Bartın University Journal of Faculty of Education
BUEFAD Volume 13, Issue 2
380-395
dergipark.org.tr/buefad
DOI: 10.14686/buefad.1424213

Evaluating the Psychometric Characteristics of Generated Visual Reading Comprehension Items

Ayfer Sayın [a*]

a* Dr., Gazi University, https://orcid.org/0000-0003-1357-5674, *ayfersayin@gazi.edu.tr

Research Article
Received: 23.1.2024
Revised: 2.2.2024
Accepted: 23.2.2024

Abstract

Reading comprehension, a crucial skill in today's information-rich environment, extends beyond text to include visual elements. Manual creation of visual reading comprehension items poses challenges, necessitating an innovative approach. This situation has led to the exploration of Automatic Item Generation (AIG) as a solution. This study aims to demonstrate the use of AIG for the creation of visual reading comprehension items. By developing cognitive and item models through expert input and utilizing computer algorithms for item generation, the study seeks to provide a time-efficient and reliable alternative for item writers. The field test involved 1,380 8th-grade students to evaluate the psychometric properties of the generated visual reading comprehension items. The AIG process starts with expert insights to develop cognitive and item models. Computer algorithms are then employed for AIG. The study utilizes a diverse sample of 8th-grade students for field testing, assessing the psychometric properties of the generated items. Field test results indicate the potential of AIG in efficiently producing a substantial item pool for visual reading comprehension. The generated items exhibit consistent difficulty levels (0.58 to 0.66), ensuring an appropriate challenge for students. High item discrimination (0.48 to 0.69) effectively distinguishes between students with varying visual reading comprehension skills. Item-total correlations (0.40 to 0.57) further validate the quality and validity of the generated items. The automated process yields efficient results in terms of item difficulty and discrimination, emphasizing the potential of AIG for high-quality assessment of visual reading comprehension items.

Keywords: Visual reading comprehension, automatic item generation, high-stake tests, parallel test.

Otomatik Üretilen Görsel Okuduğunu Anlama Maddelerinin Psikometrik Özelliklerinin Değerlendirilmesi

Öz

Günümüz bilgi dünyasında diğer becerilere ve disiplinlere de temel oluşturan okuduğunu anlama, metnin ötesine geçerek görsel unsurları da kapsamaktadır bu nedenle de ölçülmesi önemli görülmektedir. Görsel okuma maddelerinin oluşturulma süreci ise geleneksel yaklaşımla manuel bir şekilde yürütülmekte, görseller üzerinde müdahaleler kısıtlı olmakta ya da he bir tablo, grafik, diyagram madde yazarları tarafından manuel oluşturulmaktadır. Bu da zaman alıcı, yoğun çaba gerektiren bir sürece işaret etmektedir. Bu durumla başa çıkmak için bilgisayar teknolojisindeki yenilikleri kullanan bir yöntem olan Otomatik Madde Üretimi (OMÜ) yöntemi geliştirilmiştir. Bu çalışma, görsel okuduğunu anlama maddelerinin oluşturulması için OMÜ'nün nasıl kullanılacağını göstermeyi amaçlamaktadır. Model tabanlı OMÜ'ye göre gerçekleştirilen işlemlerde öncelikle uzmanlar bilişsel ve madde modelleri geliştirmiş, ardından madde üretimi için bilgisayar algoritmaları kullanılmıştır. Eş değer maddelerin üretiminin amaçlandığı bu çalışma madde yazarları için zaman açısından verimli ve güvenilir bir alternatif sağlamayı amaçlamaktadır. Üretilen görsel okuduğunu anlama maddelerinin psikometrik özelliklerini değerlendirmek için 8. sınıfta öğrenim gören 1.380 öğrenci ile ön uygulama çalışması gerçekleştirilmiştir. Araştırma sonucunda madde güçlük seviyelerinin (0,58 ila 0,66) benzer özellikler gösterdiği belirlenmiştir. Yüksek madde ayırt ediciliği (0,48 ila 0,69), farklı görsel okuduğunu anlama becerilerine sahip öğrenciler arasında etkili bir ayrım yapmaktadır. Madde-toplam korelasyonları (0,40 ila 0,57), oluşturulan maddelerin kalitesini ve geçerliliğini ek bir kanıt sunmaktadır. Bu çalışmada gerçekleştirilen otomatikleştirilmiş süreç, madde güçlüğü ve ayırt edicilik açısından verimli sonuçlar vermekte ve OMÜ'nün görsel okuduğunu anlama maddelerinin yüksek kalitede değerlendirilmesine yönelik potansiyelini vurgulamaktadır.

Anahtar kelimeler: Görsel okuduğunu anlama, otomatik madde üretimi, yüksek riskli testler, eş değer test

To cite this article in APA Style:

Sayın, A. (2024). Evaluating the psychometric characteristics of generated visual reading comprehension items. *Bartın University Journal of Faculty of Education, 13*(2), 380-395. https://doi.org/10.14686/buefad.1424213

INTRODUCTION

Reading comprehension items directly measures an individuals' ability to understand and interpret texts. These items require individuals to grasp the main idea, make inferences, draw conclusions (Fielding & Pearson, 1994; Woolley & Woolley, 2011), and engage in critical evaluation of the text, including identifying supporting evidence, analyzing the purpose and point of view, and making connections to other texts and real-life situations (Aloqaili, 2012; Brevik, 2019; Hosseini et al., 2012). Students are presented with reading comprehension items to provide data on students' curricular achievements and various aspects of their development. The results obtained from the process help to identify individual differences and learning difficulties among students and allow educators to provide targeted support and intervention. These interventions may encompass areas such as vocabulary development, inferencing skills, and understanding text structure (Chandran & Shah, 2019; Cornoldi & Oakhill, 2013; Klingner et al., 2015; Westwood, 2016). Moreover, reading comprehension is a fundamental skill for all disciplines, beyond academia, and is also crucial for professional contexts and the development of daily life skills, where individuals encounter complex multimodal texts on social media and other platforms (Barnes, 2015; Boonen et al., 2016; Kinniburgh & Shaw, 2009; Oliver, 2009; Österholm, 2006; OECD, 2021).

The development of comprehension skills involves a combination of language and cognitive abilities, ranging from literal understanding of explicit information to inferential and evaluative comprehension, which requires inference and critical analysis, respectively (Basaraba et al., 2013). Although these skills are listed hierarchically, they are interconnected. For example, lower-level language skills such as vocabulary and grammar are important for developing higher-level comprehension skills (Silva and Cain, 2015), and oral language skills and higher-level cognitive skills form the basis for improving listening and learning, reading comprehension, and managing complex texts (Lervåg et al., 2017; Eason et al., 2012). Research also distinguishes between literal and inferential comprehension; states that skilled comprehenders are successful in inferential tasks (Duncan et al., 2015) and recognizes digital stories as tools to improve comprehension at different levels (Al-Hameed and Al-Shuair, 2019). In this regard, the elements developed for reading comprehension need to change according to the feature they aim to measure. In the context of PIRLS, fourth-grade readers are evaluated based on their ability to focus on and retrieve explicitly stated information, make straightforward inferences, interpret and integrate ideas and information, and evaluate and critique content and textual elements. This framework posits four distinct levels of comprehension assessment. Similarly, the PISA framework elaborates on the comprehension process by delineating it into several cognitive processes: locating information, accessing and retrieving information within a text, searching and selecting relevant text, reading fluently, understanding (which encompasses representing literal meaning and integrating and generating inferences), and evaluating and reflecting (which includes assessing the quality and credibility of the text, reflecting on content and form, and detecting and handling conflicts). These categorizations exemplify the nuanced approaches to measuring reading comprehension, underscoring the multifaceted nature of understanding textual material in educational assessments. This research aimed to generate visual reading comprehension items required by inferential cognitive processes.

In today's information-rich society, reading comprehension items extend beyond written text and involve visual texts, which are combinations of textual information and visuals (Li et al., 2019; Woolley & Woolley, 2011). Visual texts, which are increasingly prevalent in our daily lives, also play a crucial role in reading comprehension assessments such as SAT, TOEFL, PISA, and PIRLS. These assessments specifically measure students' ability to make inferences, conclude, and critically analyze the relationship between textual and visual information, thereby assessing higher-level skills (Cahalan et al., 2002; Cohen & Upton, 2006; Unsworth, 2014; Mullis et al., 2017; OECD, 2019). Additionally, well-constructed visual texts with captivating visual stimuli enhance students' motivation for exams (Glenberg & Langston, 1992; Hoyt, 1992). However, creating visual texts and writing visual reading comprehension items can be challenging and time-consuming compared to other item types (Author, 2023).

Visual texts demand effective integration of visual elements with accompanying text, including selecting appropriate visuals that align with the content and purpose of the item. Balancing textual and visual components coherently and meaningfully can be more complex than writing text-only items (Daly & Unsworth, 2011; Sabatini et al., 2014). The images used in visual texts must accurately represent the information presented in the text be clear, appealing, and effectively convey the intended message. Additionally, factors such as layout, design, and readability of visuals should align with the objectives of the item and support comprehension for the target audience (Hoyt, 1992). Furthermore, the integration of visual and auditory elements has become a compelling feature of computer-based tests, making them highly appealing and widely used in modern educational settings.

As a result, computer-based testing is shown to be an effective method for assessing students' visual reading comprehension skills.

Computer-based tests offer flexible testing options and rapid score calculation, benefiting educators and students alike (Chen et al., 2019; Gierl et al., 2021). This flexibility is particularly advantageous in classroom practice, where traditional paper-and-pencil exams can be time-consuming to score due to large class sizes and other responsibilities (Chen et al., 2019). It provides swift feedback, allowing teachers to identify individual learning needs promptly and facilitate targeted support (Weber et al., 2003). Moreover, the use of multimedia elements, such as photos and videos, in electronic tests enhances assessment opportunities and supports diverse item types (Gierl et al., 2021; Kosh et al., 2019). However, digital assessments or computer-based tests also face challenges, particularly in the context of distance education (Arrend, 2007). Security concerns and the need to create a substantial item pool are noteworthy issues. To prevent the disclosure of items before exams, synchronous test administrations have been adopted, but this approach sacrifices the flexibility that computer-based tests can offer (ÖSYM, 2020). Furthermore, the practice effect, where repeated test performance influences scores, can compromise the validity and reliability of measurement (Hausknecht et al., 2007). To ensure diverse items for in-class follow-up tests and personalized assessments, a substantial item pool with established psychometric properties is essential (Hausknecht et al., 2007). For that, creating an item pool with scalable difficulty is crucial, and it applies not only to the textual components but also to visuals in visual reading comprehension items. Ensuring that visuals are adaptable to difficulty levels adds flexibility to computer-based tests, allowing students to take the test at different times and locations, such as over three days. However, it's worth knowing that this process is challenging and resource intensive. To address this challenge, the field of AIG has emerged, combining computer technology with cognitive and psychometric theories (Arendasy & Sommer, 2012; Embretson & Yang, 2006; Gierl & Haladyna, 2012b).

Automatic item generation

Automatic item generation (AIG) is the process of automatically generating tests, exams, or items for educational and assessment purposes. It leverages cognitive and psychometric theories along with computer technology to produce high-quality items efficiently (Embretson & Yang, 2006; Gierl et al., 2019; Gierl & Lai, 2018; Gierl et al., 2012; Irvine & Kyllonen, 2013). AIG aims to continuously generate and diversify new items to assess student's various abilities and learning styles. It ensures items meet assessment criteria such as objectivity, reliability, and validity (Gierl & Haladyna, 2012a). AIG enables the creation of item pools for individual-specific tests, facilitates adaptation to updated curricula and learning objectives, and saves time and costs compared to traditional item writing processes (Gierl et al., 2019; Kosh et al., 2019).

AIG involves two main methods: artificial intelligence and templated-based approaches. The present study employs the templated-based AIG, which consists of three stages: developing the cognitive model, creating the item model, and automatically generating items using computer technology. This framework was similarly applied in the generation process of visual reading items. Given that visual texts are comprised of images analogous to how traditional texts are composed of words, the methodology was adapted to define visual elements within these texts with the same rigour and systematic process used for word-based texts. Specifically, the visual elements were identified and characterized following a structured approach, ensuring that the generated items aligned with the cognitive and item models initially developed. This adaptation underscores the flexibility and applicability of the template-based AIG method in addressing both textual and visual information, thereby facilitating the generation of comprehensive assessment items that accurately evaluate reading comprehension across different modalities. In the first stage, subject matter experts define the content required to generate new items, forming the cognitive model that encompasses the knowledge, skills, and abilities necessary to solve problems. In the second stage, item models are created as templates for the assessment tasks, specifying the parts of the items that should be modified to generate new items. In the final stage, the content from the first step is inserted into the template described in the second step using computer-based algorithms. The final stage involves combining the content from the cognitive model with specific parts of the item model according to predefined rules and restrictions established by subject experts. This process enables quick and AIG. Following these stages, AIG can be applied in various disciplines such as medicine, dentistry, mathematics, and literature (Adji et al., 2018; Embretson & Kingston, 2018; Falcão et al., 2022; Gierl & Lai, 2012; Lai, Gierl, Byrne, et al., 2016; Author, 2023).

AIG has demonstrated successful results in verbal fields, including reading comprehension items, by generating a wide range of items with various types, formats, and difficulty levels (Holling et al., 2009; Setiawan et al., 2022; Shin & Gierl, 2022). These items can effectively integrate textual information with images, graphics, or other visual elements. AIG allows customization of reading comprehension assessments based on individual

student characteristics and performance (Shin & Gierl, 2022). It dynamically adjusts the level of difficulty or item type presented to each student, providing a personalized and engaging assessment experience. AIG can also provide feedback and personalized learning support based on students' responses, promoting self-learning, and helping students identify their strengths and weaknesses in reading comprehension items (Author, 2023). However, using automatic item creation for visual reading comprehension items presents two main challenges (Shin & Gierl, 2022). First, visual item responsiveness requires a simultaneous understanding of both visual content and natural language elements within an image. Additionally, effectively modelling the interactions between visual and textual elements in the generation of visual reading comprehension items can be challenging (Li et al., 2019). In this research, a model has been designed to overcome these challenges, and data-based evidence has been obtained.

The Present Study

Visual reading comprehension extends beyond traditional written text and involves the interpretation of visual elements in combination with textual information. As a result, visual texts are increasingly prevalent in various assessments, such as SAT, TOEFL, PISA, and PIRLS, aiming to evaluate higher-order cognitive skills. However, the creation of effective visual reading comprehension items poses significant challenges, and the process can be time-consuming compared to text-only items. To address the need for creating items with scalable difficulty, this study aims to use visual reading comprehension assessment through AIG (AIG). Leveraging cognitive and psychometric theories, AIG combines computer technology to efficiently generate high-quality and scalable items. By developing a cognitive model and item model for verbal and visual text, it was generated a vast array of visual reading comprehension items using Python codes. The items underwent testing through field trials, assessing their difficulty and discrimination levels. The findings demonstrated the potential of AIG as a reliable and innovative approach to creating diverse, engaging, and valid visual reading comprehension items. As technological advancements continue to shape the educational landscape, it is considered that AIG emerges as a powerful tool to elevate the quality and efficiency of visual reading comprehension assessments, ultimately benefitting learners across diverse educational settings.

In shortly, this study aims to automatically generate visual reading comprehension items and examine their psychometric properties. The research intends to provide item writers with an alternative method to address their difficulties when writing visual reading comprehension items (Setiawan et al., 2022; Shin, 2021). The study encompasses (i) the development of an AIG model, (ii) the generation of items with similar item difficulties, (iii) the creation of a large item pool, and (iv) the examination of the psychometric properties of the items.

METHOD

Research Design

This study was designed as descriptive survey research to automatically generate visual reading comprehension test items and to examine their psychometric properties.

Participants

The research was conducted in the 2023-2024 academic year, with 1,380 students attending the 8th grade at a private educational institution in Türkiye. The age range of the students was between 13 and 14 years, and their native language is Turkish. The field test of 5 items randomly selected from the generated items was carried out in the screening test applied to these 8th-grade students by a private institution.Within the scope of the LGS screening test, five booklets have been prepared for participants. A visually based reading comprehension question, developed specifically for this research, has been incorporated into the 19th question of each booklet. The administration of the test has been conducted in a computer-based format. Students are afforded the opportunity to complete the test between Monday and Wednesday, thereby providing flexibility in terms of both location and time.

Data Collection

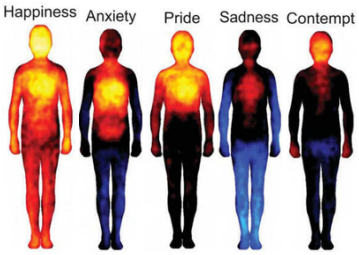
In the current study, visual reading comprehension items in the high school entrance exam (known as LGS), which is a high-risk test in Türkiye, were determined as parent items to generate items. This item's objective of the 8th-grade Turkish curriculum is "Establishes a relationship between visual elements and what they read the text". The items aimed for students to "conclude by evaluating visual and written text together." To achieve this goal, first, the five criteria to be included in an item are defined: (i) evaluating visual and verbal text together, (ii) organizing the options to be verbal or visual text, (iii) being scientifically accurate, (iv) engaging and intriguing,

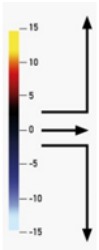
(v) establishing a relationship with daily life. Since this study aimed to generate items of similar difficulty, the semantic or proposal variable, which directly affects the item's difficulty, was added to the sixth criterion, (vi) establishing content on the same topic and main idea for both verbal and visual text. After determining the criteria for the features of the items to be produced, the generated process started. The visual reading comprehension items in this study were constructed using AIG following a three-stage process. In the first stage, a cognitive model was developed. As seen in Figure 1, the item body consists of verbal text, visual text, and a diagram. Also, idioms were used in the options as a verbal text for establishing a relationship with daily life. In this process, which was determined according to expert opinion, the cognitive model was created by listing the information sources, features, and elements for each part of the stem and the options. These variables were placed in the item model development by the experts in the second stage. In the third stage, the information in the cognitive model was placed into the item model and the items were automatically generated with computer algorithms using Python codes.

In this study, generation was performed by taking a parent item written by the researcher as a reference. As seen in Table 1, the parent item contains verbal text and two different visual texts. Since the second visual text is not generated, it is called a diagram.

Table 1. Parent item

Embarking on a captivating journey, the compelling "Bodily Maps of Emotions" study rallied more than 700 individuals from Finland, Sweden, and Taiwan, aged between 18 to 45, into a fascinating exploration. This research voyage set out to plumb the depths of emotionally charged materials and their profound ramifications on the human psyche. Participants were treated to an array of captivating stimuli, a medley encompassing words, videos, facial expressions, and stories, prompting them to introspect and identify the distinct regions of their bodies that experienced heightened arousal or, conversely, exhibited indifference. Leveraging cutting-edge computer technology (detectors), the researchers adroitly recorded the participants' bodily reactions, skilfully interwoven with their candid self-reported responses. Amidst diverse cultural backgrounds, the study yielded an enthralling revelation, unveiling strikingly similar body sensory maps among the participants. A part of the body sensation map is shown below and a comment on the interpretation of colours has been added:





From black to warm colors→ 0-15 indicated by	It expresses an increased sensation in the body. For example, after encountering material that expresses an emotion, a person's pulse rate increases and the reaction increases in certain parts of the body.
Black→ denoted by 0.	It refers to the body's normal course of sensation. For example, there is no change in a person's pulse rate when they encounter any material that expresses emotion. The body response remains the same.
Black to cool colors from → 0 indicated by -15	It expresses a decrease in sensation in the body. For example, after encountering any emotionally expressive material, the pulse rate drops and the reaction in certain parts of the body decreases.

Which of the sentences below aligns with the findings of the "Bodily Maps of Emotions"?

A) He always said that when he felt excessively contemptuous of someone, it gave him a stomachache.

B) After receiving the accident news, my feet were eager to run away from there due to the sadness.

C) Happiness is like butterflies dancing in my stomach, bringing joy to every corner of my being. *

D) When I saw my son on the television screen, I was filled with pride from head to foot toe.

In this study, as it is seen in the patent item the item generation process encompassed both verbal and visual texts within the stem, with each being generated accordingly. Additionally, a diagram incorporating both visual and verbal elements (constituting a second visual text) was created. Unlike the verbal and primary visual texts, this diagram was not generated but manually crafted. The variables from both text types were translated into a cognitive model, which informed the development of the item model. Subsequently, algorithms for generating items were devised. Furthermore, the research extended to the generation of answer choices, incorporating idiomatic expressions relevant to the context presented in the stem. The use of contextually appropriate idioms represented advanced reading comprehension skills. Following the identification of necessary features for generating plausible distractors, the generation process was executed using Python. This approach underscores the integration of computational methods with cognitive and linguistic theories to facilitate the automated generation of complex assessment items, thereby advancing the generating for evaluating comprehensive reading comprehension skills.

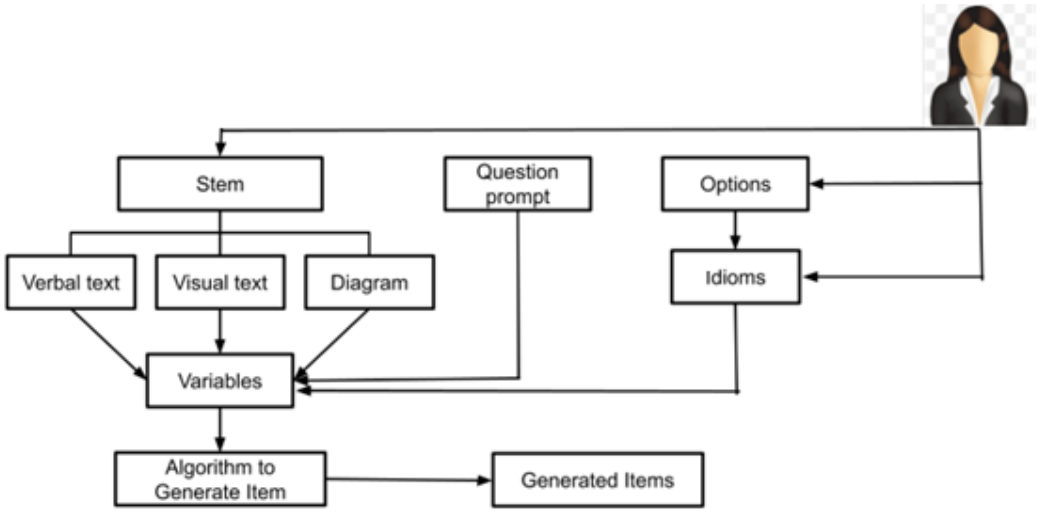


Figure 1. AIG Process

Development cognitive model

In this study, a text with a combination of text and visuals was designed to measure the acquisition of "Establishes a relationship between visual elements and what they read". In this way, the main purpose of the generated items is to interpret and integrate ideas and information. For the automatically generated items to be of similar difficulty, the item's topic and main idea features, which can be called semantic or purposive, were kept constant. As it is seen in the parent item, the focal point of this research is the exploration of bodily maps of emotions, as featured in the parent item stem. Bodily maps of emotions delineate the specific topographical distribution of bodily sensations that correspond to various emotional states. These representations have been empirically validated across diverse cultures, offering a universally consistent framework that underpins discrete emotional experiences (Nummenmaa et al., 2013; Goldstein et al., 2020). Studies, including those conducted by Hietanen et al. (2016), have elucidated those basic emotions—namely anger, fear, disgust, happiness, sadness, and surprise—are each linked to unique patterns of bodily sensations. This specificity in the bodily changes associated with distinct emotions suggests that these topographical representations are critical in differentiating between emotional states, thereby serving as an essential component of emotional experience (Goldstein et al., 2020). In essence, the study of bodily maps of emotions was selected for the item because there is a relationship between somatic sensations and emotional processes, presenting an interesting subject for students.

Following the determination of the subject matter, the stages of developing the cognitive model have defined the cognitive processes applicable for visual reading comprehension items: focusing on and retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information, and evaluating and critiquing content and textural elements. The objective of this study was to develop parallel forms by generating isomorphic items, thereby maintaining consistency in the topic and cognitive load. The subject matter focused on the bodily maps of emotions, with cognitive processing occurring at the level of interpreting and integrating ideas and information. Upon deciding the scenario for the cognitive model, the next phase involved identifying features and elements. For the verbal text's introduction (feature), information related to the purpose of the research and its participants (elements) was specified, in the development (feature) phase, the method and form (elements) were outlined, and for the conclusion (feature), details regarding the outcomes of the research (element) were defined. To achieve items of equivalent difficulty, elements were constrained. In the second phase of developing the cognitive model, the same processes applied to the visual text. As observed in the parent item, based on the research outcomes, five body shapes were identified. These shapes were categorized into increased sensation (feature) represented by warm colors (elements), decreased sensation (feature) indicated by cool colors (elements), and mixed sensation (feature) depicted using both warm and cool colors (elements). During the visual constraints process, emotions (elements) were identified in alignment with the bodily maps presented in the research, reflecting real-world knowledge encountered in everyday life. Since no visual was produced to explain the color tones within the scope of the research, this was expressed through a diagram. Options were also generated in the research; these options articulated the body-emotion relationship through idioms, with body parts determined as elements for the features. Following the development of the cognitive model, the study progressed to the phase of developing the item model. The developed cognitive model is shown in Figure 2.

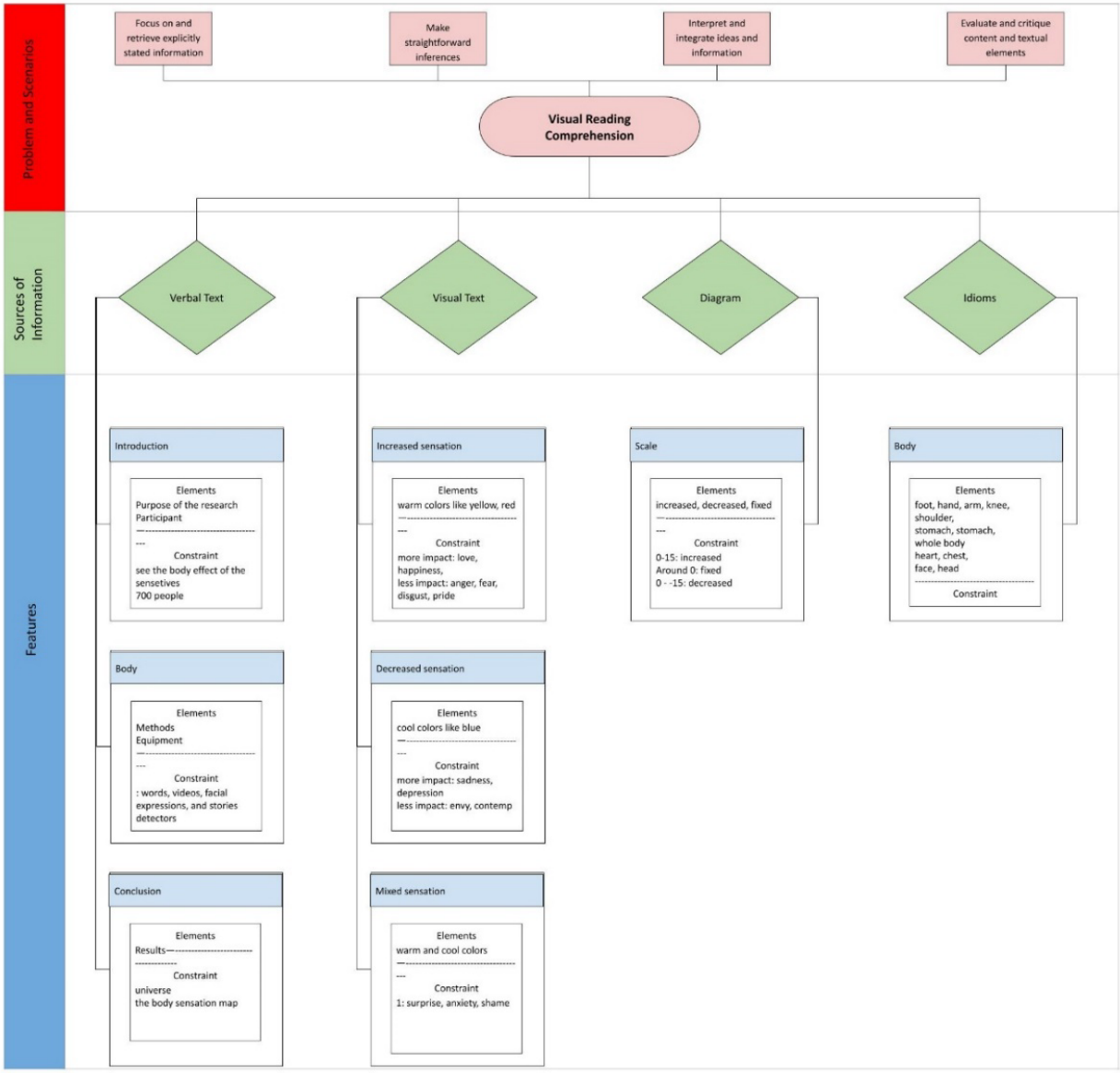


Figure 2. Cognitive Model

Development item model

In the second stage, the item model, which is the item format, was developed and shown in Table 1. As seen in Table 1, the item stem starts with a text. The introduction, development, and conclusion sections in the text were generated and rephrased using the possibilities of the language. In the research, results related to 13 emotions were included (Nummenmaa et al., 2014). Since it was carried to a multiple-choice item, 5 of these emotions were selected in line with the cognitive model. The selection criteria are shown in the item model. Finally, there is a diagram in the item stem. The options include idioms or words expressing emotion, which are commonly used in Turkish.

Table 1. Item Model

Stem	Verbal Text: Body sensation map <Introduction of the research>. Within the scope of research <methods>. The result of the research <result>.
	Visual text: Some emotions of the body sensation map created as a result of the research is shown in the figure below: <Increased1> <Increased2> <Decrased1> <Decrased2> <Mixed>
	Diagram: The colors on this map are interpreted as follows: <Scale>.
Element	*Introduction of the research:* Purpose of the research (see the body effect of the emotions),

participants (700 people) *Methods:* words, videos, facial expressions, and stories detectors *Result:* the sensation universe map Emotions: *Increased1:* love, happiness *Increased2:* anger, fear, disgust, pride *Decrased1:* sadness, depression *Decrased2:* envy, contempt *Mixed:* surprise, anxiety, shame *Part of the body:* foot, hand, arm, knee, shoulder, stomach, whole body, heart, chest, face, head.

Question prompt	In which of the following sentences does the situation expressed correspond to the results of the "body sensation map"?

Options	Idioms including <part of the body> about an <emotions>.

Data Analysis

Item difficulty, discrimination, and point biserial of the pre-applied items were estimated based on the Classical Test Theory (CTT). Since it was requested to prepare items of similar difficulty within the scope of this research, first, the item difficulty index was estimated. Item difficulty refers to the difficulty level of an item in a test or measurement tool. It usually reflects the mental or cognitive skill level required to answer the item correctly (Adedoyin et al., 2008). The item discrimination index is a statistical measure of how effectively a test item discriminates between high and low-performing test takers (Aiken, 1979). A high item discrimination index indicates that the item discriminates more clearly between performance levels, while a low index value makes the item poor at discriminating. This index is used to assess the reliability and effectiveness of the test and is usually calculated by statistical methods such as biserial correlations. Items with high item-total correlations have higher discrimination, while low correlations indicate low discrimination (Brenner, 1964). In this study, based on the data, both the item difficulties and the item discriminations were calculated to determine the level of achievement of the objective. In this study, Python was employed for item development and analysis, particularly in the item generation process, due to its robust computational capabilities and extensive library support. Python's wide array of libraries, such as NumPy for numerical computations, pandas for data manipulation, and SciPy for scientific computing, enables efficient handling and analysis of complex datasets. This facilitates the estimation of item characteristics like difficulty and discrimination indices with precision, adhering to the principles of Classical Test Theory (CTT). Moreover, Python's versatility and ease of integration with statistical packages allow for the implementation of algorithms that can automate the generation of test items while ensuring they meet specified criteria for difficulty and discrimination. This automation not only streamlines the item generation process but also enhances the reliability and validity of the items by applying consistent criteria across all items. The use of Python thus significantly contributes to the methodological rigor of the research, enabling the creation of well-calibrated assessment tools that accurately measure the constructs of interest.

Research Ethics

The ethics committee approval for research was obtained after the decision given by the Gazi University Ethics Committee on 23.05.2023 with the document number E-77082166-604.01.02-665657.

FINDINGS

In this research, the item generation process involved the creation of both verbal and visual texts, along with a manually crafted diagram that merged these elements, serving as a secondary visual text. This integrative approach led to the development of a cognitive model that guided the formulation of the item model and the subsequent algorithmic generation of items. The study further expanded to include the generation of answer choices, with a specific emphasis on the use of contextually relevant idioms to assess advanced reading comprehension abilities. The key features were identified for the creation of effective distractors. In this regard, firstly the generate process is mentioned, and then the evaluation results based on the field test results are included.

Generation Process

The initial step in the process was the selection of the subject matter, which in this case, was the bodily maps of emotions. This thematic focus provided a foundation for the cognitive processes to be explored and represented in the item generation. Following the selection of the subject matter, the study outlined the cognitive processes relevant to visual reading comprehension. Interpreting and integrating ideas and information: Synthesizing various pieces of information to form a coherent understanding. The aim was to develop parallel forms of items, known as isomorphic items, to ensure consistency in both topic and cognitive load across the generated items. Upon establishing the cognitive model, the next phase entailed deciding on the scenario to be covered by the items and identifying specific source of information, features and elements. The generation process is elaborated in three stages: verbal text, visual text, and options generation.

Verbal text generation

As is seen in the cognitive model, the verbal text consists of an introduction, development, and conclusion. The introduction includes information about the purpose of the study and the participants. The development section contains information about the methodology and the process of the study. The conclusion section contains the results of the study. To ensure that the generated items had similar item difficulty, the elements within each part (introduction, development, conclusion) were constrained. These determined elements were placed in the item model. By leveraging the capabilities of language, texts with coherent and fluent introduction, development, and conclusion sections can be generated, each containing information pertinent to the research topic's objective, participants, methodology, and outcomes. As texts are generated in alignment with the cognitive model and item model, the information and flow within the texts exhibit similarities. Here, the same information is presented in different contexts by utilizing the possibilities of language, thereby facilitating the achievement of similar item difficulty. Consequently, the developed items can be used in individually tailored booklets or administered to different students in simultaneously conducted exams. The generated verbal text is displayed in the table, and an examination of the table reveals that both generated texts share similar content. The first production contains 138 words, while the second has 144 words.

The sample verbal text generated is shown in Table 2.

Table 2. Sample Generated Verbal Texts of the Stem

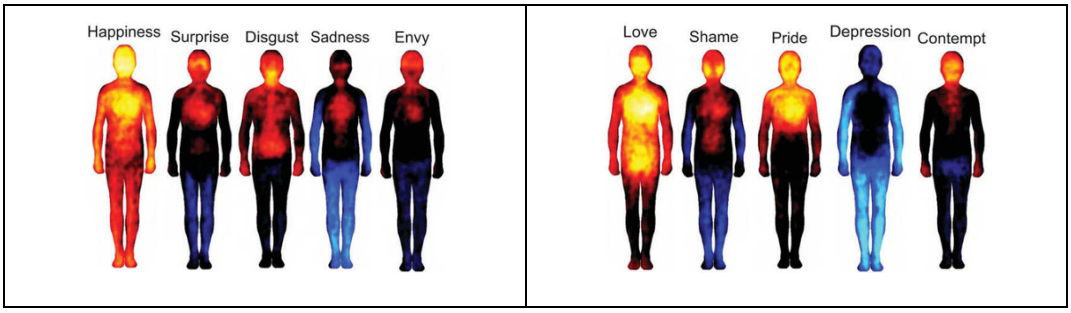
Sample 1	Sample 2
A comprehensive investigation titled "Bodily Maps of Emotions" encompassed a wide participant pool exceeding 700 individuals aged 18-45, hailing from Finland, Sweden, and Taiwan. The primary objective of this study was to delve into the impact of emotionally charged stimuli on these individuals. Diverse contents, such as words, videos, and stories, were presented to the participants, who were then tasked with identifying the specific regions of their bodies that experienced activation and those that remained unresponsive to the stimuli. To augment these findings, advanced computer technology (detectors) was utilized to monitor the participants' bodily reactions, which were subsequently correlated with their self-reported responses. The study yielded intriguing results, revealing striking similarities in the body sensory maps across participants, despite variations in their cultural backgrounds. A segment of the body sensation map is visually presented in the figure below:	In the research entitled "Bodily Maps of Emotions," a vast and diverse cohort of more than 700 individuals between the ages of 18 and 45, originating from Finland, Sweden, and Taiwan, actively participated. This pioneering study sought to uncover the profound impact of emotionally charged materials on the human body. Participants were exposed to a rich array of stimuli, including words, videos, facial expressions, and stories, and were then requested to pinpoint the specific areas of their bodies that exhibited heightened responsiveness, as well as those that remained indifferent. The participants' bodily reactions were meticulously recorded using computer technology (detectors), and these measurements were juxtaposed with the participants' accounts. Intriguingly, despite the diversity in cultural backgrounds, the study remarkably unveiled remarkable similarities in the body sensory maps among the participants. A visual segment of the body sensation map is depicted in the figure below:

Visual text generation

The results of the study are presented with a heat map including body coloring for 13 different emotions. For the visual text, these emotions are grouped in three ways in the context of the heatmap: (i) Emotions that increased sensation: Here, emotions such as love, and happiness have colours that indicate an increase in body temperature. This temperature increase is observed throughout the body in emotions such as love, while in emotions such as anger, it occurs in a certain part of the body. The body parts activated by the emotion are also categorized. (ii) Emotions that decreased sensation: Emotions such as sadness and depression have been found to

decrease the sensation in the body in general. Similarly, in emotions such as sadness, there is a decrease in sensation in the whole body in general, while in emotions such as envy, it is limited in certain areas. Emotions with reduced sensation by body region were also grouped. (iii) Mixed sensation: In emotions such as surprise, anxiety, etc., sensation increases in certain parts of the body and decreases in others. Similarly, these emotions are grouped within themselves to ensure that the images are of similar difficulty. The 13 emotions found at the end of the research were placed in the item model with Python codes in line with the features in the cognitive model (increasing, decreasing and mixed). The sample visual text generated is shown in Table 3.

Table 3. Sample Generated Visual Texts of the Stem



Option Generation

After the text was created, possible correct answers to be associated with the cognitive model were defined. At this point, idioms or commonly used words expressing emotion were chosen so that students could relate what they read to daily life. In this selection process, body parts where body sensation increases were also selected as a source of information. After that, commonly used emotion-body part idioms in Turkish were listed and categorized. In the research, 13 correct answers were prepared for each emotion. The sample options are shown in Table 4.

Table 4. Sample Generated Options

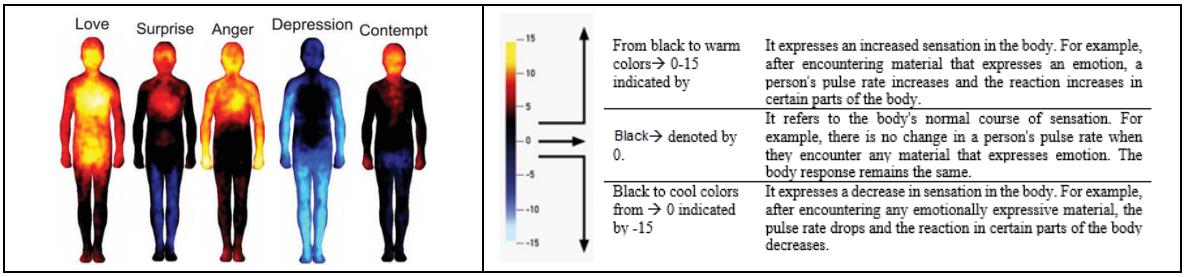
Correct options	I'm head over heels in **love** with you; every time I see you, I get butterflies in my **stomach**.
	When I received the good news, I was over the moon with joy and **happiness** from **head to toe**.
	I was so **saddened** that it felt like **my feet** had become as heavy as lead, and I could barely walk through the mud.
Distracters	I was so **surprised** by my brother's arrival that my **hands** started shaking like a leaf.
	Fear made my **knees** turn to jelly, and I couldn't figure out how to move to approach you.
	As winter approaches, she mentioned feeling blue and experiencing **depression**, and her **heart** races.

Generated items

After the generation of verbal, visual texts, and options, they were compiled in Python to give the items their final form. Although the produced items bear content similarity to the parent item, both the correct answers and visual information differ. This aims to significantly alleviate security concerns for students taking the exam simultaneously. It's important to remember that the goal here is to produce similar items. When the research topic is changed, the standardization facilitated by the cognitive and item models enables the creation of unique texts. For example, if the research topic were "Countries' Perspectives on the Use of Artificial Intelligence," it would be feasible to similarly produce new and unique texts detailing the research's objectives, participants, methodology, and outcomes. The sample item is shown in Table 5.

Table 5. Sample of the Generated Items

The study titled "Bodily Maps of Emotions" involved more than 700 participants aged 18-45 from Finland, Sweden, and Taiwan. The research aimed to explore emotionally charged materials' impact on individuals. Participants were presented with various contents, including words, videos, facial expressions, and stories, and asked to indicate the activated and unresponsive parts of their bodies in response to these stimuli. Additionally, the participants' bodily reactions were measured using computer technology (detectors) and combined with their self-reported answers. The study found that despite cultural differences, the participants exhibited similar body sensory maps. A segment of the body sensation map is presented in the figure below:



Which of the sentences below aligns with the findings of the "Bodily Maps of Emotions"?

A) I was so surprised to see my brother arrive that my hands started shaking like a leaf.

B) She would say her stomach twisted in knots whenever she felt green with envy.

C) I'm so in love with you that every time I see you, I get butterflies in my stomach.

D) She mentioned feeling blue and experiencing depression, and her heart races. *

Results of the field test item

In this study, the aim was to generate items of similar item difficulty for use in the same exam. Accordingly, out of 325 generated items, 5 items were selected for a field trial. These items were placed as the 19th item in a 5 different booklet of a screening test. The test, administered on a computer-based platform, allowed students to take it within a given 2-day period. Following the administration, the psychometric properties of the items were calculated. The item difficulty, item discrimination, and point biserial of the field test items which were generated are shown in Table 6. As can be seen in the Table 6, different emotions constitute the correct options for the five visual reading comprehension items that were generated and pre-applied in the study. In addition, the emotion-body feeling zone also differs. It is seen that the item difficulty indices of the pre-applied items vary between 0.58 and 0.66. In other words, the difficulty of the items that were automatically generated from the same cognitive and item model and had different answer keys were quite close to each other. The discrimination indices of the items were between 0.48 and 0.69. Item-total correlations also vary between 0.40 and 0.57.

Table 6. Item Statistics of the Generated Items

Items	Correct answer (in the idiom)		Item difficulty	Item discrimination	Biserial correlation
	Emotion	Part of the body			
1	Love	Abdomen	0.61	0.69	0.57
2	Pride	Breast	0.58	0.68	0.54
3	Surprise	Knee	0.59	0.57	0.51
4	Anxiety	Heart	0.66	0.48	0.40
5	Anger	Face	0.60	0.57	0.56

DISCUSSION AND CONCLUSION

This research aimed to automatically generate visual reading comprehension items used in the high school transition system in Türkiye and to evaluate the results. In the production with the template-based AIG, a cognitive model and item model were first developed for verbal and visual text, and then the items were automatically generated with the Python codes.

Five items selected among the generated items were field tests. As a result of the field test, it was determined that the difficulty indices of the items ranged between 0.58 and 0.66. In other words, a total of 1168 students answered 58-66% of the visual reading comprehension items correctly. This result shows that the item was slightly easier than the average for the students. The findings of this study align with previous research. It was determined that the average difficulty of the language test for high-school entrance including visual reading comprehension items generated in this study varied between 0.46 and 0.62 in the last five years (Author, 2023). Similarly, Freedle and Kostin (1991) found that the item difficulties of the main idea questions on the SAT ranged from 46% to 59%. In reading comprehension questions, there is research showing that items containing visual text are more difficult than those containing only verbal text (Santi et al., 2015), as well as studies indicating that visual content increases the proficiency level of the item but does not affect its difficulty (Caldwell & Pate, 2013). For instance, Khasawneh and Al-Rub (2020) stated that the difficulties of the visual reading comprehension items they developed to increase the reading skills of students with learning difficulties were still moderate or easy. In summary, the field test results demonstrate that the selected visual reading comprehension items were slightly easier than average for the students. This aligns with previous research highlighting the variation in item difficulty and the impact of visual content on reading comprehension tasks. In addition, the difficulty indices of the items were similar to each other in this study, which shows the usability of AIG for parallel test construction. Parallel questions are made so that each examinee gets a different question but has the same level of difficulty (Adji et al., 2018). The created parallel tests are useful for accurately assessing students' reading comprehension skills, especially in digital tests. Adji et al. (2018) found that with the AIG system, he was able to generate parallel math questions for 55% of the problems with the help of a flexible math editor. Similarly, Fu et al. (2022) generated isomorphic questions from the same cognitive model using AIG. It is also known that items of different difficulty can be created using AIG depending on the purpose. Gierl et al. (2016) generated items with item difficulties ranging from 19-95% from the same item model from a model they developed in the field of medical education. Similarly, Lai, Gierl, Touchie, et al. (2016) estimated the difficulty index of the 13 generated items between 0.13 and 0.91. These results show that AIG can generate items of different difficulty or similar difficulty depending on the purpose.

The discrimination indices of the visual reading comprehension items ranging from 0.48 to 0.69 showed a high level of discrimination. This shows that the items effectively discriminated against students with different levels of reading comprehension skills (Brennan, 1972). Furthermore, item-total correlations ranging from 0.40 to 0.57 provided additional evidence for the validity of these visual reading comprehension items. The research shows that in general, AIG-generated items are perceived to be of comparable or higher quality than manually written items (Falcão et al., 2023; Gierl et al., 2016; Gorin & Embretson, 2012; Harrison et al., 2017).

In conclusion, the automatic generation and evaluation of visual reading comprehension items revealed promising results in terms of difficulty and discrimination indices. The findings of this study highlight the potential of an automated generation approach using template-based AIG in the development of visual reading comprehension items. The high discrimination indices also demonstrate the effectiveness of the generated items in discriminating between students with different levels of reading comprehension skills. In addition, it is important to note that the success of the AIG approach is highly dependent on the quality and accuracy of the cognitive and item models used in the AIG process. Therefore, it is recommended that these models are continuously refined and improved to ensure the generation of high-quality items using verbal and visual text that are aligned with the desired assessment objectives.

Implications

This research demonstrates the application of AIG in efficiently generating a large pool of visual reading comprehension items. The study provides evidence of the effectiveness of AIG-generated visual reading comprehension items by examining their psychometric properties, including item difficulty and discrimination. The generated items show comparable difficulty levels, making them suitable for constructing personalized assessments for students with different reading comprehension skills. The use of visuals in reading comprehension assessments has the potential to enhance students' motivation and engagement during exams. This also shows that the items generated in using AIG can be asked to the students at different times in computer-based tests, thus avoiding the security problem.

Limitations

This study, while shedding light on the promising outcomes of employing AIG for the efficient creation of visual reading comprehension items, acknowledges certain limitations. The effectiveness of AIG is contingent on the precision and quality of the cognitive and item models, introducing a potential vulnerability should any

shortcomings exist in these models. Furthermore, the study's focus on Türkiye's high school transition system prompts a consideration of the generalizability of these findings to diverse educational systems and cultural contexts. Addressing these limitations calls for the need for further research to refine and enhance the applicability of the AIG approach in diverse educational settings.

Statements of Publication Ethics

The author has provided a clear and concise statement of their ethical practices, and their manuscript complies with the journal guidelines.

Researchers' Contribution Rate

Authors	Literature review	Method	Data Collection	Data Analysis	Results	Conclusion
Author 1	☒	☒	☒	☒	☒	☒

Conflict of Interest

The authors declare that there is no conflict of interest.

Acknowledgment

I would like to extend my gratitude to Cem Mutlu Türkseven for his assistance in the data collection process.

REFERENCES

Adedoyin, O. O., Nenty, H., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Reviews*, *3*(3), 83.

Adji, T. B., Pribadi, F. S., Prabowo, H. E., Rosnawati, R., & Wijaya, A. (2018). Generating Parallel Mathematic Items Using Automatic Item Generation. *ICEAP 2019*, *1*(1), 89-93.

Aiken, L. R. (1979). Relationships between the item difficulty and discrimination indexes. *Educational and Psychological Measurement*, *39*(4), 821-824.

Al-Hameed, F. and Al-Shuair, M. (2019). The effectiveness of using digital stories (on internet) to improve the literal, organizational and inferential reading comprehension skills of English as a second language. J*ournal of Research in Curriculum Instruction and Educational Technology, 5*(3), 45-81.

Aloqaili, A. S. (2012). The relationship between reading comprehension and critical thinking: A theoretical study. *Journal of King Saud University-Languages and Translation*, *24*(1), 35-41.

Amin, M. (2019). Developing reading skills through effective reading approaches. *International Journal of Social Science and Humanities*, *4*(1), 35-40.

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and individual differences*, *22*(1), 112-117.

Barnes, M. A. (2015). What do models of Reading comprehension and its development have to contribute to a science of comprehension instruction and assessment for adolescents? *Improving reading comprehension of middle and high school students*, 1-18.

Boonen, A. J., de Koning, B. B., Jolles, J., & Van der Schoot, M. (2016). Word problem solving in contemporary math education: A plea for reading comprehension skills training. *Frontiers in psychology*, *7*, 191.

Brennan, R. L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, *32*(2), 289-303.

Brenner, M. H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, *48*(2), 98.

Brevik, L. M. (2019). Explicit reading strategy instruction or daily use of strategies? Studying the teaching of reading comprehension through naturalistic classroom observation in English L2. *Reading and writing*, *32*(9), 2281-2310.

Cahalan, C., Mandinach, E. B., & Camara, W. J. (2002). Predictive Validity of SAT® I: Reasoning Test for Test-Takers with Learning Disabilities and Extended Time Accommodations. Research Report No. 2002-5. ETS RR-02-11. *College Entrance Examination Board*.

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American journal of pharmaceutical education*, *77*(4).

Chandran, Y., & Shah, P. M. (2019). Identifying learners' difficulties in ESL reading comprehension. *Creative Education*, *10*(13), 3372-3384.

Cohen, A. D., & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. *ETS Research Report Series*, *2006*(1), i-162.

Cornoldi, C., & Oakhill, J. V. (2013). *Reading comprehension difficulties: Processes and intervention*. Routledge.

Daly, A., & Unsworth, L. (2011). Analysis and comprehension of multimodal texts. *Australian Journal of Language and Literacy, The*, *34*(1), 61-80.

Duncan, L., McGeown, S., Griffiths, Y., Stothard, S., & Dobai, A. (2015). Adolescent reading skill and engagement with digital and traditional literacies as predictors of reading comprehension. *British Journal of Psychology, 107*(2), 209-238.

Eason, S., Goldberg, L., Young, K., Geist, M., & Cutting, L. (2012). Reader–text interactions: how differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515-528.

Embretson, S., & Yang, X. (2006). 23 Automatic item generation and cognitive psychology. *Handbook of statistics*, *26*, 747-768.

Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement*, *55*(1), 112-131.

Falcão, F., Costa, P., & Pêgo, J. M. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, *27*(2), 405-425.

Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., & Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Advances in Health Sciences Education*, 1-25.

Fielding, L. G., & Pearson, P. D. (1994). Reading Comprehension: What Works. *Educational leadership*, *51*(5), 62-68.

Freedle, R., & Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages. *ETS Research Report Series*, *1991*(1), i-52.

Fu, Y., Choe, E. M., Lim, H., & Choi, J. (2022). An Evaluation of Automatic Item Generation: A Case Study of Weak Theory Approach. *Educational Measurement: Issues and Practice*.

Gierl, M., Lai, H., & Zhang, X. (2019). Automatic item generation. In *Advanced Methodologies and Technologies in Modern Education Delivery* (pp. 192-203). IGI Global.

Gierl, M. J., & Haladyna, T. M. (2012a). Automatic item generation: An introduction. In *Automatic item generation* (pp. 13-22). Routledge.

Gierl, M. J., & Haladyna, T. M. (2012b). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International journal of testing*, *12*(3), 273-298.

Gierl, M. J., & Lai, H. (2018). Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing. *Applied Psychological Measurement*, *42*(1), 42-57. https://doi.org/10.1177/0146621617726788

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A.-P., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education*, *29*(3), 196-210.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, *46*(8), 757-765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Glenberg, A. M., & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of memory and language*, *31*(2), 129-151.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*(4), 351-373.

Gorin, J. S., & Embretson, S. E. (2012). Using cognitive psychology to generate items and predict item characteristics. In *Automatic item generation* (pp. 146-166). Routledge.

Harrison, P., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, *7*(1), 1-18.

Hill, C. A. (2003). Reading the visual in college writing classes. In *Intertexts* (pp. 134-159). Routledge.

Holling, H., Bertling, J. P., & Zeuch, N. (2009). Automatic item generation of probability word problems. *Studies in Educational Evaluation*, *35*(2-3), 71-76.

Hosseini, E., Khodaei, F. B., Sarfallah, S., & Dolatabadi, H. R. (2012). Exploring the relationship between critical thinking, reading comprehension and reading strategies of English university students. *World Applied Sciences Journal*, *17*(10), 1356-1364.

Hoyt, L. (1992). Many ways of knowing: Using drama, oral interactions, and the visual arts to enhance reading comprehension. *The Reading Teacher*, *45*(8), 580-584.

Irvine, S. H., & Kyllonen, P. C. (2013). *Item generation for test development*. Routledge.

Khasawneh, M. A. S., & Al-Rub, M. O. A. (2020). Development of reading comprehension skills among the students of learning disabilities. *Universal Journal of Educational Research*, *8*(11), 5335-5341.

Kinniburgh, L. H., & Shaw, E. L. (2009). Using question-answer relationships to build: Reading comprehension in science. *Science Activities*, *45*(4), 19-28.

Klingner, J. K., Vaughn, S., & Boardman, A. (2015). *Teaching reading comprehension to students with learning difficulties*. Guilford Publications.

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice*, *38*(1), 48-53.

Lai, H., Gierl, M. J., Byrne, B. E., Spielman, A. I., & Waldschmidt, D. M. (2016). Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of Dental Education*, *80*(3), 339-347.

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A.-P., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and learning in medicine*, *28*(2), 166-173.

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2017). Unpicking the developmental relationship between oral language skills and reading comprehension: it's simple, but complex. *Child Development, 89*(5), 1821-1838.

Li, H., Wang, P., Shen, C., & Hengel, A. v. d. (2019). Visual question answering as reading comprehension. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

MEB, (2018). Milli Eğitim Bakanlığı Ortaöğretime Geçiş Yönergesi. https://www.meb.gov.tr/meb_iys_dosyalar/2018_03/26191912_yonerge.pdf

Mullis, I., Martin, M., Foy, P., & Hooper, M. (2017). PIRLS 2016 International Results in Reading. Boston College, TIMSS & PIRLS International Study Center. In.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). *PIRLS 2021 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/

Nummenmaa, L., Glerean, E., Hari, R., & Hietanen, J. K. (2014). Bodily maps of emotions. *Proceedings of the National Academy of Sciences*, *111*(2), 646-651.

O'Neil, K. E. (2011). Reading pictures: Developing visual literacy for greater comprehension. *The Reading Teacher*, *65*(3), 214-223.

OECD (2019), "PISA 2018 Reading Framework", in *PISA 2018 Assessment and Analytical Framework, OECD Publishing, Paris, https://doi.org/10.1787/5c07e4f1-en.*

OECD. (2021). *21st-Century Readers*. https://doi.org/doi:https://doi.org/10.1787/a83d84cb-en

Oliver, K. (2009). An investigation of concept mapping to improve the reading comprehension of science texts. *Journal of Science Education and Technology*, *18*, 402-414.

Österholm, M. (2006). Characterizing reading comprehension of mathematical texts. *Educational studies in mathematics*, *63*, 325-346.

Sabatini, J. P., O'reilly, T., Halderman, L., & Bruce, K. (2014). Broadening the scope of reading comprehension using scenario-based assessments: Preliminary findings and challenges. *LAnnee psychologique*, *114*(4), 693-723.

Santi, K. L., Kulesz, P. A., Khalaf, S., & Francis, D. J. (2015). Developmental changes in reading do not alter the development of visual processing skills: an application of explanatory item response models in grades K-2. *Frontiers in psychology*, *6*, 116.

Setiawan, H., Hidayah, I., & Kusumawardani, S. S. (2022). Automatic Item Generation with Reading Passages: A Systematic Literature Review. 2022 8th International Conference on Education and Technology (ICET),

Shin, E. (2021). *Automated Item Generation by Combining the Non-template and Template-based Approaches to Generate Reading Inference Test Items* University of Alberta]. Canada.

Shin, J., & Gierl, M. J. (2022). Generating reading comprehension items using automated processes. *International journal of testing*, *22*(3-4), 289-311.

Silva, M. and Cain, K. (2015). The relations between lower and higher-level comprehension skills and their role in prediction of early reading comprehension. *Journal of Educational Psychology, 107*(2), 321-331.

Unsworth, L. (2014). Multimodal reading comprehension: Curriculum expectations and large-scale literacy testing practices. *Pedagogies: An international journal*, *9*(1), 26-44.

Westwood, P. (2016). *Teaching and Learning Difficulties 2nd ed* (Vol. 2). Acer Press.

Woolley, G., & Woolley, G. (2011). *Reading comprehension*. Springer.

Wyer Jr, R. S., Hung, I. W., & Jiang, Y. (2008). Visual and verbal processing strategies in comprehension and judgment. *Journal of Consumer Psychology*, *18*(4), 244-257.