Research Doi: 10.25282/ted.1376840

Psychometric Analysis of the First Turkish Multiple-Choice Questions Generated Using Automatic Item Generation Method in Medical Education

Tıp Eğitiminde Otomatik Soru Üretme Yöntemi Kullanılarak Oluşturulan İlk Türkçe Çoktan Seçmeli Soruların Psikometrik Analizi

Yavuz Selim Kıyak[*] (ORCID: 0000-0002-5026-3234)*

Özlem Coşkun[*] (ORCID: 0000-0001-8716-1584)*

Işıl İrem Budakoğlu[*] (ORCID: 0000-0003-1517-3169)*

Canan Uluoğlu[*] (ORCID: 0000-0003-0682-5794)*

[*]Gazi University Faculty of Medicine, Ankara, TÜRKİYE

Corresponding Author: Yavuz Selim KIYAK, E-Mail: yskiyak@gazi.edu.tr

Abstract

Aim: Automatic item generation is "a process of using models to generate items using computer technology". The use of automatic item generation typically involves one of three primary methods: syntax-based, semantic-based, and template-based. Non-template automatic item generation approaches leverage natural language processing techniques. A study showed the potential of using template-based automatic item generation to create high-quality multiple-choice questions for assessing clinical reasoning in Turkish, marking a first in the field. However, the findings of the study were based only on expert opinions, necessitating further research to examine the psychometric qualities of Turkish items. The aim of this study was to reveal psychometric characteristics of the first Turkish case-based multiple-choice questions generated by using automatic item generation in medical education.

Methods: This was a psychometric study. Three Turkish case-based multiple-choice questions generated using template-based automatic item generation on essential hypertension were included in an exam that 281 fourth-year medical students participate in. This examination was carried out in-person in classroom settings under proctor supervision. Item difficulty and item discrimination (point-biserial correlation) were calculated, and non-functioning distractors were determined.

Results: All three items had acceptable levels (higher than 0.20) of point-biserial correlation (p<0.001). The item difficulty levels indicated the presence of one easy, one moderate, and one difficult question. Each item had 2-3 non-functioning options among five options. All three items had acceptable levels (higher than 0.20) of point-biserial correlation (p<0.001). The item difficulty levels indicated the presence of one easy, one moderate, and one difficult question. Each item had 2-3 non-functioning options among five options.

Keywords:
Automatic Item Generation, Automatic Question Generation, Psychometric Analysis, Clinical Reasoning, Medical Education

Anahtar Sözcükler:
Otomatik Madde Üretimi, Otomatik Soru Üretimi, Psikometrik Analiz, Klinik Akıl Yürütme, Tıp Eğitimi

Gönderilme Tarihi
Submitted: 17.11.2023
Kabul Tarihi
Accepted: 15.12.2023

To cite this article: Kıyak YS, Coşkun O, Budakoğlu Iİ, Uluoğlu C. Psychometric Analysis of the First Turkish Multiple-Choice Questions Generated Using Automatic Item Generation Method in Medical Education. World of Medical Education. 2023;22(68):154-161

Conclusions: The results indicated that the items successfully discriminate between high and low performers, providing validity evidence on the quality of the questions in evaluating students' comprehension of the subject. Additionally, the findings suggest that it is feasible to create multiple-choice questions with different difficulty levels in Turkish using a single automatic item generation model. This study demonstrated for the first time that automatic generation of case-based multiple-choice questions in Turkish produces acceptable psychometric characteristics in an authentic assessment setting in medical education. The ability to automatically generate effective multiple-choice questions in Turkish holds promise for enhancing the efficiency of written assessment in Turkish medical education.

Özet

Amaç: *Otomatik soru (madde) üretimi "bilgisayar teknolojisini kullanarak madde üretmek için model kullanma süreci" olarak tanımlanır. Otomatik soru üretimi kullanımı tipik olarak üç temel yöntemden birini içerir: sözdizimi tabanlı, anlamsal tabanlı ve şablon tabanlı. Şablon tabanlı olmayan otomatik soru üretimi yaklaşımları doğal dil işleme tekniklerinden faydalanır. Bir çalışma, Türkçede klinik akıl yürütme becerisini değerlendirmek için yüksek kaliteli çoktan seçmeli sorular oluşturmak üzere şablon tabanlı otomatik soru üretimi yöntemi kullanımının potansiyelini göstererek bu alanda ilk çalışma olmuştur. Bununla birlikte, çalışmanın bulguları yalnızca uzman görüşlerine dayanmaktadır ve Türkçe soruların psikometrik niteliklerini incelemek için daha fazla araştırma yapılması gerekmektedir. Bu çalışmanın amacı, tıp eğitiminde otomatik soru üretimi kullanılarak oluşturulan ilk Türkçe olguya dayalı çoktan seçmeli soruların psikometrik özelliklerini ortaya koymaktır.*

Yöntem: *Bu çalışma psikometrik bir çalışmadır. Esansiyel hipertansiyon konusunda şablon tabanlı otomatik soru üretimi kullanılarak oluşturulan üç Türkçe olguya dayalı çoktan seçmeli soru, 281 dördüncü sınıf tıp öğrencisinin katıldığı bir sınava dahil edildi. Bu sınav gözetmen denetiminde sınıf ortamında yüz yüze gerçekleştirildi. Madde güçlüğü ve madde ayırt ediciliği (point-biserial korelasyon) hesaplandı ve işlevsel olmayan çeldiriciler belirlendi.*

Bulgular: *Her üç madde de kabul edilebilir düzeyde (0.20'den yüksek) point-biserial korelasyona sahipti (p<0.001). Madde güçlük düzeyleri bir kolay, bir orta ve bir zor sorunun varlığına işaret etmekteydi. Her madde beş seçenek arasında 2-3 işlevsel olmayan seçeneğe sahipti.*

Sonuç: *Bulgular, maddelerin yüksek ve düşük performans gösteren öğrencileri başarılı bir şekilde ayırt ettiğini ve öğrencilerin konuyu anlamalarını değerlendirmede soruların kalitesine ilişkin geçerlilik kanıtı sağladığını göstermiştir. Ayrıca bulgular, tek bir otomatik soru üretimi modeli kullanarak Türkçede farklı zorluk seviyelerine sahip çoktan seçmeli sorular oluşturmanın mümkün olduğunu göstermektedir. Bu çalışma, Türkçe olguya dayalı çoktan seçmeli soruların otomatik olarak oluşturulmasının tıp eğitiminde otantik bir değerlendirme ortamında kabul edilebilir psikometrik özellikler ürettiğini ilk kez göstermiştir. Kaliteli çoktan seçmeli Türkçe soruların otomatik olarak üretilebilmesi, Türkçe tıp eğitiminde yazılı ölçme-değerlendirmenin verimliliğini artırma konusunda umut vaat etmektedir.*

INTRODUCTION

Clinical reasoning stands out as a crucial skill in medical education, and evaluating this higher-order skill can be carried out through diverse assessment tools, including but not limited to multiple-choice questions (MCQs), key feature questions, script concordance tests, oral examinations, and more (1). Assessing clinical reasoning skills through case-based MCQs is both highly useful and common (2). However, writing process of high-quality items (questions) places a substantial demand on the resources of medical schools. Specifically, writing a single context-rich MCQ to evaluate higher-order skills can consume hours of a medical teacher's time (3). A vast quantity of MCQs are required in medical schools. Even a question bank designed solely for progress testing may need to incorporate thousands of questions (4). Given the extensive effort and resources involved in generating just one MCQ, the significant challenge for writing thousands of questions becomes apparent. To address this challenge, researchers developed automatic item generation (AIG) as "a process of using models to generate items using computer technology" (5).

The use of AIG typically involves one of three primary methods: syntax-based, semantic-based, and template-based (6). Non-template AIG approaches leverage natural language processing (NLP) techniques that demand advanced processing capacity and a corpus of data closely aligned with the test's purpose (7). In health professions education, a template-based method pioneered by Gierl et al. has been in use for over a decade (7). Grounded in cognitive models and question templates built by subject matter experts, this augmented intelligence approach employs software assistance to generate hundreds of MCQs at once (5).

The template-based AIG method encompasses three sequential stages (7): the formulation of a cognitive model, the establishment of an item model, and the generation of items using software. In the initial phase, a cognitive model is constructed to delineate the content for item generation. The structure of this cognitive model includes the problems and scenarios, sources of information regarding the problem, and features of the information. This model reveals the approach of subject matter experts when confronted with a clinical problem. Subsequently, in the second stage, an item model is formulated based on the previously established cognitive model. The item model comprises a stem, which is the question that students are expected to answer, and options. Within the item model, specific components of the stem are identified for manipulation by considering potential variables based on the cognitive model. Finally, in the third stage, a computer-based assembly system generates all plausible combinations via iterative processes.

The literature showed that only a few item models and a few hours of work could result in the generation of hundreds or even thousands of questions (8), not only in medical education but also in other health professions, such as pharmacy (9). This substantiates the template-based AIG's capability to produce high-quality MCQs that assess higher-order skills rather than factual recall (10). Case-based items created for clinical medicine using AIG exhibit psychometric properties, such as item difficulty and item discrimination, similar to those MCQs written in a traditional way (11). Furthermore, successful demonstrations of the template-based AIG method's efficacy in various languages, including English, French, Chinese, Spanish, and Korean, have been presented in the literature (7). Subsequently, for the first time in the literature, a study revealed that the use of AIG in Turkish to generate high-quality MCQs to assess clinical reasoning is possible (12). Subsequently, another study showed the feasibility of AIG in Turkish in a non-medical (Turkish literature exam for high-school graduates) area (13). However, these studies' findings were only based on expert opinion. Therefore, further research on Turkish items to examine psychometric characteristics is necessary.

The aim of this study is to reveal psychometric characteristics of the first Turkish MCQs generated by using AIG in medical education.

METHODS

Study Design

This was a psychometric study.

Setting and Participants

This research was conducted during the academic year 2023-2024 at Gazi University Faculty of Medicine, Ankara, Turkey. As part of the fourth year in the six-year undergraduate medical program, a series of five-day small group activities were implemented to train students on the fundamentals of rational prescribing. The training was based on the

WHO 6-Step Model (14). It used cases related to essential hypertension and type-2 diabetes mellitus in adults to teach rational drug prescribing. Subsequent to the training, students take a theoretical examination consisting of MCQs. All 281 fourth-year medical students who took the rational drug prescribing clerkship exam were eligible for participation in the study. Considering our goal to include all students, we opted not to conduct a formal sample size calculation. None of the students were excluded, 281 students who participated in the exam were included in the analysis.

Items

In a previous study, Turkish MCQs on essential hypertension were generated by using AIG (12). Among these questions, three randomly-chosen MCQs were included in rational prescribing exam, without making a single change in the questions (Figure 1). The questions were approved by the board of rational pharmacotherapy to be included in the exam. The questions can be found in the Appendix. The written exam comprised of single-best answer MCQs, including those written in a traditional way. This examination was carried out in-person in classroom settings under proctor supervision.

Statistical analysis

We carried out a psychometric evaluation based on Classical Test Theory. Utilizing Microsoft Excel, we performed item analysis to determine two main metrics: item difficulty, calculated by dividing the total score of test-takers by the maximum possible score, and item discrimination, indicated by the point-biserial correlation value. In order to calculate point-biserial correlation values, Spearman correlation was employed in SPSS 22.0 for Windows (Chicago, IL, USA). This analysis allowed us to determine whether an individual test item effectively differentiated between students who performed well overall and those who did not. Typically, large-scale standardized test developers require an item's point-biserial correlation to be at least 0.30 or higher for effectiveness (15). However, in locally developed classroom-type tests, values in the mid to high 0.20s could be considered satisfactory (15). Additionally, we computed the percentages of responses for each answer option to identify poorly functioning distractors. We considered the conventional threshold for functional distractors as those selected by participants at a rate exceeding 5% (15).

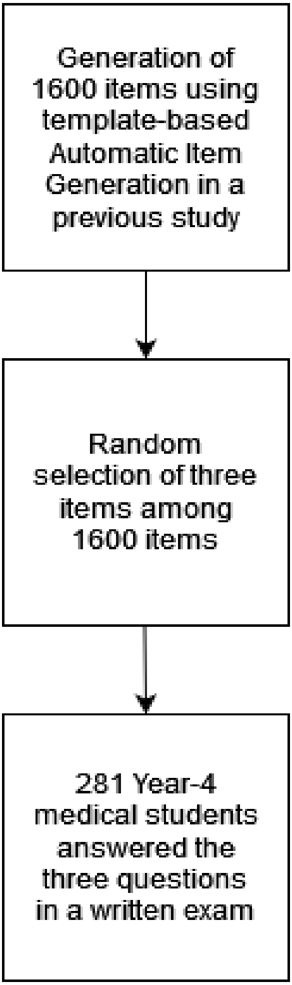
Generation of 1600 items using template-based Automatic Item Generation in a previous study
Random selection of three items among 1600 items
281 Year-4 medical students answered the three questions in a written exam

Figure 1. The Process Regarding Items

Ethical Approval

Gazi University Institutional Review Board approved the study (code: 2023-1116).

RESULTS

All three items had acceptable levels (higher than 0.20) of point-biserial correlation. The correlation values were statistically significant (p<0.001). The item difficulty levels indicated the presence of one easy, one moderate, and one difficult question. Each item had 2-3 non-functioning options among five options. Table 1 presents the values of indices and response percentages.

Table 1. Item Indices and Response Percentages

Question Number (in the Appendix)	**Indices**		**Response Percentages in Options**				
	Difficulty	Discrimination (Point-Biserial Correlation)	A	B	C	D	E
1	0.21	0.39*	0,00	**21,71**	24,56	53,02	0,71
2	0.46	0.28*	0,00	0,00	0,71	53,02	**46,26**
3	0.85	0.22*	**85,41**	4,27	9,25	0,71	0,36

**** p<0.001, Bold options are correct.***

DISCUSSION

The aim of this study was to determine whether the first automatically generated MCQs in medical education have acceptable psychometric characteristics. We found that item discrimination values (point-biserial correlation) are in a range of acceptable levels and the difficulty of the items is diverse. Additionally, each item presented two or three non-functioning options among the five provided.

The findings showed that the items effectively discriminate between high and low performers that affirms the quality of the questions in assessing students' understanding of the subject matter. The findings also indicate that it is possible in Turkish as well to generate MCQs with varying difficulty levels by using a single model in AIG, as previous studies showed the similar findings in another language (11). However, there were non-functioning options, which is an anticipated finding because subject-matter experts already recommended minor revisions to make each option more discriminative, as pointed out in the previous study (12). This issue can easily be solved by benefitting from the flexibility of AIG method. Proposed revisions by subject matter experts, informed by the literature on creating more effective distractors in AIG (16), can be universally applied to hundreds of questions at once by revising the model. This flexibility is an important advantage that eliminates the need for individual revisions of each question in traditional item writing.

This research could provide particular advantages in generating Turkish items. In medical schools globally, there is a strong need for a rich question bank evaluating higher-order skills. By mirroring the global need, creating such a high-quality resource is challenging in Turkish medical schools (17). Considering these challenges, the first advantage of template-based AIG is the efficiency (18). It allows for the rapid generation of a large number of assessment items (7,8). This is especially beneficial in medical education, where a significant volume of questions is required. The second advantage is consistency. AIG ensures a consistent format and structure for assessment items, reducing variations in question quality and format (7). Therefore, AIG enables the standardization of items, making it easier to design fair and equitable assessments for all students. It also ensures that the generated items are closely aligned with the educational objectives because the items rely on cognitive models and question templates constructed by subject matter experts. Given the successful implementation of AIG in Canada's high-stakes medical licensure exam (11), Turkish medical institutions and national exams might find significant value in adopting this useful method. However, the use of this method has some disadvantages as well (7). Initial development costs may hinder its widespread use. Developing the templates and cognitive models for AIG can be time-consuming for subject-matter experts at the beginning. However, once a model is established, it can be used to generate thousands of items. While AIG generates items quickly, there is a need for ongoing monitoring and revision to ensure that the quality of generated items meets educational standards. However, the revisions can be applied to thousands of questions with just a “click”. Another disadvantage would be the lack of creativity. Template-based AIG relies on predefined templates and models, which may limit the diversity of the generated assessment items. Finally, resistance to change would be the most prominent obstacle. Faculty and staff may be resistant to adopting AIG, particularly if they are accustomed to traditional item-writing methods.

Despite the promising findings, the results should be interpreted considering the limitations of the study. It focused only on a single disease, essential hypertension, and the generalization of the results to other medical topics requires caution. Additionally, the specific characteristics of the student population might influence the generalizability of the findings. However, since we did not collect data on demographic characteristics of the participants, we could not analyze the relationship between these characteristics and the outcomes. To build upon this first Turkish AIG study conducted in real-world educational setting, future research should explore the applicability of AIG in Turkish to a broader range of medical topics. Another limitation is that the predominant use of point-biserial correlation as the primary indicator of item quality is challenging to encompass all aspects. However, it is an important first step to lead a more common use of AIG in Turkish.

CONCLUSIONS

This study demonstrated that automatic generation of case-based MCQs in Turkish is feasible in authentic assessment settings in medical education. It contributed to the literature by showing that an AIG model is able to generate hundreds of case-based MCQs in Turkish with acceptable discrimination levels and various difficulty levels, such as easy, moderate, and difficult. Successful implementations of AIG had been demonstrated in five languages. Our study provided the first psychometric evidence from the field in the sixth language as Turkish. The ability to automatically generate effective MCQs in Turkish holds promise for enhancing the efficiency of written assessment in Turkish medical education.

REFERENCES

1. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med. 2019 Jun;94(6):902–12.

2. 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 May;41(5):569–77.

3. Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004 Sep;38(9):974–9.

4. Wrigley W, Van Der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. Medical Teacher. 2012 Sep;34(9):683–97.

5. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. Medical Education. 2012;46(8):757–65.

6. Kurdi G, Leo J, Parsia B, Sattler U, Al-Emari S. A Systematic Review of Automatic Question Generation for Educational Purposes. Int J Artif Intell Educ. 2020 Mar;30(1):121–204.

7. Gierl MJ, Lai H, Tanygin V. Advanced Methods in Automatic Item Generation. 1st ed. Routledge; 2021.

8. Falcão F, Costa P, Pêgo JM. Feasibility assurance: a review of automatic item generation in medical assessment. Adv in Health Sci Educ. 2022 May;27(2):405–25.

9. Leslie T, Gierl MJ. Using Automatic Item Generation to Create Multiple-Choice Questions for Pharmacy Assessment. American Journal of Pharmaceutical Education. 2023 May;100081.

10. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Can automated item generation be used to develop high quality MCQs that assess application of knowledge? RPTEL. 2020 Dec;15(1):12.

11. Gierl MJ, Lai H, Pugh D, Touchie C, Boulais AP, De Champlain A. Evaluating the Psychometric Characteristics of Generated Multiple-Choice Test Items. Applied Measurement in Education. 2016 Jul 2;29(3):196–210.

12. Kıyak YS, Budakoğlu Iİ, Coşkun Ö, Koyun E. The First Automatic Item Generation in Turkish for Assessment of Clinical Reasoning in Medical Education. Tıp Eğitimi Dünyası. 2023 Apr 3;22(66):72–90.

13. Sayin A, Gierl MJ. Automatic item generation for online measurement and evaluation: Turkish literature items. International Journal of Assessment Tools in Education. 2023 Jun 26;10(2):218–31.

14. Budakoğlu Iİ, Coşkun Ö, Kıyak YS, Uluoğlu C. Teaching rational prescribing in undergraduate medical education: a systematic search and review. Eur J Clin Pharmacol. 2023 Jan 9;79:341–8.

15. Downing SM, Yudkowsky R. Assessment in Health Professions Education. Routledge; 2009. 338 p.

16. Lai H, Gierl MJ, Touchie C, Pugh D, Boulais AP, De Champlain A. Using Automatic Item Generation to Improve the Quality of MCQ Distractors. Teaching and Learning in Medicine. 2016 Apr 2;28(2):166–73.

17. Cansever Z, Acemoğlu H, Avşar Ü, Hoşoğlu S. Tıp Fakültesindeki Çoktan Seçmeli Sınav Sorularının Değerlendirilmesi. Tıp Eğitimi Dünyası. 2016 Apr 28;14(44):44–55.

18. Kosh AE, Simpson MA, Bickel L, Kellogg M, Sanford-Moore E. A Cost–Benefit Analysis of Automatic Item Generation. Educational Measurement: Issues and Practice. 2019 Mar;38(1):48–53.

Appendix

Turkish multiple-choice questions generated by using template-based automatic item generation method.

1. 38 yaşında kadın hasta, bir klinikte yapılan sabah ve akşam olmak üzere bir günlük tansiyon ölçümü sonucuyla aile sağlığı merkezine başvuruyor. Daha önce hipertansiyon tanısı almamış hastanın ölçümlerindeki ortalamanın 150/95 mmHg olduğu saptanıyor. Hasta herhangi bir şikâyeti olmadığını söylüyor. Hastanın sigara kullanımı günde bir paket. Fizik muayenede vücut kitle indeksinin 34 kg/m2 olduğu hesaplanıyor.

Herhangi bir ek hastalığı olmayan bu hastada, aşağıdaki seçenekler arasında diğerlerine göre en uygun yaklaşım hangisidir?
A.Herhangi bir ek öneride bulunmadan, yılda bir kan basıncı ölçümü önerisi yeterlidir
B.Müdahale edilmeden, en az beş günlük tansiyon ölçümü yaparak tekrar başvurmalı
C.Yaşam tarzı değişiklikleri önerisinde bulunulması yeterlidir
D.Antihipertansif ilaç reçete edilerek bir ay sonra kontrole çağırılmalı
E.Gerekli ilk müdahale yapılarak hemen acil servise gönderilmeli
Doğru Cevap: B

2. 38 yaşında kadın hasta, bir klinikte yapılan bir haftalık (sabah-akşam) tansiyon ölçümü sonucuyla aile sağlığı merkezine başvuruyor. Daha önce hipertansiyon tanısı almamış hastanın ölçümlerindeki ortalamanın 190/130 mmHg olduğu saptanıyor. Hasta bazen hafif baş ağrısının olduğunu söylüyor. Hastanın sigara kullanımı günde bir paket. Fizik muayenede vücut kitle indeksinin 28 kg/m2 olduğu hesaplanıyor.

Herhangi bir ek hastalığı olmayan bu hastada, aşağıdaki seçenekler arasında diğerlerine göre en uygun yaklaşım hangisidir?
A.Herhangi bir ek öneride bulunmadan, yılda bir kan basıncı ölçümü önerisi yeterlidir
B.Müdahale edilmeden, en az beş günlük tansiyon ölçümü yaparak tekrar başvurmalı
C.Yaşam tarzı değişiklikleri önerisinde bulunulması yeterlidir
D.Antihipertansif ilaç reçete edilerek bir ay sonra kontrole çağırılmalı
E.Gerekli ilk müdahale yapılarak hemen acil servise gönderilmeli
Doğru Cevap: E

3. 56 yaşında kadın hasta, bir klinikte yapılan bir haftalık (sabah-akşam) tansiyon ölçümü sonucuyla aile sağlığı merkezine başvuruyor. Daha önce hipertansiyon tanısı almamış hastanın bu ölçümlerindeki ortalamanın 110/70 mmHg olduğu saptanıyor. Hasta herhangi bir şikâyeti olmadığını söylüyor. Hastanın sigara kullanımı yok. Fizik muayenede vücut kitle indeksinin 23 kg/m2 olduğu hesaplanıyor.

Herhangi bir ek hastalığı olmayan bu hastada, aşağıdaki seçenekler arasında diğerlerine göre en uygun yaklaşım hangisidir?
A.Herhangi bir ek öneride bulunmadan, yılda bir kan basıncı ölçümü önerisi yeterlidir
B.Müdahale edilmeden, en az beş günlük tansiyon ölçümü yaparak tekrar başvurmalı
C.Yaşam tarzı değişiklikleri önerisinde bulunulması yeterlidir
D.Antihipertansif ilaç reçete edilerek bir ay sonra kontrole çağırılmalı
E.Gerekli ilk müdahale yapılarak hemen acil servise gönderilmeli
Doğru Cevap: A