuduğunu anlama başarılarına göre sınıflandırma konusunda ise ÇST ve BDT arasındaki uyumun oldukça düşük olduğu bulunmuştur. Bu sonuç her iki testin puanları arasındaki ilişkinin sonuçlarıyla çelişmektedir. Bunun en büyük nedeni öğrencileri okur türlerine göre sınıflandırmada kullanılan kesim noktalarıdır. Örneğin bu araştırmaya katılan bir öğrenci ÇST'den 8 puan almışsa engelli, 9 puan almışsa eğitsel düzeyde bir okur olmaktadır. Benzer biçimde BDT'den 23 puan alan bir okur engelli düzeyde yer alırken 24 puan alan bir okur eğitsel düzeyde yer almaktadır. Bu durumda küçük puan farklılıkları öğrencilerin düzeyinde belirleyici olmaktadır. Bu nedenle her iki test türü okur türünü belirlemede düşük uyum göstermiştir. Okurların çeşitli sınıflara ayrılmasında yalnızca test puanlarına güvenmenin doğru olmayacağı ifade edilebilir.

Bu araştırmada bir boşluk doldurma testi kullanılmıştır. Ancak literatürde, boşluk doldurma testlerinde silme sıklığının değiştirilmesinin hem testin bilişsel zorluk düzeyini hem de puanların metin özelliklerini yansıtma biçimini etkileyebileceği ifade edilmektedir (Alderson, 1979). Bu durum, test tasarımında silme stratejisinin dikkatle belirlenmesi gerektiğine işaret etmektedir. Boşluk doldurma testlerinin geçerliliği ve güvenilirliği açısından test tasarımında silme sıklığının dikkatle seçilmesi ve belirli bağlamlarda test edilmesi gerektiğini vurgulamaktadır. Ayrıca Henk (1981), silinen sözcüklerin farklılığının testin zorluk düzeyini önemli ölçüde etkileyebileceğini belirtmiştir. Bu bulgulara ek olarak Ulusoy (2009) tarafından yapılan bir çalışma, aynı metne dayalı olarak oluşturulan iki farklı

boşluk doldurma testinin sonuçlarının farklı olduğunu ortaya koymuştur. Bilki (2011) de belirli silme yöntemleri ve metin türleri bazı beceri değerlendirmeleriyle daha güçlü bir uyum sergilediğini bulgulamıştır. Bu bağlamda, farklı sözcük türlerinin (örneğin isim, fiil, bağlayıcı vb.) sistematik olarak silindiği boşluk doldurma testleriyle yeni araştırmalar yapılması, test tasarımı ve okuduğunu anlama düzeylerinin daha ayrıntılı biçimde anlaşılmasına katkı sağlayabilir.

Sonuç olarak bu çalışma, aynı metne dayalı olarak geliştirilen çoktan seçmeli test ve boşluk doldurma testinin, okuduğunu anlama becerisini ölçme konusunda büyük ölçüde benzer örüntüler sunduğunu göstermektedir. Elde edilen bulgular, iki testten elde edilen toplam puanlar arasında yüksek ve anlamlı bir ilişki bulunduğunu, buna karşılık test puanlarına dayalı okur düzeyi sınıflandırmalarında sınırlı düzeyde uyum olduğunu ortaya koymaktadır. Bu durum, boşluk doldurma testinin çoktan seçmeli testle birlikte ele alındığında okuduğunu anlama performansına ilişkin tamamlayıcı bilgiler sunabileceğini düşündürmekle birlikte, puanlardaki ve sınıflandırmalardaki farklılıkların yalnızca madde türüne dayalı olarak açıklanamayacağını göstermektedir.

Çalışmada boşluk doldurma testi, sözcük silme tekniğine dayalı tek bir uygulama üzerinden yapılandırılmıştır. Bu nedenle elde edilen sonuçlar, boşluk doldurma testlerinin farklı silme stratejileri ya da farklı madde yapılarına dayalı uygulamalarına genellenemez. Gelecek araştırmalarda, silme sıklığı, silinen sözcük türleri ve maddelerin bilişsel düzeyleri dikkate alınarak geliştirilecek farklı boşluk doldurma uygulamalarının, okuduğunu anlama becerisini ölçme açısından sunduğu olanakların daha kapsamlı biçimde incelenmesi önerilmektedir.

References

Aitken, K. G. (1977). Using cloze procedure as an overall language proficiency test. *TESOL Quarterly*, 11(1), 59–67. https://doi.org/10.2307/3585592

Alderson, J. C. (1979). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading*, 2(2), 108–119.

Alderson, J. C. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning,* 30(1), 59–76. https://doi.org/10.1111/j.1467-1770.1980.tb00151.x

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Pearson Education Limited.

Anikin, A., & Sychev, O. (2020). *Ontology-based modelling for learning on Bloom's taxonomy comprehension level*. In A. V. Samsonovich (Ed.), *Biologically inspired cognitive architectures 2019* (pp. 22–27). Springer International Publishing.

Auphan, P., Ecalle, J., & Magnan, A. (2019). Computer-based assessment of reading ability and subtypes of readers with reading comprehension difficulties: A study in French children from G2 to G9. *European Journal of Psychology of Education*, 34(3), 641–663. https://doi.org/10.1007/s10212-018-0396-7

Baghaei, P., & Ravand, H. (2019). Method bias in cloze tests as reading comprehension measures. *SAGE Open*, 9(1), 2158244019832706. https://doi.org/10.1177/2158244019832706

Bilki, U. (2011). *The effectiveness of cloze tests in assessing the speaking/writing skills of university EFL learners* [Master's thesis, Bilkent University]. https://repository.bilkent.edu.tr

Bormuth, J. R. (1967). Comparable cloze and multiple choice comprehension test scores. *Journal of Reading,* 10(5), 291–299. http://www.jstor.org/stable/40009351

Bråten, I., Haverkamp, Y., & Anmarkrud, Ø. (2024). Gaining a deeper understanding of the deep cloze reading comprehension test: Examining potential contributors and consequences. *Reading and Writing,* 38, 425–446. https://doi.org/10.1007/s11145-024-10521-y

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2013). *Bilimsel araştırma yöntemleri [Scientific research methods].* Pegem Akademi.

Cooper, B., & Foy, J. M. (1967). Guessing in multiple choice tests. *Medical Education*, 1(3), 212–215. https://doi.org/10.1111/j.1365-2923.1967.tb01699.x

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Çetinkaya, G. (2010). *Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması [Defining and classifying the readability levels of Turkish texts]* (Thesis No. 265580) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]* (Vol. 2). Pegem Akademi.

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology*: General, 130(2), 208–223. https://doi.org/10.1037/0096-3445.130.2.208

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press.

Fotos, S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning,* 41(3), 313–336. https://doi.org/10.1111/j.1467-1770.1991.tb00609.x

Fuhrman, M. (1996). Developing good multiple choice tests and test questions. *Journal of Geoscience Education*, 44(4), 379–384. https://doi.org/10.5408/1089-9995-44.4.379

Furnham, A., & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and Individual Differences*, 37, 943–955. https://doi.org/10.1016/j.paid.2003.10.016

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning,* 66(2), 419–447. https://doi.org/10.1111/lang.12157

Gellert, A. S., & Elbro, C. (2012). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, 31(1), 16–28. https://doi.org/10.1177/0734282912451971

Gençer, S. L. (2014). İyi uykular tatlı rüyalar [Good night, sweet dreams]. *Bilim Çocuk E-Dergi*, 193, 26–29. https://bilimcocuk.tubitak.gov.tr/wp-content/uploads/sites/157/2025/09/28043779-967e-4985-aebd-0e3c2e6fb1be.pdf

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple choice tests in education: A comprehensive review.

Review of Educational Research, 87(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Gooskens, C., & van Heuven, V. J. (2017). Measuring cross-linguistic intelligibility in the Germanic, Romance and Slavic language groups. *Speech Communication,* 89, 25–36. https://doi.org/10.1016/j.specom.2017.02.008

Henk, W. A. (1981). Effects of modified deletion strategies and scoring procedures on cloze test performance. *Journal of Reading Behavior*, 13(4), 347–357. https://doi.org/10.1080/10862968109547423

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: *A Multidisciplinary Journal*, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test…or testing to teach: Exams requiring higher-order thinking skills encourage greater conceptual understanding. *Educational Psychology Review,* 26(2), 307–329. https://doi.org/10.1007/s10648-013-9248-9

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, 10(4), 363–380. https://doi.org/10.1207/s1532799xssr1004_2

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 62–69. https://doi.org/10.1177/2372732215624707

Kızılaslan Tunçer, B., & Erden, G. (2015). *Boşluk doldurma testlerinin ilkokul 4. sınıf öğrencilerinin okuduğunu anlama düzeylerini belirlemede kullanılabilirliği [Usability of cloze tests in determining 4th-grade primary students' reading comprehension levels, Special issue on XIV. International Participation Symposium of Primary School Teacher Education (21–23 May 2015)].* Bartın University Journal of Faculty of Education, 4(Special Issue), 318–324. https://doi.org/10.14686/BUEFAD.2015USOSOzelsayi13219

Kleijn, S., Pander Maat, H., & Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing*, 36(4), 553–572. https://doi.org/10.1177/0265532219840382

Kozak, S., & Recchia, H. (2019). Reading and the development of social understanding: Implications for the literacy classroom. *The Reading Teacher*, 72(5), 569–577. https://doi.org/10.1002/trtr.1760

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice*. John Wiley & Sons.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160.

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple choice tests as tools for learning. *Memory & Cognition*, 43(1), 14–26. https://doi.org/10.3758/s13421-014-0452-8

Liu, X., Zhang, L., Yu, S., Bai, Z., Qi, T., Mao, H., Zhen, Z., Dong, Q., & Liu, L. (2024). The effects of age and reading experience on the lifespan neurodevelopment for reading comprehension. *Journal of Cognitive Neuroscience*, 36(2), 239–260. https://doi.org/10.1162/jocn_a_02086

Lonigan, C. J., & Burgess, S. R. (2017). Dimensionality of reading skills with elementary-school-age children. *Scientific Studies of Reading*, 21(3), 239–253. https://doi.org/10.1080/10888438.2017.1285918

Luchkina, T., Ionin, T., Lysenko, N., Stoops, A., & Suvorkina, N. (2021). Evaluating the Russian language proficiency of bilingual and second language learners of Russian. *Languages*, 6(2), 83. https://doi.org/10.3390/languages6020083

McDonald, R. P. (2013). *Test theory: A unified treatment.* Psychology Press.

Monrad, S. U., Bibler Zaidi, N. L., Grob, K. L., Kurtz, J. B., Tai, A. W., Hortsch, M., Gruppen, L. D., & Santen, S. A. (2021). What faculty write versus what students see? Perspectives on multiple choice questions using Bloom's taxonomy. *Medical Teacher*, 43(5), 575–582. https://doi.org/10.1080/0142159X.2021.1879376

Muijselaar, M. M. L., Swart, N. M., Steenbeek-Planting, E. G., Droop, M., Verhoeven, L., & de Jong, P. F. (2017). The dimensions of reading comprehension in Dutch children: Is differentiation by text and question type necessary? *Journal of Educational Psychology*, 109(1), 70–83. https://doi.org/10.1037/edu0000120

Ono, K., Sumita, K., & Seijii, M. (1994). Abstract generation based on rhetorical structure extraction. In Proceedings of the 15th Conference on Computational Linguistics (Vol. 1, pp. 344–348). Association for Computational Linguistics.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, 25(4), 399–438. https://doi.org/10.1080/07370000701632371

Pino, M. C., & Mazza, M. (2016). The use of 'literary fiction' to promote mentalizing ability. *PLOS ONE,* 11(8), e0160254. https://doi.org/10.1371/journal.pone.0160254

Pretorius, E. J. (2002). Reading ability and academic performance in South Africa: Are we fiddling while Rome is burning? *Language Matters*, 33(1), 169–196. https://doi.org/10.1080/10228190208566183

Rankin, E. F., & Culhane, J. W. (1969). Comparable cloze and multiple choice comprehension test scores. *Journal of Reading*, 13(3), 193–198. http://www.jstor.org/stable/40017267

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–474. https://doi.org/10.1191/0265532206lt337oa

Tighe, E. L., & Schatschneider, C. (2016). Examining the relationships of component reading skills to reading comprehension in frustrated adult readers. *Journal of Learning Disabilities*, 49, 395–409. https://doi.org/10.1177/0022219414555415

Ülper, H. (2010). *Okuma ve anlamlandırma becerilerinin kazandırılması [Teaching reading and meaning-making skills]*. Nobel Yayın Dağıtım.

Ulusoy, M. (2009). Boşluk tamamlama testinin okuma düzeyini ve okunabilirliği ölçmede kullanılması [Using cloze test to measure students' readıng levels and readabılıty of texts]. *Türk Eğitim Bilimleri Dergisi*, 7(1), 105-126.

Vayre, E., & Vonthron, A. (2019). Relational and psychological factors affecting exam participation and student achievement in online college courses. *Internet and Higher Education*, 43, 100671. https://doi.org/10.1016/j.iheduc.2018.07.001

Verenna, A. M. A., Noble, K. A., Pearson, H. E., & Miller, S. M. (2018). Role of comprehension on performance at higher levels of Bloom's taxonomy: Findings from assessments of healthcare professional students. *Anatomical Sciences Education*, 11(5), 433–444. https://doi.org/10.1002/ase.1768

Vlachos, F., & Papadimitriou, A. (2015). Effect of age and gender on children's reading performance: The possible neural underpinnings. *Cogent Psychology,* 2(1), 1045224. https://doi.org/10.1080/23311908.2015.1045224

Wait, S. S. (1987). *Textbook readability and the predictive value of the Dale–Chall, comprehensive assessment program and cloze* [Unpublished doctoral dissertation]. The Florida State University, USA.

Wissman, K. T., Zamary, A., & Rawson, K. A. (2018). When does practice testing promote transfer on deductive reasoning tasks? *Journal of Applied Research in Memory and Cognition,* 7(3), 398–411. https://doi.org/10.1016/j.jarmac.2018.03.002

Yu, L., Yu, J. J., & Tong, X. (2023). Social–emotional skills correlate with reading ability among typically developing readers: *A meta-analysis. Education Sciences*, 13(2), 220. https://doi.org/10.3390/educsci13020220

Ethical and Author Declarations | Etik ve Yazar Beyanları

Authors' Contributions	Yazarların Katkı Düzeyleri
The first author contributed to the conceptualization of the study, conducted the data analysis, prepared the first draft of the manuscript, and managed the submission and publication processes. The second author contributed to the development of the theoretical framework, conducted the data collection process, and reviewed and finalized the manuscript.	Birinci yazar çalışmanın kavramsallaştırılmasına katkı sağlamış, veri analizini gerçekleştirmiş, makalenin ilk taslağını yazmış ve makalenin gönderim ile yayımlanma süreçlerini yürütmüştür. İkinci yazar çalışmanın kuramsal çerçevesinin geliştirilmesine katkı sağlamış, veri toplama sürecini yürütmüş ve makalenin gözden geçirilmesi ile son okumasını gerçekleştirmiştir.

Conflict of Interest	Çıkar Çatışması Beyanı
The authors declare that they have no competing interests relevant to the content of this article.	Yazarların, bu makalenin içeriğiyle ilgili beyan edilecek herhangi bir çıkar çatışması bulunmamaktadır.

Ethical Approve	Etik Onay
The planning and implementation of this research were approved by the Scientific Research and Publication Ethics Committee for Social and Human Sciences of Akdeniz University with the decision dated 10/06/2022 and numbered 379259.	Bu araştırmanın planlanması ve yürütülmesi, Akdeniz Üniversitesi Sosyal ve Beşeri Bilimler Bilimsel Araştırma ve Yayın Etiği Kurulu'nun 10/06/2022 tarihli ve 379259 sayılı kararıyla etik açıdan uygun bulunmuştur.

Use of Artificial Intelligence	Yapay Zeka Kullanımı
The authors used [ChatGPT 5.2] only for language editing, phrasing improvement, translation checking, and formatting assistance (including reference formatting according to the journal guidelines). No artificial intelligence tools were used in the study design, data collection, data analysis, or interpretation of the findings. The authors are fully responsible for the entire scientific content of the manuscript.	Bu çalışmada [ChatGPT 5.2] yalnızca dil düzenleme, ifade iyileştirme, çeviri kontrolü ve biçimsel düzenleme (kaynakçanın dergi yazım kurallarına uygun biçimde düzenlenmesi dâhil) amacıyla kullanılmıştır. Bunun dışında, araştırmanın tasarımı, veri toplama, veri analizi veya bulguların yorumlanması aşamalarında herhangi bir yapay zekâ aracı kullanılmamıştır. Çalışmanın tüm bilimsel içeriği ve sorumluluğu yazarlara aittir.

This study has been evaluated under double-blind peer review and verified to be free of plagiarism using iThenticate software.
Bu çalışma, çift taraflı kör hakemlik kapsamında değerlendirilmiş ve iThenticate yazılımı kullanılarak intihal içermediği teyit edilmiştir.

Ethical disclosure | Etik bildirim: ebfd@ankara.edu.tr