Araştırma Doi: 10.25282/ ted.1779377

Psychometric Evaluation of Automatically Generated Template-Based Psychiatry Questions for Medical Students: A Validity and Reliability Study

Tıp Öğrencileri İçin Şablon Tabanlı Otomatik Olarak Üretilmiş Psikiyatri Sorularının Geçerlik ve Güvenirliği: Bir Psikometrik Değerlendirme

Esra Emekli[1]
Orcid: 0000-0001-5937-6270
Rabia Soylu[2]
Orcid: 0009-0005-6971-1710
Emre Emekli[3]
Orcid: 0000-0001-5989-1897
Yavuz Selim Kıyak[4]
Orcid: 0000-0002-5026-3234
Yasemin Hoşgören Alıcı[5]
Orcid: 0000-0003-3384-8131
Özlem Coşkun[4]
Orcid: 0000-0001-8716-1584
Işıl İrem Budakoğlu[4]
Orcid: 0000-0003-1517-3169
[1]Eskişehir Şehir Hastanesi, Psikiyatri Kliniği, Eskişehir, Türkiye
[2]Gazi Üniversitesi Tıp Fakültesi, Ankara, Türkiye
[3]Eskişehir Osmangazi Üniversitesi Tıp Fakültesi Radyoloji Ana Bilim Dalı, Eskişehir, Türkiye
[4]Gazi Üniversitesi Tıp Fakültesi Tıp Eğitimi ve Bilişimi Ana Bilim Dalı, Ankara, Türkiye
[5]Başkent Üniversitesi Tıp Fakültesi Psikiyatri Ana Bilim Dalı, Ankara, Türkiye.

Sorumlu Yazar:
Emre Emekli
E-posta:
emreemekli90@gmail.com

Anahtar Sözcükler:
Çoktan seçmeli soru, Otomatik soru üretimi, Psikiyatri eğitimi, Ölçme-Değerlendirme
Keywords:
Multiple Choice Questions, Automatic Question Generation, Psychiatry Education, Educational Measurement

Gönderilme Tarihi / Submitted:
09.09.2025
Kabul Tarihi / Accepted:
26.10.2025

Künye:
Emekli E, Soylu R, Emekli E, Kiyak YS, Alici YH, Coskun O, Budakoglu II. Psychometric Evaluation of Automatically Generated Template-Based Psychiatry Questions for Medical Students: A Validity and Reliability Study. World of Medical Education, 2025;24(74):209-216

Abstract

Background: Multiple-choice questions (MCQs) are widely used in medical education due to their objectivity and broad coverage of knowledge. Case-based MCQs, in particular, offer significant advantages in assessing students' clinical reasoning and decision-making skills. Psychiatry education presents unique challenges because of subjective symptom descriptions and overlapping features among diagnoses. The aim of this study was to evaluate the psychometric properties (difficulty and discrimination indices) of psychiatry MCQs generated through template-based automatic item generation (AIG).

Materials and Methods: The study was conducted with the approval of the Ethics Committee of Gazi University. Among 224 students who completed the psychiatry clerkship during the 2023–2024 and 2024–2025 academic years, 138 (61.6%) voluntarily participated. From a pool of 1,189 previously generated questions using template-based AIG, 22 were randomly selected to create the exam. The test was administered in a classroom under supervision, and students were not informed of the origin of the items. Difficulty indices were calculated as the proportion of correct responses, while discrimination indices were determined by comparing the performance of the top 27% and bottom 27% of students. In addition, differences between upper and lower groups were tested using the Mann–Whitney U test. Corrected item–total correlation (CITC) values were computed for each item, and overall test reliability was assessed with Cronbach's alpha (equivalent to KR-20 for dichotomous items).

Results: The mean exam score was 15.21 ± 3.55 out of 22. The average difficulty index was 0.69, classifying the exam as "easy." Of the items, 63.6% were very easy, 9.1% easy, and 27.3% moderate. The most difficult item concerned somatization (0.33), while the easiest item was related to bipolar disorder (0.92). Discrimination indices ranged from 0.19 to 0.70, with an average of 0.37. Ten items (45.6%) showed very good discrimination, eleven (50%) were acceptable, and one (4.5%) was weak. The highest discrimination was observed in the schizophreniform disorder item (0.70), and the lowest in the postpartum psychosis item (0.19). Mann–Whitney U tests indicated significant differences between upper and lower groups for all items ($p < 0.05$). CITC values ranged from 0.08 to 0.49, with a mean of 0.30 and a median of 0.34. Eleven items (50.0%) were strong, eight (36.4%) acceptable, and three (13.6%) weak. The overall reliability of the test was acceptable, with Cronbach's alpha = 0.74.

Conclusion: This study represents the first implementation of template-based AIG MCQs in Turkish psychiatry education. The psychometric findings demonstrated that the items were acceptable in terms of validity and reliability. These results also show that questions previously evaluated by expert review functioned similarly when administered to students. Future studies applying template-based AIG across multiple centers, larger samples, and in comparison with faculty-authored items may provide stronger evidence for its use in medical education.

Özet

Amaç: Tıp eğitiminde çoktan seçmeli sorular (ÇSS), objektiflikleri ve geniş bilgi kapsamı nedeniyle yaygın olarak kullanılmaktadır. Özellikle olgu temelli ÇSS'ler, öğrencilerin klinik muhakeme ve karar verme becerilerini ölçmede önemli avantajlar sağlamaktadır. Psikiyatri eğitimi, subjektif semptom tarifleri ve tanılar arası semptom örtüşmeleri nedeniyle özgün zorluklar içerir. Bu çalışmanın amacı, şablon tabanlı otomatik soru üretimi (OSÜ) ile üretilen psikiyatri alanındaki ÇSS'lerin psikometrik özelliklerinin (güçlük ve ayırt edicilik indeksleri) değerlendirilmesidir.

Gereç ve Yöntem: Çalışma Gazi Üniversitesi Etik Kurulu onayıyla yürütülmüş, 2023–2024 ve 2024–2025 akademik yıllarında psikiyatri stajını tamamlayan 224 öğrenciden 138'i (%61,6) gönüllü olarak katılmıştır. Daha önce şablon tabanlı OSÜ yöntemiyle üretilen 1189 soru arasından rastgele seçilen 22 soruluk bir sınav oluşturulmuştur. Sınav sınıf ortamında gözetmen eşliğinde uygulanmış, öğrencilere soruların kaynağı açıklanmamıştır. Soruların güçlük indeksleri doğru cevaplanma oranı ile, ayırt edicilik indeksleri ise üst %27 ve alt %27'lik öğrenci gruplarının performansları karşılaştırılarak hesaplanmıştır. Ayrıca, alt ve üst gruplar arasında fark olup olmadığı Mann–Whitney U testi ile incelenmiş, her bir madde için düzeltilmiş madde–toplam korelasyonu değerleri hesaplanmış ve sınavın genel güvenirliği Cronbach alfa katsayısı (KR-20 eşdeğeri) ile değerlendirilmiştir.

Bulgular: Sınavın genel ortalama puanı 22 üzerinden 15,21 ± 3,55 bulunmuştur. Soruların ortalama güçlük indeksi 0,69 olup sınav genel olarak "kolay" kategorisinde değerlendirilmiştir. Soruların %63,6'sı çok kolay, %9,1'i kolay, %27,3'ü orta zorluktaydı. En zor soru somatizasyon (0,33), en kolay soru ise bipolar bozukluk (0,92) ile ilişkiliydi. Ayırt edicilik indeksleri 0,19–0,70 arasında değişmekte olup ortalama değer 0,37 idi. On soru (%45,6) çok iyi ayırt edicilik gösterirken, 11 soru (%50) kabul edilebilir, bir soru (%4,5) ise zayıf kategorisindeydi. En yüksek ayırt edicilik şizofreniform bozukluk (0,70), en düşük ise postpartum psikoz (0,19) sorusunda bulundu. Mann–Whitney U testleri tüm maddelerde üst ve alt gruplar arasında anlamlı fark olduğunu göstermiştir ($p<0.05$). Düzeltilmiş madde–toplam korelasyon değerleri 0,08–0,49 arasında değişmiş, ortalama 0,30 ve ortanca 0,34 bulunmuştur. On bir madde (%50) güçlü, sekiz madde (%36,4) kabul edilebilir, üç madde (%13,6) ise zayıf düzeydedir. Sınavın genel güvenirliği Cronbach alfa = 0,74 olarak hesaplanmıştır.

Sonuç: Bu çalışma, Türkçe psikiyatri eğitimi özelinde OSÜ ile üretilen ÇSS'lerin öğrenciler üzerinde ilk uygulamasını temsil etmektedir. Elde edilen psikometrik veriler, bu soruların geçerlik ve güvenirlik açısından kabul edilebilir düzeyde olduğunu göstermektedir. Bu bulgular, daha önce uzman görüşüyle değerlendirilen soruların öğrenci uygulamasında da benzer şekilde işlevsel olduğunu ortaya koymaktadır. Gelecekte, OSÜ ile üretilen soruların farklı merkezlerde, daha geniş örneklemlerde ve uzman yazımıyla karşılaştırmalı biçimde incelenmesi, yöntemin tıp eğitiminde kullanılabilirliği hakkında daha güçlü kanıtlar sağlayacaktır.

INTRODUCTION

Owing to its dynamic nature, the field of medical education has been proactive in integrating innovative approaches into educational processes to improve teaching and assessment. Today, a wide range of conventional and innovative assessment tools can be used in medical education (1–3). However, due to multiple advantages, assessment with multiple-choice questions (MCQs) frequently emerges as the most commonly used method. MCQs are recognized as a key assessment tool for their efficiency, objectivity, and capacity to succinctly cover a broad spectrum of knowledge (4). At the same time, MCQs can be designed to address broader educational objectives that include developing critical thinking skills and clinical decision-making ability (5). Over time, as emphasis has increased on the practicality of clinical knowledge and the importance of problem-solving within medical education, there has been growing interest in case-based MCQs rather than those probing lower-level skills such as rote memorization. Such questions offer learners opportunities to engage with clinical scenarios they may encounter in practice, thereby facilitating the integration of theoretical knowledge with practical application (4). The clinical decision-making process involves not only recall of facts but also multilayered cognitive operations such as assessing the patient's contextual features, prioritizing, and choosing among alternative diagnoses (6). Therefore, structured MCQs aimed at measuring clinical decision-making provide a more effective means to evaluate students' readiness for real-life practice.

In the context of psychiatry education, MCQ development entails unique challenges. Symptom overlap across psychiatric diagnoses, the absence of objective diagnostic tools such as laboratory or imaging compared to other medical disciplines, subjective symptom descriptions, and the need for contextual interpretation may affect both the validity and the discriminative power of questions developed in this domain (7). Consequently, it is important in psychiatry to develop questions based on realistic case scenarios that reflect clinical reasoning skills.

Developing MCQs—particularly when producing case-based items—is a time-consuming, labor-intensive process. Thus, approaches that can automate MCQ generation are highly valuable for medical educators. From the perspective of automatic item generation (AIG), there are essentially two approaches (8): AI-based AIG (9) and template-based AIG (10). AI-based AIG refers to the use of natural language processing or large language models to generate test items in a data-driven manner. These systems typically learn patterns from large corpora and then create novel items with minimal human input. While this approach can produce a wide variety of items, the quality, accuracy, and alignment with curricular objectives may vary, requiring extensive expert review.

In contrast, template-based AIG relies on structured item models or templates in which variable elements are systematically substituted. A template usually specifies the clinical scenario, stem, and options, while placeholders allow controlled variation (e.g., diagnosis, symptom, laboratory finding). This approach ensures that generated questions adhere to a consistent logical and psychometric framework, reduces construct-irrelevant variance, and provides a transparent method for faculty to supervise and validate the content. For example, in psychiatry, a template can define a prototypical clinical vignette of a disorder, and by systematically exchanging symptom clusters or patient characteristics, multiple parallel items of equivalent difficulty and discrimination can be generated. Although AI-based AIG offers the advantage of rapid and large-scale item generation, its outputs may vary in accuracy, clinical plausibility, and psychometric quality, often requiring substantial expert revision. Template-based AIG, by contrast, produces items through controlled variation within predefined structures, ensuring that each question aligns with curricular objectives and maintains comparable difficulty and discrimination levels. This apparent repetitiveness—changing only a few parameters across similar items—is in fact an advantage, as it guarantees consistency, transparency, and reproducibility while still allowing clinically meaningful variability. For these reasons, template-based AIG was considered the more suitable approach for the present study, particularly as an initial empirical evaluation in psychiatry education.

In Turkish medical education, there are only two studies that generated case-based questions using the template-based AIG method, both conducted on hypertension (11,12). Additionally, in a study conducted by the authors in Turkish, it was shown for the first time in psychiatry that case-based MCQs could be produced using template-based AIG, and the items were evaluated through expert review (13). However, these items were not administered to students; therefore, basic psychometric properties (difficulty and discrimination indices) were not evaluated for compliance with standards.

The aim of this study is to administer to medical students the MCQs previously developed in psychiatry using a template-based AIG method and, using the resulting data, to evaluate the psychometric properties (difficulty and discrimination indices) of these items. In this respect, the study represents both the first implementation on students of an automatically generated exam comprising Turkish psychiatry questions and aims to contribute to the validity of such items from a measurement and evaluation standpoint.

MATERIALS AND METHODS

Prior to the study, approval was obtained from the Ethics Committee of Gazi University (Date: 13.02.2024, Approval No: 2024 - 459).

Study Design

This study was designed as a descriptive psychometric study.

Participants

The study was conducted with students enrolled at the Ankara Hospital of the Faculty of Medicine, Başkent University, during the 2023–2024 and 2024–2025 academic years. In the six-year medical curriculum, the psychiatry clerkship is undertaken in the fifth year. Therefore, students who were taking the psychiatry clerkship were invited to participate, and those who agreed were included. During the 2023–2024 and 2024–2025 academic years, a total of 224 students completed the psychiatry clerkship. Of these, 138 students (61.6%) volunteered to participate in the study.

Examination Format

In a prior study conducted by the authors, Turkish MCQs related to psychiatric disorders were generated using template-based AIG. These items were reviewed by psychiatry specialists, and minor revisions were made based on the feedback received. Following these revisions and feedback, necessary changes were implemented in item generation and in the item templates, and the questions were regenerated using the same techniques. The development and production process of the items is detailed in the previous study (13). From the total of 1,189 generated questions, two questions were randomly selected for each of the 11 diagnoses on which item generation was based. A 22-item exam was constructed from the selected questions. The exam comprised single-best-answer multiple-choice questions. For each cohort, the exam was administered face-to-face in a classroom setting under supervision to students who consented to participate. The source of the questions was not disclosed to students prior to the exam. Participation was voluntary, and the exam scores did not affect the determination of students' final grades for passing/failing the clerkship.

Data Analysis

For each item in the study, difficulty and discrimination indices were calculated. The overall difficulty level of the exam was also determined. Statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 26. The difficulty index was calculated as the proportion of participants who answered the item correctly. In this method, difficulty values range from 0 to 1, with higher values indicating easier items. In this study, items with values ≤ 0.29 were classified as difficult, 0.30–0.49 as moderate, 0.50–0.69 as easy, and 0.70–1.00 as very easy (14). The discrimination index was calculated by dividing students into two groups based on total test scores: the top 27% (high-achievement group) and the bottom 27%

(low-achievement group). This value was obtained by subtracting the number of correct answers in the low-achievement group from the number of correct answers in the high-achievement group and dividing the result by the group size. Values ≥0.30 were considered very good, 0.20–0.29 acceptable, and <0.20 weak (recommended for removal) (15). In addition, to test whether the differences between the upper and lower groups were statistically significant, Mann–Whitney U tests were conducted for each item. In addition, corrected item–total correlations (CITC) were computed for each item as a complementary indicator of item discrimination. Typically, large-scale standardized test developers require an item's point-biserial correlation to be at least 0.30 or higher for effectiveness. However, in locally developed classroom-type tests, values in the mid to high 0.20s could be considered satisfactory (15). Thus, CITC values in this range were also interpreted as acceptable in the present study. Additionally, the overall reliability of the test was evaluated using Cronbach's alpha coefficient (equivalent to KR-20 for dichotomous items). Score distributions for the entire exams were reported using descriptive statistics along with mean and standard deviation.

RESULTS

The overall mean exam score was calculated as 15.21 ± 3.55 out of 22 questions. The average item difficulty index was 0.69, indicating that the exam could be evaluated overall as "easy." Based on item distribution: six items (27.3%) were moderate (0.30–0.49), two items (9.1%) were easy (0.50–0.69), and 14 items (63.6%) were very easy (≥0.70). The most difficult item was Item 14 (Somatization) (0.33), whereas the easiest was Item 4 (Bipolar disorder) (0.92).

Discrimination indices ranged between 0.19 and 0.70. The average discrimination index was 0.37. Among the items, 10 (45.6%) exhibited very good discrimination (≥0.40), 11 (50%) were acceptable (0.20–0.39), and one (4.5%) fell into the weak (recommended for removal) category (<0.20). The item with the highest discrimination was Item 8 (Schizophreniform disorder) (0.70), whereas the lowest value was observed in Item 15 (Postpartum psychosis) (0.19). Mann–Whitney U tests indicated that the differences between the upper and lower 27% groups were statistically significant for all items ($p < 0.05$), further supporting the discriminative validity of the test items. Consistent with these findings, corrected item–total correlation (CITC) values ranged from 0.08 to 0.49 (M = 0.30±0.11). Using commonly cited thresholds (≥0.30 strong; 0.20–0.29 acceptable; <0.20 weak), 11 items (50.0%) were strong, 8 (36.4%) acceptable, and 3 (13.6%) weak. (Table 1). The internal consistency of the test was acceptable, with a Cronbach's alpha (KR-20) coefficient of 0.74.

DISCUSSION

The aim of this study was to determine the capacity of the template-based AIG method to produce psychometrically robust MCQs for undergraduate psychiatry education. With 138 clerkship students responding to a 22-item test form, it was found that the items produced had a mean difficulty of 0.69 and a mean discrimination of 0.37. Corrected item–total correlation (CITC) values ranged between 0.08 and 0.49, with a mean of 0.30 and a median of 0.34. While 11 items (50.0%) had strong CITC values (≥0.30), 8 items (36.4%) were acceptable (0.20–0.29), and 3 items (13.6%) were weak (<0.20). These findings suggest that although certain items may require revision or removal, the majority met acceptable thresholds, and the overall reliability of the test (Cronbach's alpha = 0.74) remained at an acceptable level. This level of difficulty can be considered appropriate for the psychiatry clerkship context, as it ensures that the items are neither too difficult to discourage students nor too easy to prevent meaningful discrimination. Although template-based AIG was among the earliest approaches, more recent item generation systems increasingly rely on AI-based or hybrid methods. The present study focused on template-based AIG to establish an empirical baseline in Turkish psychiatry education, where such data had not previously been available. We consider this a necessary first step, providing psychometric evidence under controlled conditions and allowing subsequent studies to build on this foundation by comparing template-based items with AI-supported or hybrid approaches.

To date, only one study has examined a psychiatry-specific template-based AIG method. Emekli et al. (2025) generated 1,189 Turkish MCQs and evaluated a randomly selected sample of items solely through expert review; the items were not administered to

Table 1. Psychometric data of the exam items

Question Number	Diagnosis / Correct Answer	Difficulty Index	Discrimination Index	p-value	Corrected Item–Total Correlation
1	Brief psychotic disorder	0.74	0.45	<0.001	0.35
2	Conversion disorder	0.38	0.45	<0.001	0.22
3	Schizophrenia	0.86	0.27	0,001	0.29
4	Bipolar disorder	0.92	0.20	0,003	0.34
5	Substance-induced psychosis	0.89	0.27	0,001	0.39
6	Depression	0.83	0.25	0,002	0.22
7	Postpartum depression	0.91	0.20	0,006	0.24
8	Schizophreniform disorder	0.67	0.70	<0.001	0.49
9	Schizophrenia	0.82	0.36	<0.001	0.35
10	Depression	0.74	0.50	<0.001	0.36
11	Brief psychotic disorder	0.81	0.43	<0.001	0.47
12	Bipolar disorder	0.88	0.30	<0.001	0.43
13	Substance-induced psychosis	0.86	0.27	0,001	0.34
14	Somatization disorder	0.33	0.42	<0.001	0.17
15	Postpartum psychosis	0.77	0.19	0,045	0.21
16	Dysthymia	0.39	0.54	<0.001	0.35
17	Postpartum depression	0.83	0.20	0,027	0.18
18	Conversion disorder	0.36	0.28	0,01	0.08
19	Postpartum psychosis	0.74	0.35	<0.001	0.22
20	Dysthymia	0.38	0.47	<0.001	0.22
21	Schizophreniform disorder	0.66	0.54	<0.001	0.37
22	Somatization disorder	0.42	0.50	<0.001	0.2

students (13). Similarly, in their review examining the feasibility and validity evidence of template-based AIG, Falcão et al. (16) addressed existing studies in the literature and highlighted the lack of psychometric data. By presenting psychometric data obtained from examinees, our study extends the preliminary evidence presented by Emekli et al. (2025) and provides the first empirical validation that questions generated via template-based AIG in psychiatry function as intended under examination conditions (13).

Although there are few studies on template-based AIG in psychiatry, insights from other fields are instructive. Emekli and Karahan compared template-based and non-template-based (artificial intelligence) AIG for questions on abdominal emergencies and found that both methods produced items of acceptable quality based on expert evaluation (17). Template-based AIG items outperformed faculty-written items in clarity and content validity. Our findings reflect this robustness. Despite symptom overlap across diagnoses and the narrative-rich nature of psychiatric cases, template-based AIG produced items with discrimination properties consistent with the aforementioned study.

In psychiatry education, where diagnostic and treatment processes largely depend on contextual evaluations and history-taking, case-based questions offer an ideal method to assess not only students' knowledge but also their clinical reasoning abilities (18). Despite the frequent symptom overlap across different diagnoses in psychiatry (7,19), producing case-based items via AIG enables a systematic approach, with templates supporting the consistent generation of discriminative, cross-diagnostic items capable of assessing this feature. Moreover, diagnostic processes that necessarily rely on verbal narratives can be substantially standardized, thereby ensuring consistency in item quality. In addition, AIG can overcome the time-consuming nature of original item writing by requiring experts only to oversee the templates, thus reducing academic workload (20).

Among the strengths of this study is the psychometric analysis conducted using real student data. Regarding limitations, first, the study was carried out at a single centre with a limited sample, which restricts the generalizability of the obtained psychometric findings. Second, only items generated via the AIG technique were produced. The absence of comparison with expert-authored items limits the ability to comment on the relative effectiveness of the system. In addition, this study did not classify the cognitive levels of the items (e.g., according to Bloom's taxonomy). Because the items were randomly selected from the pool, it was not possible to determine the distribution of knowledge-, comprehension-, or application-level questions. Future research incorporating such analyses could provide a more comprehensive understanding of the pedagogical as well as psychometric value of automatically generated items. Future studies involving multicentre collaborations with larger student samples and more comprehensive exams, as well as comparisons of template-based AIG items with faculty-authored items, may provide more robust evidence.

In conclusion, this study represents one of the first direct student implementations of an automatic question generation system in the context of Turkish psychiatry education. The findings obtained for both difficulty and discrimination demonstrate the potential of the system in the domain of assessment; however, they also indicate a need for further development to generate items that measure higher-order cognitive processes. Especially in disciplines with high cognitive complexity such as psychiatry, meeting such qualitative evaluation criteria with case-based MCQs highlights the potential of question generation systems beyond traditional methods in the educational assessment of clinical reasoning skills.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding

No financial support was received.

References

1. Rohlfsen CJ, Sayles H, Moore GF, Mikuls TR, O'Dell JR, McBrien S, et al. Innovation in early medical education, no bells or whistles required. BMC Med Educ. 2020;20:39.
2. Gordon M, Farnan J, Grafton-Clarke C, Ahmed N, Pelly T, Roberts M, et al. Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide No. 54. Med Teach. 2019;41:732–45.
3. Daniel M, Rencic J, Durning SJ, Torre D, King A, Gordon M, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med. 2019;94:902–12.
4. Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: making a continued case for the use of MCQs in medical education. Med Teach. 2019;41:569–77.
5. Zaidi NLB, Grob KL, Monrad SM, Schroeder R, Santen SA, Hughes DT, et al. Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? Acad Med. 2018;93:856–9.
6. Corrao S, Argano C. Rethinking clinical decision-making to improve clinical reasoning. Front Med (Lausanne). 2022;9:900543.
7. Rejón AC. Logic structure of clinical judgment and its relation to medical and psychiatric semiology. Psychopathology. 2012;45:344–51.
8. Gierl MJ, Lai H, Tanygin V. Advanced Methods in Automatic Item Generation. 1st ed. New York: Routledge; 2021. p.42–66.
9. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J. 2024;100:858–65.

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

11. Kıyak YS, Budakoğlu İİ, Coşkun Ö, Kaya S. The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. Tıp Eğitimi Dünyası. 2023;22:72–90.

12. Kıyak YS, Coşkun Ö, Budakoğlu İİ, Kaya S. Psychometric analysis of the first Turkish multiple-choice questions generated using automatic item generation method in medical education. Tıp Eğitimi Dünyası. 2023;22:154–61.

13. Emekli E, Emekli E, Kiyak YS, Alici YH, Coşkun Ö, Budakoğlu İİ. Assesment of Clinical Reasoning in Psychiatric Education: Development of Multiple-Choice Questions with Automatic Item Generation in Turkish. Turk Psikiyatri Derg. 2025;36:336-43.

14. Çalık M, Ayas A. Çözeltilerde kavram başarı testi hazırlama ve uygulama. Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 2003;14:1-17.

15. Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York: Routledge; 2009.

16. Falcão F, Costa P, Pêgo JM. Feasibility assurance: a review of automatic item generation in medical assessment. Adv Health Sci Educ Theory Pract. 2022;27:405–25.

17. Emekli E, Karahan S. A comparison of template-based and non-template-based automatic item generation for abdominal emergencies in medical education. [In press]

18. Messineo L, Allegra M. An educational model for undergraduate psychiatry students to promote clinical diagnostic reasoning. Procedia Soc Behav Sci. 2014;141:1309–14.

19. Gomes AI, Jesus S, Simões G, Vicente S. Symptomatological transversality and the absence of pathognomonic symptoms in psychiatry. Eur Psychiatry. 2023;66(Suppl 1):S998–9.

20. Vie JJ, Popineau F, Bruillard É, Bourda Y. A review of recent advances in adaptive assessment. In: Métais E, Meziane F, Saraee M, Sugumaran V, Vadera S, editors. Natural Language Processing and Information Systems. NLDB 2017. Lecture Notes in Computer Science, vol 10260. Cham: Springer; 2017. p.17–30. https://doi.org/10.1007/978-3-319-59569-6_2