Received Date: 04 March 2025 I Accepted Date: 25 March 2025

RESEARCH ARTICLE

OPUS Journal of Society Research

Comparison of Psychometric Properties of Turkish as a Foreign Language Listening Tests Composed of Independent and Common-Stem Items*

Ceren Tunaboylu Demir[1], Havva Gökçe Çavdar Paksoy[2], Duygu Anıl[3]

[1] PhD. Candidate, Hacettepe University, Ankara/Türkiye
ORCID: 0000-0001-8090-8913
E-Mail: cerentunaboylu@gmail.com

[2] PhD. Candidate, Bolu Abant İzzet Baysal University, Bolu/Türkiye
ORCID: 0000-0003-1813-2725
E-Mail: gkccavdar@gmail.com

[3] Prof. Dr., Hacettepe University, Ankara/Türkiye
ORCID: 0000-0002-1745-4071
E-Mail: duygu.anil73@gmail.com

Corresponding Author:
Ceren Tunaboylu Demir

March 2025
Volume:22
Issue:2
DOI: 10.26466/opusjsr.1651342

Citation:
Tunboylu Demir, C., Çavdar Paksoy, H. G. & Anıl, D. (2025). Comparison of psychometric properties of Turkish as a foreign language listening tests composed of ındependent and common-stem ıtems. *OPUS–Journal of Society Research*, *22*(2), 265-279.

Abstract

This study comparatively examined the psychometric properties of Turkish listening tests composed of independent and common-stem multiple-choice items that measure the same construct in the context of teaching Turkish as a foreign language. Additionally, students' perceptions of these two different item formats were evaluated. The research was designed as a descriptive study using a relational survey model. The study group consisted of 201 international students enrolled at the B1 level in Turkish Language Teaching Centers during the 2024–2025 academic year. The assessment tools were developed on the Concerto platform, and the tests were administered online in a single session across three stages. According to the findings, comparisons of item difficulty indices revealed statistically significant differences and moderate to large effect sizes in some item pairs, indicating that item format may influence student performance. Similarly, item discrimination indices showed significant differences in several item pairs, although both tests overall consisted of highly discriminative items. Analysis of test mean scores showed that the independent item test yielded significantly higher performance at the 0.05 level. The reliability coefficients of both tests were high, with no statistically significant difference between them. Based on students' perceptions, the independent item test was found to be more advantageous in terms of time management and online usability. In contrast, the common-stem item format was perceived as more cognitively demanding. Both item types received similarly positive feedback regarding item clarity and ease of answering listening questions. These results suggest that in teaching Turkish as a foreign language, independent items may contribute to higher student achievement and improved item discrimination in listening assessments.

Keywords: *Turkish as a foreign language, listening skills test, independent item, common-root item, psychometric properties*

Öz

Bu araştırmada, yabancı dil olarak Türkçe öğretiminde, aynı özelliği ölçen, çoktan seçmeli madde türünde bağımsız ve ortak köklü maddelerden oluşan Türkçe dinleme testlerinin psikometrik özellikleri karşılaştırmalı olarak incelenmiştir. Ayrıca, öğrencilerin bu iki farklı madde yapısına yönelik algıları da değerlendirilmiştir. Araştırma, betimsel araştırma türünde, ilişkisel tarama modelinde tasarlanmıştır. Araştırmanın çalışma grubunu, 2024-2025 eğitim-öğretim yılında Türkçe Öğretim Merkezlerinde B1 düzeyinde öğrenim gören 201 yabancı uyruklu öğrenci oluşturmaktadır. Ölçme araçları, Concerto platformu üzerine inşa edilmiş ve test uygulaması çevrimiçi ortamda, tek oturumda ve üç aşamalı olarak gerçekleştirilmiştir. Araştırma bulgularına göre, madde güçlük indeksleri karşılaştırıldığında, bazı madde çiftlerinde madde formatına bağlı anlamlı farklılıklar ve orta ile büyük düzeyde etki büyüklükleri saptanmış, bu durum madde formatının öğrenci başarısını etkileyebileceğini ortaya koymuştur. Madde ayırt edicilik indeksleri incelendiğinde de benzer şekilde bazı madde çiftlerinde anlamlı farklar görülmüş; her iki testin genel olarak yüksek ayırt ediciliğe sahip maddelerden oluştuğu belirlenmiştir. Test puan ortalamaları incelendiğinde, bağımsız madde testi, öğrenciler tarafından 0,05 hata düzeyinde anlamlı olarak daha yüksek başarı ile tamamlandığı görülmüştür. Her iki test puanlarının güvenirlik katsayıları yüksek bulunmuş ve aralarında istatistiksel olarak anlamlı bir fark tespit edilmemiştir. Madde yapısına yönelik öğrenci algılarına göre, bağımsız madde testi süre yönetimi ve çevrimiçi uygulama açısından daha avantajlı bulunmuştur. Buna karşılık, ortak köklü madde testi madde formatının zorlayıcılığı daha fazla algılanmıştır. Her iki madde formatı da soru anlaşılırlığı ve dinleme sorularını yanıtlama kolaylığı açısından benzer düzeyde olumlu geri bildirim almıştır. Bu sonuçlar, yabancı dil olarak Türkçe öğretiminde dinleme becerilerini ölçmeye yönelik testlerde bağımsız madde yapısının, öğrencilerin daha yüksek başarı elde etmesine ve madde ayırt ediciliğinin artmasına katkı sağlayabileceğini göstermektedir.

Anahtar Kelimeler: *Yabancı dil olarak Türkçe, dinleme becerisi testi, bağımsız madde, ortak köklü madde, psikometrik özellikler.*

* This study was presented as an oral paper at the International Symposium on Measurement, Selection, and Placement, held in Ankara, Turkey, on October 4–6, 2024.

Introduction

As a social being, humans learn languages by nature to interact with their environment. The realization of communication in language learning depends on the development of listening, speaking, reading, and writing skills. Listening and reading skills are classified as receptive language skills, while speaking and writing skills are categorized as productive language skills (MEB, 2013). Receptive skills enable internalization, whereas productive skills ensure effectiveness in communication. Among the four language skills in both native and foreign language acquisition, listening, which serves as the initial and fundamental step of communication, plays a critical role in the development of other skills in language learning. As one of the cornerstones of the language learning process, listening enables individuals to make sense of messages from their surroundings and to communicate effectively. In the Turkish Ministry of National Education (MoNE)'s Foreign Language Teaching Program, listening is defined as "one of the fundamental ways of communication and learning, involving the ability to accurately comprehend, interpret, and evaluate a given message." Listening is the ability to correctly understand the message that the speaker intends to convey and to respond accordingly (Demirel, 2021). The listener first perceives the spoken text. The perception stage refers to the mental processing of sounds and the act of hearing. The second stage involves the cognitive process of meaning-making. These perception and meaning-making stages do not occur sequentially but simultaneously. Field (2008) described the perception of spoken text as lower-level skills within the framework of listening skills, whereas the meaning-making process is considered an advanced skill. Listening is a complex process that goes beyond mere hearing, involving comprehension and interpretation. In this context, listening serves as a fundamental component in language learning, integrating the simultaneous processes of perception and meaning-making and supporting the development of other language skills.

In the language learning process, there is a need for the balanced development of all four language skills. Monitoring this development and the progress of individuals' skill levels in a reliable manner requires the use of effective and accurate assessment tools. The accurate evaluation of skills is directly related to the quality of the assessment instruments used. In measuring listening skills, it is essential for the stimulus to be conveyed correctly and for the individual to respond accurately to the stimulus. The proper development and use of listening materials are crucial in assessing listening skills. Every stage, from preparing listening texts in accordance with grammatical structures to ensuring high-quality audio recordings, must be carefully planned and presented.

The multiple-choice item format stands out as a widely used assessment method, particularly in foreign language teaching, due to its objectivity and its ability to comprehensively assess language skills. In this context, how multiple-choice items are constructed, especially in content-based areas such as listening, becomes a crucial factor that directly affects the quality of assessment. In the following section, after outlining the fundamental characteristics of the multiple-choice item format, explanations are provided regarding the independent and common-stem item types, which constitute the focus of this study.

Multiple-Choice Items

The multiple-choice item format is widely used to assess various levels of cognitive skills, such as recall, comprehension, and application (Haladyna & Downing, 1989). It is a frequently preferred assessment tool due to its practical applicability and the possibility of objective scoring (Rodriguez, 2005). However, to enhance its validity and reliability, careful item design is essential. This item format consists of four well-structured fundamental components: the stem, options, the correct answer, and distractors. The stem forms the main part of the question and provides information to the student. It should be written in clear and comprehensible language. The stem generally consists of two elements: the situation and the prompt. The situation provides the necessary background information for the skill being assessed and may include a scenario, graph, table, or visual aid. The prompt explicitly defines the expected response from the student.

The options include both the correct answer and the distractors. Each option should be clear and unambiguous. The correct answer represents the only accurate response and should be easily identifiable by high-performing students while minimizing the likelihood of random guessing by low-performing students. Distractors, which are incorrect answer choices, directly influence the validity and reliability of the test. High-quality distractors reduce the probability of guessing and mitigate the impact of random responses (Haladyna, 2004). Additionally, distractors should effectively distinguish between knowledgeable and unknowledgeable students (i.e., high- and low-performing groups) and attract lower-performing individuals in a balanced manner. Distractors that are easily eliminated can compromise item quality and should therefore be carefully constructed.

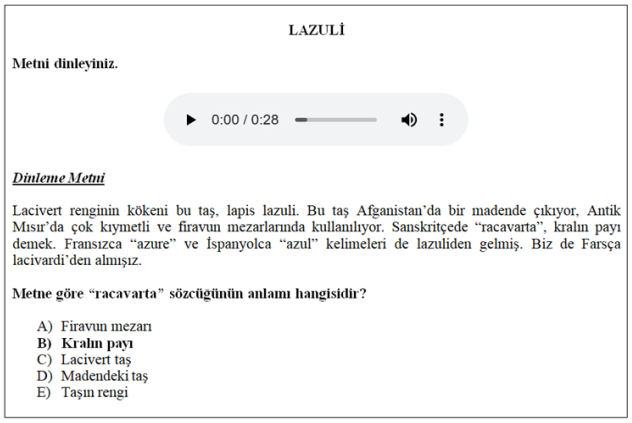
Multiple-choice items are a widely preferred tool in measurement and evaluation. Objective scoring systems simplify both the administration and scoring processes while specifically minimizing human-induced scoring errors. This item type can be applied at all educational levels and offers advantages such as the ability to assess large groups within a short period. Additionally, the high content validity of multiple-choice items allows for a broad range of knowledge to be assessed. However, there are also some limitations. Multiple-choice items may be insufficient in assessing higher-order cognitive skills such as analytical thinking. Individual differences in reading speed can negatively impact test validity. Moreover, the development of multiple-choice items is a complex process that requires expertise. Another limitation is the influence of guessing, as responses are selected from given options. Considering all these factors, the effectiveness of multiple-choice items as an assessment tool depends on their proper construction and appropriate application.

Multiple-choice items can be classified based on their structural characteristics. In general, they are categorized according to the type of correct answer (items with an absolute correct answer, items requiring the most accurate response, items requiring composite answers, and items that conceal the correct answer), the format of the stem (question-based stems, incomplete sentence stems, and negatively worded stems), and the organization of the items (common-stem items and common-option items) (Haladyna, 2004; Turgut & Baykul, 2012). This study compares independent items and common-stem items.

Independent items

Independent items are multiple-choice questions that are completely separate from one another and must be evaluated within their own context. In the literature, this item format is also referred to as a traditional multiple-choice item without a common stem (Turgut & Baykul, 2012) and as a "stand-alone item" (Haladyna, Downing & Rodriguez, 2002). These items are typically based on a short paragraph, dialogue, or passage, requiring students to respond to a single question based on the given context. In other words, one text corresponds to one question, and one listening passage corresponds to one question. Independent items measure students' ability to comprehend and recall specific, isolated pieces of information. Since each question evaluates an independent unit of knowledge, independent items allow for item-by-item assessment of student performance (Haladyna et al., 2002). This item type is particularly preferred when assessing individual knowledge components and covering a broad range of topics within a test (Haladyna, 2004). Independent items require recalling specific information rather than interpreting a general context. They offer the advantage of focusing on different topics and knowledge domains within each section of a test while making it more difficult for test-takers to guess answers by ensuring that each question has only one correct response (Rodriguez, 2005). These items are commonly used in general knowledge tests, proficiency exams, and various knowledge assessments.

An example of an independent item is presented in Figure 1. Although the listening texts are provided in written form in the given example, in actual test applications, only the audio recordings of the listening texts are available.



LAZULİ

Metni dinleyiniz.

0:00 / 0:28

Dinleme Metni

Lacivert renginin kökeni bu taş, lapis lazuli. Bu taş Afganistan'da bir madende çıkıyor, Antik Mısır'da çok kıymetli ve firavun mezarlarında kullanılıyor. Sanskritçede "racavarta", kralın payı demek. Fransızca "azure" ve İspanyolca "azul" kelimeleri de lazuliden gelmiş. Biz de Farsça lacivardi'den almışız.

Metne göre "racavarta" sözcüğünün anlamı hangisidir?

A) Firavun mezarı
B) Kralın payı
C) Lacivert taş
D) Madendeki taş
E) Taşın rengi

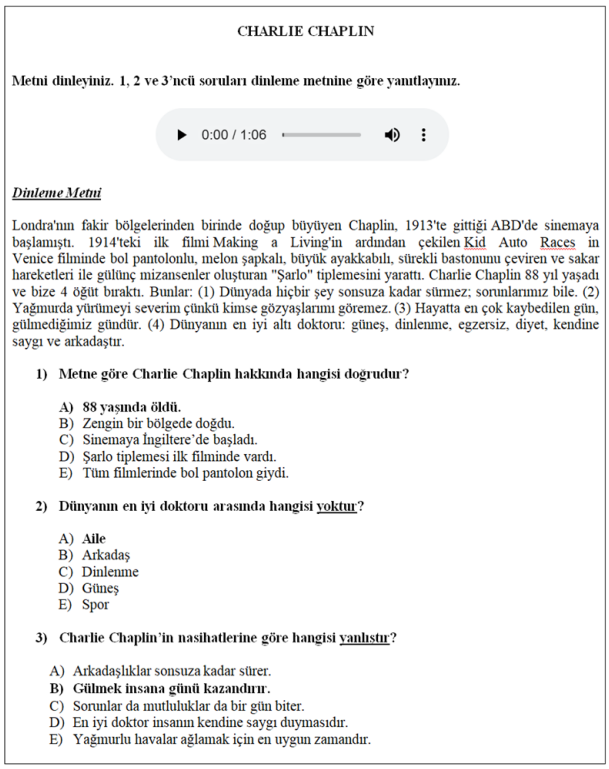
Figure 1. Example of an Independent Item

Common-stem items

Common-stem items are a type of multiple-choice question in which multiple items are based on a single context or passage (Turgut & Baykul, 2012). These items consist of a series of questions that follow the same context or passage. They are structured around a text, graph, table, or any piece of information and include multiple questions referencing this context. Common-stem items assess how students access and comprehend information within a given context. Students are expected to answer multiple questions using the same contextual information.

Common-stem items offer advantages such as efficient use of test time and the ability to assess reading comprehension skills in an integrated manner (Wainer & Thissen, 1993). They enable tests to contain more information, allowing for a more comprehensive evaluation (Bridgeman, 1992). This item type is particularly common in reading comprehension assessments, mathematical problem-solving, and situational evaluations. Additionally, common-stem items are effective in assessing students' ability to organize information within a text and derive meaning from the context.

An example of a common-stem item is presented in Figure 2. Although the listening texts are provided in written form in the given example, in actual test applications, only the audio recordings of the listening texts are available. Additionally, the common stem and the related questions for each item are displayed on the same page without requiring page transitions.



CHARLIE CHAPLIN

Metni dinleyiniz. 1, 2 ve 3'ncü soruları dinleme metnine göre yanıtlayınız.

0:00 / 1:06

Dinleme Metni

Londra'nın fakir bölgelerinden birinde doğup büyüyen Chaplin, 1913'te gittiği ABD'de sinemaya başlamıştı. 1914'teki ilk filmi Making a Living'in ardından çekilen Kid Auto Races in Venice filminde bol pantolonlu, melon şapkalı, büyük ayakkabılı, sürekli bastonunu çeviren ve sakar hareketleri ile gülünç mizansenler oluşturan "Şarlo" tiplemesini yarattı. Charlie Chaplin 88 yıl yaşadı ve bize 4 öğüt bıraktı. Bunlar: (1) Dünyada hiçbir şey sonsuza kadar sürmez; sorunlarımız bile. (2) Yağmurda yürümeyi severim çünkü kimse gözyaşlarımı göremez. (3) Hayatta en çok kaybedilen gün, gülmediğimiz gündür. (4) Dünyanın en iyi altı doktoru: güneş, dinlenme, egzersiz, diyet, kendine saygı ve arkadaştır.

1) Metne göre Charlie Chaplin hakkında hangisi doğrudur?

A) 88 yaşında öldü.
B) Zengin bir bölgede doğdu.
C) Sinemaya İngiltere'de başladı.
D) Şarlo tiplemesi ilk filminde vardı.
E) Tüm filmlerinde bol pantolon giydi.

2) Dünyanın en iyi doktoru arasında hangisi yoktur?

A) **Aile**
B) Arkadaş
C) Dinlenme
D) Güneş
E) Spor

3) Charlie Chaplin'in nasihatlerine göre hangisi yanlıştır?

A) Arkadaşlıklar sonsuza kadar sürer.
B) Gülmek insana günü kazandırır.
C) Sorunlar da mutluluklar da bir gün biter.
D) En iyi doktor insanın kendine saygı duymasıdır.
E) Yağmurlu havalar ağlamak için en uygun zamandır.

Figure 2. Example of a Common-Stem Item

Multiple-choice tests are widely preferred in foreign language education due to their practicality, objectivity, and adaptability to digital platforms. With the increasing prevalence of computer-based and individualized test applications, it has become essential to evaluate assessment tools not only in terms of validity and reliability but also with respect to variables such as time management, cognitive load, and learner perception. The item format used in these tests directly affects the measurement power, usability, and user experience of the assessment. In particular, independent and common-stem item structures represent two distinct approaches in the design of multiple-choice tests, each with its own advantages and limitations. Common-stem items generally involve longer texts, which can increase cognitive load in listening assessments by placing excessive demands on memory, thereby negatively affecting student performance (Tozlu, 2017).

A review of the literature reveals numerous studies comparing the psychometric properties of different item types, their impact on student performance, and their role in the assessment process. Many of these studies focus on how multiple-choice, open-ended, fill-in-the-blank, matching,

and mixed-format items differ in terms of item difficulty, discrimination, test mean scores, response time, and overall reliability (Kan & Kayapınar, 2006; Akyıldız & Karadağ, 2018; Öksüz & Demir, 2019; Öney, 2023; Sayın & Orbay, 2024; Koçdar, Karadağ & Şahin, 2017). Some studies indicate that mixed-format tests, which incorporate more than one item type, yield balanced results in terms of item difficulty and discrimination, thereby offering more comprehensive assessment opportunities for students with varying skill levels (Gültekin, 2011; Eren, 2015; Gürdil Ege & Demir, 2020). Others show that the way in which item formats are perceived by students and the cognitive levels they target can lead to performance differences, particularly in item types requiring higher-order thinking such as text-highlighting (Taşkıran, 2022; Demirkol & Karagöz, 2023). Similarly, international research has demonstrated that test performance can be significantly influenced by item format (Kobayashi, 2002; In'nami & Koizumi, 2009; Buck, 2001; Ghonsooly & Fatemi, 2013). However, most of these studies have been conducted in the context of English language education, and research focusing on learners of Turkish as a foreign language remains limited. Furthermore, other variables such as the use of visual elements (Özsu & Can, 2020) or scoring methods (Özdemir, 2004) have also been found to affect test psychometrics. In addition, recent literature emphasizes that item formats should be evaluated not only in terms of technical validity, but also in terms of user experience and practical applicability (Fulcher, 2010; Wilson, 2005; Boone, 2022). Nonetheless, there is a significant gap in the literature regarding empirical studies that systematically compare item formats from both technical and experiential perspectives, especially in the context of Turkish as a foreign language.

This study aims to fill this gap by comparing independent and common-stem item formats used in listening comprehension tests administered to learners of Turkish as a foreign language, focusing on both psychometric properties (item difficulty, item discrimination, test mean scores, reliability) and learner perceptions. In addition, evaluating students' experiences and perceptions of both item formats will provide a descriptive perspective on how item structure impacts test-takers. In this regard, the study aims to contribute to the development of more qualified assessments in the field of teaching Turkish as a foreign language, while also offering deeper insight into the strengths and limitations of different item formats. The findings are expected to provide practical recommendations for both test developers and practitioners working in Turkish language education.

Research Aim

This study aims to compare the psychometric properties of Turkish listening tests consisting of multiple-choice items in independent and common-stem formats, which are designed to measure the same construct in the context of teaching Turkish as a foreign language. Additionally, it seeks to examine student perceptions regarding these two test types. For this purpose, B1-level listening tests were administered simultaneously, and it was analyzed whether there were statistically significant differences between the tests in terms of item difficulty indices, item discrimination indices, test score means, and test reliability. Furthermore, student perceptions related to independent and common-stem item structures were descriptively examined.

Research Question

Is there a significant difference in test and item statistics between two multiple-choice tests—one composed of independent items and the other of common-stem items—designed to assess the same competencies? How do student perceptions differ in relation to these two types of test formats?

1. What are the descriptive statistics of the independent-item test and the common-stem-item test, both designed to assess the same competencies?
2. Do the independent-item test and the common-stem-item test show significant differences in:
   i. Item difficulty indices
   ii. Item discrimination indices
3. Do the test scores obtained from the independent-item test and the common-stem-

item test show significant differences in:
i. Total test score means
ii. Test reliability

4. How do student perceptions differ in terms of independent and common-stem item formats?

Method

Research Design

This study was designed within the framework of a descriptive research approach and employed a correlational survey model. In the correlational survey model, the presence and degree of co-variation between two or more variables are examined (Karasar, 2011). In this study, the psychometric properties of the independent item test and the common stem item test were compared, and the relationships between these tests were analyzed. In addition, a questionnaire was administered to identify students' perceptions regarding independent and common-stem item formats. In this respect, the study also carries a descriptive nature.

Study Group

The study group consists of 201 foreign students enrolled at Turkish Language Teaching Centers at the B1 level during the 2024-2025 academic year. The sample was selected using the convenience sampling method, one of the purposive sampling techniques. This method allows researchers to collect data from individuals or groups that are easily accessible (Yıldırım & Şimşek, 2006; Creswell & Poth, 2018). Participation in the study was voluntary, and students accessed the test online via turkcetest.net/test/b1.

Data Collection Instruments

In this study, two different listening tests developed at the B1 level for teaching Turkish as a foreign language (the Independent Item Test and the Common-Stem Item Test), along with a questionnaire designed to assess students' perceptions of these item formats, were used as data collection instruments. The listening skill tests were developed based on the "Turkish Language Teaching Program for Foreigners" by the Turkish MoNE. This program, which follows the Common European Framework of Reference for Languages, is designed for learners studying Turkish in formal and non-formal educational settings both in Turkey and abroad (MEB, 2020). The test items were developed in line with the action-oriented approach, in which the student is considered a language user as a social actor.

For the purpose of this research, two listening tests containing independent and common-stem items—both measuring the same construct—were designed by the researchers based on five main learning objectives. Independent items consist of individual, self-contained questions following a one text, one question format, while common-stem items consist of one text followed by three questions, forming a sequence of interrelated items within the same context. Both tests included parallel items that served as alternatives to one another.

In terms of cognitive level, the test items were designed based on Bloom's Revised Taxonomy and were limited to the comprehension level. Table 1 presents the structure of the listening tests used in the study according to learning objectives.

According to Table 1, the independent-item test consists of 15 short listening passages, with each passage corresponding to one independent item (a total of 15 items). In contrast, the common-stem-item test consists of 5 long listening passages, each of which contains three items, resulting in a total of 15 items. The independent and common stem items have entirely different text content; however, they exhibit parallelism in question structures and the traits being measured. The item numbers indicate the sequence within the test. The narration speed of the listening passages was adjusted to be suitable for B1-level learners, ensuring clarity and familiarity in accent and pronunciation. The duration of the listening passages in the independent-item test ranges from 20 to 57 seconds, whereas in the common-stem-item test, it ranges from 54 to 153 seconds.

Table 1. Framework of Listening Skills Learning Outcomes at the B1 Level

Learning Outcome	Independet Item Test	Common-Stem Item Test
B1.D.63.Selects the required information from what is listened to or watched.	- Squirrels (M5) - Tarantella Dance (M6) - Mauritania Train (M8) - Borrowed Book (M15)	- Charlie Chaplin (M2) - Unforgettable Moments of the Olympics (M6) - University of Oxford (M9) - The Genius Who Solved the Millennium Problem (M10)
B1.D.20.Identifies the main idea and supporting ideas in what is listened to or watched.	- What Will Happen? (M2) - Our Dear Friends (M4) - Mehmet Kuşman (M10) - The One Who Sleeps Wins (M12)	- Charlie Chaplin (M3) - Unforgettable Moments of the Olympics (M4) - The Genius Who Solved the Millennium Problem (M11) - Jules Verne (M15)
B1.D.30.Recognizes statements that include observations and impressions.	- 42,000-Year-Old Foal (M3) - Flamingo Birds (M11) - Travel Tour (M14)	- Unforgettable Moments of the Olympics (M5) - The Genius Who Solved the Millennium Problem (M12) - Jules Verne (M14)
B1.D.45.Identifies key words in texts.	- Lazuli (M1) - Pomegranate Mother (M7)	- University of Oxford (M7) - Jules Verne (M13)
B1.D.1.Understands texts/conversations related to needs and situations in social life contexts.	- Adana Flavor Festival (M9) - Grand Bazaar (M13)	- Charlie Chaplin (M1) - University of Oxford (M8)
Total	15	15

All test items were designed in the multiple-choice format, with five answer choices per item.

In order to ensure content and face validity of the items, expert opinions were obtained from five subject-matter experts in Turkish language education and four experts in measurement and evaluation. Experts were asked to evaluate each item based on three main criteria: (1) alignment with the intended learning outcome, (2) appropriateness for the cognitive level, and (3) linguistic clarity. The evaluations were carried out using the options "appropriate," "inappropriate," and "needs revision," and were supported with open-ended comments. Additionally, experts were asked to evaluate whether the independent and common-stem item formats measured the same underlying competence. Based on the feedback received, certain items were revised for linguistic simplicity, instructions were clarified, and some items were modified or rewritten to better align with the intended outcomes. Experts indicated that both types of items—independent and common-stem—measured the B1-level listening outcomes with similar cognitive demands and based on the same underlying construct. As a result, the necessary structural consistency between the two item formats was achieved, and the assessment instrument was finalized accordingly.

The final section of the data collection instrument includes a five-item questionnaire designed to evaluate students' perceptions of independent and common-stem item formats. The questionnaire was developed to allow students to compare and reflect on their experiences with the two different test types. It covers five dimensions: time management, clarity of the question format, cognitive challenge of the question format, applicability in an online environment, and ease of answering listening questions. For each item, participants were asked to choose one of the following options: "Independent Item Test," "Common-Stem Item Test," "Both," or "Neither." This questionnaire aims to contribute to a comparative evaluation of the two test formats based on students' perspectives and to provide insight into their perceptions regarding the structure of the tests.

The data collection tool was implemented using the Concerto Platform (concertoplatform.com), an open-source software developed by the Psychometrics Centre at the University of Cambridge. The Concerto Platform, with its user-friendly interface, enables the development and administration of

online tests without requiring coding skills. Students participating in the study accessed the tests voluntarily via the online link provided (turkcetest.net/test/b1).

Data Collection Process

The test administrations, including the Independent Item Test, the Common-Stem Item Test, and the questionnaire, were completed in a single session across three stages. Participants who did not complete the entire test administration were excluded from the analysis; however, no restrictions were applied to the questionnaire responses. In the Turkish listening comprehension tests, students were allowed to listen to each passage twice. The total test duration varied between 50 to 70 minutes, depending on the participants' individual pace.

Data Analysis

In this study, the Independent-Item Test and the Common-Stem-Item Test were analyzed comparatively based on Classical Test Theory. Within this framework, item and test statistics were examined in detail. For both tests, item means, standard deviations, number of items, skewness, and kurtosis values were calculated. The normality of data distribution was assessed based on descriptive statistics by examining skewness and kurtosis values. Skewness indicates whether the distribution of scores is concentrated in lower or higher score ranges, whereas kurtosis (peakedness) evaluates whether the distribution is more or less clustered around the center compared to a normal distribution (Özçelik, 2013).

Item difficulty refers to the proportion of test-takers who correctly answer an item. Item difficulty indices (pj) were calculated for each item, and a Z-test for the difference between two proportions was conducted to determine whether there was a significant difference between the difficulty indices of the Independent-Item Test and the Common-Stem-Item Test. Effect size was interpreted using Cohen's h. Item difficulty is a fundamental measure used to determine how easy or difficult a test item is. According to Crocker and Algina (1986), items with a difficulty index between 0.00 and 0.20 are classified as "very difficult", meaning they are typically only answered correctly by high-ability individuals. Items with a difficulty index between 0.20 and 0.40 are considered "difficult", indicating that a relatively small number of test-takers answer them correctly. Items with a difficulty index between 0.40 and 0.60 are classified as "moderately difficult" and can differentiate between high- and low-performing individuals. Items with a difficulty index between 0.60 and 0.80 are classified as "easy", meaning most test-takers answer them correctly. Finally, items with a difficulty index between 0.80 and 1.00 are considered "very easy", as nearly all participants answer them correctly.

A range of item difficulties is important for detecting individual differences and accurately measuring targeted abilities. Therefore, achievement tests should be structured to include a wide distribution of difficulty levels to cover all skill levels effectively. This classification is essential in creating a balanced test that includes items of varying difficulty levels in the test development process.

The item discrimination index represents the correlation between item scores and total test scores. In this study, point-biserial correlation values were calculated as the discrimination index for test items. The significance of the differences between correlation coefficients was assessed using a Z-test, and effect size was interpreted using Cohen's h. Item discrimination values are a crucial measure for determining how well an item assesses success on the test. According to Ebel (1972), if the discrimination index falls between -1.00 and -0.20, the item is considered negatively discriminating and should be removed from the test. If the discrimination index is between -0.19 and 0.19, the item does not differentiate between high- and low-performing students and fails to adequately measure the intended skill; therefore, such items should also be excluded from the test. A discrimination index between 0.20 and 0.29 indicates a partially discriminating (valid) item, which requires revision before inclusion in the test. If the discrimination index falls between 0.30 and 0.39, the item is moderately discriminating and may be included in the test with minor modifications. Finally, a discrimination index between 0.40 and 1.00 suggests that

the item is highly discriminating, requiring no modifications, and should be included in the test as it significantly contributes to overall performance. The effect size for differences in item difficulty and item discrimination indices was calculated using Cohen's h, which measures the magnitude of differences between proportions (Cohen, 1988).

According to Table 2, the skewness and kurtosis values of both the Independent-Item Test and the Common-Stem-Item Test fall within the range of -1.5 to +1.5. Based on Tabachnick & Fidell (2013), this suggests that the data follow a normal distribution.

Table 2. Descriptive Statistics of the Independent-Item Test and the Common-Stem-Item Test

	N	Min	Mak.	$\bar{X}$	Std. Dev.	Variance	Skewness	Kurtosis
Independent Item Test	201	1	15	8,53	0,29	17,13	-0,04	-1,43
Common-Stem-Item Test	201	0	15	6,82	3,68	13,57	0,18	-1,21

According to Cohen (1988), a small effect size (h ≈ 0.2) suggests that the difference between two groups is minimal or insignificant. A medium effect size (h ≈ 0.5) indicates a meaningful difference between groups, although not a strong one. Such differences are generally considered important in practical applications. Finally, a large effect size (h ≈ 0.8) represents a substantial and strong difference between groups, implying that the effect is significant and should be taken into account in practical applications. The reliability coefficients for both test applications were calculated using the KR-20 (Kuder-Richardson Formula 20) reliability coefficient. The significance of the differences between reliability coefficients was analyzed using Fisher's Z-transformation and the Z-test. Additionally, the mean scores of the Independent-Item Test and the Common-Stem-Item Test were compared using a paired-samples t-test to assess statistical significance. Student feedback on the usability of the tests was analyzed using descriptive statistics, including frequency and percentage calculations. All statistical analyses were conducted using Microsoft Excel Office and SPSS 20 software.

Findings

This section presents the findings obtained from the data analyses conducted in line with the research questions.

Descriptive Statistics

The descriptive statistics of the Independent-Item Test and the Common-Stem-Item Test, both designed to assess the same competencies, are presented in Table 2.

Comparison of Item Difficulty Indices

The differences in item difficulty indices between the test types and item formats were analyzed using the Z-test, while effect sizes were interpreted using Cohen's h coefficient. The results are presented in Table 3.

Examining Table 3, the item difficulty indices for the Independent-Item Test range between 0.38 and 0.77, suggesting that the test does not contain very easy or very difficult items but consists of items of moderate difficulty. The item difficulty indices for the Common-Stem-Item Test range between 0.38 and 0.67, indicating that the test includes both difficult and moderately difficult items. The mean difficulty index for the Independent-Item Test is 0.57, whereas for the Common-Stem-Item Test, it is 0.45. This suggests that both tests have similar levels of difficulty, with an overall moderate difficulty level.

The Z-test was conducted to examine whether there were significant differences in item difficulty indices between the Independent-Item Test and the Common-Stem-Item Test, and the effect sizes were assessed using Cohen's h coefficient.

The analyses revealed that for the M4-M4, M3-M5, M14-M14, M7-M13, and M9-M1 item pairs, no significant differences were found. However, for the remaining item pairs, significant differences were detected at the 0.01 significance level, indicating that the item difficulty indices varied significantly between the two test formats.

Table 3. Item Difficulty Indices of the Independent-Item Test and the Common-Stem-Item Test

Item No	Independent Item No	Common Stem Item No	$p_{independent}$	$p_{commonstem}$	Z	Cohen's h
1	M5	M2	0,56	0,43	2,61**	0,37
2	M6	M6	0,38	0,56	-3,62**	0,52
3	M8	M9	0,65	0,23	8,48**	1,32
4	M15	M10	0,65	0,36	5,81**	0,86
5	M2	M3	0,52	0,47	1,00**	0,14
6	M4	M4	0,58	0,37	4,22	0,61
7	M10	M11	0,68	0,38	6,03**	0,89
8	M12	M15	0,49	0,50	-0,20**	0,03
9	M3	M5	0,77	0,45	6,58	0,98
10	M11	M12	0,56	0,56	0,00**	0,00
11	M14	M14	0,57	0,12	9,49	1,52
12	M1	M7	0,49	0,53	-0,80**	0,11
13	M7	M13	0,51	0,52	-0,20	0,03
14	M9	M1	0,72	0,74	-0.45	0,06
15	M13	M8	0,42	0,59	-3.41**	0,49

Note: Independent and common stem items are different items targeting the same content. The item numbers indicate the sequence within the test.
**p<0.05; **p<0.01*

In terms of effect size, the differences in M8-M9, M15-M10, M10-M11, and M14-M14 pairs were found to be very large, suggesting that these items show substantial differences depending on the test format. Moderate effect sizes were observed for the M5-M2, M6-M6, M4-M4, and M13-M8 item pairs, indicating that their difficulty levels varied significantly based on the test format, though the differences were not exceptionally large. For item pairs with small effect sizes, the differences can be considered negligible in practical terms.

As a result, these findings suggest that the format of multiple-choice items (independent vs. common-stem) may influence students' success rates, indicating that test design choices could impact performance outcomes.

Comparison of Item Discrimination Indices

The differences in item discrimination indices between the test types and item formats were analyzed using the Z-test, while effect sizes were interpreted using Cohen's h coefficient. The results are presented in Table 4.

Examining Table 4, the item discrimination indices for the Independent-Item Test range between 0.38 and 0.67, indicating that the test consists of discriminating and highly discriminating items. Similarly, the item discrimination indices for the Common-Stem-Item Test range between 0.35 and 0.72, suggesting that this test also contains discriminating and highly discriminating items. However, the M2 and M14 items in the Common-Stem-Item Test were found to have insufficient discrimination indices and were therefore removed from the test. For the same reason, the M5-M2 and M14-M14 item pairs were excluded from the evaluation.

Table 4. Item Discrimination Indices of the Independent-Item Test and the Common-Stem-Item Test

Item No	Independent Item No	Common Stem Item No	$r_{pb\text{-}independent}$	$r_{pb\text{-}commonstem}$	Z	Cohen's h
1	M5	M2	0,63	0,14	10,10**	1,65
2	M6	M6	0,38	0,42	-0,82	0,12
3	M8	M9	0,62	0,35	5,42**	0,79
4	M15	M10	0,67	0,65	0,42	0,06
5	M2	M3	0,64	0,63	0,21	0,03
6	M4	M4	0,43	0,61	-3,61*	0,52
7	M10	M11	0,49	0,45	0,80	0,11
8	M12	M15	0,57	0,58	-0,20	0,03
9	M3	M5	0,51	0,55	-0,80	0,11
10	M11	M12	0,61	0,71	-2,12	0,30
11	M14	M14	0,58	0,13	9,43*	1,51
12	M1	M7	0,49	0,72	-4,72**	0,68
13	M7	M13	0,64	0,68	-0,85**	0,12
14	M9	M1	0,64	0,52	2,44*	0,35
15	M13	M8	0,56	0,44	2,41*	0,34

Note: Independent and common stem items are different items targeting the same content. The item numbers indicate the sequence within the test.
**p<0.05; **p<0.01*

The Z-test was conducted to examine whether there were significant differences in item discrimination indices between the Independent-Item Test and the Common-Stem-Item Test, and the effect sizes were assessed using Cohen's h coefficient. The analysis results indicate that there was a significant difference in favor of the Independent-Item Test for the M8-M9, M7-M13, M9-M1, and M13-M8 item pairs at the 0.01 significance level. Additionally, a significant difference in favor of the Common-Stem-Item Test was found for the M4-M4 and M1-M7 item pairs at the 0.05 significance level.

In terms of effect size, large effect sizes were observed particularly in the M5-M2 and M14-M14 item pairs, where the discrimination indices were low, and removal from the test was deemed appropriate. Moderate effect sizes were found in M8-M9,

M4-M4, and M1-M7 item pairs, which showed significant differences. While some other item pairs exhibited statistically significant differences, their effect sizes were generally small or negligible in practical terms.

The mean item discrimination index for the Independent-Item Test was calculated as 0.56, whereas for the Common-Stem-Item Test, it was 0.51. The fact that the mean discrimination indices for both tests exceed 0.40 suggests that the items are highly discriminating, demonstrating that both tests have a strong capacity to differentiate individual differences among students.

t-Test for the Comparison of Test Means

A paired-samples t-test was conducted to determine whether there was a significant difference between the mean scores of students on the Independent-Item Test and the Common-Stem-Item Test. The results are presented in Table 5.

Table 5. t-Test for the Comparison of Test Means

	N	$\bar{X}$	S	SD	t	p
Independent Item Test	201	8,53	4,14	200	9,950	0,00**
Common Stem Item Test	201	6,82	3,68			

**p < 0.01

According to Table 5, there is a statistically significant difference between the mean scores of the Independent-Item Test and the Common-Stem-Item Test at the 0.01 significance level. The mean score of students in the Independent-Item Test was calculated as 8.53, while the mean score in the Common-Stem-Item Test was 6.82. This finding indicates that students performed significantly better on the Independent-Item Test, suggesting that item format influences students' test performance. Accordingly, students achieved higher success in the Independent-Item Test.

Test Reliability

The reliability levels of the Independent-Item Test and the Common-Stem-Item Test were calculated using the KR-20 reliability coefficient. The reliability coefficients of both tests were compared using Fisher's Z transformation, and the significance of the difference between them was evaluated using the Z-test. The results are presented in Table 6.

Table 6. Test Reliability and Z-Test Results

	N	K	KR-20	Z
Independent Item Test	201	15	0,85	1,58
Common-Stem Item Test	201	15	0,80	

Examining Table 6, the KR-20 reliability coefficient was calculated as 0.85 for the Independent-Item Test and 0.80 for the Common-Stem-Item Test. In achievement tests, a minimum reliability value of 0.70 is expected (Fraenkel & Wallen, 1993; Kline, 2000). According to Cohen, Swerdlik, and Sturman (2013), KR-20 reliability coefficients within the range of 0.80 to 0.89 indicate high reliability. Based on this, both tests can be considered to have high reliability. Furthermore, the Z-test results indicate that there is no statistically significant difference in reliability levels between the two tests at the 0.01 significance level. This finding suggests that the reliability of the test results is not affected by item format, confirming that both tests provide reliable measurements.

Students' Perception of Independent and Common-Stem Item Formats

A questionnaire was administered to evaluate students' perceptions regarding independent and common-stem item formats. A total of 201 students participated in the test application; however, 11 of them chose not to respond to the questionnaire. Therefore, the findings are based on the responses of 190 participants. Descriptive statistics related to the results are presented in Table 7.

Findings from Table 7 indicate that students' perceptions regarding independent and common-stem item formats differ across dimensions such as time management, clarity of the question format, and practicality of administration on an online platform. In terms of time management, 53.7% of the students found the independent item test more practical, while 16.8% indicated that the common-stem item test was more advantageous. Additionally, 20.5% of the students stated that both tests were equally practical, whereas 8.9% believed that

neither test had an advantage in terms of time management. Regarding the clarity of the question format, 27.9% of the students found the independent item test easier to understand, while 16.3% preferred the common-stem item test in this respect. However, nearly half of the participants (47.9%) reported that both tests were equally clear. When asked which test had a more challenging question format, 58.4% of the students reported the common-stem item test as more challenging. The proportion of students who found the independent item test more challenging was 13.7%. Concerning ease of administration on an online platform, 49.5% of the students considered the independent item test easier to administer, while 17.4% favored the common-stem item test in this regard.

Table 7. Students' Perceptions of Independent and Common-Stem Item Formats

		Independent Item Test		Common Stem Item Test		Both		None	
		f	%	f	%	f	%	f	%
1	Which test was more practical in terms of time management?	102	53,7	32	16,8	39	20,5	17	8,9
2	In which test was it easier to understand the question format?	53	27,9	31	16,3	91	47,9	15	7,9
3	Which test had a more challenging question format?	26	13,7	111	58,4	28	14,7	25	13,2
4	Which test was eaier to administer in an online platform?	94	49,5	33	17,4	42	22,1	21	11,1
5	Which test's question format made it easier for you to answer the listening questions?	46	24,2	30	15,8	96	50,5	18	9,5

Finally, 50.5% of the students stated that both test formats similarly facilitated answering the listening questions, while 24.2% perceived the independent item test and 15.8% the common-stem item test as more advantageous in this respect. Overall, students perceived the independent item test as more advantageous in terms of time management and ease of administration on an online platform, whereas the common-stem item test was considered more challenging in terms of question format.

However, both test formats received similarly positive feedback regarding the clarity of the questions and their role in facilitating responses to listening tasks.

Discussion, Conclusion and Recommendations

This study compared the psychometric properties and student perceptions of independent and common-stem item formats used in B1-level Turkish listening tests designed for learners of Turkish as a foreign language. The findings revealed that the independent item test was more advantageous in terms of item difficulty and discrimination, suggesting a more favorable impact on student performance. This result aligns with Haladyna's (2004) assertion that independent items are more effective in assessing discrete pieces of knowledge. It also parallels the findings of Koçdar et al. (2017), who emphasized that item types may significantly influence item difficulty and discrimination indices.

Students perceived the independent-item format as easier, clearer, and more manageable. This is consistent with Tezel's (2020) view that long listening passages may increase cognitive load and hinder comprehension. Similarly, Tozlu (2017) argued that such long texts may measure memory retention rather than actual listening skills, thereby reducing construct validity in common-stem item formats. The findings of this study support these concerns, suggesting that the shorter and more focused structure of independent items reduces cognitive burden.

These results are also in line with studies conducted in other language teaching contexts. Kobayashi (2002) and In'nami & Koizumi (2009) reported significant performance differences based on item format in reading and listening tests. Buck (2001) and Ghonsooly & Fatemi (2013) likewise

emphasized the critical role of item format in listening assessment. Moreover, Fulcher (2010), Wilson (2005), and Boone (2022) highlighted the need to evaluate test formats not only in terms of psychometric features but also with regard to learner experience, time management, and usability.

Within this context, the findings indicate that the independent item format offers greater usability, particularly in computer-based and individualized testing environments. The ability to present each item with a separate audio file enhances focus and allows for more flexible technical implementation.

Although both tests in this study demonstrated high reliability, no statistically significant difference was observed in reliability indices. This suggests that test format may not directly influence reliability; however, it can play a decisive role in determining item difficulty and discrimination (Akyıldız & Karadağ, 2018; Öksüz & Demir, 2019). Additionally, findings based on student perceptions indicated that common-stem items required more effort, were harder to understand, and posed challenges in time management. These findings are consistent with Tezel's (2020) identification of psychological and language-level-related factors—such as lack of concentration, fatigue, and comprehension difficulties—that negatively affect listening performance.

Overall, the results emphasize the importance of selecting appropriate item formats based on the purpose of the test, the test environment, and the characteristics of the target group. Independent items, particularly when accompanied by short and focused audio prompts, appear to be a more suitable alternative for computer-based listening assessments due to their reduced cognitive load and ease of administration. This study contributes to the limited body of research comparing item formats in the context of Turkish as a foreign language and provides context-specific empirical evidence to inform test design.

In an effort to enhance test development and language assessment practices, several recommendations can be made for relevant stakeholders. Test developers are encouraged to place greater emphasis on independent item formats in listening assessments, particularly considering factors such as passage length and cognitive load. Independent items offer notable technical and pedagogical advantages in computer-based environments, whereas common-stem formats may be more appropriate for measuring integrative comprehension skills that require contextual continuity. For practitioners, employing assessment tools that account for student perceptions may contribute to increased motivation and reduced test anxiety. Therefore, it is advisable to gather post-test feedback from students and integrate this data into future test design processes. For researchers, it is important to conduct similar comparative studies across different language skills, such as reading and speaking, and at varying proficiency levels ranging from A1 to C1. In addition, deeper investigation is needed into how item formats relate to learners' cognitive processes, test-taking strategies, and motivation. Longitudinal research should also be undertaken to examine the long-term impact of item types on learning outcomes.

References

Akyıldız, M., & Karadağ, N. (2018). Farklı soru türlerinin güçlük ve ayırt edicilik düzeylerinin incelenmesi. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 4*(1), 112-122.

Boone, W. J. (2022). *Rasch analysis for instrument development: Why, when, and how?* Routledge. https://doi.org/10.4324/9780429295954

Bridgeman, B. (1992). A comparison of quantitative questions in different formats. *Educational and Psychological Measurement*, *52*(4), 913-918.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th ed.). McGraw-Hill Education

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). Thousand Oaks, CA: Sage Publications.

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

Demirkol, S., & Karagöz, M. A. (2023). PISA 2015 Okuma Becerisi Maddelerinin Güçlük İndeksini Etkileyen Madde Özelliklerinin İncelenmesi. *Ana Dili Eğitimi Dergisi*, *11*(3), 567-579. Doi: 10.16916-aded.1212049-2802742

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Ege, H. G., & Demır, E. (2020). Examining of Internal Consistency Coefficients in Mixed-Format Tests in Different Simulation Conditions. *Eurasian Journal of Educational Research*, *20*(87), 101-118. DOI: 10.14689/-ejer.2020.87.5

Eren, B. (2015). *Çoktan seçmeli ve karma test uygulamalarına ilişkin öğrenci başarıları ile öğrenci ve öğretmen görüşlerinin karşılaştırılması* (Unpublished master's thesis). Ankara Üniversitesi, Ankara.

Field, J. (2008). *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Fraenkel, R. J. & Wallen, E. N. (1993). *How to design and evaluate research in education*. McGraw-Hill Inc.

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Ghonsooly, B., & Fatemi, A. H. (2013). The effect of item format on listening comprehension: Multiple-choice versus open-ended formats. *Journal of Language Teaching and Research, 4*(3), 563–569. https://doi.org/10.4304/jltr.4.3.563-569

Gültekin, S. (2011). *Çoktan seçmeli, açık uçlu ve karma testlerin psikometrik özelliklerinin madde tepki kuramına dayalı olarak değerlendirilmesi* (Unpublished master's thesis). Ankara Üniversitesi, Ankara.

Haladyna, T. M. (2004). Developing and validating multiple-choice test items. Lawrence Erlbaum Associates, Publishers.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 51-78.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 15(3), 309-334.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi.org/10.1177-/0265532208101006

Kan, A., & Kayapınar, U. (2006). Yabancı dil eğitiminde aynı davranışları yoklayan çoktan seçmeli ve kısa cevaplı iki testin madde ve test özelliklerinin karşılaştırılması. *Eğitim ve Bilim*, *31*(142).

Karasar, N. (2011). *Bilimsel Araştırma Yöntemi.* Nobel Yayıncılık.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge. https://doi.org/10.43-24/9780203224342

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*,19(2),193–220. https://doi.org/10.-1191/0265532202lt227oa

Koçdar, S., Karadağ, N., Şahin, M. D., & Karadeniz, A. (2017). Uzaktan eğitimde çoktan seçmeli soruların güçlük ve ayırt edicilik değerlerinin soru türlerine göre incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *32*(1), 168-184.

Millî Eğitim Bakanlığı (2013). *Diller için Avrupa Ortak Başvuru Metni: Öğrenme, öğretme ve değerlendirme.* MEB Yayınları.

Millî Eğitim Bakanlığı. (2020). *Türkçenin Yabancı Dil Olarak Öğretimi Programı.* MEB Yayınları.

Öksüz, Y., & Demir, E. G. (2019). Açık uçlu ve çoktan seçmeli başarı testlerinin psikometrik özellikleri ve öğrenci performansı açısından karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *34*(1), 259-282.

Öney, S.S. (2023). *Aynı özelliği ölçmeye yönelik olarak hazırlanan çoktan seçmeli ve karma testlerin psikometrik özelliklerinin karşılaştırılması.* (Unpublished master's thesis) Hacettepe Üniversitesi, Ankara.

Özçelik, D. A. (2013). *Test hazırlama kılavuzu* (5. Baskı). Pegem Akademi Yayınları.

Özdemir, D. (2004). Çoktan seçmeli testlerin klasik test teorisi ve örtük özellikler teorisine göre hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlanması yönünden karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *26*(26).

Özsu, M., & Can, N. (2020). İngilizce Dersi Test Maddelerinde Resim Kullanımının Test ve Madde Psikometrik Özelliklerine

Etkisi. *Kahramanmaraş Sütçü İmam Üniversitesi Sosyal Bilimler Dergisi*, *17*(1), 85-103. doi: 10.33437-ksusbd.693800-1079141

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13.

Sayın, V. & Orbay, K. (2024). Geometride Ölçme Değerlendirme: Çoktan Seçmeli ve Açık Uçlu Sınavların Karşılaştırılması, *Üçüncü Sektör Sosyal Ekonomi Dergisi*, 59(4), 2918-2931. doi: 10.15659/3.sektor-sosyal-ekonomi.24.12.2564

Tabachnick, B.G., Fidell, L.S. (2013) Using Multivariate Statistics (sixth ed.) Pearson, Boston (2013)

Taşkıran, E. (2022). *Üniversite öğrencilerinin okuduğunu anlama düzeylerinin belirlenmesinde farklı madde formatlarının karşılaştırılması* (Master's thesis) Hasan Kalyoncu Üniversitesi, Gaziantep.

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

Tezel, V.K. (2020) Türkçenin Yabancı Dil Olarak Öğretiminde Dinleme Eğitimi, Türkçenin Yabancı Dil Olarak Öğretimi El Kitabı, Pegem Akademi, 9.Bölüm, 271-310.

Tozlu, E. (2017). *The development of a listening test for learners of turkish as a foreign language*. (Master Thesis). Boğaziçi Üniversitesi, İstanbul.

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme.* Pegem Akademi.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, 6(2), 103-118.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Yıldırım, A., & Şimşek, H. (2006). *Sosyal bilimlerde nitel araştırma yöntemleri* (6. baskı). Ankara: Seçkin Yayıncılık.