RESEARCH ARTICLE

• 2025 Article No: 39

Turkish Journal of Psychiatry (2025) **36**:39
https://doi.org/10.5080/u27540

Clinical Reasoning in Psychiatric Education: Development of Multiple-Choice Questions with Automatic Item Generation in Turkish

Esra EMEKLİ[1], Emre EMEKLİ[2], Yavuz Selim KIYAK[3], Yasemin HOŞGÖREN ALICI[4], Özlem COŞKUN[5], Işıl İrem BUDAKOĞLU[6]

ABSTRACT

Objective: This study aims to evaluate the suitability of the automatic item generation (AIG) for producing Turkish case-based multiple-choice questions (MCQs) in psychiatry.

Method: The study was planned as a descriptive study. In the first stage, topics were determined and a cognitive model was created by subject matter experts. In the second stage, a question template was created, variables were determined, the format of answer options was organized, and two equivalent templates of question content with different combinations were created. In the final stage, questions were generated using Python-based software based on these models. Following the question generation, random samples were selected and evaluated by experienced educators using a structured form.

Results: A total of 1189 questions were generated, with 11 questions sampled for each diagnosis. In the evaluation conducted by experts, six of the questions were deemed appropriate for each parameter, while minor corrections were suggested for five questions. It was stated that all the questions assess clinical reasoning skills rather than factual recall.

Conclusion: The template-based AIG method allows for the rapid and effective production of high-quality questions needed in medical education. The study demonstrated that AIG in the Turkish language for generating MCQs that assess clinical reasoning is applicable in the field of psychiatry. This method enables the production of a large number of questions in a short time, enriched with various combinations.

Keyword: Automated Item Generation, Clinical Reasoning, Medical Education, Multiple Choice Question, Psychiatry Education

INTRODUCTION

Clinical reasoning is defined as the process in which a physician acquires information, synthesizes it, generates hypotheses, and subsequently develops a clinical method, prognosis, diagnosis, treatment plan, and patient care plan (Durning et al. 2013). In addition, clinical reasoning can also be described as the mental processes a physician undertakes when encountering a problem in the diagnosis or treatment of a disease (Cate et al. 2018). Given its critical nature, it is essential for medical students to improve clinical reasoning skills during their education (Schmidt and Mamede 2015). To assess these skills, various methods have been developed in medical education (Daniel et al. 2019).

Clinical reasoning skills can be assessed using the evaluation methods at the "knows how," "shows how," and "does" levels of Miller's widely accepted four-level (knows - knows how - shows - does) pyramid in medical education (Miller 1990). However, assessment activities targeting Miller's "shows how" and "does" levels for preclinical students are less frequently applied because they require a clinical environment. Therefore, many methods aiming to measure clinical reasoning focus on the clinical phase. Written exams, test assessments, and problem-based learning (PBL) methods come to the forefront

How to cite: Emekli E, Emekli E, Kıyak YS, et al. (2025) Clinical Reasoning in Psychiatric Education: Development of Multiple-Choice Questions with Automatic Item Generation in Turkish. *Turk Psikiyatr Derg* **36**:39. https://doi.org/10.5080/u27540

Received: 12.06.2024, **Accepted:** 29.11.2024, **Available Online Date:** 07.04.2025

[1]Psychiatrist, [4]Assoc. Prof., Başkent University, Faculty of Medicine, Department of Psychiatry, Ankara; [2]Assis. Prof., Eskişehir Osmangazi University, Faculty of Medicine, Department of Radiology, Eskişehir; [3]Assis. Prof., [5]Assoc. Prof., [6]Prof., Gazi University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Türkiye.

e-mail: emreemekli90@gmail.com

for evaluating the clinical reasoning skills of preclinical students before they encounter patients (Hawks et al. 2023). In PBL, students face complex problems taken from real life or fictional scenarios; this student-centered educational method contributes to the development of clinical reasoning skills while attempting to solve these problems. With this method, it is possible to evaluate clinical reasoning skills more deeply and practically (Manuaba et al. 2022). However, issues such as the need for more instructors for small group sessions, instructor shortages, and insufficient student motivation lead to the more frequent use of test assessments for evaluating clinical reasoning in our country (Rochmawati and Wiechula 2010). Test assessments allow students to be tested simultaneously, standards to be established, and reliable scores to be obtained.

To create multiple-choice questions (MCQs) aimed at measuring higher-level skills like clinical reasoning instead of rote memorization, the most up-to-date medical information should be considered by integrating it with clinical practice. However, the half-life of medical knowledge is continuously and rapidly increasing (Densen 2011). Therefore, it may be difficult for medical educators to keep up with the latest developments in question production. Producing MCQs is also time-consuming, and applications that can automate this process are highly valuable for medical educators. There are two main methods in the automation of MCQ production: template-based automatic question generation (AIG) and non-template-based AQG (Gierl et al. 2021). Both of these methods provide a more efficient process than the traditional MCQ writing process.

Non-template-based methods rely more on artificial intelligence (AI) tools. Although promising results have been achieved in terms of efficiency in generating MCQs using AI (Cheung et al. 2023, Emekli and Karahan 2024, Kıyak and Emekli 2024, Laupichler et al. 2023, Zuckerman et al. 2023), significant problems arise due to AI's ability to produce low-quality and inconsistent information (Deng et al. 2023, Walker et al. 2023) and even occasionally fabricate scientific sources (Masters 2023). Moreover, because AI is a "black box"—meaning that humans do not yet fully understand how outputs generated by complex algorithms are formed—it is not possible to have complete control over its outputs. This leads to errors and inconsistencies that cannot be easily corrected. Although it requires more effort than methods using AI, the problem of misinformation is less when generating MCQs using template-based methods, and because the mechanism of creating MCQs through templates is directly observable and fully controllable by humans, errors can be easily corrected (Gierl et al. 2021). It is stated that if appropriate cognitive models are used, the efficiency of this method and the psychometric properties of the generated questions are similar to traditional methods (Pugh et al. 2020). However, this method requires more effort and time compared to AI-based methods.

Gierl et al. (2012) describe the template-based AIG process in three stages. In the first step, the content required for question generation is determined by subject matter experts. This content is expressed as a cognitive model that emphasizes the knowledge, skills, and problem-solving processes that physicians need to reach a specific medical diagnosis. In the second step, a question model (template) is developed, into which the cognitive model content is to be incorporated to create new questions. This question model is created in the form of a template containing the variables (question content and answer choices) for each new question generated. The words or phrases to be assigned to these variables in the template are also determined. In the final step, computer-based algorithms can generate hundreds of questions from a single item model (Gierl and Lai 2013, Gierl and Lai 2015). The successful use of AIG has been demonstrated in various languages (Gierl et al. 2021). In medical education, there are only two studies focused on generating MCQs in Turkish using the template-based AIG method, and both are on the topic of hypertension (Kıyak et al. 2023a, Kıyak et al. 2023b).

Psychiatry, like other medical fields, requires a thorough knowledge of semiology for accurate diagnosis and appropriate treatment. However, psychiatric semiology is particularly complex, and the diagnostic systems (Diagnostic and Statistical Manual of Mental Disorders - DSM-5, International Classification of Disease - ICD-11) are continually updated, leading to constant changes in diagnoses and terminologies (Rejon 2012). Unlike other specialties, biomarkers such as laboratory tests and imaging are less frequently used as diagnostic tools in psychiatry. This makes diagnosis and differential diagnosis more challenging. Psychiatric interviews and case evaluations are crucial for diagnosis. One of the current approaches adopted in medical education in recent years is the teaching model that progresses from symptom to disease (Moghadami et al. 2021). In this model, students are first expected to identify symptoms and mentally visualize possible diagnoses that could present with these symptoms. In our country, this approach is also supported within the framework of the National Core Education Program (UÇEP), which forms the basis of medical school curricula (National Working Group 2020). In this respect, case-based education is even more important in imparting diagnostic and differential diagnostic skills to students in the field of psychiatry. As far as is known, there is no study conducted in this field that allows for question generation in the field of psychiatry.

This study aims to accelerate the process of producing high-quality questions needed in medical education and to evaluate the feasibility of automatically generating questions that assess clinical reasoning skills.

METHOD

Study Design

The study was planned as a descriptive study. Since it does not involve human participants or sensitive personal data, it does not require ethics committee approval. The study focused on the development of automatically generated questions, and the evaluation was conducted through anonymous expert opinions.

Question Generation

First Stage: Development of the cognitive model.

First, it was decided that psychiatry questions aimed at medical students would be prepared by a physician with a doctorate degree in medical education (author 2) and a psychiatry resident physician (author 1). At this stage, the topics of the questions to be prepared were determined based on the National Core Education Program (National Working Group, 2020) and the curriculum of the Faculty of Medicine at Başkent University. The topics were composed of a total of 15 diagnoses included in the DSM-5 (American Psychiatric Association, 2013). Three diagnostic groups were determined according to their common and prominent clinical features (containing mood episodes / accompanied by somatic symptoms / containing psychotic symptoms) as the first, second, and third diagnostic groups:

First diagnostic group (depression, dysthymia, bipolar disorder, postpartum depression, schizoaffective disorder)

Second diagnostic group (conversion disorder, somatic symptom disorder, illness anxiety disorder, factitious disorder, factitious disorder imposed on another)

Third diagnostic group (brief psychotic disorder, schizophreniform disorder, schizophrenia, substance-induced psychosis, postpartum psychosis)

The symptom clusters are also based on the core symptoms specified in the diagnostic criteria of the DSM-5. After listing the common symptoms for each diagnostic group, a core set of symptoms was created through random assignments from these symptoms. Then, questions aimed at finding the most appropriate diagnosis were designed by adding suitable sociodemographic data, duration, and additional distinguishing features. This approach aims to develop students' clinical reasoning skills in line with the symptom-to-disease teaching model adopted in medical education (Moghadami et al. 2021). Diagnoses and symptom groupings were made by bringing together diagnoses that share similar and related symptoms to facilitate question generation. Therefore, although difficulties were encountered in grouping some diagnoses, these groups were pragmatically formed to produce questions appropriate for fifth-year medical students that facilitate moving from symptom to diagnosis, rather than creating an absolute classification.

In the next stage, patient information (age, gender, alcohol/substance use status) and disease information (symptoms, history, duration) were determined as information sources. Based on these determined information sources, appropriate age ranges, suitable gender, and alcohol/substance use status (does not use substances/socially consumes alcohol and does not use substances/consumes alcohol daily and frequently uses substances) were specified for each diagnosis. Possible symptoms were listed for each diagnosis. Similar symptoms were grouped to create symptom clusters (Symptom A, B, C and D) for each diagnosis. Similarly, disease histories were written for the diagnoses with the aim of covering more than one diagnosis. Histories were grouped (History A, B, C, D). Finally, taking into account the DSM-5 diagnostic criteria for each disease, information was added regarding how long the disease has been present in the scenario to be medically appropriate.

At the end of this stage, the diagnosis of schizoaffective disorder was excluded from question generation because, although it has mood symptoms like the diagnoses in its group, it stands out with its psychotic features and is difficult to include commonly in the question template. Similarly, factitious disorder, illness anxiety disorder, and factitious disorder imposed on another were also excluded from question generation because the determinants leading to these diagnoses cause complexity that hinders AIG in the question stem. These four diagnoses were used only as distractor options in the answer choices. As a result, it was decided to generate questions from the remaining 11 diagnoses. After the initial draft of the cognitive model was created, it was reviewed by a psychiatry faculty member (author 4) and a medical doctor with a doctorate in medical education (author 3). The symptoms and histories were reviewed by the two authors, and some corrections were made. With these adjustments, the final version of the cognitive model was finalized (Table 1).

Second Stage: Creation of the Question Model (Template)

The aim was to format the cognitive model into a suitable format for question generation. Initially, a sample question (parent item) was created. From this sample question, a question template was developed. The content of the question template was divided into five parts (reason for consultation, history, duration, alcohol/substance use information, question sentence). To reduce the similarity between the generated questions, two different ways of phrasing were written for each part of the question content, ensuring they conveyed the same meaning. This allowed for the same expression to be obtained using different words. For each question to be generated, it was aimed to use one of these templates to create the questions (Table 2).

Table 1. Components of the Cognitive Model

Disease Group	Diagnosis	Age	Sex	Symptom A	Symptom B	Symptom C	Symptom D	History A	History B	History C	History D	Duration
Group 1	Depression	18-65	Female/ Male	No	No	Yes	No	Yes	No	No	No	1-12 month
	Dysthymia	18-65	Female/ Male	No	No	Yes	No	Yes	No	No	No	2-4 years
	Bipolar Disorder	18-65	Female/ Male	No	No	Yes	No	No	No	No	Yes	1-24 month
	Postpartum Depression	18-35	Female	No	No	Yes	No	No	Evet	No	No	2-3 weeks
Group 2	Somatization	18-65	Female/ Male	Yes	No	No	No	No	No	Yes	No	1-24 month
	Conversion Disorder	18-65	Female/ Male	No	No	No	Yes	No	No	Yes	No	1-24 month
Group 3	Brief Psychotic Disorder	18-65	Female/ Male	No	Yes	No	No	Yes	No	No	No	1-29 days
	Schizophreniform Disorder	18-65	Female/ Male	No	Yes	No	No	Yes	No	No	No	1-5 month
	Schizophrenia	18-65	Female/ Male	No	Yes	No	No	Yes	No	No	No	6-24 mounth
	Substance-Induced Psychosis	18-65	Female/ Male	No	Yes	No	No	Yes	No	No	No	1-24 month
	Postpartum Psychosis	18-35	Female	No	Yes	No	No	No	Yes	No	No	2-3 weeks
Symptom A	sweating, palpitations, hot flashes, shortness of breath, feeling of faintness, trembling in the hands, nausea, feeling of distress, abdominal pain											
Symptom B	thoughts being stolen, thinking he is a prophet, being told what he should do, thoughts being inserted into his brain, hearing things being whispered into his ear, hearing voices from somewhere, thinking he will be harmed, talking to himself											
Symptom C	feeling sad, frequently crying, feeling helpless, feeling empty, experiencing hopelessness, taking no pleasure in anything, not wanting to engage in daily activities, having no appetite, lacking expectations for the future, feeling pessimistic, not wanting to get out of bed, thinking they are worthless, and wanting to die.											
Symptom D	inability to speak, fainting, weakness in the legs, weakness in both arms, numbness in the hands, inability to swallow, difficulty speaking, inability to walk, inability to see, inability to hear, inability to make sounds, and inability to maintain balance											
History A	for some time, they have not wanted to leave the house, and their self-care has declined.											
History B	she recently gave birth.											
History C	She/he presented to the emergency department with similar complaints and was told that she does not have any medical illness.											
History D	At times, she/he sleeps very little, feels so energetic that she/he cannot sit still, and goes through periods of excessive spending.											

In the study, three diagnostic groups were initially determined based on common clinical features: first diagnostic group, second diagnostic group, and third diagnostic group. In the questions created, it was aimed to assign the five diagnoses from the respective diagnostic group as answer choices for each scenario.

Third Stage: Question Generation Using Software

At this stage, it was necessary to create software to generate the questions based on the information presented in the previous steps. Therefore, the authors, with the help of a software developer, created a one-time code based on Python 3.1 (Rossum and Drake 1995). In this way, the words describing the symptoms and histories given in Table 1 were automatically assigned to the relevant places in the templates provided in Table 2, generating the questions. The answer choices were also determined as expressed in the previous stage (Table 3).

Evaluation of Generated Questions

Due to the large number of generated questions, a sample from the generated questions was selected using Statistical Package for the Social Sciences (SPSS version 22.0) to include each diagnosis. The selected questions were shared with experienced psychiatrists using an evaluation form (see the supplementary file) created by the authors. This evaluation form was based on existing forms in the literature (Pugh et al. 2020). The evaluation criteria consisted of six statements with "Yes/No" options in the first five:

1. The question text is clear.
2. The question is clinically appropriate.

Table 2. Question Model

QUESTION CONTENT TO BE GENERATED: [Reason for Admission] [History] [Duration] [Alcohol] [Question Sentence]		
	Template 1	Template 2
Reason for Admission	<AGE> yaşındaki <SEX> hasta, <SYMPTOM1> ve <SYMPTOM2> şikayetleri ile psikiyatri polikliniğine başvuruyor.	<SYMPTOM1> ve <SYMPTOM2> yakınmaları nedeniyle ruh sağlığı ve hastalıkları uzmanına başvuran <AGE> yaşındaki <SEX> hasta.
History	Alınan öyküsünde hasta < HISTORY> ifade ediyor.	Öyküsü, hastanın <HISTORY> ortaya çıkarıyor.
Duration	Bahsettiği şikayetlerin <DURATION> gibi bir süredir olduğunu söylüyor.	Hasta, yakınmalarının < DURATION > gibi bir süredir mevcut olduğunu belirtiyor.
Alcohol and Substance Use	Alkol ve madde kullanımı sorgulandığında <ALCOHOL> ifade ediyor.	Alkol ve madde kullanıp kullanmadığı sorulduğunda <ALCOHOL> belirtiyor.
Question Sentence	Bu hastada, aşağıda verilen tanılardan hangisi diğerlerine göre daha muhtemeldir?	Aşağıdakilerden hangisi bu hasta için diğerlerine göre en olası tanıdır?

Table 3. Example Questions Generated for the Diagnosis of Depression

Question 1 25 yaşındaki kadın hasta, üzüntülü hissetme ve sık sık ağlama şikayetleri ile psikiyatri polikliniğine başvuruyor. Öyküsü, hastanın bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ortaya çıkarıyor. Hasta, yakınmalarının 6 ay gibi bir süredir mevcut olduğunu belirtiyor. Alkol ve madde kullanımı sorgulandığında kullanmadığını ifade ediyor. Bu hastada, aşağıda verilen tanılardan hangisi diğerlerine göre daha muhtemeldir?
Panik bozukluk
Distimi
Postpartum depresyon
Bipolar bozukluk
Depresyon
Question 2 İştahsızlık ve karamsar olma yakınmaları nedeniyle 19 yaşındaki kadın hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Öyküsü, hastanın bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ortaya çıkarıyor. Hasta, yakınmalarının 10 ay gibi bir süredir mevcut olduğunu belirtiyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Aşağıdakilarden hangisi bu hasta için diğerlerine göre en olası tanıdır?
Panik bozukluk
Postpartum depresyon
Distimi
Depresyon
Bipolar bozukluk
Question 3 Ölmek isteme ve günlük aktiviteleri yapmak istememe yakınmaları nedeniyle 49 yaşındaki kadın hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Alınan öyküsünde hasta bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ifade ediyor. Hasta, yakınmalarının 2 ay gibi bir süredir mevcut olduğunu belirtiyor. Alkol ve madde kullanımı sorgulandığında kullanmadığını ifade ediyor. Aşağıdakilarden hangisi bu hasta için diğerlerine göre en olası tanıdır?
Depresyon
Distimi
Panik bozukluk
Bipolar bozukluk
Postpartum depresyon

3. The question has a single correct answer.
4. The information provided is sufficient to find the correct answer.
5. The distractors are plausible.
6. The question primarily aims to assess (please mark one of the following options): Factual knowledge ( ) Clinical reasoning skills ( )

Additionally, the evaluators were asked to provide any comments or suggestions they had about the questions.

RESULTS

A total of 1189 questions were generated. From these, 11 questions (see the supplementary file) were sampled, one for each diagnosis. Three psychiatrists with experience in writing multiple-choice questions for medical schools evaluated the questions. The psychiatrists had an average of 7 years of clinical experience.

Questions 2, 3, 5, 9, 10, and 11 were rated as appropriate by all three evaluators for each parameter. The second

evaluator made suggestions for questions 1, 4, 6, and 8, while the third evaluator made suggestions for questions 4 and 7. For questions 1 and 8, the second evaluator answered "no" to the questions, "Does the question have only one correct answer?" and "Are the given details sufficient to find the correct answer?". The evaluators stated that all questions assess clinical reasoning skills.

Expert Comments:

Question 1: Evaluator 2 mentioned, "Considering the patient's age and gender, postpartum could also be considered a correct answer. Therefore, a higher age should be specified for female patients."

Question 4: Two evaluators made similar comments. Evaluator 3 noted, "To diagnose substance-induced psychosis, it would be more appropriate to clearly state the relationship between the timing of substance use and the onset of psychotic symptoms."

Question 6: Evaluator 2 stated, "Clearer information should be provided regarding the preoccupation with health and somatic symptoms."

Question 7: Evaluator 3 suggested, "To reach a conversion disorder diagnosis, more explicit exclusion criteria should be included."

Question 8: Evaluator 2 commented, "For a patient being evaluated for dysthymia and depression, it is suitable to specify whether there has been any previous manic episode."

DISCUSSION

Case-based learning holds a significant place in both practical and theoretical learning in psychiatric education (Morreale 2019). Various methods, including written scenarios, video demonstrations, patient simulations, or observing real patients through a one-way mirror, are employed to allow medical students to observe psychiatric cases. This helps them develop diagnostic and differential diagnostic skills (Hassoulas 2017). At the same time, case examples are also used as an assessment tool in theoretical and practical exams. These skills are further assessed using case examples in both theoretical and practical exams. Given the psychiatric education structure in medical schools in Turkey (frequency of internships, number of questions, etc.), this creates a substantial demand for questions.

This study aimed to generate Turkish case-based MCQs that assess clinical reasoning in psychiatry. To achieve this, a cognitive model was first established, followed by the creation of a question template. Using a Python-based script (Rossum and Drake 1995), 1189 questions were generated automatically. This approach has made it possible for the first time to produce a large number of case-based questions for different clinical scenarios, which can be used both in exams and for students' exam preparations in psychiatry training.

The evaluation of the 11 sampled questions was conducted by psychiatrists with experience in question preparation. As a result, six questions were deemed completely suitable for use in exams based on the evaluation form parameters. One evaluator suggested revisions for four questions, and two evaluators suggested revisions for one question. Although some questions required modifications, this study demonstrated for the first time that it is feasible to generate Turkish case-based MCQs that assess clinical reasoning in psychiatry using template-based AIG. Moreover, the suggestions provided by evaluators can be addressed by modifying any part of the questioning the template. This finding indicates that the fundamental algorithm for question generation is functioning correctly. In the literature, successful implementation of AIG for question generation has been shown in English, French, Chinese, Spanish, and Korean (Gierl et al. 2021). However, no studies utilizing template-based AIG in generating multiple-choice questions in psychiatry have been found. Previous studies have focused on surgery, pharmacology, neonatal jaundice, upper gastrointestinal bleeding, liver disease in adults, and emergency medicine. This study aimed to overcome this challenge by generating a template and having experienced psychiatry physicians evaluate the generated questions. In conclusion, this study is significant as it demonstrates the applicability of AIG in the field of psychiatry, thus opening new avenues for efficient question generation in psychiatry training.

Evaluators unanimously indicated that all questions assessed clinical reasoning skills. Writing high-quality MCQs is a resource-intensive task for medical faculties. The literature states that developing MCQs aimed at high-level assessments, particularly those measuring clinical reasoning skills, consumes a significant portion of medical educators' time (Schuwirth and van der Vleuten 2004). Given the vast number of exams and course content in medical schools, the demand for questions from question banks is very high (Wrigley et al. 2012). Additionally, considering the clinical responsibilities of doctors in educator positions, especially in Turkey, the time spent on preparing questions adds a substantial workload. Therefore, the ability to produce a large number of exam-ready questions in a short time is crucial. This study paves the way for generating a large volume of high-quality questions quickly.

Another significant contribution of this study to the literature is the diversification of question stems by using different combinations in the question content. In studies conducted on generating Turkish MCQs using AIG methods, it is observed that the question stems and answer options are always the same (Kıyak et al. 2023a, Kıyak et al. 2023b). This study also aimed to reduce the likelihood of questions

querying the same diagnosis being similar to each other. This similarity was minimized by using variations in question templates and different symptom/history combinations. Additionally, two equivalent templates were created, each divided into five subheadings containing two variations (reason for admission, history, duration, alcohol-substance use, question sentence). Thus, even for a question querying the same diagnosis, a total of 32 different formats of question content were created. However, it is inevitable that questions belonging to the same diagnosis share some common features. No special algorithm was used to measure the similarity of the questions. In future studies, it would be beneficial to develop algorithms that measure and minimize question similarity using advanced statistical techniques and natural language processing methods.

The question groups in this study were based on core symptoms such as the presence of mood disorder symptoms, somatic symptoms, and psychotic features. With this basic logic and question generation technology, it was tested over only three symptom groups that case-based question generation is possible in psychiatry. However, considering the rich diagnostic groups and common symptomatology in psychiatry, these groups can be expanded in future comprehensive studies.

In addition to the strengths of the study, there are some limitations. First, only 11 questions were randomly selected and evaluated to represent each diagnosis among the generated questions. This approach allowed us to make a general evaluation using a sample question for each diagnosis. However, this limited number of questions may not fully represent the entire question pool. This is due to limited resources during the research process and the time-consuming nature of the evaluation process. Especially, due to the time constraints of expert evaluators, it was not practically possible to examine more questions. Additionally, since the study serves as a methodological preliminary assessment, it was intended in the initial phase to determine the general suitability and quality level of the questions with a smaller sample. Secondly, variations in question templates and different symptom/history combinations were used to minimize the similarity of the generated questions. Nevertheless, it is inevitable that questions belonging to the same diagnosis share some common features. In future studies, algorithms that measure and minimize question similarity using advanced statistical techniques and natural language processing methods can be developed to further reduce these similarities.

Another limitation is the small number of evaluators; however, all evaluators are in teaching positions at universities and have experience in question preparation. Also, a specific algorithmic method (Lai et al. 2016) was not used to enhance the quality of distractors in forming the answer choices. However, this issue was addressed by forming groups for psychiatric diseases with common symptom sets and assigning other diagnoses in the common symptom set to the answer choices. Additionally, anxiety disorders and mood disorders with psychotic features were not included in this study. Anxiety disorders increase the complexity of the cognitive model and question templates due to the overlap of their symptoms with other diagnostic groups and containing numerous subtypes. Similarly, the inclusion of mood disorders with psychotic features was excluded from the scope of the study because it would significantly expand symptom combinations and possible diagnosis options. This is an important limitation of our study, and including these diagnostic groups in future research would be beneficial. Moreover, the diagnosis and symptom groupings were made for pragmatic purposes and may not fully comply with specific classification systems; however, this approach was adopted in line with educational objectives. Lastly, the questions were not administered to students. Therefore, psychometric properties such as the difficulty index—which indicates the proportion of correct answers by students—and the discrimination index—which shows how well the question can differentiate between high-performing and low-performing students (Tavakol et al. 2024)—were not evaluated.

This study has made a significant contribution to the Turkish medical education literature in psychiatry by demonstrating that the template-based AIG method can be used to rapidly and effectively generate case-based questions that assess clinical reasoning in the field of psychiatry. Furthermore, providing diversity in question content by creating question stems with different combinations allows for the enrichment of question banks and the adoption of a more dynamic and comprehensive approach in the evaluation of students.

REFERENCES

American Psychiatric Association. (2013) Diagnostic and statistical manual of mental disorders (5th ed.).

Cate O, Custers E, Durning SJ (2018) Principles and Practice of Case-based Clinical Reasoning Education: A Method for Preclinical Students. Cham, Springer Copyright s. 75-83.

Cheung BHH, Lau GKK, Wong GTC et al. (2023) ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLOS ONE 18: e0290691.

Daniel M, Rencic J, Durning SJ et al. (2019) Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med 94: 902-12.

Deng J, Zubair A, Park Y-J (2023) Limitations of large language models in medical applications. Postgrad Med J 99: 1298–9.

Densen P (2011) Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 122: 48-58.

Durning SJ, Artino AR, Jr., Schuwirth L et al. (2013) Clarifying assumptions to enhance our understanding and assessment of clinical reasoning. Acad Med 88: 442-8.

Emekli E, Karahan BN (2024) Comparison of Automatic Item Generation Methods in the Assessment of Clinical Reasoning Skills. Revista Española de Educación Médica 6: (1).

Gierl MJ, Lai H, Turner SR (2012) Using automatic item generation to create multiple-choice test items. Med Educ 46: 757–65.

Gierl MJ, Lai H (2013) Evaluating the quality of medical multiple-choice items created with automated processes. Med Educ 47: 726-33.

Gierl M, Lai H (2015) Using Automated Processes to Generate Test Items And Their Associated Solutions and Rationales to Support Formative Feedback. Interact Des Archit(s) 25: 9-20.

Gierl MJ, Lai H, Tanygin V (2021) Advanced Methods in Automatic Item Generation, 1. Baskı. New York, Routledge, s. 42-66.

Hassoulas A, Forty E, Hoskins M et al. (2017) A case-based medical curriculum for the 21st century: The use of innovative approaches in designing and developing a case on mental health. Med Teach 39: 505-11.

Hawks MK, Maciuba JM, Merkebu et al. (2023) Clinical Reasoning Curricula in Preclinical Undergraduate Medical Education: A Scoping Review. Acad Med 98: 958-65.

IBM Corporation Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Kıyak YS, Coşkun Ö, Budakoğlu Iİ et al. (2023a) Psychometric Analysis of the First Turkish Multiple-Choice Questions Generated Using Automatic Item Generation Method in Medical Education. Tıp Eğitimi Dünyası 22: 154–61.

Kıyak YS, Budakoğlu Iİ, Coşkun Ö et al. (2023b) The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. Tıp Eğitimi Dünyası. 22: 72-90.

Kıyak YS, Emekli E (2024) ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J 6: qgae065.

Lai H, Gierl MJ, Touchie C et al. (2016) Using Automatic Item Generation to Improve the Quality of MCQ Distractors. Teach Learn Med 28: 166-73.

Laupichler MC, Rother JF, Grunwald Kadow IC et al. (2023) Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions. Acad Med 99: 508–12.

Manuaba IBAP, No Y, Wu CC (2022) The effectiveness of problem based learning in improving critical thinking, problem-solving and self-directed learning in first-year medical students: A meta-analysis. PLoS One 22: e0277339.

Masters K (2023) Medical Teacher's first ChatGPT's referencing hallucinations: Lessons for editors, reviewers, and teachers. Med Teach 45: 673–5.

Miller GE (1990) The assessment of clinical skills/competence/performance. Acad Med 65 (Suppl 9): 63-7.

Moghadami M, Amini M, Moghadami M et al. (2021) Teaching clinical reasoning to undergraduate medical students by illness script method: a randomized controlled trial. BMC Med Educ 21: 87.

Morreale MK (2019) Teaching and Learning Through the Use of Patient Cases Approach to the Psychiatric Patient: Case-Based Essays, Second Edition. By John W. Barnhill. Acad Psychiatry 43: 311.

Pugh D, De Champlain A, Gierl M et al. (2020) Can automated item generation be used to develop high quality MCQs that assess application of knowledge? Res Pract Technol Enhanc Learn 15: 12.

Rejón AC (2012) Logic structure of clinical judgment and its relation to medical and psychiatric semiology. Psychopathology 45: 344-51.

Rochmawati E, Wiechula R (2010) Education strategies to foster health professional students' clinical reasoning skills. Nurs Health Sci 12: 244-50.

Schmidt HG, Mamede S (2015) How to improve the teaching of clinical reasoning: a narrative review and a proposal. Med Educ 49: 961-73.

Schuwirth LW, van der Vleuten CP (2004) Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 38: 974-9.

Tavakol M, O'Brien DG, Sharpe CC et al. (2024) Twelve tips to aid interpretation of post-assessment psychometric reports. Medical Teacher 46: 188-95.

Ulusal Çalışma Grubu Ulusal Cep-2020, Ulusal Yetkınlık Ve Yeterlıkler Calısma Grubu Ulusal Cep-2020, Davranıs Sosyal Beserı Bılımler Calısma Grubu Ulusal Cep-2020 (2020) Medical Faculty - National Core Curriculum 2020. Tıp Eğitimi Dünyası 19: 1-146.

Van Rossum G, Drake FL (1995) Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Walker HL, Ghani S, Kuemmerli C et al. (2023) Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 25: e47479.

Wrigley W, van der Vleuten CP, Freeman A et al. (2012) A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. Med Teach 34: 683-97.

Zuckerman M, Flood R, Tan RJB et al. (2023) ChatGPT for assessment writing. Med Teach 45: 1224–7.

Question Evaluation Form

Question: 1

Takip edildiği ve beynine fikir sokulduğu yakınmaları nedeniyle 42 yaşındaki kadın hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Öyküsü, hastanın bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ortaya çıkarıyor. Bahsettiği şikayetlerin 14 gün gibi bir süredir olduğunu söylüyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Aşağıdakilerden hangisi bu hasta için diğerlerine göre en olası tanıdır?
Şizofreni
Kısa Psikotik Bozukluk
Postpartum Psikoz
Şizofreniform Bozukluk
Madde Kullanımına Bağlı Psikoz

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		

Question Evaluation Form

Question: 2

39 yaşındaki erkek hasta, bir yerlerden sesler duyma ve peygamber olduğunu düşünme şikayetleri ile psikiyatri polikliniğine başvuruyor. Alınan öyküsünde hasta bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ifade ediyor. Hasta, yakınmalarının 2 ay gibi bir süredir mevcut olduğunu belirtiyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Bu hastada, aşağıda verilen tanılardan hangisi diğerlerine göre daha muhtemeldir?
Madde Kullanımına Bağlı Psikoz
Şizofreniform Bozukluk
Kısa Psikotik Bozukluk
Postpartum Psikoz
Şizofreni

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		

Question Evaluation Form

Question: 3

Kulağına bir şeyler söylendiği ve beynine fikir sokulduğu yakınmaları nedeniyle 25 yaşındaki erkek hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Öyküsü, hastanın bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ortaya çıkarıyor. Bahsettiği şikayetlerin 15 ay gibi bir süredir olduğunu söylüyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Bu hastada, aşağıda verilen tanılardan hangisi diğerlerine göre daha muhtemeldir?
Madde Kullanımına Bağlı Psikoz
Kısa Psikotik Bozukluk
Postpartum Psikoz
Şizofreni
Şizofreniform Bozukluk

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		

Question Evaluation Form

Question: 4

Bir yerlerden sesler duyma ve takip edildiği yakınmaları nedeniyle 59 yaşındaki kadın hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Alınan öyküsünde hasta bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ifade ediyor. Bahsettiği şikayetlerin 8 ay gibi bir süredir olduğunu söylüyor. Alkol ve madde kullanımı sorgulandığında her gün alkol tükettiğini ve sıklıkla madde kullandığını ifade ediyor. Aşağıdakilerden hangisi bu hasta için diğerlerine göre en olası tanıdır?
Şizofreniform Bozukluk
Şizofreni
Postpartum Psikoz
Madde Kullanımına Bağlı Psikoz
Kısa Psikotik BozuklukPsikoz

	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question text is understandable.		

The question primarily aims to assess (please mark one of the two options below):
Factual learning ( )
Clinical reasoning skills ( )

If you have any additional comments regarding the question, please specify:

Question Evaluation Form

Question: 5

Bir yerlerden sesler duyma ve beynine fikir sokulduğu yakınmaları nedeniyle 31 yaşındaki kadın hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Alınan öyküsünde hasta yakın zamanda doğum yaptığını ifade ediyor. Bahsettiği şikayetlerin 2 hafta gibi bir süredir olduğunu söylüyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Aşağıdakilarden hangisi bu hasta için diğerlerine göre en olası tanıdır?
Postpartum Psikoz
Madde Kullanımına Bağlı Psikoz
Kısa Psikotik Bozukluk
Şizofreni
Şizofreniform Bozukluk

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		

The question primarily aims to assess (please mark one of the two options below):
Factual learning ( )
Clinical reasoning skills ( )

If you have any additional comments regarding the question, please specify:

Question Evaluation Form

Question: 6

47 yaşındaki erkek hasta, ellerde titreme ve terleme şikayetleri ile psikiyatri polikliniğine başvuruyor. Öyküsü, hastanın benzer yakınmalarla acil servise başvurduğunu ve kendisine tıbbi bir hastalığının olmadığının söylendiğini ortaya çıkarıyor. Hasta, yakınmalarının 6 ay gibi bir süredir mevcut olduğunu belirtiyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Bu hastada, aşağıda verilen tanılardan hangisi diğerlerine göre daha muhtemeldir?
Bakım Verenin Yapay Bozukluğu
Hastalık Kaygısı Bozukluğu
Yapay Bozukluk
Konversiyon Bozukluğu
Bedensel Belirti Bozukluğu

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		

The question primarily aims to assess (please mark one of the two options below):
Factual learning ( )
Clinical reasoning skills ( )

If you have any additional comments regarding the question, please specify:

Question Evaluation Form

Question: 7

Yürüyememe ve ellerde hissizlik yakınmaları nedeniyle 20 yaşındaki erkek hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Alınan öyküsünde hasta benzer yakınmalarla acil servise başvurduğunu ve kendisine tıbbi bir hastalığının olmadığının söylendiğini ifade ediyor. Bahsettiği şikayetlerin 16 ay gibi bir süredir olduğunu söylüyor. Alkol ve madde kullanımı sorgulandığında sosyal içici olarak alkol tükettiğini ve madde kullanmadığını ifade ediyor. Aşağıdakilerden hangisi bu hasta için diğerlerine göre en olası tanıdır?

Konversiyon Bozukluğu
Yapay Bozukluk
Hastalık Kaygısı Bozukluğu
Bedensel Belirti Bozukluğu
Bakım Verenin Yapay Bozukluğu

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		

Question Evaluation Form

Question: 8

Hiçbir şeyden zevk almama ve iştahsızlık yakınmaları nedeniyle 20 yaşındaki kadın hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Alınan öyküsünde hasta bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ifade ediyor. Bahsettiği şikayetlerin 5 ay gibi bir süredir olduğunu söylüyor. Alkol ve madde kullanımı sorgulandığında kullanmadığını ifade ediyor. Aşağıdakilerden hangisi bu hasta için diğerlerine göre en olası tanıdır?

Bipolar bozukluk
Postpartum depresyon
Panik bozukluk
Depresyon
Distimi

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		

Question Evaluation Form

Question: 9

43 yaşındaki erkek hasta, üzüntülü hissetme ve yataktan çıkmak istememe şikayetleri ile psikiyatri polikliniğine başvuruyor. Öyküsü, hastanın bir süredir evden dışarı çıkmak istemediğini, özbakımının azaldığını ortaya çıkarıyor. Hasta, yakınmalarının 2 yıl gibi bir süredir mevcut olduğunu belirtiyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Aşağıdakilerden hangisi bu hasta için diğerlerine göre en olası tanıdır?

Panik bozukluk
Distimi
Bipolar bozukluk
Depresyon
Postpartum depresyon

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		

Question Evaluation Form

Question: 10

Sık sık ağlama ve çaresiz hissetme yakınmaları nedeniyle 53 yaşındaki kadın hasta ruh sağlığı ve hastalıkları uzmanına başvuruyor. Alınan öyküsünde hasta bazı zamanlar az uyuduğunu, yerinde duramayacak kadar çok enerjik hissettiğini ve aşırı para harcadığı dönemlerin olduğunu ifade ediyor. Hasta, yakınmalarının 6 ay gibi bir süredir mevcut olduğunu belirtiyor. Alkol ve madde kullanımı sorgulandığında kullanmadığını ifade ediyor. Bu hastada, aşağıda verilen tanılardan hangisi diğerlerine göre daha muhtemeldir?
Bipolar bozukluk
Distimi
Postpartum depresyon
Depresyon
Panik bozukluk

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		

Question Evaluation Form

Question: 11

22 yaşındaki kadın hasta, günlük aktiviteleri yapmak istememe ve gelecekten bir beklentisi olmadığı şikayetleri ile psikiyatri polikliniğine başvuruyor. Öyküsü, hastanın yakın zamanda doğum yaptığını ortaya çıkarıyor. Bahsettiği şikayetlerin 2 hafta gibi bir süredir olduğunu söylüyor. Alkol ve madde kullanıp kullanmadığı sorulduğunda kullanmadığını belirtiyor. Aşağıdakilerden hangisi bu hasta için diğerlerine göre en olası tanıdır? Depresyon
Bipolar bozukluk
Postpartum depresyon
Distimi
Panik bozukluk

Evaluation Criteria	Yes	No
The question text is understandable.		
The question is clinically appropriate.		
The question has only one correct answer.		
The information provided in the question is sufficient to find the correct answer.		
The distractors are logical.		
The question primarily aims to assess (please mark one of the two options below): Factual learning ( ) Clinical reasoning skills ( )		
If you have any additional comments regarding the question, please specify:		