International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 131–147

https://doi.org/10.21449/ijate.1359348

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Automatic item generation for non-verbal reasoning items

Ayfer Sayin[1,*], Sabiha Bozdag[1], Mark J. Gierl[2]

[1]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]University of Alberta, Faculty of Education, Department of Educational Psychology, Alberta, Canada

ARTICLE HISTORY

Received: Sep. 13, 2023
Revised: Oct. 29, 2023
Accepted: Oct. 31, 2023

Keywords:
Automatic item generation,
Non-verbal items,
Visual reasoning test,
BİLSEM.

Abstract: The purpose of this study is to generate non-verbal items for a visual reasoning test using templated-based automatic item generation (AIG). The fundamental research method involved following the three stages of template-based AIG. An item from the 2016 4th-grade entrance exam of the Science and Art Center (known as BİLSEM) was chosen as the parent item. A cognitive model and an item model were developed for non-verbal reasoning. Then, the items were generated using computer algorithms. For the first item model, 112 items were generated, and for the second item model, 1728 items were produced. The items were evaluated based on subject matter experts (SMEs). The SMEs indicated that the items met the criteria of one right answer, single content and behavior, not trivial content, and homogeneous choices. Additionally, SMEs' opinions determined that the items have varying item difficulty. The results obtained demonstrate the feasibility of AIG for creating an extensive item repository consisting of non-verbal visual reasoning items.

1. INTRODUCTION

Computer-based testing (CBT) presents several advantages, including paperless administration, flexible scheduling, and a diverse range of item types. However, CBT encounters challenges in developing continuous, content-specific items, relying on traditional item development approaches that involve experts in writing, editing, and reviewing items. To address this limitation, automatic item generation (AIG) streamlines the process through a structured workflow, ensuring a consistent supply of new, high-quality items for CBT.

The inception of AIG traces back to Bormuth's 1970 concept, which aimed to generate test items representing the intended learning outcomes (Gierl & Haladyna, 2012, p. 14). Items crafted by experts are often deemed subjective, as they reflect the experiences and personal skills of these experts. In response, Bormuth proposed automating the item writing process to eliminate subjectivity. He posited that two test developers employing the same content and item features should be capable of producing similar high-quality items (Gierl & Haladyna, 2012). AIG integrates this perspective with computer technology, marking a pioneering research field that amalgamates cognitive and psychological theories within a digital framework to generate assessment tasks (Gierl et al., 2015; Ryoo et al., 2022). The overarching goal of AIG is to

*CONTACT: Ayfer Sayın ✉ ayfersayin@gazi.edu.tr Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

standardize test item design significantly. By removing subjectivity from the assessment process, AIG strives to manage assessments scientifically and efficiently (Gierl et al., 2015; Leighton, 2012).

1.1. Automatic Item Generation (AIG)

AIG can be defined as a method of item generation that combines content expertise and computer technology through models, enabling the rapid creation of extensive and efficient item banks (Gierl et al., 2021). Another definition characterizes AIG as an approach to item development through augmented intelligence. Augmented intelligence is an artificial intelligence domain where computer systems model and replicate human cognitive abilities to enhance task performance (Gierl et al., 2021). The general operation of AIG necessitates the convergence of the cognitive processes shaped by subject matter experts' (SMEs) experiences and the processing power or intelligence of modern computational systems. If we conceive intelligence in the broadest sense as "problem-solving ability," AIG, with its ability to generate a vast item pool with a limited number of SMEs, demonstrates significant problem-solving capacity. AIG is based on two approaches: template-based and artificial intelligence-based (Shin, 2021). Non-verbal reasoning items were developed in the current study using template-based AIG. Template-based AIG involves a three-step standardized process. This process is explained as follows (Gierl & Lai, 2013):

- Cognitive model development: In the first step, SMEs define the content, which is referred to as the cognitive model. The cognitive model emphasizes the information, skills, and abilities required for problem solving by learners. It provides a concise depiction of subject-specific knowledge, interactions within the information, and simulates the problem-thinking/problem-solving process. It can be used not only as a template containing the relevant information, but also to provide appropriate feedback to students following exam administration.

- Item model development: In the second step, specialists decide which components of the item should be changed to establish a template for creating new items. Variables in the item model can be altered in areas such as the item's body, the question sentence, and the alternatives (right response and distractors). At this point, auxiliary elements such as photos, tables, graphs, and diagrams, as well as random variables that can be changed but are not required to answer the problem, can be introduced to the item model.

- Generating items using computer technology The content from the cognitive model is placed into the item model developed in the second phase using computer-based algorithms in the third step. In this step, computer algorithms generate objects based on the rules and limits established by SMEs. AIG has developed a variety of software, the majority of which is not open source.

Template-based AIG can be defined as generating extensive and efficient item pools by encoding content derived from the cognitive model into the item model using computer algorithms (Gierl et al., 2013). By following the three stages, AIG allows for the creation of heterogeneous item pools with similar or different item difficulties. In essence, AIG has two primary purposes: firstly, generating items with similar item difficulty with comparable psychometric properties, and secondly, constructing item pools with varying difficulty ranges (Sinharay & Johnson, 2005). This approach enables the production of items with the desired attributes and a scalable range of item difficulty.

To assess the effectiveness, performance, and suitability of AIG in response to evolving needs, it is meaningful to compare it with a conventional method, namely the traditional item writing process. From the past century to the present day, the item writing process has remained the most time-consuming and costly aspect of test development (Gierl & Haladyna, 2012). Particularly for significant tests like selection, placement, and certification, a continuous need for new items exists, leading to a demand for extensive item pools in psychometric and educational measurements (Embretson & Yang, 2007). The traditional item writing process entails multiple steps, including item creation, item revision, and empirical testing (Embretson & Kingston, 2018). For instance, when 1000 items are required for an exam, each item must be individually authored, formatted, and developed. The elimination of items with inadequate psychometric properties at this stage further escalates costs (Arendasy & Sommer, 2012). By way of contrast, the AIG process typically commences with a well-established anchor item, which provides a robust reference point for newly generated items (Embretson & Yang, 2007). This valid anchor item contributes to the economic feasibility of AIG by satisfying a high item demand from a small number of SMEs. In short, while the traditional item writing approach ensures the creation of high-quality items, its time-consuming and cost-intensive nature renders it insufficient for meeting the increasing item demand (Choi & Zhang, 2019). Kosh et al. (2019) also highlighted the significant cost-saving potential of AIG. Moreover, items written through the traditional item writing process are limited and updating or modifying them poses challenges (Gierl et al., 2021). In our contemporary era where knowledge constantly evolves and updates, test developers require more flexible approaches. In such a context, AIG allows for the updating of items in the pool by making appropriate changes and adjustments to the previously developed cognitive model. It can be observed that the traditional item creation method is limited due to its repetitive stages, the inability to predict the psychometric properties of items without testing, the difficulty in updating generated items, and the challenge of constructing large item pools. Especially for non-verbal items, the creation of drawings and graphics is often integrated into the item writing process. This current study exemplifies the first research on the AIG process in Türkiye, which entails the generation of non-verbal items that can be used to assess students' visual reasoning skills.

1.2. Non-Verbal Reasonings

The concept of reasoning has been regarded as an ability within the domain of thinking skills (Mercan, 2021). Building upon this notion, reasoning can be defined as a cognitive process wherein an individual identifies patterns and relationships in a given problem, formulates rules, and solves the problem (Horn & Catell, 1966; Kurtz, et al., 1999). According to Mullis et al., (2019), reasoning encompasses skills such as analysis, generalization, synthesis, verification, and solving non-routine problems. Reasoning skills are considered fundamental cognitive competencies utilized in the process of accessing justified information (Kocagül & Çoban, 2022), or abstract methods and approaches used to acquire information and draw conclusions (Lawson, 2004). Reasoning skills are classified into three dimensions: mathematical/numerical, auditory/verbal, and visual-spatial/non-verbal reasoning skills (Lohman & Hagen, 2003; Mercan, 2021). The focus of the current study is on non-verbal reasoning skills, which aim to assess individuals' cognitive abilities in reasoning, independently of their verbal and language aptitudes (Balboni et al., 2010; DeThorne & Schaefer, 2004). Well-known non-verbal intelligence tests include the Universal Non-verbal Intelligence Test (UNIT), Raven Progressive Matrices (RPM), and Naglieri Non-verbal Ability Test (NNAT) (DeThorne & Schaefer, 2004). Furthermore, non-verbal reasoning items are integrated into other widely used intelligence scales in Türkiye. For instance, the Standford Binet Intelligence Test 5, the CAS Cognitive Assessment System Non-verbal Matrices subtest, and the perceptual reasoning subtest of the Weschler Intelligence Scale for Children, all employed in Guidance and Research Centers in Türkiye, incorporate items evaluating non-verbal reasoning capabilities (Gibbons & Warne, 2019; Kemer & Çakan, 2020; Naglieri et al., 2004; Weiss et al., 2016). Bildiren (2021) brought the National Non-verbal Cognitive Ability Test, a collection of non-verbal reasoning items, into the national literature. Similarly, non-verbal reasoning items were extensively used in the Visual-Perceptual Flexibility and Visual-Analogical Reasoning subtests of the Anadolu-Sak Scale, developed in Türkiye (Sak et al., 2019; Tamul et al., 2020).

Science and Art Centers (known as BİLSEM) entrance exams are conducted annually to assess candidates and identify exceptionally talented students in Türkiye. Gifted individuals are defined as children who exhibit high levels of intelligence, motivation, creativity, leadership capacity, or exceptional performance in specific academic fields compared to their peers (Bilgiç et al., 2017; MoNE, 2022a). Students nominated by their teachers for BİLSEM undergo a preliminary evaluation through a talent test determined by the BİLSEM committee for that year, administered via tablet computers (BİLSEM Online, 2023a; MoNE, 2022a). However, one of the fundamental challenges of computer-based tests is the risk of item exposure after the exam. Candidates who excel in the preliminary evaluation are subsequently subjected to individual assessment (MoNE, 2022b). Yet, especially for students nominated in the general aptitude field, the number of SMEs capable of administering intelligence tests in RAMs is limited. Moreover, many of the intelligence tests used in Türkiye lack alternative forms. Some of these tests are also outdated, which undermines the reliability of intelligence tests (Kurnaz & Ekici, 2020). Each of these factors poses a risk of item exposure in the BİLSEM entrance exam. Familiarity with the items by students who have accessed them beforehand can create a testing effect known as the practice effect, potentially affecting the results (Hausknecht et al., 2007). To mitigate this, computer-based test applications can develop personalized tests using different items for each individual or utilize adaptive applications. However, all these processes necessitate a broad repository of psychometrically sound items (Gierl & Lai, 2015). Template-based AIG can be used to generate non-verbal reasoning items quickly, economically and with high quality.

1.3. Present Study

Templated-based AIG has begun to spread across psychology, education, and computer science disciplines in recent times (Lai et al., 2016). In the literature, it has been observed that template-based AIG has been applied intensively in fields such as medicine (Falcão et al., 2022; Gierl & Lai, 2012) and dentistry (Lai et al., 2016); it has also been found to generate automated items in diverse disciplines like mathematics (Adji et al., 2018; Embretson & Kingston, 2018) and literature (Sayın & Gierl, 2023). Notably, the studies have identified verbal expressions and numerical values within mathematical items. However, the utilization of AIG in non-verbal reasoning items is limited. Gierl et al. (2015) employed template-based AIG to create 1,340 visual reasoning items involving finding the middle position and possessing heterogeneous item difficulty for undergraduate students. Ryoo et al. (2022) developed a cognitive ability test called "MOCA" that is compatible with the Cattell-Horn-Carroll (CHC) model, encompassing two of CHC's ten ability domains (Gf and Gv). MOCA, a two-form test, was designed for 6th to 9th-grade students. In contrast to both studies, the current research selected a sample group of 4th-grade elementary students and generated reasoning rotation (mental rotation) items by modifying the item format to assess their visual reasoning skills. This is because this study focuses on the Turkish sample. Visual reasoning items are used in the entrance test to BİLSEM, a school for gifted students in Türkiye, which includes visual reasoning items. The age group for the entrance exam is determined each year by the BİLSEM commission. However, considering that screening tests and diagnostic procedures have predominantly been administered to students in the 1st to 4th grades of primary school (e.g., MoNE, 2015; 2021; 2022a), 4th-grade students were prioritized when designing non-verbal reasoning items with AIG. Additionally, cognitive models were developed to create other visual reasoning items (e.g., matching, sequencing), and item generation was implemented based on these models in previous studies (Gierl et al., 2015; Ryoo et al., 2022). Unlike other studies, this research employs a rotation problem and scenario to assess visual reasoning.

AIG was achieved by utilizing a BİLSEM entrance exam item from 2016 as the primary item. In other words, the purpose of the study is to generate non-verbal reasoning items using template-based AIG. Item writing during the assessment and evaluation process is the costliest and labour-intensive stage. Particularly in the context of visual reasoning, developing items that measure cognitive levels is a complex process requiring effort and attention. Generating items through template-based AIG will facilitate the rapid and cost-effective creation of an extensive item repository. The current study is important in terms of modeling a BİLSEM entrance exam item and serving as an example for widely used items. It also contributes to the literature and holds the distinction of creating an extensive item repository for non-verbal visual reasoning items, which is a first in Turkish literature.

2. METHOD

2.1. Research Design

This study was fundamental research, as it encompasses the automatic item generation of non-verbal visual reasoning items and their evaluation by SMEs' opinions. Fundamental research refers to investigations conducted to scrutinize, examine, reinforce, or establish a theory about a specific field (Karasar, 2022). The current study was conducted with the approval of the Gazi University Ethics Committee under the reference number E-77082166-604.01.02-686103, dated 22.06.2023.

2.2. Participants

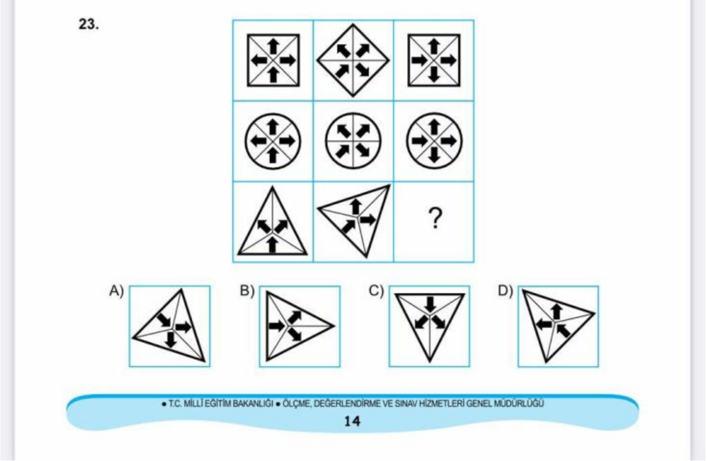
There were six participants who had previously examined BILSEM items and dealt with non-verbal items. Among the SMEs, four were female and two were male: two SMEs specializing in assessment and evaluation, one in classroom teaching, two in gifted education, and one in psychological counselling and guidance. The engagement of SMEs in the assessment and evaluation field was taken due to the test's nature and focus. In the Turkish education system, student participation in the entrance examination for a gifted education school necessitates nomination by a classroom teacher. Therefore, input from a classroom teacher was included. Given the inherent character of the test as an aptitude assessment, insights were also garnered from SMEs in the domain of gifted education. In consideration of the administration falling within the jurisdiction of psychological counsellors, the perspective of a psychological counsellor was incorporated. The SMEs, apart from classroom teachers, hold positions as university faculty members. Their professional experience varied, ranging from 5 to 17 years collectively, while their specific experience within the test development related spans 1 to 12 years.

2.3. Process

As part of the research, items were generated using AIG's three-step process. AIG generally starts with a parent item. In our study, a parent item was selected from the entrance exam for 4th-grade BİLSEM 2016 (Figure 1). BİLSEM items and data are not openly accessible. Therefore, this study concentrated on a sample exam item released by BİLSEM. While the validity evidence for the parent item could not be provided, its selection by experts in the BİLSEM commission and inclusion in the test is deemed a significant reference source.

In accordance with the parent item, the first step of AIG is the development of a cognitive model. A cognitive model represents the knowledge, skills, and abilities required to solve a specific problem within a domain. It comprehensively encompasses all the information, skills, and processes underlying test performance (Gierl & Lai, 2013). The second step of AIG focused on the development of an item model. Item models are templates that define where content needs to be placed (Gierl & Lai, 2013). The concept of an item model at AIG involves restructuring the guidelines and standards in traditional item writing using computer coding (Ryoo et al., 2022).

Figure 1. *Parent item for AIG.*


23.
?
A)
B)
C)
D)
• T.C. MİLLİ EĞİTİM BAKANLIĞI • ÖLÇME, DEĞERLENDİRME VE SINAV HİZMETLERİ GENEL MÜDÜRLÜĞÜ
14

Within the current study, two item models were developed, and items were generated following these templates. The first and second steps of AIG were developed by SMEs’ opinions. In the third step, computer algorithms are employed to place the content from the cognitive model into the item model, adhering to the elements and constraints defined in the cognitive model (Gierl & Lai, 2013). The prominent aspect of this process was the utilization of technology, specifically computer technology, for AIG. In our study, non-verbal visual reasoning items were generated through the utilization of the Python programming language. The codes, written in the PyCharm interface, were employed to accomplish the AIG for both developed models within the study. When the items were generated in Python, the prompt asked for the correct answers to be mixed among the options. For this reason, the correct answers were added to the bottom of each generated item and printed to an Excel file.

2.4. Data Collection Tool

The validity of the generated items was evaluated through SMEs' opinions. To facilitate this, an SME opinion form was created. A total of 20 items, 10 from each model, were presented to the SMEs for their assessment. The item-writing guidelines proposed by Haladyna et al. (2002) were utilized for the thorough examination of items by SMEs. Given the utilization of non-verbal reasoning items in our research, some criteria from the guidelines, such as 'Minimize reading, Simple vocabulary' were not used. Instead, four specific criteria were established to facilitate the comprehensive evaluation of the items: 'One right answer (Scientific Accuracy), Single content and behavior (Grade-Level Suitability Important), Not trivial content (Alignment with Purpose), homogeneous choices (Equitable challenge among distractors).' Experts assessed each item within the context of these four criteria, thus enabling the acquisition of broader and more detailed insights from the SMEs regarding the items. SMEs were requested to assess each item according to these criteria, using the following scale: 1-Accept, 2- Minor Revision, 3- Major Revision, 4- Reject. Additionally, SMEs' predictions about the difficulty of each item were obtained on a 1-5 scale, ranging from 1 as very easy to 5 as very hard.

2.5. Analysis

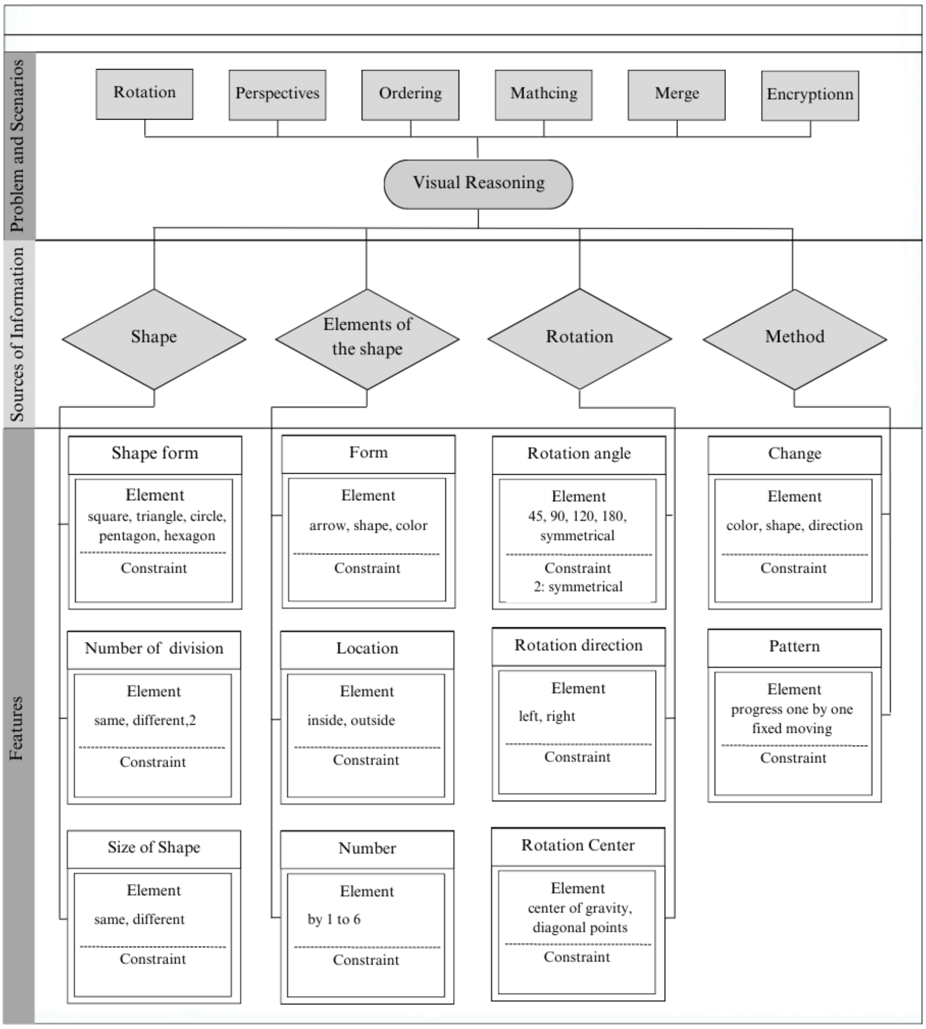
From the generated items, a random selection of 10 items was made for each model. SMEs’ opinions were then collected for a total of 20 items. Frequency and percentage were calculated for the SMEs’ opinions of the items.

3. RESULTS

3.1. Cognitive Model Development

The first step involved the examination of non-verbal visual reasoning items both at national and international exams, primarily focusing on BİLSEM entrance exams. The fundamental characteristics (problem and scenarios) underlying non-verbal visual reasoning items were determined as rotation, perspectives, ordering, matching, merging, and encryption. The sources of information for measuring these problems and scenarios were identified. Accordingly, the creation of distinct shapes, the incorporation of elements within or outside these shapes, and the formation of patterns through rotation and/or other methods were initially deemed essential. Once each source of information was determined, features and elements were selected. For the shape, various shapes such as square, triangle, circle, pentagon, and hexagon, among others could be chosen (as elements). These shapes could be divided into different numbers of parts, equal parts, or a fixed number like 2 to accommodate the placement of internal elements. The shapes might vary in size based on the pattern or remain consistent. Similarly, features and elements were determined for other sources of information in a manner analogous to the shape source. Afterwards, constraints were defined after the identification of elements. For instance, a triangle should be divided into 2 or 3 equal parts, while a hexagon could be divided into 6 equal parts. Nevertheless, no constraints were imposed on internal element shapes. For example, an arrow could be used in all problems and scenarios as a shape and could be incorporated within all shapes. Following these definitions, the cognitive model was developed and presented in Figure 2.

Figure 2. *A cognitive model developed for non-verbal visual reasoning items.*



In the present research, the generated items were based on the "rotation" of problems and scenarios. In this context, the developed cognitive model was structured within the framework of the Montreal Cognitive Assessment (MoCA) cognitive theory. MoCA measures visual reasoning by exploring the ability to use simulated mental images and employing the skill of rotation. In other words, it assesses students' visual reasoning skills by asking them to simulate how the movement of one shape affects another or how shapes rotate at different angles (Ryoo et al., 2022). In the current research, square, triangle, circle, and hexagon shapes were selected from the cognitive model. The square and circle were divided into four equal parts, a triangle into three equal parts, and a hexagon into six equal parts. The sizes of the shapes were constrained to the same size. Two inside elements (plus sign and square) were chosen and for these symbols, four different colors were selected: transparent, blue, green, and red. Five different angles were defined for rotation: 45, 60, 90, 120, and 180 degrees and constraints were defined for the angles according to the rhythmic logic of the shapes. For instance, the triangle shape was constrained to rotations of 60, 90, 120, and 180 degrees, while the square was constrained to rotations of 45, 90, and 180 degrees. In the first item model, rotations were carried out to the right, while in the second item model, rotations were executed to the left. All rotations were performed from the center of gravity. Elements are shown in Table 1.

3.2. Item Model Development

The parent item had a grid consisting of 3 columns * 3 rows. To showcase various item models within the study, two different item models were developed (Table 1). The question prompt was consistent for all items and was stated as "Mark the shape that should be in the place indicated by the question mark".

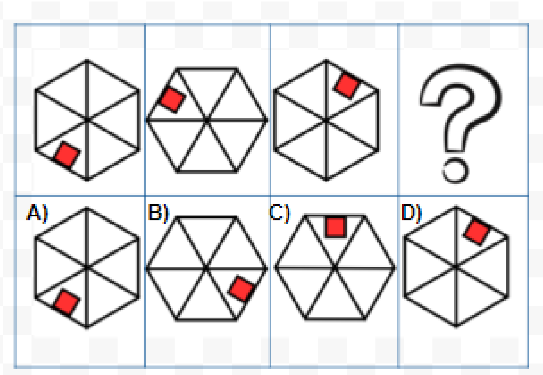
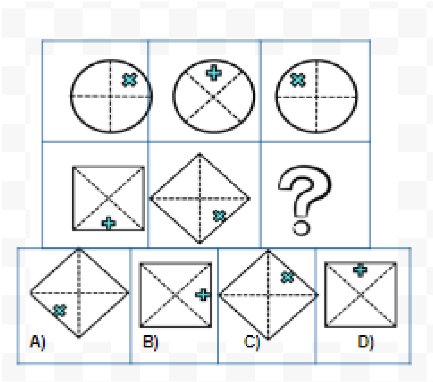
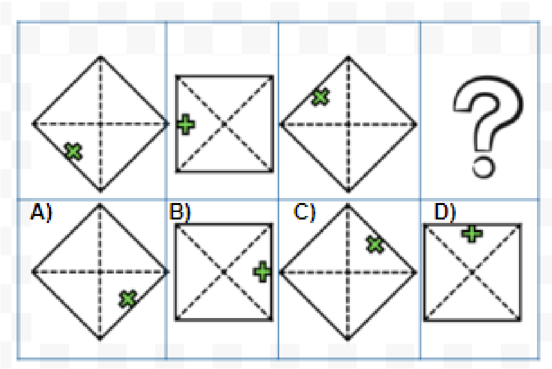
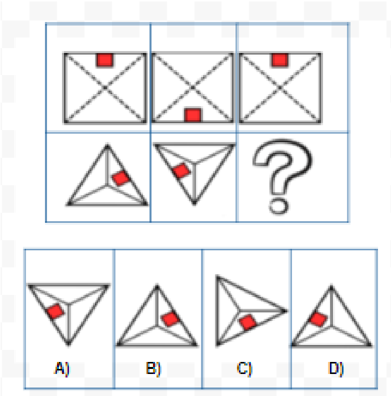
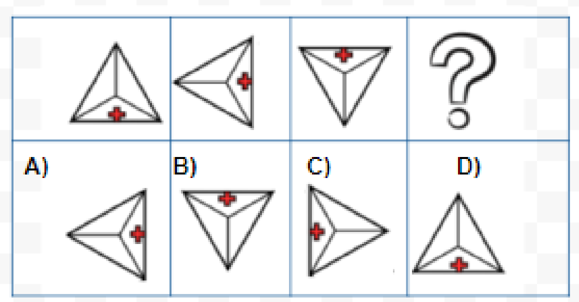
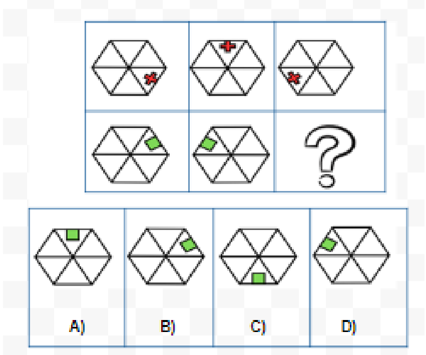
Table 1. *Item model for non-verbal reasoning items.*

Model 1	1 (column) * 4 (row) Shape x – Shape x – Shape x - ? (rotation angle and rule)
Model 2	2 (column) * 3 (row) Shape x – Shape x – Shape x (rotation angle and rule) Shape y – Shape y - ? (rotation angle and rule)
Elements	Shape_x: square, triangle, circle, hexagon Shape_y: square, triangle, circle, hexagon Rotation angle: 1. square: 45, 90, 180; 2. triangle: 60, 90, 120, 180; 3. circle: 45, 90, 180; 4. hexagon: 60, 90, 120, 180 Rotation rule: 1. right; 2. left Number of divisions: 1. square: 4, 2. triangle: 3, 3. circle: 4, 4. hexagon: 6 Internal element form: crosshair, small square Internal element color: transparent, blue, green, red
Key	Option 1, Option 2, Option 3, Option 4

3.3. Generating items using computer technology

Once the elements from the cognitive model were placed into the item model, the process of AIG for the items was initiated. At this step, Python codes were generated for each item model. 112 items from the first model and 1728 items from the second model were generated. The generated sample items are shown in Figure 3.

Figure 3. *Generated sample non-verbal reasoning items.*

Sample items from Model 1	Sample items from Model 2
1.	1.
	
Correct option: B	Correct option: B
27.	568.
	
Correct option: D	Correct option: B
51.	1098.
	
Correct option: C	Correct option: C

3.4. Review of SMEs’ opinions

A random selection of 10 items was made from the generated items of each model. Opinions from 6 SMEs were gathered for the selected 20 items. The results of the SMEs’ opinions were presented in Table 2 (for Model 1) and Table 3 (for Model 2). In only three items - 2, 3, and 8- minor revision suggestions had been proposed by two SMEs for Model 1. The minor revision in the 2nd item pertains to the perception that the item's difficulty was below that of the student's grade level. It had been indicated that rotating the circle 90 degrees clockwise (to the right) was considered quite manageable for 4th-grade students. The suggested minor revision for the 3rd item was about the potential challenge for students to comprehend a 60-degree rotation angle of the triangular shape. The minor revision suggested for the 8th item was oriented toward distractors. It has been recommended to insert a gap between options B and C. In the second model, for 5 items - 2, 4, 6, 7 and 8 - there exist minor revisions. For items 2 and 4, one SME has provided a visual minor revision proposal, suggesting the inclusion of gaps between distractors with rotation angles of 60 degrees each. One SME indicated the necessity for a minor revision at grade-level suitability in items 6, 7, and 8. The SME suggested that one of the distractors is relatively easy, and altering the rotational angle of this distractor had been recommended. For the items in the first model, the SMEs indicated that the difficulty ranged from very easy (1) to hard (4). Similarly, for the items in the second model, the SMEs expressed that the difficulty varied from moderately easy (2) to hard (4). In the first model, experts' opinions on the item difficulty varied between very difficult (1) and easy (5). There was no opinion suggesting that the generated items were very easy (5) in the first model. In the second model, the item difficulties were assessed by experts within the range of difficult (2) to easy (4).

Table 2. *SMEs' opinions_Model 1.*

Items	Difficulty		One right answer				Single content and behavior				Not trivial content				Choices homogeneous			
	Median	Average	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej
I1	2	1.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I2	2	1.7	6	0	0	0	5	1	0	0	6	0	0	0	6	0	0	0
I3	4	3.7	6	0	0	0	4	2	0	0	6	0	0	0	6	0	0	0
I4	3	3.0	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I5	3	2.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I6	3	2.8	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I7	2	2.0	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I8	2	1.8	6	0	0	0	6	0	0	0	6	0	0	0	4	2	0	0
I9	1	1.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I10	3	2.5	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0

Table 3. *SMEs' opinions_Model 2.*

Items	Difficulty		One right answer				Single content and behavior				Not trivial content				Choices homogeneous			
	Median	Average	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej	Acpt.	Minor	Major	Rej
I1	3	3.2	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I2	3	3.3	6	0	0	0	6	0	0	0	6	0	0	0	5	1	0	0
I3	3	3.3	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I4	3	3.3	6	0	0	0	6	0	0	0	6	0	0	0	5	1	0	0
I5	4	3.5	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0
I6	2	2.3	6	0	0	0	5	1	0	0	6	0	0	0	5	1	0	0
I7	2	2.5	6	0	0	0	5	1	0	0	6	0	0	0	5	1	0	0
I8	3	2.8	6	0	0	0	5	1	0	0	6	0	0	0	5	1	0	0
I9	3	2.7	6	0	0	0	5	1	0	0	6	0	0	0	6	0	0	0
I10	4	3.7	6	0	0	0	6	0	0	0	6	0	0	0	6	0	0	0

4. DISCUSSION and CONCLUSION

In the digital measurement and assessment era which is becoming increasingly widespread, the role and significance of visual aptitude tests are becoming even more pronounced. This is primarily due to the ease of using visual and auditory tools in digital measurement. Visual tests are used for measuring individuals' visual intelligence and problem-solving skills. They find applications in a wide range of areas, including identifying individuals with learning difficulties or gifted students, as well as in recruitment processes and career guidance (Atli, 2007; Cohen & Swerdlik, 2015). Additionally, the use of non-visual items also contributes to the validity of the test. Navigating a test in a language different from one's native tongue can pose challenges for students, particularly impacting performance. Socio-economic factors further affect students' achievement with the verbal itemsOpting for non-verbal item types to assess the special abilities of individuals from lower socio-economic backgrounds can enhance the accuracy of predictions (Bildiren et al., 2021; Lewis et al., 2007). In this case, AIG, an innovative approach to the process of creating non-verbal items, stands out. Rather than creating visuals for each item manually, utilizing computer technology can make the process more efficient and cost-effective. Therefore, AIG is used, which combines the expertise of professionals with computer technology. It has been observed in the literature that template-based AIG studies have been used in various fields such as medicine (Falcão et al., 2022; Gierl & Lai, 2012), dentistry (Lai et al., 2016), mathematics (Adji et al., 2018; Embretson & Kingston, 2018), literature (Sayin & Gierl, 2023). Also, limited studies in the existing literature, such as those by Gierl et al. (2015) and Ryoo et al. (2022), have shown that non-verbal items can be generated using AIG. Our study aimed to introduce how AIG can be used to create a comprehensive item pool focused on non-text-based items, especially for fields such as the BİLSEM entrance examination used in Türkiye (MoNE, 2022a). In this context, a cognitive model was initially developed for non-verbal visual reasoning. From the developed model, the "rotation" problem and scenario were chosen. The selected scenario was aligned with the MoCA scale, determining features and elements. Subsequently, two item models were developed. In the third step, the elements from the cognitive model were integrated into the item models using computer technology. For the first item model, 112 items were generated, while 1728 items were produced for the second item model using Python codes. This study aimed to demonstrate the applicability of non-verbal visual reasoning items with AIG. To achieve this, the range of shapes was limited by using four shapes as examples for the generated items. By increasing the number of shapes and including other elements, it is possible to create items with different similarities. Mental rotation tasks have been recognized as a measure of visuospatial ability (Cooper, 1975) and have attracted a great deal of interest in research on predicting abilities (Nolte et al., 2022). As a result, it appears as a preferred item type in BİLSEM exams. Since intelligence tests such as BNV and ASIS were integrated into the last BİLSEM entrance exams (BİLSEM Online, 2023b), the items were not opened. However, six mental rotation items were identified in the 50-item test (2016 test), which included only parent items, indicating that these items were used by changing the shape-rotation angle. This suggests that the items created in this study may find application in tests from different years.

In our study, the generated items were evaluated based on the SMEs` opinions. SMEs examined randomly selected 10 items from each model based on four different criteria and predicted the item difficulty. As a result of SMEs' opinions, it was determined that the items have varying item difficulty. This outcome was anticipated since the positions of different shapes at the same rotation angle exhibit variation. For instance, a square manifests a more pronounced 45-degree rotation angle than a circle due to its four sides. This discrepancy in rotation angle/speed is attributed to the size of the central mass (shape) and the congruence of sides and angles. Given that the main body size of a quadrilateral surpasses that of a triangle, the perceived rotation speed is heightened (Pylyshyn, 1979). Consequently, experts rated triangle-related items as more challenging than their quadrilateral counterparts. Additionally, the obtained results suggest the feasibility of generating items by constructing item pools with varying difficulty ranges by AIG (Sinharay & Johnson, 2005). The findings indicated that item difficulty, as perceived by experts, also varied based on the models. In the parent item, showcasing the 3*3 item model, instances of the desired pattern were presented in the final line in two distinct forms. These instances eased problem-solving by providing additional information. Similarly, items generated with the 2*3 item model in this study offer more information about problem resolution compared to items created with the 1*4 item model. Because it includes two lines for the solution. This clarifies why experts considered items from the 2*3 item model easier than those from the 1*4 item model. Furthermore, the uniform used of a single rotation rule and one internal element in both item models contributed to a general evaluation of items easily. Ultimately, item difficulties varied based on the item model and the elements. It shows that introducing new cognitive features to the item model has the potential to yield more intricate items.

Generated items were appropriate for one right answer, single content and behavior, not trivial content, and choices homogeneous by SMEs. The obtained results demonstrated the applicability of AIG for a comprehensive pool of items consisting of non-verbal visual reasoning items. Throughout the process of generating non-verbal visual items using AIG, it was noticed that the role of SMEs in shaping the scientific and item model is critically crucial. The contributions of SMEs had aided in ensuring the accuracy of item content. And it showed that the innovations brought about by utilizing computer technology had shown that it could efficiently and cost-effectively create a large item pool. This technological advancement has the potential to make it more efficient and accessible. Based on the findings of the current research, we recommend the creation of a comprehensive item pool using the results obtained. This item pool can be effectively utilized in computer-based tests, offering the advantages of personalized testing, and adaptive testing, and allowing multiple test administrations within a year. These recommendations are crucial for enhancing the evaluation of student performance and supporting more effective learning processes. Furthermore, we suggest future research initiatives, such as conducting field research and exploring equivalence for test equating. These advanced studies can further optimize the process of AIG and enhance the existing knowledge base in this field. In conclusion, the current study emphasizes the significance of visual aptitude tests in meeting the demands of contemporary digital assessments and highlights the feasibility of generating such tests using AIG. By demonstrating how AIG can facilitate the creation of a comprehensive item pool, especially for assessments used in in Türkiye, the current research aims to lay the groundwork for future research and applications in the realms of education and assessment.

Limitations

Acknowledgments of people, grants, and funds should be placed in a separate section before the References. If the study has been previously presented at a conference or a scholarly meeting, it should be mentioned here. The present study focused on exploring the viability of generating non-verbal reasoning items through AIG, with item evaluation conducted based on expert opinions. For future investigations, it would be beneficial to conduct field tests on the AIG-generated test items and estimate validity evidence by analyzing the data coming from field tests. The potential for a testing effect arises when items from the same pool are employed at different times, particularly within short-term intervals. To mitigate this, diversifying the item pool by varying elements and item models could be considered. Additionally, since this study exclusively utilized rotation, it is advisable to incorporate item samples that assess other problem situations in future research endeavours.

Acknowledgments

The authors would like to thank the blind reviewers and SMEs for their useful comments and insightful suggestions.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Gazi University Ethics Committee, E-77082166-604.01.02-686103. 22.06.2023.

Authorship Contribution Statement

Ayfer Sayin: Design, Data Collection and/or Processing, Analysis and Interpretation, Literature Review, Writing. **Sabiha Bozdag:** Materials, Data Collection and/or Processing, Literature Review, Writing. **Mark J. Gierl:** Conception, Design, Supervision, Writing, Critical Review

Orcid

Ayfer Sayin https://orcid.org/0000-0003-1357-5674
Sabiha Bozdag https://orcid.org/0000-0002-2039-8066
Mark J. Gierl https://orcid.org/0000-0002-2653-1761

REFERENCES

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. https://doi.org/10.26499/iceap.v1i1.78

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. https://doi.org/10.1016/j.lindif.2011.11.005

Atli, S. (2007). *Matematiksel mantıksal yetenek ile ritimsel yetenek arasındaki ilişkiler* [*Relations between matematical-logical talent and rhythmic intelligence*] [Unpublished master's thesis]. Gazi University.

Balboni, G., Naglieri, J.A., & Cubelli, R. (2010). Concurrent and predictive validity of the raven progressive matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment, 28*(3), 222-235. https://doi.org/10.1177/0734282909343 76

Bildiren, A., Bıkmaz Bilgen, Ö., & Korkmaz, M. (2021). National non-verbal cognitive ability test (BNV) development study. *SAGE Open, 11*(3). https://doi.org/10.1177/2158244021 104694

Bilgiç, N., Taştan, A., Kurukaya, G., Kaya, K., Avanoğlu, O., ve Topal, T. (2017). *Özel yetenekli bireylerin eğitimi strateji ve uygulama kılavuzu [Education of specially gifted individuals' strategy and implementation guide].* MEB Özel Eğitim ve Rehberlik Hizmetleri Genel Müdürlüğü. https://orgm.meb.gov.tr/meb_iys_dosyalar/2013_11/2503 4903_zelyeteneklibireylerineitimistratejiveuygulamaklavuzu.pdf

BİLSEM Online (2023a). *Sıkça sorulan sorular: BİSEM sınav soruları yeteneğe göre değişir mi? [Frequently asked questions: Do BİLSEM exam questions vary depending on ability?].* https://www.bilsemonline.com/sss

BİLSEM Online (2023b). BNV Zeka Testi Nedir? *[What is BNV Intelligence Test?].* https://bilsemonline.com/blog/bnv-zeka-testi-nedir

Choi, J., & Zhang, X. (2019). Computerized item modeling practices using computer adaptive formative assessment automatic item generation system: A tutorial. *The Quantitative Methods for Psychology, 15*(3), 214-225. https://doi.org/10.20982/tqmp.15.3.p214

Cohen, R.J., & Swerdlik, M.E. (2015). *Psychological testing and assessment*. McGraw-Hill Education.

Cooper, L.A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive Psychology, 7*(1), 20-43. https://doi.org/10.1016/0010-0285(75)90003-1

DeThorne, L.S. & Schaefer, B.A. (2004). A guide to child nonverbal IQ measures. *American Journal of Speech-Language Psychology, 13*(4), 275-290. https://doi.org/10.1044/1058-0360(2004/029)

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. https://doi.org/10.1111/jedm.12166

Embretson, S., & Yang, X. (2007). 23 Automatic item generation and cognitive psychology. *Handbook of statistics*, *26*, 747-768. https://doi.org/10.1016/S0169-7161(06)26023-1

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. https://doi.org/10.1007/s10459-022-10092-z

Gibbons, A., & Warne, R.T. (2019). First publication of subtests in the Stanford-Binet 5, WAIS-IV, WISC-V, and WPPSI-IV. *Intelligence*, *75*, 9-18. https://doi.org/10.1016/j.intell.2019.02.005

Gierl, M.J., & Haladyna, T. (2012). *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. https://doi.org/10.1080/15305058.2011.635830

Gierl, M.J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*(3), 36-50. https://doi.org/10.1111/emip.12018

Gierl, M.J. & Lai, H. (2016). Automatic item generation. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd edition, pp. 410-429). Routledge.

Gierl, M.J., Ball, M.M., Vele, V., & Lai, H. (2015). A method for generating nonverbal reasoning items using n-layer modeling. *In Computer Assisted Assessment. Research into E-Assessment: 18th International Conference, CAA 2015*, Zeist, The Netherlands, June 22–23, 2015. https://doi.org/10.1007/978-3-319-27704-2_2

Gierl, M., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.

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), 309-333, https://doi.org/10.1207/S15324818AME1503_5

Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., & Moriarty Gerrard, M.O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Horn, J.L., & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270. https://doi.org/10.1037/h0023816

Karasar, N. (2022). *Bilimsel araştırma yöntemleri (37. Basım)*. Nobel Yayıncılık.

Kemer, B., & Çakan, M. (2020). Examining the validity of the psychological scales frequently used in guidance and research centers with respect to measurement standards of validity. *Journal of Research in Education and Society, 7*(1), 323-348. https://dergipark.org.tr/en/pub/etad/issue/55359/731549

Kocagül, M., & Çoban, G.Ü. (2022). An evaluation on science teachers' scientific reasoning skills. *Cumhuriyet International Journal of Education, 11*(2), 361-373. https://doi.org/10.30703/cije.1017938

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. https://doi.org/10.1111/emip.12237

Kurnaz, A., & Ekici, S.G. (2020). BİLSEM tanılama sürecinde kullanılan zeka testlerinin psikolojik danışmanların ve BİLSEM öğretmenlerinin görüşlerine göre değerlendirilmesi [Evaluation of intelligence tests used in BİLSEM diagnostic process according to the opinions of psychological counselors and BİLSEM Teachers]. *Çocuk ve Medeniyet, 5*(10), 365-399. https://dergipark.org.tr/en/pub/cm/issue/59377/850922

Kurtz, K., Gentner, D., & Gunn, V. (1999). Reasoning. In B. M. Bly & D. E. Rumelhart (Eds), *Cognitive science*, pp. 145-200. California: Academic Press.

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. PMID: 26933110.

Lawson, A.E. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education*, *2*(3), 307-338. https://doi.org/10.1007/s10763-004-3224-2

Leighton, J.P. (2012). Learning sciences, cognitive models, and automatic item generation. In M.J. Gierl, & T.M. Haladyna (Eds), *Automatic item generation: Theory and practice*, pp. 121-135. Routledge.

Lewis, J.D., DeCamp-Fritson, S.S., Ramage, J.C., McFarland, M.A., & Archwamety, T. (2007). Selecting for ethnically diverse children who may be gifted using Raven’s Standard Progressive Matrices and Naglieri Nonverbal Abilities Test. *Multicultural Education*, *15*(1), 38-42.

Lohman, D.F., & Hagen, E. (2003). *Interpretive guide for teachers and counselors: cognitive abilities test Form 6-all levels*. ITASCA, Illinois: Riverside Publishing.

Mercan, Z. (2021). Studies on early childhood reasoning skills in Turkey. *Journal of Muallim Rıfat Faculty of Education*, *3*(2), 104-120.

Ministry of National Education (MoNE) (2015). Bilim ve Sanat Merkezleri Yönergesi [Science and Art Centers Directive]. MoNE General Directorate of Special Education and Guidance Services. https://orgm.meb.gov.tr/meb_iys_dosyalar/2015_10/26091626_blsemkilavuz26.10.2015.pdf

Ministry of National Education (MoNE) (2021). Bilim ve Sanat Merkezleri Yönergesi [Science and Art Centers Directive]. MoNE General Directorate of Special Education and Guidance Services]. https://orgm.meb.gov.tr/meb_iys_dosyalar/2021_12/30144032_2021-2022_YILI_BILIM_VE_SANAT_MERKEZLERI_OGRENCI_TANILAMA_VE_YERLESTIRME_KILAVUZU.pdf

Ministry of National Education (MoNE) (2022a). *Bilim ve Sanat Merkezleri Yönergesi [Science and Art Centers Directive]*. MoNE General Directorate of Special Education and Guidance Services. https://orgm.meb.gov.tr/meb_iys_dosyalar/2016_10/07031350_bilsem_yonergesi.pdf

Ministry of National Education (MoNE) (2022b). *Bilim ve Sanat Merkezleri öğrenci tanılama ve yerleştirme kılavuzu [Science and Art Centers student identification and placement guide]*. MoNE General Directorate of Special Education and Guidance Services. https://orgm.meb.gov.tr/www/bilsem-ogrenci-tanilama-ve-yerlestirme-kilavuzu-yayimlandi/icerik/2154

Mullin, I.V.S., Martin, M.O., & Foy, P. (2005). *IEA’s TIMSS 2003 international report on achievement in the mathematics cognitive domains*. TIMSS & PIRLS Internatonal study

Center. Lynch School of Education, Boston College. https://files.eric.ed.gov/fulltext/ED494652.pdf

Naglieri, J.A., & Ford, D.Y. (2005). Increasing minority children's representation in gifted education: A response to Lohman. *Gifted Child Quarterly*, *49*(1), 29-36. https://doi.org/10.1177/001698620504900104

Nolte, N., Schmitz, F., Fleischer, J., Bungart, M., & Leutner, D. (2022). Rotational complexity in mental rotation around cardinal and skewed rotation axes. *Intelligence*, *91*, 101626. https://doi.org/10.1016/j.intell.2022.101626

Pylyshyn, Z. W. (1979). The rate of "mental rotation" of images: A test of a holistic analogue hypothesis. *Memory & Cognition*, *7*(1), 19-28. https://doi.org/10.3758/BF03196930

Ryoo, J.H., Park, S., Suh, H., Choi, J., & Kwon, J. (2022). Development of a new measure of cognitive ability using automatic item generation and its psychometric properties. *SAGE Open*, 1-13. https://doi.org/10.1177/21582440221095016

Sak, U., Sezerel, B.B., Dulger, E., Sozel, K., & Ayas, M.B. (2019). Validity of the Anadolu-Sak Intelligence Scale in the identification of gifted students. *Psychological Test and Assessment Modeling*, *61*(3), 263-283. https://psycnet.apa.org/record/2020-53108-001

Sayın, A., & Gierl, M.J. (2023). Automatic item generation for online measurement and evaluation: Turkish literature items. *International Journal of Assessment Tools in Education*, *10*(2), 218-231. https://doi.org/10.21449/ijate.1249297

Shin, E. (2021). *Automated item generation by combining the non-template and template-based approaches to generate reading inference test items.* [Doctoral dissertation, University of Alberta]. Education and Research Archive. https://doi.org/10.7939/r3-75wr-hc80

Sinharay, S., & Johnson, M. (2005). Analysis of data from an admissions test with item models. *ETS Research Report Series*, *2005*(1), 1-32. https://files.eric.ed.gov/fulltext/EJ1111287.pdf

Tamul, Ö.F., Sezerel, B.B., Sak, U., & Karabacak, F. (2020). Social validity study of the Anadolu-SAK intelligence scale (ASIS). *PAU Journal of Education*, *49*, 393-412. https://doi.org/10.9779/pauefd.575479

Weiss, L.C., Saklofske, D.H., Holdnack, J.A., & Prifitera, A. (2016). WISC-V: Advances in the assessment of intelligence. In L.G. Weiss, D.H. Saklofske, J.A. Holdnack, & A. Prifitera (Eds.), *WISC-V assessment and interpretation: Scientist-practitioner perspectives* (pp. 3–23). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-404697-9.00001-7