International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 218–231
https://doi.org/10.21449/ijate.1249297
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Automatic item generation for online measurement and evaluation: Turkish literature items

Ayfer Sayin[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: Feb. 09, 2023
Revised: May 24, 2023
Accepted: May 30, 2023

Keywords:
Automatic item generation,
Large item pool,
Computer-based testing,
Online measurement, and evaluation,
Turkish literature.

Abstract: Developments in the field of education have significantly affected test development processes, and computer-based test applications have been started in many institutions. In our country, research on the application of measurement and evaluation tools in the computer environment for use with distance education is gaining momentum. A large pool of items is required for computer-based testing applications that provide significant advantages to practitioners and test takers. Preparing a large pool of items also requires more effort in terms of time, effort, and cost. To overcome this problem, automatic item generation has been widely used by bringing together item development subject matter experts and computer technology. In the present research, the steps for implementing automatic item generation are explained through an example. In the research, which was based on the fundamental research method, first a total of 2560 items were generated using computer technology and SMEs in field of Turkish literature. In the second stage, 60 randomly selected items were examined. As a result of the research, it was determined that a large item pool could be created to be used in online measurement and evaluation applications using automatic item generation.

1. INTRODUCTION

Automatic item generation (AIG) is a research area where cognitive and psychometric theories and computer technology are used to create content for tests (Gierl & Haladyna, 2012; Gierl & Lai, 2013; Irvine & Kyllonen, 2002). AIG, which was first introduced by Bormuth (1969), helps to generate large numbers of items to item pools quickly and economically (Gierl & Haladyna, 2012; Gierl & Lai, 2013; Irvine & Kyllonen, 2002; Sinharay & Johnson, 2005). In this research, it is aimed to introduce the steps of AIG in detail, based on an item in the Field Proficiency Test (original acronym: AYT), the second session of the Higher Education Institutions Examination (original acronym: YKS) in Türkiye. In addition, the process of the examination of automatically generated items is explained.

It has been determined that there are difficulties in measurement and evaluation in the reports regarding the distance education process, which was rapidly passed in 2019 with the COVID-19 pandemic. Turkish Education Association Think Tank (original acronym: TEDMEM) stated that in this process, there were differences in testing practices and policies, especially in higher

*CONTACT: Ayfer Sayin ayfersayin@gazi.edu.tr Gazi Faculty of Education, Department of Educational Sciences, Division of Assessment and Evaluation in Education, Ankara, Türkiye

e-ISSN: 2148-7456 /

education institutions; while some universities made pass-fail decisions, some wanted grades not to be included in the academic achievement grade point average (TEDMEM, 2020). These differences in testing practices between universities indicate that there is no universal agreement on measurement and evaluation practices for distance education. Similarly, although the courses are held remotely by the Ministry of National Education (original acronym: MEB) in Türkiye, mid-term exams were carried out face-to-face in Türkiye at the some grades. (MEB, 2021). Research have shown that most teachers cannot measure and evaluate in the distance education process (Saygı, 2021); shows that they also have a negative attitude toward online measurements and evaluation (Sarı & Nayır, 2020; Saygı, 2021). Teachers stated that they had difficulties in measurement and evaluation practices in the distance education process, especially due to security problems such as cheating (Adıgüzel, 2020; Kınalıoğlu & Güven, 2011). Cheating in test applications, insufficient evidence for the psychometric properties of the items; situations such as students plagiarizing in their homework, sharing the load by one or more students in group work; it also constitutes a problem of face-to-face measurements and evaluations. To prevent these problems, which are even more prominent in measurement and evaluation practices in distance education, individualized tests, different tests consisting of items of equivalent quality to each student can be applied. Measurement and evaluation problems in distance education are not limited to summative tests in which students' achievement is measured.

Traditional measurement and assessments and training focus on leveling to determine the extent to which students have achieved their goals rather than process assessments. Modern measurement and assessments, together with the process and the generation, come to the fore. At this point, it is necessary not only to apply for an exam to the students at the end of a certain period but also to make applications so that the students can see their deficiencies in the process. This process, in which formative tests are used, aims to support students' learning process by giving effective feedback (Bennett, 2011; Gierl & Lai, 2018). Similarly a more systematic and continuous assessment and evaluation approach should be adopted in the distance education process (Gaytan & McEwen, 2007). Balta and Türel (2013) state that if measurement and assessments are carried out continuously, students can be given feedback by considering their differences, which will significantly increase the quality of learning in distance education. For this, it is necessary both to implement practices to improve teachers' online measurement and evaluation skills and to create a digital infrastructure in schools.

Within the scope of the Movement to Increase Opportunities and Improve Technology (original acronym: FATİH) Project carried out by the MEB, smart boards were placed in the classrooms, and the way for students to use tablets in the education environment was paved (Artırma & Hareketi, 2020). Similarly, the Education Information Network (original acronym: EBA) created by the MEB has been widely used both inside and outside the school since 2012 (MEB, 2020). Adaptive learning environments added to the network in 2019, it is becoming even more effective. Measurement Selection and Placement Center (original acronym: ÖSYM) has been conducting some applications of the Foreign Language Exam in the computer environment as called e-YDS since 2014 in Türkiye (ÖSYM, 2022). It is considered that these applications will become more widespread in the coming years. However, in this electronic test, all candidates are presented with a single booklet (all items are the same even if the items and options' place are different) at the same time. This is essentially a computerized version of a paper-and-pencil test and therefore does not provide the advantages of computer-based testing.

In today's world of rapidly developing technology, test development, and administration processes have begun to integrate with computer technology. The advantages of computer-based applications such as portability, usability, and the absence of time and space constraints have paved the way for the widespread use of computer applications in the field of measurement

and evaluation (Chen et al., 2019; Gierl & Haladyna, 2012; Lane et al., 2015; Weber et al., 2003). For instance, the American College Test (ACT), which is one of the entrance exams for higher education institutions in the United States, Graduate Management Admission Test (GMAT) and Graduate Record Examinations (GRE) applied for graduate education; Transition tests (MCCQE, NCLE) in medical education applied in Canada are computer-based. The fact that computer-based tests provide flexible administration time and place for all students and that scores are calculated quickly, it is attractive not only for large-scale but also for in-class applications (Chen et al., 2019; Gierl et al., 2021). It is a rapid transition to computer-based tests in all classroom applications from preschool to secondary education, called K-12 (Gierl et al., 2022).

In classroom practices, teachers tend to require a long time to score the tests because of the large number of students in the classroom, the intensity of the lesson, and other works. When tests require a long time to score, students do not have enough time to follow their individual developments, to see their shortcomings, and to make applications to make up for them (Choi & Zhang, 2019; Clark & Rust, 2006). This reduces the effectiveness of feedback. Since computer-based test applications quickly explain the results to the students, they can follow their individual progress and manage their own learning processes (Corbett & Anderson, 2001; Zhu et al., 2020). In short, by taking advantage of digital education, it is necessary to establish a measurement and evaluation system that prioritizes the individual differences of the student, increases the motivation of the student as well as academic success, supports group work and communication, and allows interventions for the problems of distance education. For this, it is necessary to utilize technology not only in the application infrastructure but also in the creation of a large item pool.

While the test development process is difficult on its own, creating a large item pool requires much effort in terms of labor, time, and cost (Gierl & Lai, 2013; Kosh et al., 2019). Studies also demonstrate that teachers already have difficulties in writing items (Karatay & Dilekçi, 2019; Özyalçın & Kana, 2020). Especially in the distance education process, the teachers wanted to apply different tests consisting of equivalent items to the students for test security reasons. However beside to the cost of test development, an important problem is encountered in the process of writing an equivalent item to measure the same outcome: "subjectivity." An item is an expression of the knowledge, skills and competence of the SMEs who created the item. For this reason, the items created carry traces of the item author, and for this reason, item creation is expressed as "art" according to some researchers (Rodriguez, 2005). However, equivalent items, especially those in different forms, should have comparable content and psychometric properties. However, while there are differences even among the items created by the same item developer, it is challenging for different item developers to ensure the equivalence of the items. Haladyna and Rodriguez (2013) states that test development procedures are applied by most item writer, therefore, asking for items is a science, but the items still contain subjectivities from their authors (Gierl & Haladyna, 2012). Therefore, it is necessary to combine experts and technology in test development and to take advantage of the innovations that computer-based testing systems bring to the content, administration environment and assessment aspects of tests, as well as to the way items are written (Davey, 2011). As one of them, automatic item generation can be used as an alternative item development process to create a large item pool. In addition, in classroom applications, a large item pool can be obtained by overcoming this problem with AIG. Simultaneously, the reliability of the test can be ensured by conducting content coding studies at AIG (Gierl et al., 2022).

1.1. Automatic Item Generation (AIG)

AIG, is an item generation method that combines models, content expertise and computer technology to create large and efficient test banks in a short time (Gierl et al., 2021). AIG which

has started to spread to psychology, education, and computer sciences, is mainly carried out for two purposes: The first is to generate equivalent items with similar difficulty and psychometric properties, and the second is to create an item pool with items with different item difficulty ranges (Sinharay & Johnson, 2005). In other words, it is aimed to generate the desired quality and number of items in the item pool. Considering that time, effort, and cost must develop a test (Gierl et al., 2016), AIG provides significant cost savings while creating a large item pool (Kosh et al., 2019). Another advantage of AIG is that it allows quick and effective feedback on test results to students through practice (Gierl & Lai, 2018). In this way, students can follow the development of their individual learning.

Generally, the AIG process is carried out in two different ways: artificial intelligence-based and template-based (Gierl & Lai, 2013). In this research, the stages of template-based automatic item generation are introduced in an applied way through an example. Template-based AIG has three stages: In the first stage (i) a "cognitive model" is developed for the AIG, in the second stage (ii) an "item model" is developed in which the cognitive model content is embedded to generate new items. In the third and final stage (iii), computer algorithms are used to place the content of the cognitive model developed in the first stage into the item model developed in the second stage (Gierl & Lai, 2013). As a result, the template-based automatic item generation method, which includes these three stages, if the generated items have an equivalent and predictable difficulty level, the result can be considered successful.

2. METHOD

2.1. Design

In this research was conducted within the framework of the qualitative research model. Since the research will be based on experiments and theory aiming to acquire new information about the foundations of facts and observable facts, it can be considered fundamental research. Fundamental research is studies conducted to examine, analyze, strengthen a theory related to a certain field or to put forward a new theory (Kaptan, 1998).

2.2. Participants

In the research, the opinions of five experts, including a measurement and evaluation expert and four academicians in the field of Turkish language and literature education, were taken in collecting the validity and reliability evidence of the models developed for AIG.

2.3. Data Tools

In the research items were generated in the field of Turkish literature with AIG. AIG usually starts with a parent item. To demonstrate its applicability in this study, the YKS application with the highest number of candidates in Türkiye was chosen, and the 12th item of the AYT in the second session of YKS in 2018 was identified as the parent item. It is shown in Figure 1.

Figure 1. *12th item for 2018 in the field proficiency test.*

12. He is an 18th century Classic Turkish literature poet. He is considered one of the important representatives of the Sebk-i Hindî. He became the owner of a divan at a young age. He proved his success in mesnevi writing with his allegoric work Hüsnü Aşk.

Which of the following is the poet mentioned in this passage?

A) Süleyman Çelebi B) Ahmet Pasha

C) Sehi Bey D) Taşlıcalı Yahya

E) Âşık Pasha

The test items in the YKS application are open access. The main item in Figure 1 was accessed from https://dokuman.osym.gov.tr/pdfdokuman/2018/YKS/TSK/ayt_yks_2018.pdf OSYM's

own website. Based on the parent item, first the "cognitive model" and "item model" were created. This process is introduced in the findings section.

2.4. Analysis

The generated items were analyzed according to SMEs opinion and similarity index. SMEs' opinions on the models and items

- scientific accuracy,
- language and expression,
- item difficulty

examined in this regard.

Cosine Similarity Index (CSI) also was used to determine the similarity of automatically generated items (Gierl & Lai, 2013). CSI refers to the similarity between the vectors of the texts in the two items. It is calculated using the cosine of the angle between two vectors in the multidimensional space of unique words. In other words, CSI is a word similarity measure calculated using an algorithm based on the text-vector indexing technique. CSI is calculated with the formula in 1.1 (Bayardo et al., 2007).

$$\cos(\overrightarrow{A},\overrightarrow{B}) = \frac{\overrightarrow{A} \cdot \overrightarrow{B}}{\|\overrightarrow{A}\| \, \|\overrightarrow{B}\|} \quad (1.1)$$

In the equation in 1.1, A, and B are expressed in a binary vector of word formations. First, the length of the binary vector is found by determining the total unique number of words in texts A and B. Then, vectorization is performed for each text, depending on whether the words are in the A and B texts. The CSI value takes a value between 0 and 1; the closer to 0, the less similarity; it increases as it approaches 1. When CSI is 0, it means that there is no common word in the items generated, and when it is 1, it means that all words are common.

3. FINDINGS

The results of the research are reported in line with AIG's steps: (i) development of a cognitive model, (ii) development of an item model, (iii) generation of the items with computer technology.

3.1. Cognitive Model Development

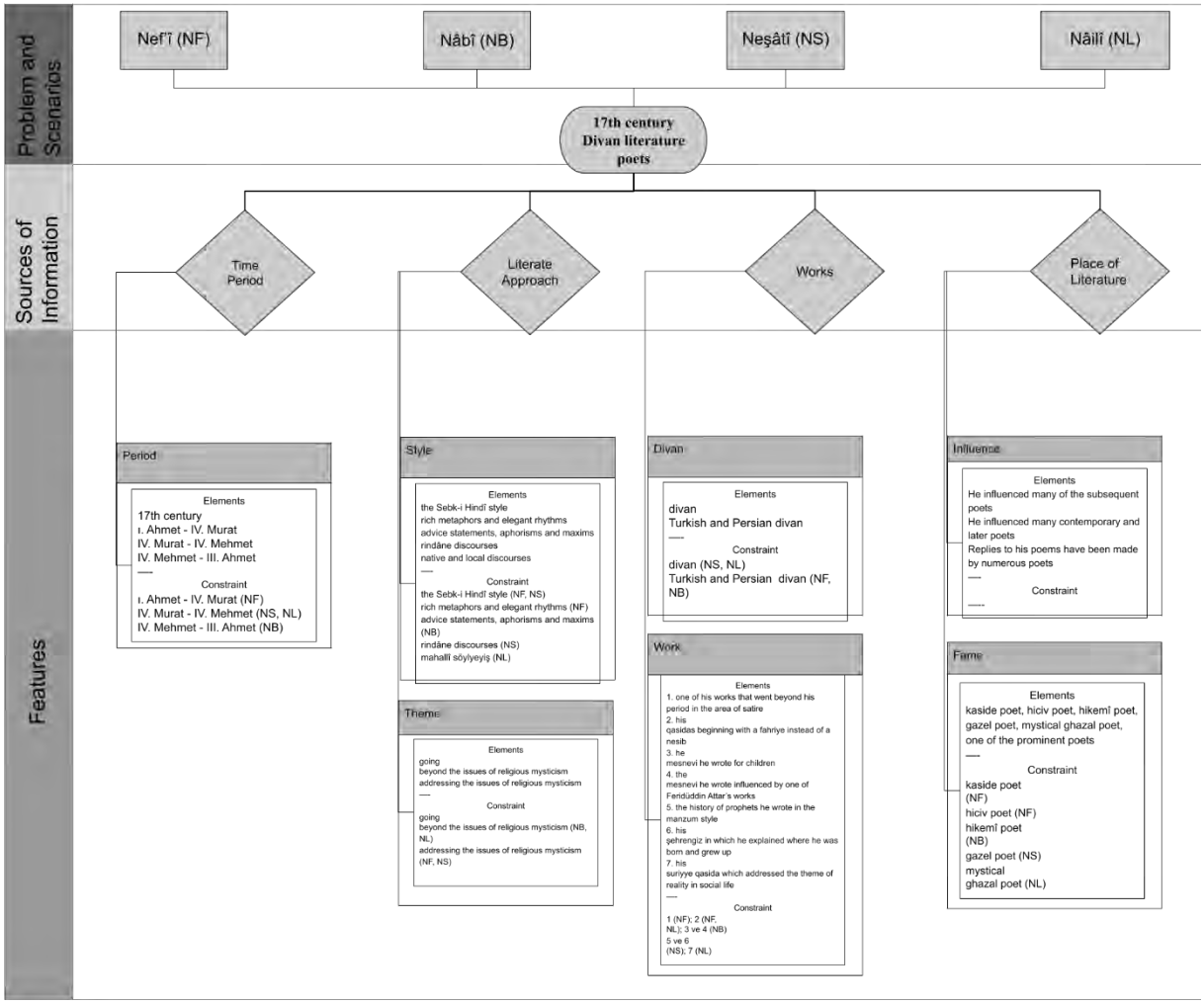
A cognitive model is an organizational diagram that includes the essential information, skills, and proficiencies and how they are used to solve a certain item. In this organizational diagram, the sources of information necessary to answer an item and the elements within each source of information are identified (Gierl et al., 2021). Thus, not only does the information needed to establish a cognitive template be identified, but effective feedback can also be given to students after an exam administration (Chen et al., 2019; Gierl et al., 2016).

At the first stage, the feature aimed at measuring the 12th item of the 2018 AYT (ÖSYM, 2018), was determined. Within the scope of the Turkish language and literature course, the item that focuses on "divan literature poets" is "Knows the poet's life and view to evaluate the relationship between poet and poetry." at A.1.10 in the curriculum (MEB, 2018). After the general features were identified, the next stage was to define the necessary information, skills, and content essential to answer the 12th item and for this reason the item was examined. The item is answered based on the key features of the "period" in which the poet lived, the "place" of the poet within the divan literature, the situation of being a "divan", the poet's "fame" in the tradition of poems, and the prominent "work" of the poet. The expert opinions of four academicians in Turkish language and literature education revealed that the features of poet's "place" in the literature and the poet's "fame" were very similar and overlapping. Considering that Sebk-i Hindî in the sample item is a feature of style, the "place" feature was expressed as

two separate features: “style” and “theme”. Moreover, the “influence” feature in similar items in the AYT was added to finalize the key features.

After the key features to form the cognitive model were identified, then the correct responses and the elements and limitations for each correct response were defined (Gierl et al., 2021), so that the cognitive model was developed. In the present study, four poets from the 17th century divan literature were chosen as samples for the cognitive model; however, another model in which all the centuries are taken into consideration, for example, can also be formed. The elements and limitations for the key features previously mentioned for each poet were defined and placed into an organizational structure. For example, although all poets lived in the 17th century, the rulers of their time differed. While Nef’î lived during the period of four padishahs from Ahmet I to Murat IV, Nâbî lived in the periods of Mehmet IV to Ahmet III. The Encyclopedia of Islam of the Religious Foundation of Türkiye (orig. TDV) was used as the main source for determining the characteristics of Divan literature poets (TDV, 2022). After revisions were made in the cognitive model developed based on the expert opinions obtained from four academicians in Turkish language and literature education, the cognitive model was finalized (see Figure 2).

Figure 2. *Cognitive model development in the first step of AIG.*



3.2. Item Model Development

In the second stage, an "item model" is developed in AIG. An item model defined where the key features will be placed. It includes basic properties and elements that can be changed to generate new items (Gierl et al., 2021). This place can occur in the stem of the item, in the

question prompt and in the options (correct answer and distracters). In this stage, ancillary features including text, pictures, tables, graphs, and diagrams can also be added to the item model, and random variables that can be manipulated, although not necessary, to solve the problem (Gierl & Lai, 2013).

The item model can be created in a one-layer or n-layered manner (Gierl & Lai, 2013). The purpose of item generation using the one-layer model is to generate few items by manipulating only the basic variables in the cognitive model. While fewer items are generated according to the one-layer item model, many more items can be generated compared to the n-layered model. Since the manipulations in a layered model are limited only by the number of elements, the similarity of the generated items is very high. This causes the generated items to be called "clones." In the n-layered model, the language used in the body, item sentence and options by using the syntactic structures of the language; structured hierarchically. Using the possibilities of the language, the content, key features, and elements can be manipulated by embedding them into each other. In this way, much more items can be generated from a cognitive model, and less similar items can be obtained (Gierl et al., 2021). An example of both one-layer and n-layer models can be seen in Table 1.

Table 1. *Examples of one and n-layer models.*

Stem (one layer)	He is a Classical Turkish literature poet who lived in the <**period**>. He is considered among important poets who skilfully used the <**style**>. The poet has <**divan**>. <**Influence**> He is considered as a <**fame**> in the divan poetry tradition. His works were <**theme**> of religious mysticism and it was proved his skill in writing with <**work**>.
Stem (n-layer)	*Period, style and theme + work, divan and fame* **Period, style and theme** 1. He is a Classical Turkish literature poet who lived during the <**Period**>. He is shown to be among the important poets who used the <**theme**> religious mysticism and <**style**> skilfully in his poems. 2. He is among the poets who used the <**theme**> religious mysticism and <**style**> skilfully in their poems. He is one of the divan literature poets who lived in the <**period**>. **Work, divan and fame** 1. The poet has the <**divan>** divan. <**Influence**> In the divan poetry tradition is referred to as a <**fame**> . He proved his skill in writing with his work <**work**>. 2. <**Influence**> He gained fame in the period he lived and in divan literature as a <**fame**>. The poet has the <**divan**> divan. The poet demonstrated his skill in <**work**>. 3. The poet has the <**divan**> divan and he has become identified with <**work**>. <**Influence**> He is referred to as <**fame**> in divan poetry. 4. <**Influence**> He displayed his skill in writing with <**work**> . The poet has the <**divan**> divan. In divan literature, it is known as <**fame**>.
Item Promtp	**Which of the following is the poet mentioned in this passage?**
Correct Options	Nef'î, Nâbî, Neşâtî, Nâilî

3.4. Generating Items Using Computer Technology

In the third and final stage of AIG, the model content created in the first step using computer technology is placed in the item model developed in the second stage, and item generation is carried out (Gierl et al., 2021). Different software are developed for item generation in the literature: Math Test Creation Assistant (Singley & Bennett, 2002), ModelCreator (Higgins et al., 2005), Item Distiller (Higgins, 2007), IGOR (Gierl & Lai, 2012), EAQC (Gutl et al., 2011), MARTEN (https://www.mghlpartners.com/software). In the present study, items were generated using scripts written in Phyton. After the third stage of AIG was successfully completed, 320 items in a one-layer model and 2560 items in a n-layer model were

automatically generated. Samples of the generated items are presented in Table 2. The original items in Turkish are added in the Appendix.

Table 2. *Samples of the generated items.*

Sample items generated with the one-layer model	Sample items generated with the n-layer model
1. He is a Classical Turkish literature poet who lived in the 17th century. He is considered among important poets who skilfully used the Sebk-i Hindî style. The poet has Turkish and Persian divan. He influenced many contemporary and later poets. He is considered as a kasida poet in the divan poetry tradition. His works were going beyond the issues of religious mysticism and it was proved his skill in writing with his qasidas beginning with a fahriye instead of a nesib. **Which of the following is the poet mentioned in this passage?** A) Nef'î* B) Nesîmî C) Nedim D) Şeyh Gâlip E) Nâbî	**1.** He is one of the divan literature poets who lived in the periods of four padishahs between IV. Murad, Sultan İbrâhim & IV. Mehmet. He is among the poets who used the going beyond the issues of religious mysticism and Sebk-i Hindî style skilfully in their poems. The poet has a divan where he collects his poems and is identified with the history of the prophets, which he wrote in verse. He influenced many contemporary and later poets. In divan literature, it is known as one of the prominent poets. **Which of the following is the poet mentioned in this passage?** A) Bâkî B) Nâbî C) Şeyh Galip D) Neşatî* E) Nedim
127. He is a Classical Turkish literature poet who lived in the periods of four padishahs between Mehmet IV and Ahmet III. He is considered among important poets who skilfully used the native andd local discourses. The poet has Turkish and Persian divan. He is considered as a hikemî poet in the divan poetry tradition. His works were addressing the issues of religious mysticism and it was proved his skill in writing with the mesnevi he wrote influenced by one of Feridüddin Attar's works. **Which of the following is the poet mentioned in this passage?** A) Neşati B) Necati Bey C) Nâbî* D) Gülşehrî E) Nef'î	1368. The poet is shown to be among the important poets who used the addressing the issues of religious mysticism and hikemî poet skilfully in his poems. He is a Classical Turkish literature poet who lived during the 17th century. Replies to his poems have been made by numerous poets. He became famous as a mystical ghazal poet in the period he lived and in divan literature. The poet has a divan where he collects his poems. He showed his poetic skill with his syria eulogy, which is about the reality in social life. **Which of the following is the poet mentioned in this passage?** A) Fuzûlî B) Nâilî* C) Nedim D) Şeyh Galip E) Nef'î
205. He is a Classical Turkish literature poet who lived in the 17th century. He is considered among important poets who skilfully used the rich metaphors and elegant rhythms. The poet has Turkish and Persian divan. He is considered as a hiciv poet in the divan poetry tradition. His works were going beyond the issues of religious mysticism and it was proved his skill in writing with his qasidas beginning with a fahriye instead of a nesib. **Which of the following is the poet mentioned in this passage?** A) Nesîmî B) Nedim C) Şeyh Galip D) Nâbî E) Nef'î*	2054. He is shown to be among the important poets who used the addressing the issues of religious mysticism and the native andd local discourses skilfully in his poems. He is the one of the Classical Turkish literature poet who lived in the 17th century. He showed his mastery in writing with his masnavi written for children. The poet has Turkish and Persian divan. He is known as a hikemi poet in Divan literature. **Which of the following is the poet mentioned in this passage?** A) Neşati B) Necati Bey C) Gülşehrî D) Nâbî* E) Nef'î

3.4. Examining the generated items by AIG

After the three steps of AIG were carried out, SMEs' opinion was used to evaluate the generated items, and then the similarity ratios of the generated items were calculated. At this stage, the opinions of four SMEs in the field of Turkish and literature education were taken. 60 items randomly selected among the automatically generated items, and they were examined by SME. In the items the experts examined, they reported that there were no errors in the linguistic accuracy and that the information given to answer the items was sufficient and clear. The experts whose opinions were received within the scope of the study revised the wordings of some of the automatically generated items, and the item model was redeveloped by taking these revisions into consideration. According to the four academicians, who stated that the difficulty level of the items varied, the easiest item was the one with the correct answer Nef'î and the most difficult item was the one with the correct answer Nâilî. The experts stated that the difficulty level of the items with the same correct answer did not significantly vary, and they were equivalent.

In the second stage of examination, Cosine Similarity Index (CSI) was used to determine the similarity of automatically generated items. The CSI value of the items generated with the one-layer model within the scope of the research was between 0.66 and 0.99; the CSI value of the items generated with the n-layered model was calculated between 0.27 and 0.99.

4. DISCUSSION and CONCLUSION

The advantages of computer-based applications, such as portability, usability, and lack of time and space restrictions, have paved the way for the widespread use of computer applications in the field of measurement and evaluation (Weber et al., 2003). Computer-based test systems bring innovations in the form of item preparation as well as the content they offer, the application environment and the evaluation dimension of the tests (Parshall et al., 2002). In the test development process with the help of computers, automatic systems that can generate different items from the same item pattern have gained importance (Gierl & Haladyna, 2012). One of these innovations is AIG, where computers are integrated into the test development process. In this research, we introduce how to automatically generate high-stakes test items for university entrance exams in Türkiye. Although AIG practices are largely in English (Embretson & Yang, 2006; Kosh et al., 2019; Lai et al., 2016), there are also practices in Korean (Choi et al., 2018; Gierl et al., 2021) and German (Arendasy & Sommer, 2012). The results of this research are important in terms of showing that AIG is both suitable for Turkish language items and can be used in high-stakes tests.

Additionally, in the literature, it is seen that there are studies in the field of medicine and mathematics in template-based automatic item generation (Colvin, 2014; Lai et al., 2016; Singley & Bennett, 2002; Sun et al., 2019). AIG should be used in psychological testing areas where cognitive models are involved, and individuals' reasoning skills should also be measured (Hommel et al., 2022; Sun et al., 2019; Yang et al., 2021) and cognitive ability items can be developed in the reasoning areas (Freund et al., 2008; Poinstingl, 2009). In the previous study, AIG was used to generate items in the field of Turkish literature. To indicate that automatic item generation is suitable for test development in different disciplines, automatic item generation is introduced through a literature item within the scope of 2018 AYT in this study. AIG can also be used in different disciplines.

A cognitive model was created in the first stage of AIG, and an item model was developed in the second stage. After the third stage, 320 items in a one-layer model and 2560 items in a n-layer model were automatically generated. In the first and second stages of template-based AIG, experts such as item writer, SME, measurement, and evaluation expert, as in traditional item writing, took part. This study demonstrates that unlike the traditional item writing process, the

SMEs were not involved in writing or examining the item, but in the process of creating the models that would form the item and examining these models. Gierl et al. (2016) stated that experts are still needed in the AIG process. However, it takes a very long time and intensive labor for an expert group of 5 experts to create 2560 items. Studies show that AIG significantly reduces the cost in terms of time, labor, and cost in the test development process (Alves et al., 2010; Gierl et al., 2021; Kosh et al., 2019).

Since evaluating the generated items is important, items were examined based on SMEs opinions and the CSI in this study. According to SMEs opinion, items with the same correct answer show equivalent characteristics; in the items in which different poets were asked, it was determined that items with different difficulty indexes were generated. Similarly based on the results obtained after the pre-application, Gierl et al. (2016) determined that the generated items are in different difficulty ranges. Ryoo et al. (2022) also stated that the item difficulties of the generated items are the similar and different. Therefore the result of this study is in consistency with the fundamental aim of AIG (Sinharay & Johnson, 2005). This shows that the items generated with AIG can be used in tests prepared for different purposes (summative, formative and diagnostic assessment or in-class and large-scale etc.), in other words, it is a common area of use.

It has been determined that the CSI value of the items generated with the 1-layered model CSI value from 0,66 to 0,99, while the n-layered model is calculated between 0,27 and 0,99. It means that the items generated by the n-layer model are less similar to each other and they are not clones. This result is consistent with other research results (Gierl & Lai, 2012). It is expected that the CSI values of the items generated with the n-layered model are low, and it is recommended to use the n-layered model for AIG (Gierl & Lai, 2013).

In this study, the examination of the automatically generated items was carried out by SMEs' opinions and by calculating the cosine similarity index. Pre-tests of the automatically generated items can be made, and the results of the pre-application and examination can also be carried out by calibrating. Using them is just as important as creating a large pool of items. Further studies can be conducted on the use of the item pool. In this study, template-based AIG approach is introduced. Artificial intelligence-based AIG studies can also be conducted. Simultaneously, international studies in automatic item generation depend on the changing data type (Chen et al., 2019), calibration shapes (Bai, 2019), and difficulty estimation methods (Chen et al., 2019; Gierl et al., 2016).

Effectiveness in classroom tests can be increased by creating item pools based on AIG on the EBA platform, which is widely used in schools created by the MoNE. A measurement and evaluation system can be created that all teachers can contribute and use in this process. In large-scale test applications, mobile computer-based applications can be disseminated, and AIG can be used in applications.

Acknowledgments

This research was supported by 2219-International Postdoctoral Research Fellowship Program for Turkish Citizens and the Scientific and Technological Research Council of Türkiye.

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, E-77082166-604.01.02-609178.

Authorship Contribution Statement

Ayfer Sayin: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Mark J. Gierl:** Methodology, Supervision, and Validation.

Orcid

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

REFERENCES

Adıgüzel, A. (2020). Teachers' views on distance education and evaluation of student success in the pandemic process. *Milli Eğitim Dergisi*, *49*(1), 253-271. https://doi.org/10.37669/milliegitim.781998

Alves, C.B., Gierl, M.J., & Lai, H. (2010). Using automated item generation to promote principled test design and development. *American Educational Research Association, Denver, CO, USA*.

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.

Artırma, F., & Hareketi, T.İ. (2020). *FATİH projesi.* Retrieved from: http://fatihprojesi.meb.gov.tr

Bai, Y. (2019). *Cognitive Diagnostic Models-based Automatic Item Generation: Item Feature Exploration and Calibration Model Selection*. Columbia University.

Balta, Y., & Türel, Y. (2013). An examination on various measurement and evaluation methods used in online distance education. *Turkish Studies-International Periodical For The Languages, Literature and History of Turkish or Turkic*, *8*(3), 37-45. http://dx.doi.org/10.7827/TurkishStudies.427

Bayardo, R.J., Ma, Y., & Srikant, R. (2007). Scaling up all pairs similarity search. Proceedings of the 16th international conference on World Wide Web.

Bennett, R.E. (2011). Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, *18*(1), 5-25.

Chen, B., Zilles, C., West, M., & Bretl, T. (2019). Effect of discrete and continuous parameter variation on difficulty in automatic item generation. Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, Proceedings, Part I 20,

Choi, J., Kim, H., & Pak, S. (2018). Evaluation of Automatic Item Generation Utilities in Formative Assessment Application for Korean High School Students. *Journal of Educational Issues*, *4*(1), 68-89.

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.

Clark, C.M., & Rust, F.O.C. (2006). Learning-centered assessment in teacher education. *Studies in Educational Evaluation*, *32*(1), 73-82.

Colvin, K.F. (2014). *Effect of automatic item generation on ability estimates in a multistage test*. University of Massachusetts Amherst.

Corbett, A.T., & Anderson, J.R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. Proceedings of the SIGCHI conference on Human factors in computing systems,

Davey, T. (2011). A Guide to Computer Adaptive Testing Systems. *Council of Chief State School Officers*.

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

Freund, P.A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied psychological measurement*, *32*(3), 195-210.

Gaytan, J., & McEwen, B.C. (2007). Effective online instructional and assessment strategies. *The American journal of distance education*, *21*(3), 117-132.

Gierl, M.J., & Haladyna, T.M. (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.

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.

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.

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

Gierl, M.J., Shin, J., Firoozi, T., & Lai, H. (2022). Using content coding and automatic item generation to improve test security. Frontiers in Education,

Gutl, C., Lankmayr, K., Weinhofer, J., & Hofler, M. (2011). Enhanced Automatic Question Creator--EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education. *Electronic Journal of e-Learning*, *9*(1), 23-38.

Haladyna, T.M., & Rodriguez, M.C. (2013). Developing and validating test items.

Higgins, D. (2007). Item Distiller: Text retrieval for computer-assisted test item creation. *Educational Testing Service Research Memorandum (RM-07-05). Princeton, NJ: Educational Testing Service*.

Higgins, D., Futagi, Y., & Deane, P. (2005). Multilingual generalization of the ModelCreator software for math item generation. *ETS Research Report Series*, *2005*(1), i-38.

Hommel, B.E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *psychometrika*, *87*(2), 749-772.

Irvine, S., & Kyllonen, P. (2002). *Generating items for cognitive tests: Theory and practice*. Erlbaum.

Kaptan, S. (1998). *Bilimsel araştırma teknikleri ve istatiksel yöntemleri.* Tekışık Ofset Tesisleri.

Karatay, H., & Dilekçi, A. (2019). Competencies of turkish teachers in measuring and evaluating language skills. *Milli Eğitim Dergisi*, *48*(1), 685-716.

Kınalıoğlu, İ.H., & Güven, Ş. (2011). Issues and solutions on measurement of student achievement in distance education. *XIII. Akademik Bilişim Konferansı Bildiriler*, 637-644.

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., 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.

Lane, S., Raymond, M.R., Haladyna, T.M., & Downing, S.M. (2015). Test development process. In *Handbook of test development* (pp. 19-34). Routledge.

MEB. (2018). *Ortaöğretim Türk dili ve edebiyatı dersi (9, 10, 11 ve 12. sınıflar) öğretim programı*. http://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=353

MEB. (2020). *EBA Yeni Dönem*. https://yegitek.meb.gov.tr/www/egitim-bilisim-aginin-eba-yeni-donem-lansmani-gerceklesti/icerik/2999

MEB. (2021). *Çevrim içi sınav*. https://www.meb.gov.tr/yuz-yuze-egitime-bir-haftalik-aradan-sonra-devam/haber/24621/tr

ÖSYM. (2018). *AYT örnek soruları*. https://www.osym.gov.tr/TR,13680/2018.html

ÖSYM. (2022). *E-YDS*. https://www.osym.gov.tr/TR,25238/2023.html

Özyalçın, K.E., & Kana, F. (2020). An evaluation on the skills of writing sub-text questions of teachers of Turkish as a foreign language. *Çukurova University Journal of Turkology Research (ÇÜTAD)*, *5*(2), 488-506.

Parshall, C.G., Spray, J.A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. Springer Science & Business Media.

Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychological Test and Assessment Modeling*, *51*(2), 123.

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

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*, *12*(2), 21582440221095016.

Sarı, T., & Nayır, F. (2020). Education in the pandemic period: Challenges and opportunities. *Electronic Turkish Studies*, *15*(4). http://dx.doi.org/10.7827/TurkishStudies.44335

Saygı, H. (2021). Problems encountered by classroom teachers in the covid-19 pandemic distance education process. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *7*(2), 109-129. https://doi.org/10.51948/auad.841632

Singley, M.K., & Bennett, R.E. (2002). *Item generation and beyond: Applications of schema theory to mathematics assessment*. Generating Items for Cognitive Tests: Theory and Practice., Nov, 1998, Educational Testing Service, Princeton, NJ, US; This chapter was presented at the aforementioned conference.,

Sinharay, S., & Johnson, M. (2005). Analysis of Data from an Admissions Test with Item Models. Research Report. ETS RR-05-06. *ETS Research Report Series*.

Sun, L., Liu, Y., & Luo, F. (2019). Automatic generation of number series reasoning items of high difficulty. *Frontiers in Psychology*, *10*, 884.

TDV. (2022). Türk İslam Ansiklopedisi. In https://islamansiklopedisi.org.tr/

TEDMEM. (2020). *2020 eğitim değerlendirme raporu* (TEDMEM Değerlendirme Dizisi [2020 education evaluation report], Issue.

Weber, B., Schneider, B., Fritze, J., Gille, B., Hornung, S., Kühner, T., & Maurer, K. (2003). Acceptance of computerized compared to paper-and-pencil assessment in psychiatric inpatients. *Computers in Human Behavior*, *19*(1), 81-93.

Yang, A.C., Chen, I.Y., Flanagan, B., & Ogata, H. (2021). Automatic generation of cloze items for repeated testing to improve reading comprehension. *Educational Technology & Society*, *24*(3), 147-158.

Zhu, M., Liu, O.L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668.

APPENDIX: SAMPLES OF THE GENERATEDITEMS IN TURKISH

Tek katmanlı modelle üretilen maddeler	Çok katmanlı modelle üretilen maddeler
1. XVII. yüzyılda yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde Sebk-i Hindî üslubunu kullanan önemli şairler arasında gösterilir. Şairin Türkçe ve Farsça divanı bulunmaktadır. Kendisinden sonra gelen pek çok şairi etkilemiştir. Divan şiiri geleneğinde kaside şairi olarak kabul edilen şair; şiirlerinde dinî tasavvufî konuların dışına çıkmış, yazarlık gücünü nesib yerine fahriye ile başlayan kasideleriyle kanıtlamıştır. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Nef'î* B) Nesîmî C) Nedim D) Şeyh Gâlip E) Nâbî	**1.** IV. Murad, Sultan İbrâhim ve IV. Mehmet arasındaki dört farklı padişah döneminde yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde dinî tasavvufî konuların dışına çıkan şair, Sebk-i Hindî üslubunu ustalıkla kullanan önemli şairler arasında gösterilir. Şairin şiirlerini topladığı bir divanı vardır. Manzum tarzda kaleme aldığı peygamberler tarihi eseri ile özdeşleşen şair, dönemindeki ve kendisinden sonra gelen pek çok şairi etkilemiştir. Divan şiirinde gazel ustası olarak adından söz ettirmiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Bâkî B) Nâbî C) Şeyh Galip D) Neşatî* E) Nedim
127. IV. Mehmet'ten III. Ahmet'e dört padişah döneminde yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde hikmet ve darbımeselleri ustalıkla kullanan önemli şairler arasında gösterilir. Şairin Türkçe ve Farsça divanı bulunmaktadır. Divan şiiri geleneğinde hikemî şairi olarak kabul edilir. Şiirlerinde dinî tasavvufî konuları ele alan şair, yazarlık gücünü Feridüddin Attar'ın bir eserinden esinlenerek kaleme aldığı mesnevisi ile ispat etmiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Neşati B) Necati Bey C) Nâbî* D) Gülşehrî E) Nef'î	1368. Şiirlerinde dinî tasavvufî konuları ele alan şair, şarkılarında yerli ve mahallî söyleyişleri ustalıkla kullanan şairler arasında yer alır. XVII. yüzyılda yaşamış divan edebiyatı şairlerinden biridir. Pek çok şair tarafından şiirlerine nazireler yazılmıştır. Yaşadığı dönemde ve divan edebiyatında tasavvufî gazel şairi olarak ün yapmıştır. Şairlik maharetini sosyal yaşamdaki gerçekliği konu alan suriyye kasidesi ile göstermiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Fuzûlî B) Nâilî* C) Nedim D) Şeyh Galip E) Nef'î
205. XVII. yüzyılda yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde zengin mecazları ve ihtişamlı ahenkleri ustalıkla kullanan önemli şairler arasında gösterilir. Şairin Türkçe ve Farsça divanı bulunmaktadır. Divan şiiri geleneğinde hiciv şairi olarak kabul edilir. Şiirlerinde dinî tasavvufî konuların dışına çıkan şair, yazarlık gücünü hiciv alanında dönemini aşan bir eseri ile ispat etmiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Nesîmî B) Nedim C) Şeyh Galip D) Nâbî E) Nef'î*	2054. Şiirlerinde dinî tasavvufî konuları ele alan şair, öğüt ifadelerini ustalıkla kullanan şairler arasında gösterilir. XVII. yüzyılda yaşamış divan edebiyatı şairlerinden biridir. Yazmadaki ustalığını çocuklar için yazdığı mesnevisi ile göstermiştir. Türkçe ve Farsça divanı bulunan şair, divan edebiyatında hikemî şairi olarak bilinir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Neşati B) Necati Bey C) Gülşehrî D) Nâbî* E) Nef'î