Review Article

Best Practices in Designing Extended Matching Questions for Evaluating Higher Order Cognitive Skills in Medical Education

Salam A[1], Majumder MAA[2], Kamarudin MA[3], Besar MNA[4], Abdelhalim AT[5], Zakaria H[6], Zainol J[7].

ABSTRACT

Applying best practices in designing extended matching questions (EMQs) is fundamental for accurately assessing the higher order cognitive skills of medical students. EMQs, when well-constructed, can evaluate critical thinking, clinical reasoning, and decision-making skills, which are vital in the medical field. Prioritizing clarity, relevance, and complexity in question design, educators can create assessments that align more closely with real-world medical problem-solving. Implementing these strategies is a key step towards uplifting the standards of medical education. This paper offers a comprehensive guide in designing high quality EMQs that test higher order cognitive skills of medical students, aiming to promote best assessment practices in medical education and uphold high standards of student evaluation.

Keywords

Student assessment, Design, EMQ, Evaluate, cognitive skills, Medical education.

INTRODUCTION

Medical curricula are planned with precise content to ensure high quality medical education in order to produce competent medical doctors with the intention of providing high quality medical care to the communities and clients globaly[1]. Assessment is an important step in medical education that validates teaching objectives while benefiting both students and educators[2]. Assessment drives learning, and learning drives practice[3-7]. There are three main educational domains: cognitive (thinking), affective (feeling), and psychomotor (doing). In medical education, different assessment tools are employed based on the specific domains being evaluated. For example, the cognitive domain is tested using written formats such as multiple-choice questions (MCQs), essays and oral examinations, and the psychomotor domain is tested using OSPE/ OSCE or direct observation[8].

Among the different tools, MCQs are the most frequently used tools in written assessment. However, many in-house MCQs are found to be faulty in assessing knowledge in isolated

1 Abdus Salam, Medical Education Unit, Faculty of Medicine, Widad University College (WUC), Kuantan, Pahang, Malaysia.
2. Md. Anwarul Azim Majumder, Director, Medical Education, Faculty of Medical Sciences, The University of the West Indies, Cave Hill Campus, Barbados.
3. Mohammad Arif Kamarudin, Head of the Department of Medical Education, Faculty of Medicine, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia.
4. Mohd Nasri Awang Besar, Faculty of Medicine, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia.
5. Abdelbaset Taher Abdelhalim, Faculty of Medicine, Paediatric and Pharmacology Unit, WUC, Kuantan, Pahang, Malaysia.
6. Hasbullani Zakaria, Biochemistry Unit, Faculty of Medicine, and Deputy Vice Chancellor (Research & Postgraduate), WUC, Kuantan, Pahang, Malaysia.
7. Jamaludin Zainol, Surgery Unit and Dean, Faculty of Medicine and Deputy Vice Chancellor (Academic) WUC, Kuantan, Pahang, Malaysia.

Correspondence

Dr Abdus Salam, Medical Educationalist and Public Health Specialist, Associate Professor and Head of Medical Education Unit, Faculty of Medicine, Widad University College, Bandar Indera Mahkota (BIM) Point, 25200 Kuantan, Pahang, Malaysia, Email: abdussalam.dr@gmail.com

Bangladesh Journal of Medical Science Vol. 24 No. 01 January'25

DOI: https://doi.org/10.3329/bjms.v24i1.78701

facts and assessing lower levels of thinking skills[9]. It is difficult to assess clinical reasoning using this tool as it involves a doctor applying knowledge and experience to diagnose and manage a patient's problem[8,10]. A good assessment is a test that uses the tools to assess the higher order cognitive or thinking skills of the students[3]. Extended matching questions (EMQ) represents a variant of MCQ now used in medical education, which has proved to be able to assess higher-order thinking skills like clinical reasoning such as evaluating and creating[8]. This paper describes EMQs and outlines the process of designing high quality EMQs that test higher order cognitive skills of medical students, with the ultimate aim of promoting best assessment practices in medical education and uphold high standards of student evaluation.

EXTENDED MATCHING QUESTIONS (EMQ)

The EMQ was introduced in 1990 by Case and Swanson as a different MCQ format to be used in the United States Medical Licensing Examination (USMLE)[11]. In EMQs, the problem is presented as a short case called 'vignettes'. For example, a case is presented in a few sentences outlining the patient's symptoms and lab results, and the student is required to reach a diagnosis. They need to choose from a long list of choices rather than the traditional five choices used in the MCQ of single best answer (SBA) or one best answer (OBA) type[11,12]. The EMQ is a form of MCQ, organised into sets that use (1) a theme; (2) an extended option list varies from six to 25 or more potential answers; (3) a lead-in statement that directs the students or instructs what to do; and (4) at least two, or more patient-based vignettes or item stems related to theme requiring the student to specify a clinical decision for each item stem[10,11].

Theme

The theme is a topic or title of the EMQ set addressed by the set of item stems. Based on the theme, the EMQs examined all the disorders under the selected theme[11,13]. Themes can be related to basic science, such as anatomic sites, cell types, functions, pathogens, pathophysiological states, etc. It can be clinical signs or a symptom or chief complaint, for example, back pain, chest pain, dyspnoea, fatigue, etc.; a class of drugs like antibiotics, corticosteroids etc., investigation, laboratory data, diagnosis or treatment; and so on[9,14].

Options

Candidates are given an extended number of answer choices called options. The list of answer options varies from six to 25 or more choices or potential answer options including the correct answer [8,10,11,15]. The options are presented as single words or short phrases and arranged either alphabetically or numerically as appropriate. The options should be homogeneous, for example, all anatomical sites, all diagnoses, all therapies, etc. Students need to select one option for each numbered problem /case that most closely answers to the question[11].

Lead-in Statement

The lead-in is a question or statement that directs or instructs the students about their task or what they need to do. In most circumstances, it is preferred to ask in one of the best-answer response formats. The lead statements may be written as: 'For each patient described below, choose the single most likely management plan from the above list of options'; 'For each of the following patients, select the single most likely diagnosis or most appropriate investigation or most accurate figure', etc.[10]. However, lead-in can be structured to require the examinee to select more than one response, such as, 'For each patient with fatigue described below, select three laboratory studies'. There is a single lead-in statement used for all items in a set[14].

Item-stem

The item stems comprise the questions to be answered by students. It is a list of numbered clinical vignettes, which should be at least two and describe patients in clinical situations related to the theme. The vignettes can be described briefly or elaborated depending on the clinical scenario[10,11]. Usually, it is described in 2-6 sentences, for example, describing the patient's symptoms, the laboratory test results, etc. and the student is instructed to conclude a diagnosis to be chosen from several options rather than only 4-5 five options as in SBA/OBA. By understanding the information provided in the vignette, students come to a conclusion and answer by choosing from the list[12]. Vignettes containing relevant information are a crucial part of EMQ, as they enable students to effectively respond to the questions. Consequently, careful consideration should be paid to constructing EMQs with good patient vignettes[16].-

Stems within a set or the descriptions of the patients should be similar in structure in all the sets. For

illustration, if ethnicity or occupation is included in one item, it should be included in all items; if laboratory data are included in one item, include them in all items. It is advisable not to mix adult and paediatric cases in the same set as age is an important indicator for selecting or eliminating options[17].

TYPES OF EMQ

The EMQs are mainly of two types: R-type and N-type. In the R-type format, students are instructed to identify a single answer from the option list while in the N-type format, students are instructed to identify 2, 3, 4 or even 5 answers from the options list[11,13]. The rationale for decision to specify exactly how many options (2, 3, 4 or 5) to identify is derived from the main difference between multiple true/false and one-best-answer type of questions, where true/false items require the examinee to indicate all responses that are appropriate, and one-best-answer items require the examinee to indicate a specific number of responses[11]. The N-type EMQs are suitable in questions having a variety of possibilities or the answer is not clear-cut such as differential diagnosis, drug side effects, etc.[13]. Identifying the sspecific number of options changes the task from a multiple true/false task to a best-answer task[11]. -

The following are a few (six) examples of R-type (five) and N-type (one) of EMQs, three for basic science /pre-clinical phase and three for clinical phase, taken from Case and Swanson 1993 and 2001[11,17].

R-TYPE EMQ

EMQ 1.

Theme: Arterial involvement (Anatomy)[17]

Options:

A. Left anterior cerebral artery.
B. Right anterior cerebral artery
C. Left middle cerebral artery
D. Right middle cerebral artery
E. Left posterior cerebral artery
F. Right posterior cerebral artery
G. Left lenticulostriate
H. Right lenticulostriate arteries

Lead-in:

For each patient below with neurologic abnormalities, select the artery that is most likely to be involved. Each option can be used once, more than once or not at all

Stems:

1. A 72-year-old right-handed man has weakness and hyperreflexia of the right lower limb, an extensor plantar response on the right, normal strength of the right arm, and normal facial movements. **Ans: A**

2. A 68-year-old right-handed man has right spastic hemiparesis, an extensor plantar response on the right, and paralysis of the lower two-thirds of his face on the right. His speech is fluent, and he has normal comprehension of verbal and written commands. **Ans: G**

EMQ 2.

Theme: Arterial blood gas (Physiology)[11,17]

Options	pH	PO_2 (mmHg)	PCO_2 (mmHg)	HCO_3 (mEq/L)
A.	7.15	98	33	11
B.	7.15	98	24	8
C.	7.30	56	80	38
D.	7.40	100	40	25
E.	7.50	100	33	25
F.	7.50	100	24	18
G.	7.50	56	33	25

Lead-in:

For each patient described below, select the most likely arterial blood gas findings. Each option can be used once, more than once or not at all.

Stems:

1. A 22-year-old man with a 3-week history of polyuria and polydipsia has had nausea, vomiting, and decreased responsiveness for the past 12 hours. Urinalysis (dipstick) shows 4+ glucose and 4+ ketones. **Ans: B**
2. A 25-year-old woman is brought to the emergency department 12 hours after a suicide attempt. She took approximately 100 500-mg aspirin tablets. **Ans: F**

EMQ 3.

Theme: Adverse effect of drug (Pharmacology) [17]

Options:

A. Acetaminophen
B. Amiodarone
C. ACE inhibitors
D. Aspirin
E. Atenolol
F. Bleomycin
G. Cytosine arabinoside
H. Furosemide
I. Metronidazole
J. Nalidixic acid
K. Nitrofurantoin
L. Penicillin
M. Prednisone
N. Procainamide
O. Propranolol
P. Sulfasalazine
Q. Tetracycline
R. Verapamil

Lead-in:

For each patient, select the drug most likely to have caused the adverse effect. Each option can be used once, more than once or not at all.

Stems:

1. A 56-year-old man with recurrent ventricular arrhythmias began taking an antiarrhythmic drug 5 months ago. He now has progressive dyspnoea, cough, and low-grade fever. Erythrocyte sedimentation rate is increased. X-ray film of the chest shows a diffuse interstitial pneumonia. Pulmonary function tests show that diffusing capacity for carbon monoxide is decreased.
Ans: B

2. A 62-year-old man with chronic obstructive pulmonary disease begins therapy with an antihypertensive drug. Two weeks later, he has marked worsening of dyspnoea and clearly audible wheezing. **Ans: O**

EMQ 4.

Theme: Back pain (Clinical)[1]

Options:
A. Ankylosing spondylitis
B. Aortic dissection
C. Intervertebral disc infection
D. Lumbar spondylosis
E. Metastatic malignancy
F. Pars interarticularis defect
G. Prolapsed intervertebral disc
H. Vertebral fracture

Lead-in:
For each patient with back pain, select the most likely diagnosis. Each option can be used once, more than once or not at all

Stems:
1. A 23-year-old man has a 6-month history of lower back pain. His pain is predominantly at the thoracolumbar junction and in the right buttock. The pain is worse in the morning and he has difficulty in getting out of bed. There is some improvement during the day. Examination shows restriction of lumbar spinal movements, particularly lateral flexion. **Answer: A**

2. A 32-year-old lady presents with acute onset of low back pain. The pain is constant and is not significantly affected by posture. All spinal movements are painful and difficult. Three weeks earlier, she had a urinary tract infection which had been treated with amoxicillin. **Answer: F**

STEPS IN WRITING EMQ

The steps for writing EMQ are as follows[9,11,18]

(1) First, identify the 'theme' for the EMQ set. The theme is a topic or a title for a set addressed by the set of matters. It can be chief complaints (e.g. chest pain, fatigue), a drug class (antibiotics), or basic science themes such as anatomical sites, cell components, pathophysiological states, pathogens, etc.

(2) Prepare the list of relevant and realistic 'options'. A comprehensive options list includes all relevant options, rather than requiring students to guess the three or four distractors they think will be most appealing[11]. Depending on the nature of the options and the emphasis on the basic science versus clinical sciences, the vignettes are short and focused in some cases and long in some cases[11]. The EMQ format aids in specifying and organising exam content. Here, the options list flows from the theme, lead-in, and stem flows from the option list[11]. The options list should be in single words or very short phrases and homogeneous, for example, all diagnoses, all management, all anatomical sites, all vitamin options, etc.). The options, especially those involving laboratory values, are often expressed in tabular form. Eight to ten options are usually used, followed by at least two or more patient-based scenarios requiring the examinee to indicate a clinical decision for each item[9,18].

(3) Write the 'lead-in'. Example; 'For each of the following patients, select the most likely diagnosis'. Lead-in designates the relationship between the stems and options, clarifying the question for examinees.

(4) Write the 'item stems'. Item stems are the patient vignettes or clinical scenarios. There should be a minimum of two vignettes, while three vignettes with 8-10 options are preferable. The length of the stem or vignettes affects reading time, as longer vignettes take more time to read than shorter ones. The EMQ set that test the application of knowledge require more time than those test recall of isolated facts.

(5) Review the items for 'quality control'. After complete preparation of the EMQ, the written question, especially the stem /patient vignettes /clinical scenario, needs to be checked to make sure that there is a 'single best' answer for each question. Also to be confirm that at least four realistic distractors are present for each question. Then, for the final check, ask the help from one or more colleagues to review the items without

EMQ 5.

Theme: Fatigue (Clinical)[11,17]

Options:

A. Acute leukemia
B. Anemia of chronic disease
C. Congestive heart failure
D. Depression
E. Epstein-Barr virus infection
F. Folate deficiency
G. G6-phosphate dehydrogenase deficiency
H. Hereditary spherocytosis
I. Hypothyroidism
J. Iron deficiency
K. Lyme disease
L. Microangiopathic hemolytic anemia
M. Miliary tuberculosis
N. Vitamin B12 (cyanocobalamin) deficiency

Lead-in:

For each patient with fatigue, select the most likely diagnosis

Stems:

1. A 19-year-old woman has had fatigue, fever, and sore throat for the past week. She has a temperature of 38.3 C (101 F), cervical lymphadenopathy, and splenomegaly. Initial laboratory studies show a leukocyte count of 5000/mm3 (80% lymphocytes, with many lymphocytes exhibiting atypical features). Serum aspartate aminotransferase (AST, GOT) activity is 200 U/L. Serum bilirubin concentration and serum alkaline phosphatase activity are within normal limits. **Ans: E**

2. A 15-year-old girl has a two-week history of fatigue and back pain. She has widespread bruising, pallor, and tenderness over the vertebrae and both femurs. Complete blood count shows hemoglobin concentration of 7.0 g/dL, leukocyte count of 2000/mm3, and platelet count of 15,000/mm3. **Ans: A**

N-TYPE EMQ

EMQ 6.

Theme: Diagnostic Testing (Clinical)[17].

Options:

A. Analysis and culture of cerebrospinal fluid
B. Blood culture
C. Complete blood count
D. Examination of the stool for leukocytes
E. Measurement of serum electrolyte levels
F. Urinalysis
G. Urine culture
H. X-ray film of the abdomen
I. X-ray film of the chest

Lead-in:

For each child with fever, select the appropriate initial diagnostic studies.

Stems:

1. A previously healthy 1-year-old girl is brought to the emergency department because of fever for 1 day. Her temperature is 41 C (105.8 F). She is otherwise asymptomatic. Physical examination shows no abnormalities. (SELECT 4 STUDIES). **Ans: B, C, G, I**

2. A 7-year-old boy with sickle cell disease is brought to the emergency department because of fever for 1 day and chest pain for 1 hour. His temperature is 39.5 C (103.1 F). Breath sounds are slightly decreased in the right lower lung; he is not in respiratory distress. (SELECT 3 STUDIES). **Ans: B, C, I**

the correct answers indicated. If it is difficult for the colleague to determine the correct answer, then the option list or item needs to be modified to eliminate the ambiguity.

(6) Time allocation: One minute per stem /vignette appears to be sufficient generally for some sets of EMQ. In contrast, less than 20-40 seconds may be appropriate for others depending on the attitudes toward "speededness"[11]. This is a rough guideline for timing, and more research is needed. Faculty also need to determine the time allocation based on the characteristics of vignettes in consensus.

THE ROLE AND BENEFITS OF EMQ IN MEDICAL EDUCATION

The EMQs are now increasingly practising in medical education for undergraduate and postgraduate courses[10]. The assessment of clinical reasoning is appropriate with EMQs as the question is constructed as a clinical scenario with relevant questions. Moreover, having several options reduces the guessing of the correct answer[10]. The EMQs are reliable assessment tools to test the core knowledge and clinical reasoning of students[15], and are feasible, reliable, valid, authentic, and able to discriminate good from poor performers[19]. The EMQs can be stored in the question bank, modified and recycled, and used repeatedly[13]. They can be electronically scored; results can be statistically analysed and problematic questions can be identified[13]. However, constructing a high quality EMQ presents challenges and can be time-consuming for the faculties.

Medical education, though constantly changing, traditional curricula, inadequate funding, lack of objectivity, varying standards of assessment methods with weak quality assurance in higher education are a big problem[4,20,21]. Underestimation of assessment blueprint and lack of formal training on question and blueprint construction are added to the problem[3]. An assessment blueprint is a plan or test specification table /template (TST) that fixes the learning outcomes with related content to be tested and lists the number of questions and type of questions across the content with relative weightage among three educational domains[22,23,24]. As teaching and assessment are regarded as two sides of the same coin[4,20], there needs to be a well constructive alignment between teaching and assessment. Teachers must understand the process of teaching-learning and procedures of assessment to ensure the reliability and validity of the assessment[25]. The 21st century is a rapid development of science and technology, which makes people continuously improve, to add knowledge and ability[7]. It is necessary to train the faculties to construct high quality EMQs that assess higher level of cognitive skills of the students[13].

Effective training is always an important factor for the competency of the staff to fill the gap between desired performance and actual staff performance[26]. Faculty members are the valuable scholarly assets within institutions, and faculty development is an integral part of an institution's educational advancement[27,28]. Hence, well-trained trainers should implement faculty developmental programmes regularly for a sustainable institutional development[3]. Faculty development serves as the foundation for any curriculum innovation; without it, curriculum improvements would be impossible[29]. Therefore, institutions should prioritize regular faculty development programmes led by well-trained facilitators to ensure sustainable growth and development through the effective use of high-quality EMQs.

CONCLUSION

This paper offers a comprehensive guide for medical educators worldwide on how to design high quality EMQs for best practices in the assessment of higher-order thinking skills of students. The EMQs are a variant of MCQs suitable for testing medical students' clinical reasoning and problem-solving skills. Here, the students resolve a problem rather than recall isolated pieces of information. Using the EMQ can greatly reduce technical flaws, where students are instructed to choose an option from several options rather than only from 4-5 options, as in SBA/OBA. There is less chance of examinees guessing the correct response because of more relevant options. By integrating a well-defined theme, an extended list of relevant options, a clear lead-in statement, and two or more item stems, educators can create assessments that truly reflect the cognitive capabilities of their students. It is essential for EMQs to align with the learning objectives of the medical curriculum to ensure that they accurately measure the competencies required for clinical practice. Additionally, EMQs should be integrated into a broader assessment strategy that includes diverse evaluation methods to comprehensively assess medical student's competencies.

Creating a high-quality EMQ is a challenging task, and well-trained, experienced examiners can construct it. Training among the faculties is necessary to construct a high-quality EMQ. Medical schools should consider the needs of the faculty while organising faculty developmental training workshops to forge links between education and practice to ensure sustainable development. By adhering to above mentioned best practices and recommendations, medical educators can significantly enhance the quality of assessments, thereby fostering a deeper understanding of medical knowledge and improving the readiness of future healthcare professionals.

Funding

No funding was received for this paper.

Conflict of Interest

The authors declare no conflicts of interest.

Authors' Contribution

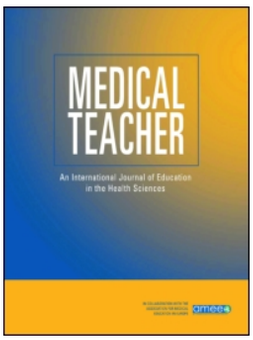
All authors participated well in the preparation of this paper and approved the final version to submit to the Journal for publication.

REFERENCES

1. Asani M. Assessment methods in undergraduate medical schools. *Niger J Basic Clin Sci.* 2012; **9**:53-60
2. Tabish SA. Assessment methods in medical education. *Int J Health Sci (Qassim).* 2008; **2**(2):3-7.
3. Salam A, Kamarudin MA, Algantri KR, Saghir FSA, Yusoof MBA, Zakaria H, Zainol J. Creating High Quality Single Best Answer Questions: A Guide for Medical Educators. *Int J Hum Health Sci.* 2024; **08**(03):223-227. DOI: http://dx.doi.org/10.31344/ijhhs.v8i3.715
4. Salam A, Yousuf R, Allhiani RF, Zainol J. Continuous Assessment in Undergraduate Medical Education Towards Objectivity and Standardization. *Int J Hum Health Sci.* 2022; **06(**03):233-236. DOI: http://dx.doi.org/10.31344/ijhhs.v6i3.453
5. Salam A. Input, Process and Output: system approach in education to assure the quality and excellence in performance. *Bangladesh J Med Sci.* 2015; **14**(01):1-2. *http://dx.doi.org/10.3329/bjms.v14i1.21553*
6. Haque M, Yousuf R, Abu Bakar SM, Salam A. Assessment in Undergraduate Medical Education: Bangladesh Perspectives. *Bangladesh J Med Sci.* 2013; **12**(04):357-363. http://dx.doi.org/10.3329/bjms.v12i4.16658
7. Salam A. Issues of objective, content, method and assessment in the development of relevant curriculum in medical schools. *Malaysian Medical Association (MMA) News.* April 2010; **40**(4):22-24. www.mma.org.my | info@mma.org.my
8. Nalini YC, Manivasakan S, Pai DR. Comparison between MCQ, Extended matching questions (EMQ) and Script concordance test (SCT) for assessment among first-year medical students -A pilot study. *J Edu Health Promot.* 2024; **13**:52.
9. Salam A, Yousuf R, Bakar SMA. Multiple Choice Questions in Medical Education: How to Construct High Quality Questions. *Int J Hum Health Sci*. 2020; **04**(02):79-88.
10. George S. Extended matching items (EMIs): solving the conundru m. *Psychiatric Bulletin.* 2003; **27**:230-232. https://doi.org/10.1192/pb.27.6.230.
11. Case SM, Swanson DB. Extended matching items: a practical alternative to free response questions. *Teach Learn Med*. 1993; **5**:107–1115. DOI: 10.1080/10401339309539601
12. Wood EJ. What are Extended Matching Sets Questions? *Bioscience Education.* 2003; **1** (1):1-8, DOI: 10.3108/beej.2003.01010002
13. Samuels A. Extended Matching Questions and the Royal Australian and New Zealand College of Psychiatrists written examination: an overview. *Australasian Psychiatry.* 2006; **14** (1):63-66. https://doi.org/10.1111/j.1440-1665.2006.02247.x.
14. Wilson RB, Case SM. Extended Matching Questions: An Alternative to Multiple-choice or Free-response Questions. *JVME* 1993; **20**(3): https://scholar.lib.vt.edu/ejournals/JVME/V20-3/wilson.html
15. Eijsvogels TMH, van den Brand TL, Hopman MTE. Multiple choice questions are superior to extended matching questions to identify medicine and biomedical sciences students who perform poorly. *Perspect Med Educ*. 2013; **2**:252-63.
16. van Bruggen L, van Woudenbergh MM, Spierenburg E, Vos J. Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspect Med Educ*. 2012; **1**:162-171. DOI 10.1007/s40037-012-0024-1.
17. Case SM and Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. Third Edition,

2001. *National Board of Medical Examiners*. 3750 Market Street, Philadelphia, PA 19104.

18. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. *J Family Communiy Med*. 2006; **3**(3):125-133.
19. Frey A, Leutritz T, Backhaus J, Hörnlein A, König S. Item format statistics and readability of extended matching questions as an effective tool to assess medical students. *Sci Rep*. 2022; **12**(1):20982. doi: 10.1038/s41598-022-25481-y.
20. Salam A, Rahim AFAB, Aziz RA, Fakri NMRM, Ja'afar, R. Assessment of the students: Tools used in University Sains Malaysia. *Bangladesh Med J* 2005; **34**(1):11-13.
21. Majumder MAA, Haque M and Razzaque MS. Editorial: Trends and challenges of medical education in the changing academic and public health environment of the 21st century. *Front. Commun.* 2023; **8**:1153764. doi: 10.3389/fcomm.2023.1153764
22. Patel T, Saurabh MK, Patel P. Perceptions of the Use of Blueprinting in a Formative Theory Assessment in Pharmacology Education. *Sultan Qaboos University Med J*. 2016; **16**(4): e475-481. doi: 10.18295/squmj.2016.16.04.012
23. Adkoli BV, Deepak KK. Blueprinting in assessment. In: Singh T, Anshu, Eds. Principles of Assessment in Medical Education, 1st ed. New Delhi, India: Jaypee Brothers Medical Publishers Ltd., 2012. Pp. 205–13. doi: 10.5005/jp/books/11647_19.
24. Patil SY, Gosavi M, Bannur HB, Ratnakar A. Blueprinting in assessment: A tool to increase the validity of undergraduate written examinations in pathology. *Int J Appl Basic Med Res*. 2015; **5**:S76-79. doi: 10.4103/2229-516X.162286.
25. Salam A, Hashim R, Zakaria H. Techniques of Teaching to Achieve the Highest Educational Outcomes in Medical Education. *Int J Hum Health Sci*. 2023; **07**(04):273-276. DOI: http://dx.doi.org/10.31344/ijhhs.v7i4.586
26. Islam N, Salam A. Evaluation of Training Session Applying Gagne's Events of Instructions. *Bangladesh J Med Sci.* 2019; **18**(03):552-556. DOI: https://doi.org/10.3329/bjms.v18i3.41625
27. Salam A, Wahab MKBA, Ahamad A, Aziz NBA. Faculty perspectives on "Foundation in Teaching and Learning" training workshop. *Australas Med J.* 2017; **10**(7):645-646.
28. Salam A, Mohamad N, Siraj HH, Kamarudin MA, Yaman MN, Bujang SM. Team-based learning in a medical centre in Malaysia: Perspectives of the faculty. *The Natl Med J India.* 2014; **27**(6):350-351.
29. Salam A, Mohamad MB. Teachers Perception on What Makes Teaching Excellence: Impact of Faculty Development Programme. *International Medical Journal*. 2020; **27**(1):1-4.



Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning

Marla Nayer, Susan Glover Takahashi & Patricia Hrynchak

To cite this article: Marla Nayer, Susan Glover Takahashi & Patricia Hrynchak (2018): Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning , Medical Teacher, DOI: 10.1080/0142159X.2018.1481281

To link to this article: https://doi.org/10.1080/0142159X.2018.1481281

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2018.1481281

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

TWELVE TIPS

Twelve tips for developing key-feature questions (KFQ) for effective assessment of clinical reasoning

Marla Nayer[a], Susan Glover Takahashi[a] and Patricia Hrynchak[b]

[a]University of Toronto, Toronto, ON, Canada; [b]University of Waterloo, Waterloo, ON, Canada

ABSTRACT

Clinical reasoning is the cognitive process that makes it possible for us to reach conclusions from clinical data. "A key feature (KF) is defined as a significant step in the resolution of a clinical problem. Examinations using key-feature questions (KFQs) focus on a challenging aspect in the diagnosis and management of a clinical problem where the candidates are most likely to make errors." KFs have been used at different levels of medical education and practice, from undergraduate to certification examinations. KFQs illuminate the strengths and limits of an individual's clinical problem-solving ability. These types of items are more likely than other forms of assessment to discriminate among stronger or weaker candidates in the area of clinical reasoning. The 12 tips in this article will provide guidance to faculty who wish to develop KFQs for their tests.

Introduction

Clinical reasoning is the cognitive process that makes it possible for us to reach conclusions from clinical data, and come to a clinical decision. "A key feature (KF) is defined as a significant step in the resolution of a clinical problem. Examinations using key-feature questions (KFQs) focus on a challenging aspect in the diagnosis and management of a clinical problem where the candidates are most likely to make errors" (Hrynchak et al. 2014). KFQs have been used for undergraduate medical education, graduate medical education, and licensure examinations (Farmer and Hinchy 2005; Fischer et al. 2005; Leung et al. 2016). KFQs, by their nature, are focused on clinical reasoning and move away from the assessment of rote knowledge or comprehension towards synthesis and evaluation of information in Bloom's cognitive taxonomy (Armstrong 1956; Anderson and Krathwohl 2001; Krathwohl 2002).

Some authors use the terms clinical reasoning and clinical decision making and problem solving interchangeably (Van der Vleuten and Newble 1995; Page 1999 Introduction), or have different definitions of these terms (van Bruggen, Manrique-van Woudenbergh et al. 2012; Durning et al. 2013). For our purposes, clinical reasoning is a concept that reflects the cognitive process. It can include the assessment, diagnosis, and management of a patient. This includes, but is not limited to, clinical decision making (Hrynchak et al. 2014; Escudier et al. 2018). KFQs measure clinical reasoning (Eva 2005; Ilgen et al. 2012).

Research suggests that clinical reasoning skills are specific to the case or problem encountered (case specificity, also referred to as context or content specificity) (Norman et al. 2006). Successful clinical reasoning is contingent on understanding and using the few elements of the problem that are crucial to its successful resolution. KFs represent the critical information needed in the identification or management of a clinical problem. KFQs are focused on case scenarios, often with two to five items for each scenario, and illuminate the strengths and limits of an individual's clinical reasoning. This enables the instructor to have accurate information about the learner's clinical decision making ability. For example, a KFQ will focus on those key elements in a case history that are most likely to lead to a correct diagnosis, either by ruling in or ruling out specific differential diagnoses. These types of items are more likely than other forms of assessment to discriminate among stronger or weaker candidates in the area of clinical reasoning (Schuwirth et al. 2001; Leung et al. 2016).

KFQs have been validated by being administered to practicing clinicians, with positive results. These include physicians (Bordage et al. 1997), and physical therapists and occupational therapists (Glover Takahashi et al. 2012). These types of items appear to have predictive ability for future regulatory complaints (Tamblyn et al. 2007) as well as for quality of care (Wenghofer et al. 2009; Tamblyn et al. 2010). They have been used successfully with clinical clerks (Hatala and Norman 2002; Fischer et al. 2005; Lang et al. 2014), and junior doctors (Leung et al. 2016), as well as in licensure or certification examinations and maintenance of competence programs (Bordage, Brailovsky, et al. 1995; Page and Bordage 1995; Page et al. 1995; Farmer and Hinchy 2005; Lawrence et al. 2011; Glover Takahashi et al. 2012; Brailovsky et al. 2014). They have also been used for jurisprudence content, as well as various intrinsic CanMEDS roles (Royal College of Physicians and Surgeons of Canada 2005): e.g. Communicator, Collaborator, Health Advocate, Scholar, and Professional (Glover Takahashi et al. 2012). Incorporating KFQs into assessment programs will enhance the assessment programs and provide additional information to faculty on learner abilities (Hrynchak et al. 2014).

CONTACT Marla Nayer Marla.nayer@utoronto.ca 500 University Avenue 6th Floor, Toronto, ON M5G 1V7, Canada

Supplemental data for this article can be accessed here.

As with any type of assessment, developing strong items will be central to how well the test functions.

Tip 1

Define the key competencies related to decision making that are to be assessed and create a blueprint

The first step in any examination development is to create an examination blueprint (Downing and Haladyna 2006; Haladyna and Rodriguez 2013). Normally a program of instruction will have established exit-level competencies that each graduate should achieve. Each instructional component will have established learning objectives that are seen to contribute toward the exit-level competencies. These objectives may include professional standards and ethics, as well as diagnosis and management. In most health professions, clinical reasoning (sometimes referred to as problem solving) is a key component of the instructional content, whether it is clinical or addressing professional standards. The frequency of use and importance of each objective will help drive the weighting process of content development and the number of KFQs needed. This will establish content validity of the examination.

The blueprint should be based on the instructional content for the course or program and, for a KF examination, should address the key reasoning areas to be covered. For a very basic example of a blueprint, see Table 1. It is not necessary to fill in every cell in the table, though the *totals* for the rows and columns are important in the creation of an examination. Examples of examinations using blueprints include the Medical Council of Canada (2014), the Medical Council, Ireland (University College Cork Ireland 2015), and the Royal College of Obstetricians and Gynecologists, England. For further information on blueprint development, see the "12 Tips" article on that subject by Coderre et al. (2009).

Tip 2

Choose a clinical presentation or situation

The type of case scenario will depend on the content area and the level of the learner. For a more junior learner, it might be a focused problem or a complaint related to a single system with a typical presentation. For a more advanced learner, it might be an undifferentiated problem or complaint or an atypical presentation, or it might include multisystem involvement.

Many organizations that have developed their own milestones or competency documents [e.g. ACGME milestones (Accreditation Council for Graduate Medical Education (ACGME) and American Board of Pediatrics 2012), the United Kingdom (General Medical Council 2014), Australian Society of Pharmacists (Pharmaceutical Society of Australia 2010), Royal Australian College of General Practitioners (2015), or the Royal College of Physicians and Surgeons of Canada (Frank et al. 2014)]. When such a document is available consider aligning or linking different KFQs to the different stages, milestones or competency statements.

Tip 3

Select the "key feature" level of difficulty that is appropriate for the learners

This is the focus for a KFQ: make sure that the KF is at the appropriate level of difficulty for the level of the learner. KF exams have been used for learners at many levels (Bordage, Brailovsky, et al. 1995; Page and Bordage 1995; Page et al. 1995; Bordage et al. 1997; Hatala and Norman 2002; Farmer and Hinchy 2005; Fischer et al. 2005; Lawrence et al. 2011; Glover Takahashi et al. 2012; Brailovsky et al. 2014; Lang et al. 2014; Leung et al. 2016). Is the learner an undergraduate medical student, a trainee in Internal Medicine, a subspecialty trainee in Cardiology? Each level would require a different KF.

It is necessary to identify the elements or steps most likely to result in errors, the challenging aspects of the identification and management of the problem in clinical practice, or the common misconceptions about the clinical scenario. This is where the writer must differentiate between decisions or steps that are appropriate but not critical, and the steps that *must* be taken to identify and manage the patient's problem. Where are the learners most likely to make an error? What is the challenge in identifying or managing this situation? It is best to make sure that each question deals with a single KF.

An understanding of the common "real-life" misunderstandings and/or errors made by the learners at the different levels comes from experience in teaching and assessing learners at a certain level. This may come out of clinical teaching or from common errors seen on other types of assessments.

Tip 4

Focus the key feature

A KF may pertain to history, physical examination results, other investigations, clinical decision making, management, or the application of professional standards (Page and Bordage 1995; Page et al. 1995; Page 1999; Glover Takahashi et al. 2012, 2013).

The KF should be stated in a single sentence. Some examples: a fourth-year clinical clerk will be able to recognize an anterior ST segment elevation MI on ECG; a junior doctor will recognize the substitute decision-maker hierarchy when a patient is unable to make decisions about

Table 1. Sample blueprint.

	Dimension of Care					
Competency Area	Assessment	Diagnosis	Management	Communication	Professional Behaviour	% of exam
Behavioural Medicine						20%
Surgical Skills						15%
Care of the Elderly						20%
Paediatrics						30%
Obstetrics						15%
% of exam	25%	20%	25%	15%	15%	100%

their own health; a practitioner will recognize inappropriate advertising and know what follow-up actions are needed.

Tip 5

Develop the scenario

To develop the scenario, think of real cases from practice. The authors' experience is that cases from practice will ground the scenarios in the realities of "real" practice. The Medical Council of Canada uses five clinical situations, which can be used in selecting cases (Page et al. 1995). These include: an undifferentiated problem or complaint; a single typical or atypical problem; a multiple problem or multisystem involvement; a life-threatening situation; and preventive care and health promotion.

Include the relevant specific case information, such as age, gender, setting, presenting condition, and any other details that are appropriate. An easy template to start off an item is: A (xx-year-old) (man/woman/child) presents to the (location) with a complaint of (chief complaint). While it is appropriate to include information that the candidate must recognize as not being relevant in this particular case, it is best to avoid extraneous data that is completely irrelevant to the question that is presented.

Tip 6

Develop the item: stem, question (lead-in), and options (correct answer and distractors)

Many different response formats can be used with KFQs. The one that is used most often, particularly as it fits well with computer administration or scan sheets, is "Pick N" or Multiple Select. In these types of items, there is a long list of options and a number of them, perhaps three options, are correct answers (Farmer 1998; Farmer and Page 2005; Fischer et al. 2005)

A variation of the Pick N is a short menu format, also called an extended-matching item, where there is a longer list of options (10-45 options) however only one answer is correct. This might be a list of potential diagnoses, where only one is the correct answer, or a list of investigations where one is the critical investigation to allow for the correct diagnosis to be made (Case and Swanson 1998, p. 69, Fischer et al. 2005; Rotthoff et al. 2006; Haladyna and Rodriguez 2013, p. 75).

Another common format for KFQs is the Long Menu. In this format, the list of options is extremely long, perhaps over 500 items (Fischer et al. 2005; Rotthoff et al. 2006; Cerutti et al. 2016; Huwendiek et al. 2017). For example, for a question related to diagnosis the option list could be the entire International Classification of Diseases (World Health Organization (WHO) 2016). Only one answer is correct and the candidate must type it into a field in the computer, at which point the program provides for all options that match the spelling provided.

Other options for the response format include multiple choice, short answer, matching, or multiple true/false (Case and Swanson 1998; Downing and Haladyna 2006; Haladyna and Rodriguez 2013).

While focusing on the KF identified is most important, the different answer formats are also of relevance; see Supplemental Table for pros and cons, and examples. Table 2 provides clinical examples of key feature questions.

Some of the formats match well to clinical decision-making activities in practice. For example, it is unlikely that a single blood test is ordered; more likely a number are ordered at the same time. A Pick-N format could ask for the three most important investigations to order. A matching scenario would work well for connecting specific clinical presentations with a specific disease, or drugs with a specific class of medication. True/false items could work well in determining what medications are appropriate and those that are contraindicated.

Items may stand alone or there may be three to five items for each case scenario. In a series, the questions might ask about what information to elicit in the history, interpreting symptoms, identifying key physical findings, making a diagnosis or coming up with a differential diagnosis, or selecting appropriate treatments. When creating a series of items that go with one scenario, it is important to make sure that each question stands alone, i.e. there should be no cueing from one question to the next and it should be possible to answer one question incorrectly and yet still get the others correct. If this is not possible, it may be necessary to create a different case scenario for one or more of the KFs.

Tip 7

Focus the question

It is appropriate to focus the question. This could mean using specific qualifiers (Paniagua and Swygert 2016), e.g. What would you do **FIRST**? What are the three **MOST** important questions to ask in the history? What are the two **MOST LIKELY** differential diagnoses? Which of the following are the three most appropriate **INITIAL** goals?

Clear and unambiguous phrasing of answers is important in the preparation of items (Rotthoff et al. 2006).

Tip 8

Develop the options, both correct answer and distractors

As noted in the National Board of Medical Examiners manual, as well as other publications, options should be "plausible and attractive to the uninformed" (Bordage, Carretier, et al. 1995; Case and Swanson 1998, p. 41; Paniagua and Swygert 2016 Chapter 4). Use common misconceptions that the learners have expressed in teaching sessions to develop the incorrect options (Case and Swanson 1998, p. 41). All the options should be about the same length and use the same grammatical structure. A simple guideline might be to have two incorrect options for each correct option (e.g. four incorrect and two correct). There is no hard and fast rule about this ratio; however, given that there is ample evidence that three-option multiple choice questions are just as good as, if not better than, four-option multiple choice questions, this ratio seems appropriate (Haladyna and Downing 1993; Rodriguez 2005; Piasentin 2010; Schneid et al. 2014; Kilgour and Tayyaba 2016) and is what was originally recommended by Farmer (1998).

Table 2. Sample key feature questions in different formats.

Example 1 – Pick-N item
Which of the following are most appropriately considered 'interests' rather than 'positions'? (Pick 2)
A. "We feel that junior doctors should respond to pages in less than 10 minutes"
B. "We want to provide the best care—sometimes we can't wait for a page return."
C. "Junior doctors do not respond to pages from the ward so we call repeatedly."
D. "We all would like the best communication system we can get."
E. "We wait by the phone until calls are returned."
Answers: B & D

Example 2 – Extended Matching item
For the following patients, select the vitamin that is most likely deficient in the patient's diet:
Scenario 1 A 24-year-old woman presents with complaints of fatigue, heart palpitations and a pricking sensation in her toes. She follows a strict vegan diet.
Scenario 2 A 65-year-old patient who is alcoholic presents with difficulty seeing at nighttime. He has dry irritated eyes and keratinized growths (metaplasia) on the conjunctivae.
a. Vitamin A (retinoids)
b. Vitamin B1 (Thiamine)
c. Vitamin B12 (Cobalamin)
d. Vitamin B2 (Riboflavin)
e. Vitamin B3 (Niacin)
f. Vitamin B5 (Pantothenic acid)
g. Vitamin B6 (Pyridoxine)
h. Vitamin B9 (Folic acid)
i. Vitamin C (Ascorbic Acid)
j. Vitamin D (Calciferol, 1,25-dihydroxy vitamin D)
k. Vitamin E (tocopherol)
l. Vitamin H (Biotin)
m. Vitamin K
Answer Scenario 1: d
Answer Scenario 2: a

Example 3 – Fill-in-the-blank
A 78-year-old woman presents to the office on a Friday afternoon at 4:00 pm for an urgent appointment. She is complaining of a sudden onset of blurred and decreased vision in her right eye with distortion. She says that there is no redness or pain in the eye. She has not had any trauma. She has hypertension that is under control but denies any other health conditions.
What is the most likely diagnosis in this case?
Answer: age-related macular degeneration

Example 4 – Matching
Match each drug with the most common side-effect:

a. Drug 1	1. Side effect 1
b. Drug 2	2. Side effect 2
c. Drug 3	3. Side effect 3
d. Drug 4	
e. Drug 5	

Example 5 – Multiple True/False
Indicate whether each of the following are recommendations from Choosing Wisely Canada? (T/F)
a. Recommend routine daily self-glucose monitoring in adults with stable type 2 diabetes (F)
b. Don't routinely order a thyroid ultrasound in patients with abnormal thyroid function tests unless there is a palpable abnormality of the thyroid gland. (T)
c. Use Free T4 or T3 to screen for hypothyroidism or to monitor and adjust levothyroxine (T4) dose in patients with known primary hypothyroidism. (F)
d. Only prescribe testosterone therapy when there is biochemical evidence of testosterone deficiency. (T)
e. Routinely test for Anti-Thyroid Peroxidase Antibodies (anti – TPO). (F)

When using a long menu format (Rotthoff et al. 2006) it is important that the options are single terms and synonyms are accounted for, as well as common misconceptions.

Tip 9

Develop instructions for answering

For each item, there must be clear instructions for how the candidate is to answer the question. Is there one answer? Three? Can they pick as many as they like? Some options include:

- Select up to four
- Which one of the following …
- Select as many as appropriate
- Fill in the blank

"Which one of the following … " works best with the one-best-answer multiple-choice question. The challenge with "select up to … " or "select as many as appropriate" is that candidates find the uncertainty unsettling—they like to know *how many* they should be looking for and selecting. On the other hand, "select as many investigations as appropriate" might work well in assessing resource usage, where a candidate may be penalized for selecting too many investigations. Focus the instructions for answering the KF—what is the main concept/knowledge/skill being assessed? The instructions to be used will often be clear when viewed in reference to the KF listed.

Tip 10

Develop the scoring guideline for each item

Various scoring options have been described for KFQs (Page and Bordage 1995; Page, 1995, p. 162, Farmer and Page 2005; Rotthoff et al. 2006). It is possible to penalize critical errors. Some suggest only scoring if all correct options are selected (Rotthoff et al. 2006); however, part marks can also be used. The part mark approach, as well as the summative versus average scoring approach, have both been shown to provide higher reliability than using a dichotomous score (Page and Bordage 1995).

The various types of scoring include (Hrynchak et al. 2014):

- Dichotomous scoring: 0/1; Partial credit score: number between 0 and 1
- Part mark approach: takes into account the number of incorrect as well as the number of correct responses
- Summative problem scoring: the problem score is the sum of the question scores within a problem
- Averaging problem scoring: the problem score is the average of the question scores within a problem
- Summative approach: each problem score is weighted by the number of questions it contains
- Averaging approach: all problem scores are equally weighted

Examples of scoring might be:

Lead-in: Write down the most important differential diagnosis to rule out.

Scoring: Score 1 for the correct differential. (Note: different terms that refer to the same condition may be granted the same scores.)

Lead-in: Select three steps in the management of this patient.

Scoring: Score 1 point for each correct management; however, if option C is selected, then score the whole item as 0 points, as C is contraindicated for this patient.

Lead-in: Select seven questions to ask on the history.

Scoring: Score 1 for up to five of the following seven options. (Note: full option list includes 15 options; seven options are most important however the item is to be weighted for only 5 correct answers.)

Lead-in: Select as many as appropriate.

Scoring: Score 1 point for up to 5 options; 0 if more than 5 options are selected.

Tip 11

Make sure item-writing guidelines are followed

There are books and articles that outline item-writing guidelines. Case and Swanson (Case and Swanson 1998) is an excellent starting point as is the recently updated version of this guide (Paniagua and Swygert 2016), which is available on line through the National Board of Medical Examiners (NBME) web site.

There are also books and journal articles that address item-writing (Haladyna and Downing 1989a,b, Jozefowicz et al. 2002; Haladyna 2004; Downing and Haladyna 2006; Haladyna and Rodriguez 2013) and there is evidence that faculty development in this area is successful (Abdulghani et al. 2015, 2017; Abozaid et al. 2017; Alamoudi et al. 2017).

Here are some key points for developing items. Always pose a question in a way that allows the candidate to decide on the correct answer without looking at the options. This approach is often called the "hand over" technique (i.e. it is possible to answer even if the options are covered by a hand). Following this tip will prevent having unfocused questions. Avoid, or use extremely sparingly, negatively worded questions; these questions encourage measurement error when able candidates become confused, they are challenging to respond to, and disadvantage those who are writing the examination in a language other than their mother tongue. Avoid frequency terms, such as rarely (how rare is rare?), usually (how often is usually?), or sometimes (once a day? once a week? once a month?).

Tip 12

Consider the words/language used in the items

There is some research that indicates that the language used in items may affect how the learners respond.

Weaker students will perform better when items use medical terminology rather than lay language (Norman et al. 2003; Eva et al. 2010). In some situations, it may be quite appropriate to use lay language (e.g. "a 55-year-old patient comes in to the clinic complaining of coughing up blood"; rather than "a 55-year-old patient comes in to the clinic complaining of haemoptysis"). When reasonable, use the language that the patient would use in solving a patient interaction, and more technical language if interpreting diagnostic findings or reviewing a case with supervisor.

Conclusions

KFQs are a valuable validated assessment approach to assessing the complex knowledge and clinical reasoning that takes place in real-life practice. Developing KFQs requires sophisticated thinking, a deep understanding of candidates' likely responses to questions, an awareness of candidates' perceptions about content, and the ability to write with a high degree of precision. Additional resources on writing KFQs may be found on line (e.g. Medical Council of Canada's guide (Medical Council of Canada 2012), Page's guide (Page 1999), and the Royal Australian College of General Practitioners guide (Farmer 1998; Farmer and Page 2005)).

Integrating KFQs into current systems of assessment would add value by promoting clinical reasoning, as well as identifying learners who have gaps in their ability to apply content knowledge.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

Dr. Marla Nayer, PhD, is an assessment consultant, working in postgraduate medical education and teaches a graduate level assessment course at University of Toronto.

Dr. Glover Takahashi, PhD, works in postgraduate medical education and teaches a graduate level assessment course at University of Toronto.

Dr. Patricia Hrynchak, OD, is a clinical professor at the School of Optometry and Vision Science, University of Waterloo.

ORCID

Marla Nayer http://orcid.org/0000-0002-3249-3140
Susan Glover Takahashi http://orcid.org/0000-0003-0722-7876
Patricia Hrynchak http://orcid.org/0000-0002-3187-0338

References

Abdulghani HM, Ahmad F, Irshad M, Khalil MS, Al-Shaikh GK, Syed S, Aldrees AA, Alrowais N, Haque S. 2015. Faculty development programs improve the quality of Multiple Choice Questions items' writing. Sci Rep. 5:9556.

Abdulghani HM, Irshad M, Haque S, Ahmad T, Sattar K, Salah Khalil M. 2017. Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. Plos One. 12: e0185895.

Abozaid H, Park YS, Tekian A. 2017. Peer review improves psychometric characteristics of multiple choice questions. Med Teach. 1–5.

Accreditation Council for Graduate Medical Education (ACGME) and American Board of Pediatrics. 2012. The Pediatrics Milestone Project; [accessed 2017 Aug 4]. https://acgme.org/Portals/0/PDFs/Milestones/PediatricsMilestones.pdf.

Alamoudi AA, El-Deek BS, Park YS, Al Shawwa LA, Tekian A. 2017. Evaluating the long-term impact of faculty development programs on MCQ item analysis. Med Teach. 39(sup1):S45–S49.

Anderson LW, Krathwohl D, editors. 2001. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York, NY: Longman Publishers.

Armstrong P. Bloom's Taxonomy; [accessed 2017 Jul 12]. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/.

Bloom BS. 1956. Taxonomy of educational objectives: The classification of educational goals. New York, NY: McKay.

Bordage G, Brailovsky C, Cohen T, Page GG. 1997. Maintaining and enhancing key decision-making skills from graduation into practice: An exploratory study. Seventh Ottawa Conference on Medical Education and Assessment: Advances in Medical Education, Maastricht, The Netherlands: Kluwer Academic Publishers.

Bordage G, Brailovsky C, Carretier H, Page G. 1995. Content validation of key features on a national examination of clinical decision-making skills. Acad Med. 70:276–281.

Bordage G, Carretier H, Bertrand R, Page G. 1995. Comparing times and performances of French- and English-speaking candidates taking a national examination of clinical decision-making skills. Acad Med. 70:359–365.

Brailovsky C, Allen T, Lawrence K, Crichton T, Laughlin T, Van der Goes T. 2014. Short answer questions based on Key Features have higher discrimination indices on a certification examination in family medicine Ottawa Conference. Ottawa, ON.

Case SM, Swanson DB. 1998. Writing written test questions for the basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners.

Cerutti B, Blondon K, Galetto A. 2016. Long-menu questions in computer-based assessments: a retrospective observational study. BMC Med Educ. 16:55.

Coderre S, Woloschuk W, McLaughlin K. 2009. Twelve tips for blueprinting. Med Teach. 31:322–324.

Downing S, Haladyna TA. 2006. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Assoc. Inc.

Durning SJ, Artino AR, Schuwirth L, van der Vleuten C. 2013. Clarifying assumptions to enhance our understanding and assessment of clinical reasoning. Acad Med. 88:442–448.

Escudier M, Woolford M, Tricio J. 2018. Assessing the application of knowledge in clinical problem solving: The structured professional reasoning exercise. Eur J Dent Educ. 22:e269–e277.

Eva K. 2005. What every teacher needs to know about clinical reasoning. Med Educ. 39:98–106.

Eva KW, Wood TJ, Riddle J, Touchie C, Bordage G. 2010. How clinical features are presented matters to weaker diagnosticians. Med Educ. 44:775–785.

Farmer E. 1998. Writing key feature problems. Australia: Royal Australian College of General Practitioners. [accessed 2018 June 12]. https://www.academia.edu/1749144/Writing_Key_Features_Problems.

Farmer EA, Hinchy J. 2005. Assessing general practice clinical decision making skills: the key feature approach. Austr Fam Phys 34: 1059–1061.

Farmer EA, Page G. 2005. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 39:1188–1194.

Fischer MR, Kopp V, Holzer M, Ruderich F, Junger J. 2005. A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. Med Teach. 27:450–455.

Frank JR, Snell LS, Sherbino J. 2014. The Draft CanMEDS 2015 Milestones Guide; [accessed 2017 Aug 4]. http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/canmeds_milestone_guide_sept2014_e.pdf.

General Medical Council. 2014. Good medical practice. United Kingdom: General Medical Council.

Glover Takahashi S, Herold J, Clark M, Nayer M, Beggs C, Corbett C, Drynan D, Cho N, Dignum T, Hudson B, Corbett K. 2012. The use of key features cases to assess clinical decision-making, CanMEDS roles & competence First Montreal Conference on Clinical Reasoning. Montreal, QC.

Glover Takahashi S, Herold J, Clark M, Nayer M, Drynan D, Cho N, Dignum T, Corbett K, Hudson B, Hynes M. 2013. Building better written exams – The use of key features cases to assess clinical decision-making, CanMEDS roles and competence International Conference on Residency Education (ICRE). Calgary, Alberta.

Haladyna TA, Downing S. 1993. How many options is enough for a multiple-choice teste item. Educ Psychol Meas. 53:999–1010.

Haladyna TM. 2004. Developing and validating multiple-choice test items. Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna TM, Downing SM. 1989a. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 2:37–50.

Haladyna TM, Downing SM. 1989b. Validity of a taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 2:51–78.

Haladyna TM, Rodriguez MC. 2013. Developing and validating test items. New York, NY: Routledge Taylor & Francis Group.

Hatala R, Norman GR. 2002. Adapting the Key Features Examination for a clinical clerkship. Med Educ. 36:160–165.

Hrynchak P, Glover Takahashi S, Nayer M. 2014. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ. 48:870–883.

Huwendiek S, Reichert F, Duncker C, de Leng BA, van der Vleuten CPM, Muijtjens AMM, Bosse HM, Haag M, Hoffmann GF, Tönshoff B, Dolmans D. 2017. Electronic assessment of clinical reasoning in clerkships: A mixed-methods comparison of long-menu key-feature problems with context-rich single best answer questions. Med Teach. 39:476–485.

Ilgen J, Humbert A, Kuhn G, Hansen M, Norman G, Eva KW, Charlin B, Sherbino J. 2012. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. Acad Emerg Med. 19:1454–1461.

Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. 2002. The quality of in-house medical school examinations. Acad Med. 77:156–161.

Kilgour JM, Tayyaba S. 2016. An investigation into the optimal number of distractors in single-best answer exams. Adv Health Sci Educ Theory Pract. 21:571–585.

Krathwohl D. 2002. A revision of Bloom's taxonomy: An overview. Theory Pract. 41:212–218.

Lang AJ, Bronander K, Harrell H, Kovach R, Monteiro S, Bordage G. 2014. Validity evidence for a key features examination to assess clinical decision making in the internal medicine clerkship 16th Ottawa Conference. Ottawa, ON.

Lawrence K, Allen Brailovsky T, Crichton C, Bethune T, Donoff C, Laughlin M, Wetmore TS, Carpentier M-P, Visser S. 2011. Defining competency-based evaluation objectives in family medicine: key-feature approach. Canadian Family Physician 57:e373–e380.

Leung F-H, Herold J, Iglar K. 2016. Family medicine mandatory assessment of progress: results of a pilot administration of a family medicine competency-based in-training examination. Can Fam Physician 62:e263–e267.

Medical Council of Canada. 2012. Guidelines for the Development of Key Feature Problems and Test Cases; [accessed 2017 Jan 26] http://mcc.ca/wp-content/uploads/cdm-guidelines.pdf.

Medical Council of Canada. 2014. Blueprint Project: Qualifying examinations blueprint and content specifications. Ottawa, ON, Medical Council of Canada.

Norman GR, Arfai B, Gupta A, Brooks LR, Eva KW. 2003. The privileged status of prestigious terminology: impact of "medicalese" on clinical judgments. Acad Med. 78:S82–S84.

Norman G, Bordage G, Page G, Keane D. 2006. How specific is case specificity? Med Educ. 40:618–623.

Page GG. 1999. Writing key feature problems for the clinical reasoning skills examination; [accessed 2017 Jan 26]. http://www.idealmed.org/workshop/SectionD-KeyFeatures.pdf.

Page GG, Bordage G. 1995. The Medical Council of Canada's Key Features Project: A more valid written examination of clinical decision-making skills. Acad Med. 70:104–110.

Page GG, Bordage G, Allen T. 1995. Developing key-feature problems and examinations to assess clinical decision-making skills. Acad Med. 70:194–201.

Paniagua MA, Swygert KA. 2016. Writing written test questions for the basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners.

Pharmaceutical Society of Australia. 2010. National Competency Standards Framework for Pharmacists in Australia. Australia: Pharmaceutical Society of Australia.

Piasentin KA. 2010. Exploring the optimal number of options in multiple-choice testing. CLEAR Exam Rev (Winter). 18–22.

Rodriguez MC. 2005. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. Educ Meas Issues Prac (Summer). 24:3–13.

Rotthoff T, Baehring T, Dicken HD, Fahron U, Richter B, Fischer MR, Scherbaum WA. 2006. Comparison between Long-Menu and Open-Ended Questions in computerized medical assessments. A randomized controlled trial. BMC Med Educ. 6:50.

Royal Australian College of General Practitioners. 2015. Competency profile of the Australian general practitioner at the point of Fellowship. Australia: Royal Australian College of General Practitioners. [accessed 2018 June 12] https://www.racgp.org.au/download/Documents/VocationalTrain/Competency-Profile.pdf

Royal College of Physicians and Surgeons of Canada. 2005. CanMEDS 2005 Framework. Ottawa, ON: Royal College of Physicians and Surgeons of Canada.

Schneid SD, Armour C, Park YS, Yudkowsky R, Bordage G. 2014. Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. Med Educ. 48:1020–1027.

Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. 2001. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ. 35:348–356.

Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Blackmore D, Winslade N, Girard N, et al. 2007. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA 298:993–1001.

Tamblyn R, Abramowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Eguale T, Winslade N, Girard N, Bartman I, Buckeridge D, Hanley J. 2010. Influence of physicians' management and communication ability on patients' persistence with antihypertensive medication. Arch Intern Med. 170:1064–1072.

The Royal College of Obstetricians and Gynecologists. Blueprint Grid for the Membership of the Royal College of Obstetricians and Gynecologists Examination; [accessed 2018 Feb 8]. https://www.rcog.org.uk/globalassets/documents/careers-and-training/mrcog-exam/part-1/ex-part-1-blueprinting-grid-new.pdf.

University College Cork Ireland. 2015. How to Use the Draft Blueprint for the Pre-Registration Examinations (PRES) Level 3; [accessed 2018 Feb 8]. https://www.medicalcouncil.ie/Information-for-Doctors/Examinations-/How-to-use-the-Blueprint-for-the-PRES.pdf.

van Bruggen L, Manrique-van Woudenbergh M, Spierenburg E, Vos J. 2012. Preferred question types for computer-based assessment of clinical reasoning: a literature study. Perspect Med Educ. 1:162–171.

Van der Vleuten C, Newble D. 1995. How can we test clinical reasoning? Lancet. 345:1032–1034.

Wenghofer E, Klass D, Abrahamowicz M, Dauphinee D, Jacques A, Smee S, Blackmore D, Winslade N, Reidel K, Bartman I, Tamblyn R. 2009. Doctor scores on national qualifying examinations predict quality of care in future practice. Med Educ. 43:1166–1173.

World Health Organization (WHO). 2016. International Classification of Diseases, ICD10; [accessed 2018 Mar 28]. http://apps.who.int/classifications/icd10/browse/2016/en#/V.

NBME®

NBME® ITEM-WRITING GUIDE

Constructing Written Test Questions for the Health Sciences

FEBRUARY 2021

CHAPTER 2: MULTIPLE-CHOICE ITEM FORMATS

One of the most crucial aspects of a multiple-choice test item or question (MCQ) is its type or structure. Different item types can be used for different topic areas, and each item type carries with it advantages and disadvantages. A critical aspect to consider when choosing an item type is the inclusion of potential flaws that might benefit the savvy test-taker or introduce irrelevant difficulty. This chapter covers the basics of several multiple-choice item formats and introduces some potential flaws that are common to specific formats, while Chapter 3 will discuss specific item flaws in more detail.

ONE-BEST-ANSWER VS. TRUE-FALSE ITEMS

The universe of multiple-choice items can be divided into two families:

- Items that require test-takers to indicate a single, most accurate response (one-best-answer)
- Items that require test-takers to indicate all responses that are appropriate (true-false)

NBME has used multiple item formats within each family in the past, listed below by designating letter.

One-best-answer item formats that require test-takers to select the single best response:

- A-type (4 or more options, single items or sets)
- F-type (2 to 3 items grouped into a set around specific content or case scenario where test-takers **cannot** return to previously seen items in the set)
- G-type (2 or 3 items grouped into a set around specific content where test-takers **can** return to previously seen items in the set)

True-false item formats that require test-takers select some set of options that are true:

- C-type (A/B/Both/Neither response items)
- K-type (complex true-false items)
- X-type (simple true-false items)

The letters used to label the item formats hold no intrinsic meaning; letters were assigned more or less sequentially to new item formats as they were developed. For an extended list of item types formerly used by NBME, ordered by their designated letters, see Appendix C: NBME Retired Item Formats.

The One-Best-Answer Family

In contrast to true-false questions, one-best-answer questions are designed to make explicit that only one option is to be selected. These items are the most widely used multiple-choice item format. They consist of a stem, which most often includes a vignette (eg, a clinical case presentation) and a lead-in question, followed by a series of option choices, with one correct answer and anywhere from three to seven distractors. The incorrect option choices should be directly related to the lead-in and be homogeneous with the correct answer. This item describes a situation (in this instance, a patient scenario) and asks the test-taker to indicate the most likely cause of the problem.

Sample Stem (Vignette & Lead-in) With Option Set

VIGNETTE:

A 32-year-old man comes to the office because of a 4-day history of progressive weakness in his extremities. He has been healthy except for an upper respiratory tract infection 10 days ago. His temperature is 37.8°C (100.0°F), pulse is 94/min, respirations are 42/min and shallow, and blood pressure is 130/80 mm Hg. He has symmetric weakness of both sides of the face and the proximal and distal muscles of the extremities. Sensation is intact. No deep tendon reflexes can be elicited. Babinski sign is absent.

LEAD-IN:

Which of the following is the most likely diagnosis?

OPTION SET:

A. Acute disseminated encephalomyelitis
B. Guillain-Barré syndrome*
C. Myasthenia gravis
D. Poliomyelitis
E. Polymyositis

Note that the incorrect options are not wholly wrong. The options can be diagrammed as follows:

D C A E B

Least Likely Diagnosis — Most Likely Diagnosis

Even though the incorrect options are not completely wrong, they are less correct than the “keyed answer” (indicated by the asterisk in the option set). The test-taker is instructed to select the “most likely diagnosis.” Experts would all agree that the most likely diagnosis is B; they would also agree that the other diagnoses are somewhat likely, but less likely than B. As long as the options can be laid out on a single continuum, in this case from “Least Likely Diagnosis” to “Most Likely Diagnosis,” distractors in one-best-answer items do not have to be totally wrong.

"Cover-the-Options" Rule

This leads us to another important guideline for writing good one-best-answer items—the "cover-the-options" rule. If a lead-in is properly focused, a test-taker should usually be able to read the vignette and lead-in, cover the options, and guess the correct answer without seeing the option set. For example, in this next item, after reading the lead-in, the test-taker should be able to answer the item without seeing the options. When writing items, covering the options and attempting to answer the item is a good way to check whether this rule has been followed.

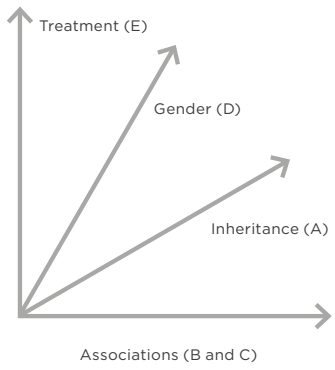
Example of "Cover-the-Options" Rule

A 58-year-old man comes to the office because of pain in the right knee for the past 3 days. He has a history of type 2 diabetes mellitus, hypertension, and hyperlipidemia controlled with daily glyburide, lisinopril, and atorvastatin. There is no family history of similar disorders. On physical examination, the knee is swollen, passive motion produces pain, and ballottement discloses an effusion. Synovial fluid is cloudy and contains positive birefringent crystals and no bacteria. X-ray shows chondrocalcinosis. Which of the following is the most appropriate pharmacotherapy?

A. Allopurinol
B. Betamethasone
C. Ibuprofen*
D. Infliximab
E. Levofloxacin

Homogeneous Options

Along with a focused lead-in, a good item will have a keyed answer and distractors that are homogeneous. They all directly address the lead-in in the same manner and can be rank ordered along a single dimension. The one-best-answer example below is a flawed item that can occur when options are not listed on a single dimension. After reading the lead-in, the test-taker has only the vaguest idea what the question is about. In order to determine the "best" answer, the test-taker must decide whether "it occurs frequently in women" is more or less true than "it is seldom associated with acute pain in a joint." The diagram of these options might look like the figure to the left of the sample item below. The options are heterogeneous and deal with miscellaneous facts; they cannot be rank ordered from least to most true along a single dimension. Although this item appears to assess knowledge of several different points, its inherent flaws preclude this. The item by itself is not clear; the item cannot be answered without looking at the options.



Which of the following is true about pseudogout?

A. It is clearly hereditary in most cases
B. It is seldom associated with acute pain in a joint
C. It may be associated with a finding of chondrocalcinosis
D. It occurs frequently in women
E. It responds well to treatment with allopurinol

General Rules for One-Best-Answer Items

Because test-takers are required to select the single best answer, one-best-answer items must satisfy the following rules (for more detail, see Chapter 6):

- Item and option text must be clear and unambiguous. Avoid imprecise phrases such as "is associated with" or "is useful for" or "is important"; words that provide cueing such as "may" or "could be"; and vague terms such as "usually" or "frequently."
- The lead-in should be closed and focused and ideally worded in such a way that the test-taker can cover the options and guess the correct answer. This is known as the "cover-the-options" rule.
- All options should be homogeneous so that they can be judged as entirely true or entirely false on a single dimension.
- Incorrect options can be partially or wholly incorrect.

Recommendations for Using One-Best-Answer Items

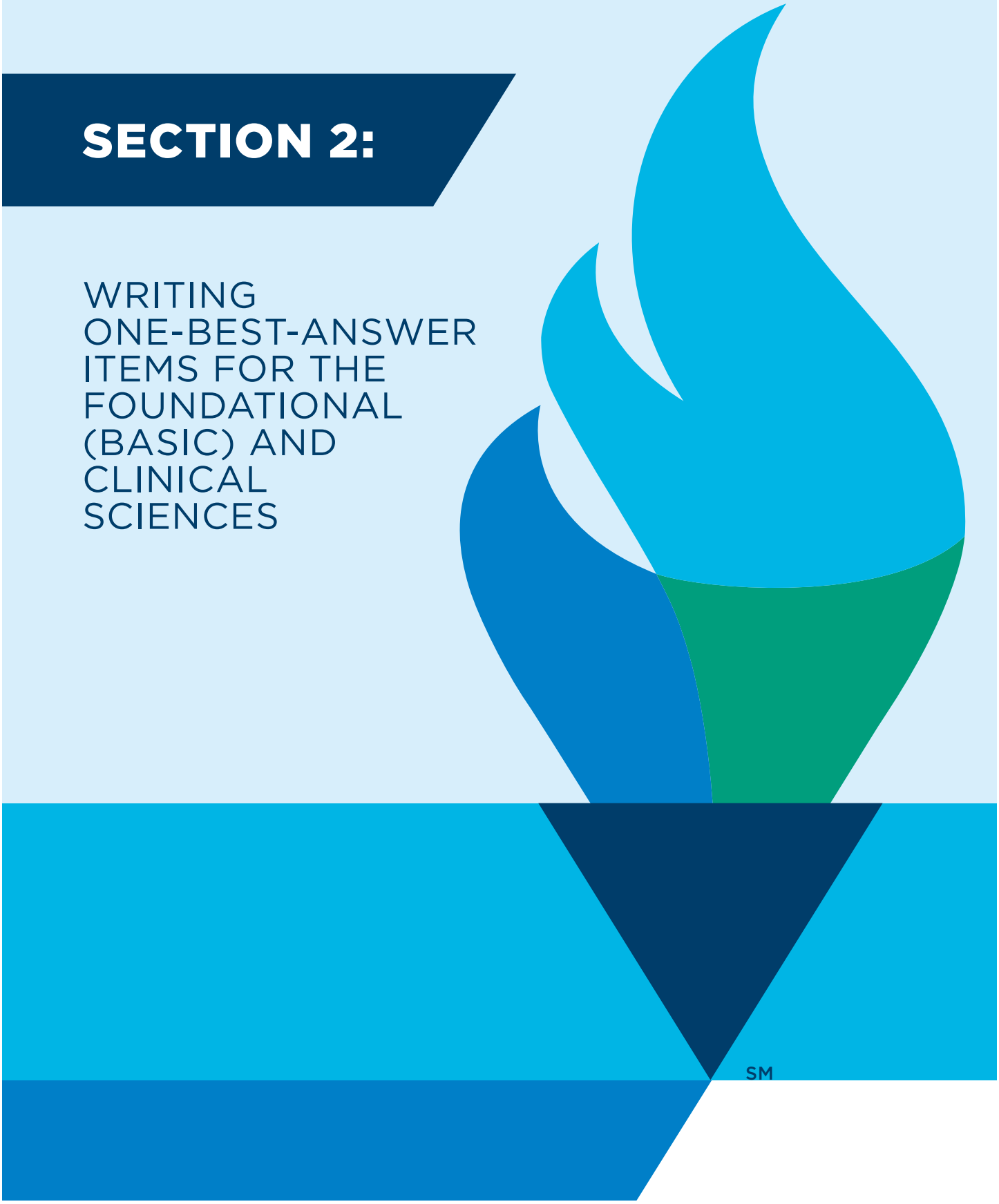
We recommend using one-best-answer items whenever possible. This format helps prevent confusion on the part of the test-taker from having to guess the author's intent. In addition, this format can also be easier and more efficient to write because wrong options do not have to be entirely incorrect, and different lead-ins can be paired with the same stems (a patient scenario stem can include items with lead-ins about diagnosis and management) to create item sets. NBME currently uses only one-best-answer format items on exams.

See Appendix C: NBME Retired Item Formats for a historic list (and examples) of these retired item formats.

NOTES

SECTION 2:

WRITING ONE-BEST-ANSWER ITEMS FOR THE FOUNDATIONAL (BASIC) AND CLINICAL SCIENCES



CHAPTER 5: BASIC RULES FOR WRITING ONE-BEST-ANSWER ITEMS

RULE 1: Each item should focus on an important concept or testing point.

As a health care provider or educator involved in the development of an examination, you may be asked to write items to assess test-taker knowledge of a particular domain. What do you want the test-taker to know or demonstrate? The topic of the item usually results from the examination blueprint, which is the outline of the major topics to be covered. For instance, if an examination is intended to assess knowledge of the cardiovascular system, the blueprint might have two dimensions: 1) disease-based (eg, hypertension, ischemic heart disease, systolic heart failure), and 2) task-based (eg, assessment of foundational science principles, diagnosis, history, prognosis). The blueprint would likely include topics along both dimensions and might call for six items on hypertension, four on systolic heart failure, two on diastolic heart failure, ten on ischemic heart disease, and so on. Along the task dimension there might be a similar distribution of topics. A clear and comprehensive blueprint or other set of test specifications should always be available so that item writers can stay focused on the important topics and write a sufficient number of items for each topic.

RULE 2: Each item should assess application of knowledge, not recall of an isolated fact.

The first step in writing an item is to develop an appropriate stimulus to introduce the topic, such as a clinical or experimental vignette, to provide context to the question being asked. If there is no such stimulus, the resulting item will generally be assessing knowledge recall. Recall items make it difficult for the educator to assess any higher level within Bloom's taxonomy, such as "application of knowledge." For instance, an item consisting of one sentence, "Which of the following medications is used to decrease preload in systolic heart failure?" would assess only the recall on the mechanisms of action of a list of pharmacotherapeutic agents.

It can be helpful to use actual patient scenarios that you previously encountered as a source of ideas for items and vignettes. However, you should avoid relying on or adhering too closely to patient cases because these often have atypical features that may divert from a typical or representative case and lead to confusion. Additionally, in some instances, such as the example with systolic heart failure, there will be an additional step that you must keep in mind: you should consider the underlying cause of the heart failure. Patient demographics, past medical history, and other factors will differ depending on the cause of the condition. Patients with systolic heart failure from a viral cardiomyopathy versus from ischemic heart disease may have different demographics and a different history (eg, a younger patient with a viral illness preceding the onset of heart failure symptoms as compared to an older patient with risk factors for ischemic heart disease).

The details of the vignette should be guided by the level of the test-taker. Here are two examples for test-takers with two levels of education/experience:

Test-taker with Less Education/New Experience: A systolic heart failure vignette for a second-semester first-year medical student would include very typical features and classic symptoms: shortness of breath with physical activity that improves with rest; awakening at night short of breath, relieved by sitting up; pedal edema; and pertinent negatives such as the absence of chest pain. Risk factors might include an upper respiratory illness two weeks ago or a history of heavy alcohol ingestion over 20 years.

Test-taker with More Advanced Education/Skills: A test-taker, such as one sitting for a specialty certifying examination, would be able to work through a vignette that included some atypical features, as is the case with many actual patients. The demographic information may or may not be significant for the more advanced test-takers. For instance, every patient lives somewhere, and many will have a current or past occupation that may or may not be related to the cause of their illness. In a vignette for a 30-year-old man with shortness of breath and wheezing in which the diagnosis is asthma, the demographic information might or might not be related to the diagnosis. The patient might be a farmer, but the most likely diagnosis is still asthma and not farmer's lung or silo-filler's lung.

RULE 3: The item lead-in should be focused, closed, and clear; the test-taker should be able to answer the item based on the vignette and lead-in alone.

The next step in item writing is to phrase the question with the use of a lead-in, where the accompanying vignette allows the lead-in to be focused on the patient, such as, "Which of the following is the most appropriate next step in management?" or "Which of the following is the most likely diagnosis?" An open-ended lead-in such as, "The diagnosis in the patient is:" should be avoided. The lead-in should be a single, closed, clear question. Ideally, after reading the vignette and the lead-in, a test-taker should be able to answer the item without seeing the options. Another reason to use a closed lead-in is because it helps to avoid certain item flaws, such as grammatical cueing.

RULE 4: All options should be homogeneous and plausible to avoid cueing to the correct option.

Homogeneity:

At this point in item writing, the patient-based, closed lead-in created with Rule 3 in mind will direct the focus and grammatical form of the answer options. Maintaining a consistent focus and parallel format among the answer options results in homogeneity, which allows test-takers to weigh each option within a single mindset without construct-irrelevant distractions. For example, in response to "Which of the following is the most likely cause of this patient's condition?," a list of answer options in which all choices are diagnoses (eg, tuberculosis, meningitis, etc.), is easier to process than a list containing both diagnoses and underlying pathogens (eg, tuberculosis, *Neisseria meningitidis*, etc.).

Plausibility:

The correct answer should always be the "most" correct of the answer options, but the distractors should be plausible enough to entice test-takers who do not know the correct answer. Otherwise, test-takers can arrive at the correct answer by eliminating distractors based on their improbability within the context of the patient scenario.

When writing answer options, start by generating the correct answer for the lead-in. Generating parallel and plausible yet incorrect distractors is more challenging. For questions regarding diagnosis, the topic area may be the answer—if you are assigned to write two items on community-acquired pneumonia (CAP), one item on

the diagnosis and one item on management, the assignment has already generated the keyed answer for the lead-in, “Which of the following is the most likely diagnosis?” Examples of reasonable distractors in an item in which the correct diagnosis is CAP could include pulmonary embolus, lung cancer, and pneumothorax.

RULE 5: Each item should be reviewed to identify and remove technical flaws that add irrelevant difficulty or benefit savvy test-takers.

Once you have written your item, take a step back and look closely at its structure. The bulk of the text (vignette or case information) should precede, rather than follow, the lead-in. The clinical or experimental vignette should make sense and follow a logical sequence: first list the patient demographics, then history, physical examination, laboratory data, and so on. The use of a template to ensure all of these sections are in place and correctly structured is highly recommended. As you review your item, ask yourself the following questions. If the options were removed, could a knowledgeable test-taker answer the question correctly? Is there anything in the phrasing or text that would confuse the knowledgeable test-taker? Are there any clues to help a testwise student guess the item correctly? Finally, you should ask a colleague to review the items you have written, particularly for content, clarity, and appropriateness for your test-taker population.

NOTES

CHAPTER 6: TESTING APPLICATION OF FOUNDATIONAL (BASIC) AND CLINICAL KNOWLEDGE

CHOOSING THE TOPICS TO TEST

The content of an exam should be driven by the purpose of that exam and the test-taker population. Who is being tested and how will the scores be used? For example, the USMLE system is designed for use by state medical licensing authorities in their decision to grant a general licensure for allopathic and international graduate physicians or other providers in the United States. The focus is to assess knowledge of content that is necessary for the practice of medicine by the undifferentiated physician or other provider; items might be included on USMLE that assess knowledge not uniformly taught in medical school. Conversely, topics within some medical schools' curriculum might be omitted from the exam. The analogy for individual schools and courses within schools is to determine the student test-taker population and purpose of the scores. If the purpose of the content is to test for specialty or subspecialty certification for physicians or other providers, the content and its inferences should be geared toward the required minimum competence in that specialty upon beginning practice. An exam that is intended for formative feedback at a midpoint of a course will have a different focus and different content from an exam to determine end-of-clerkship grades.

DETERMINING LEVEL OF COGNITION TO ASSESS

Items can be grouped into two general categories, based on the cognitive task required of the test-taker:

1. **Recall of a Fact:** An item that assesses rote memory of a fact (without requiring its application).
2. **Application of Knowledge:** An item that requires a test-taker to apply knowledge to reach a conclusion, make a prediction, or select a course of action that does not depend on memory alone.

Items that test recall of a fact require test-takers to read an item and to recall isolated facts, concepts, and principles or to recognize previously encountered situations (eg, experiments, patient encounters, case studies). These items often begin by citing a disease and then asking what patient findings are expected. For example, "Which of the following findings is most likely to be seen in postsurgical patients with pulmonary embolism?" is an item structured similarly to most textbook questions. The test-taker could look up the disease and find the answer in a single paragraph. From a practical standpoint, these items also seem clinically backwards—patients would not tell their provider what disease they have and then ask the provider to determine the signs and symptoms.

Application of knowledge items, on the other hand, require test-takers to read an item and identify relevant information, interpret that information in a certain context, integrate that information with what they already know, and then answer the question posed. Vignette-based items (items that include a detailed patient or experimental scenario) often provide a vehicle for eliciting the demonstration of these higher-order thinking skills. Some examples of these application of knowledge items can be found throughout this book.

Determining the cognitive task for an item – recall vs application of knowledge – depends on the intended end-use of the item. The use of recall items may be best for formative assessment purposes or the evaluation of simpler concepts that might not lend themselves to clinical or experimental scenarios. For a medium-to high-stakes summative examination, use of vignette-based items that require higher-order thinking skills and application of knowledge would be preferable to simple recall items.

Figure 1 provides a few guidelines to determine which cognitive task might be most suitable for your purposes. Figure 2 shows how the same subject matter could be tested as either a recall of a fact item or an application of knowledge item.

Figure 1. Promoting Versatility in Item Creation: Recall vs Application of Knowledge

COGNITIVE TASK	FORMATIVE ASSESSMENT	SUMMATIVE ASSESSMENT
Recall	▶ Can easily detect skill deficits ▶ Can read questions quickly ▶ Helpful in classroom instruction ▷ "Rapid fire" simulation of learning ▷ Attention-keeping	▶ Best for testing simple concepts ▶ Support quantity over quality ▶ May seem artificial and lack realism ▶ May promote shallow study habits ▶ No longer consistent with NBME exams
Application of Knowledge	▶ Promote clinical reasoning ▶ Encourage problem-based learning ▶ Support team-based learning ▶ Provide clinical/experimental context	▶ Can test recall and reasoning ▶ Can test integration of skills but more difficult to detect specific deficiencies ▶ Better approximates real life ▶ May need more testing time to adequately sample the domain

NOTES

Figure 2. Example of a clinical scenario as a recall item vs an application of knowledge item

A woman is diagnosed with venous thromboembolism. Which of the following is the most appropriate treatment?

- This is an example of a recall of a fact item, which requires examinees to simply recall the treatment of venous thromboembolism.

USE A VIGNETTE AS A VEHICLE TO TEST APPLICATION OF KNOWLEDGE

A 47-year-old woman comes to the emergency department because of shortness of breath and left lower extremity pain with ambulation. Yesterday she returned from Europe after a 10-hour flight. She has no remarkable medical history and takes only an oral contraceptive daily. Vital signs are within normal limits. Physical examination shows asymmetrical swelling of the left calf of greater than 2 cm compared to the right side. The remainder of the physical examination discloses no abnormalities.

- This vignette provides more clinical context, requiring the examinee to recognize the historical and physical presentation of venous thromboembolism.
- From this, a variety of questions could be asked that require the examinee to interpret data, including the following:
  - Diagnosis
  - Next step in determining the diagnosis (which requires clinical suspicion of diagnosis and the next test to order)
  - Treatment (which requires clinical suspicion of diagnosis and determination of the most appropriate next step in management)
  - Mechanism (which requires clinical suspicion of diagnosis, the most appropriate treatment, and the mechanism of that treatment)

Benefits of Application of Knowledge Item Type

Items with a clinical vignette to assess application of knowledge have several benefits:

1. The authenticity of the examination is greatly enhanced by using items that require test-takers to integrate information to "solve" clinical problems.
2. The items are more likely to focus on important information, rather than trivia.
3. These items help to identify those test-takers who have memorized a substantial body of factual information but are unable to use that information effectively in clinical situations. Test-takers need to differentiate relevant from irrelevant information in the item.

GUIDELINES FOR CLINICAL VIGNETTE CONTENT

- Test application of knowledge using clinical vignettes to pose medical decisions in patient care situations
- Focus items on common or potentially catastrophic problems; avoid “zebras” and esoterica
- Pose clinical decision-making tasks that would be expected of a successful test-taker
- Pose/craft clinical situations that would be handled by a provider/specialist
- Focus on specific tasks that the successful test-taker must be able to undertake at the next stage of training or upon commencement of specialty practice
- Focus on areas in which clinical reasoning mistakes are often made

The following can be used as a template for a patient vignette; not all of the following components are necessary, but when present should be in the order indicated:

- Age, gender (eg, 45-year-old woman) (add option to indicate self-identified)
- Site of care (eg, the emergency department)
- Presenting symptoms (eg, headache)
- Duration of symptoms (eg, 2 days)
- Patient history, including past medical history, family history, psychosocial history, and review of systems if important and plausible for the scenario
- Physical findings
- Results of diagnostic studies
- Initial treatment, subsequent findings

NOTES

The Shape of a Good Item

A well-constructed one-best-answer item will have a particular silhouette as shown in the illustration below. A rich clinical scenario serves as the stem, and all of the options are listed in a concise and uniform manner. The stem should include all relevant facts; no additional information should be provided in the options.

Tell your story here in the **vignette**.

Pose your question here in the **lead-in**.

A.
B. Insert your answer **option set**
C. here, making sure it follows
D. the **"cover-the-options"** rule.
E.

Make sure the item stem adheres to the following rules:

- Focuses on important concepts rather than trivial facts
- Can be answered without looking at the options
- Includes all relevant data should be provided in the options
- Is not "tricky" or overly complex
- Is not negatively phrased (eg, avoid using "except" or "not" in the lead-in)

NOTES

Patient Characteristics in Item Creation

Characteristics of a patient such as age, sex, gender identity, disability, socioeconomic status, native language, country of origin, and/or occupation are sometimes mentioned within case vignettes in test items. Some patient characteristics (PC) may be important inputs into the diagnostic reasoning process. Others may lead to incorrect conclusions and misdiagnoses. Among the latter are characteristics that could potentially be associated with harmful patient stereotypes.

When creating items, be mindful of the notion that race is a social construct not linked to biology or susceptibility to disease. This is similarly true of ethnicity and culture, heritage, or even country of origin. Ancestry, if known, may be biologically important, and thus may be relevant to factors relating to health and disease. In addition, when and if these characteristics are to be considered for inclusion in your items, they should be considered based on patient self-report, not the assumption of a health care provider.

PC can be described and included in vignettes if they:

- are clinically relevant and/or could aid in distractor quality.
- are necessary for the examinee to better understand the context in which the patient is being seen (the item would be unreasonably difficult if excluded).
- add to the overall exam-level representativeness of the referenced patient population.
- increase the probability of detection, diagnosis, or recognition of an otherwise rare condition.
- do not contain negative stereotypes.

Test items should be carefully designed to measure meaningful and plausible testing points (eg, diagnosis, management, etc.), without the influence of assumptions, bias, or stereotypes. When examinees select the correct (keyed) response, they are given credit because they are demonstrating what the examination item is designed to measure. Health professionals and educators creating assessments should follow guidelines that encourage thoughtful consideration of PC, while at the same time strive to promote diversity and present patients who reflect the population served by your examinees.

Examples of Challenging Patient Characteristics Scenarios

EXAMPLE A:

A 2-year-old girl is brought to the office because of bilateral hand swelling and splenomegaly……. Her family identifies as African American.

In this example, the lack of inclusion of any patient characteristics (PC) could make clinical reasoning very difficult. However, given sickle cell disease's prevalence in Latin America, the Middle East, and Southern European regions such as Turkey, Greece, and Italy, it would be reasonable to change the PC to one that represents one of the regions where this disease association exists, but is less well known.

"... SOME PATIENT CHARACTERISTICS (PC) MAY BE IMPORTANT INPUTS INTO THE DIAGNOSTIC REASONING PROCESS. OTHERS MAY LEAD TO INCORRECT CONCLUSIONS AND MISDIAGNOSES..."

STRUCTURING ITEMS TO FIT TASK COMPETENCIES

A set of defined task competencies will assist the item writer in focusing his or her intended testing point. Each competency requires a slightly different approach to item writing. Some sample lead-ins and example items to guide item-writing efforts for each physician (or other provider) task competency are provided below. Additional lead-ins can be found in Appendix B: Sample Lead-ins Based on Provider Task Competencies.

Foundational (Basic) Science

Foundational science comprises items that require understanding and application of basic science. These items should require clinical knowledge as well as knowledge of one or more foundational science principles that would likely have been learned during preclinical study and reinforced during clinical rotations. Mechanisms of disease is an example of a competency within the foundational science category. Items in this competency should evaluate test-takers' knowledge of pathophysiology in its broadest sense, including etiology, pathogenesis, natural history, clinical course, associated findings, complications, severity of illness, and intended or unintended effects of therapeutic interventions. These items should be framed in a clinical context. In general, the writer should open items on mechanisms of disease with a clinical vignette featuring a patient with specified symptoms, signs, history, and lab study findings. The following lead-ins are examples of those used to test foundational science principles:

- Which of the following is the most likely cause/mechanism of this effect?
- Which of the following is the most likely infectious agent?
- Which of the following is the most likely explanation for these findings?
- Which of the following is the most likely location of this patient's lesion?
- Which of the following is the most likely pathogen? (Interpretation of foundational-science–based information, such as Gram stain results, should be required in the vignette to differentiate the competency being tested from that of a clinically based most likely diagnosis item.)
- Which of the following findings is most likely to be increased/decreased in this patient?
- A biopsy specimen is most likely to show which of the following?
- This patient most likely has a defect in which of the following?
- This patient most likely has a deficiency in which of the following enzymes?
- Which of the following cytokines is the most likely cause of this condition?
- Which of the following structures is at greatest risk for damage during this procedure?
- The most appropriate medication for this patient will have which of the following mechanisms of action?

A 10-year-old girl develops gross hematuria 14 days after a sore throat. She has a blood pressure of 170/100 mm Hg and 2+ pedal and pretibial edema. Serum urea nitrogen concentration is 3.2 mg/dL. Which of the following is the most likely cause?

A. Acute postinfectious glomerulonephritis*
B. Microscopic polyangiitis
C. Minimal change disease
D. Thin basement membrane nephropathy

SELECTING MEDIA

In multiple-choice examinations, media should be purposefully selected to help the student answer the question; otherwise, it is simply extraneous information. Do not describe with text that which can be easily demonstrated in the media itself. In the example below, three similar items are shown with differing lead-ins and media; no media (Example A), a static image depicting heart rhythms or sounds (Example B), and an avatar simulating placement of the stethoscope (Example C). Other possibilities include showing both the ECG and the avatar or presenting the audio file of the corresponding heart sounds with or without a live patient video.

Consider the following stem for a cardiology multiple-choice question:

> A 27-year-old man who is a US Army veteran comes to the office because of periodic dizziness, palpitations, and chest tightness during the past 3 weeks. The episodes occur when he remembers "the roadside bomb that took my friend." He has had difficulty sleeping and drinks 1 pint of vodka daily to help with "nerves." He has no documented medical history and takes no medications. Temperature is 36.7°C (98.1°F), pulse is 90/min, respirations are 20/min, and blood pressure is 128/80 mm Hg.

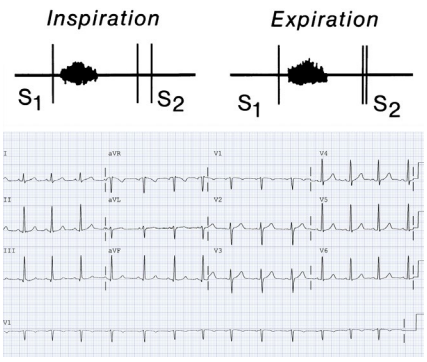
Below are three possible lead-ins and media selections for the above stem.

EXAMPLE A (no image)

Which of the following is the most likely finding on cardiac auscultation of this patient?

A. Normal examination*
B. Opening systolic snap
C. S4 gallop
D. S3 gallop
E. Systolic flow murmur

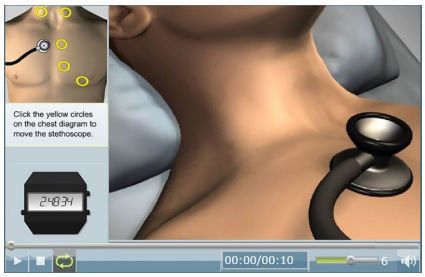
EXAMPLE B (with ECG image)



An ECG is shown. Which of the following is the most likely finding on cardiac auscultation?

(same options as above)

EXAMPLE C (with avatar that allows auscultation of actual heart sounds through headphones)



An avatar is shown. Click the yellow circles to hear the cardiac examination. Which of the following is the most likely finding on cardiac auscultation?

(Same options as the preceding example)

CONTENT AREAS CONDUCIVE TO THE USE OF MEDIA

Certain content areas lend themselves well to the use of media, such as:

- Dermatologic and musculoskeletal examination findings
- Cardiology (such as heart sounds)
- Neurologic examination findings
- Ethical and communication scenarios

Examples of two of these areas follow.

Dermatologic and Musculoskeletal Examination Findings

Dermatologic and musculoskeletal examination findings in particular benefit from the use of media. Showing findings, rather than describing the findings with text, simulates real clinical practice. Further, research has shown that response time is faster with the use of media compared with text for dermatologic findings. Consider the two following examples.

Example Item Using Text

A 79-year-old woman comes to the office 8 weeks after noticing a nontender nodule on the back of her left hand. She initially thought it was an insect bite, but it has grown in size over the past week. It bleeds when she picks at it. She has no history of serious illness. She lives in a retirement community in Texas and is an avid gardener. Examination of the dorsum of the left hand shows a 2-cm lesion that is well-demarcated, raised, and flesh-colored at the margins, with a necrotic center. Which of the following is the most appropriate next step in management?

A. Cryotherapy
B. Electrocautery ablation
C. Excision of the lesion*
D. Topical ketoconazole
E. Observation

Example Item Using Media (screenshot from an approximately 30-second interaction)

An 83-year-old woman is hospitalized for pneumonia and renal failure. She has a history of dementia, Alzheimer type, and resides in a nursing care facility. She has been offered but has refused dialysis. The patient has not designated a durable power of attorney, but she does have an advance directive that states, "No CPR, no intubation, no dialysis, and no surgery." The patient's niece, who is her closest relative, has a discussion with the provider about her aunt's refusal of treatment. Play the video to view the conversation. Which of the following is the most appropriate response to the niece?

(Same options as the preceding example)

ACQUIRING AND CREATING MEDIA

When determining new media needs, a subject matter expert group can be helpful as part of the process to oversee and monitor the acquisition process. This group can develop a list of diseases, conditions, and/or provider tasks and skills that are best illustrated with media. Once media are acquired, this group can develop exemplars to distribute for item-writing assignments. A good media image is one to which multiple test items can be written; this allows for the highest chance of the image being suitable for the exam and helps address the issue of memorability. It can also offset the cost of acquiring media. We also encourage diversity in portrayals of common conditions (eg, different skin tones and types). See "patient characteristics" section on page 42.

When acquiring media, two important considerations are patient confidentiality and metadata (ie, the information that accompanies and identifies each media image). If actual patient images or videos are used, it is important to maintain patient confidentiality. Ensure that neither the patient nor the institution can be identified from any clues in the media. For guidance, refer to your institution's patient confidentiality policy and HIPAA guidelines (http://www.hhs.gov/hipaa/for-professionals/index.html).

Metadata is the identifying information that accompanies each piece of media. It is important to obtain as much metadata as possible about the media to help with indexing/searching and reuse in the future. Think about search terms and metadata that will help with identification of images that will be used more than once. It is advisable to set standards for media and copyright of media and to use a form to record as much metadata as possible during the acquisition stage. The following is a sample list of information to collect and record for media material:

- Administrative details
- Age of the patient
- Diagnosis
- Keywords
- Description of the test being performed
- Normal or abnormal results
- Descriptive file name
- Patient ID/name
- Indication patient signed a consent form
- In/out cut points for individual clips
- Whether the clip contains important audio

REMEMBER: YOUR MEDIA ARE ONLY AS GOOD AS THEIR METADATA!

Media have little value in item writing if they cannot be retrieved easily in search results.

APPENDIX A: A QUICK REFERENCE GUIDE TO APPROACHING ITEM WRITING

Getting Started	▶ Consider the curriculum: What needs to be covered? At what level of learner? ▶ Consider classic and more common presentations you've encountered in your own practice/setting as a starting point ▶ Think of what you would want your test population to recognize (or not miss)	"...just saw a classic case of renal artery stenosis last week..."
The Testing Point	▶ Think of what you want to test (the "testing point") as you start writing	Principles of therapy in renal artery stenosis...
The Site of Care	▶ Where do you envision the site of care for the case?	Would I treat this in the clinic? Hospital?
The Case	▶ Start outlining the case that frames your testing point (see Chapter 6: Testing Application of Foundational [Basic] and Clinical Knowledge) ▶ View sample items as a guide to the style (see Chapter 6 under "Structuring Items to Fit Task Competencies") ▶ Consider if an image or other media could work as well as or better than a text description (keeping in mind that the question should not be answerable based on the image alone) (see Chapter 7: Using Media as Part of Clinical Vignettes)	"A 65-year-old man comes to the clinic because of a 2-week history of swelling of his ankles and feet....He has an 8-year history of type 2 diabetes mellitus and hyperlipidemia. Medications are...."
The Lead-in	▶ Ask: What will you have the provider do? ▶ See Appendix B: Sample lead-ins; focus on the section that matches your testing point objective (management, diagnosis, etc.)	Management lead in: "Which of the following is the most appropriate pharmacotherapy at this time?"
The Option Set	▶ Aim to have enough information in your vignette to support the key (correct answer) and reasonably link to your distractors (wrong answers) ▶ Ensure your option set is free of flaws (see Chapter 3: Technical Item Flaws)	A. Drug B. Drug C. Drug D. Drug E. Drug

NBME®

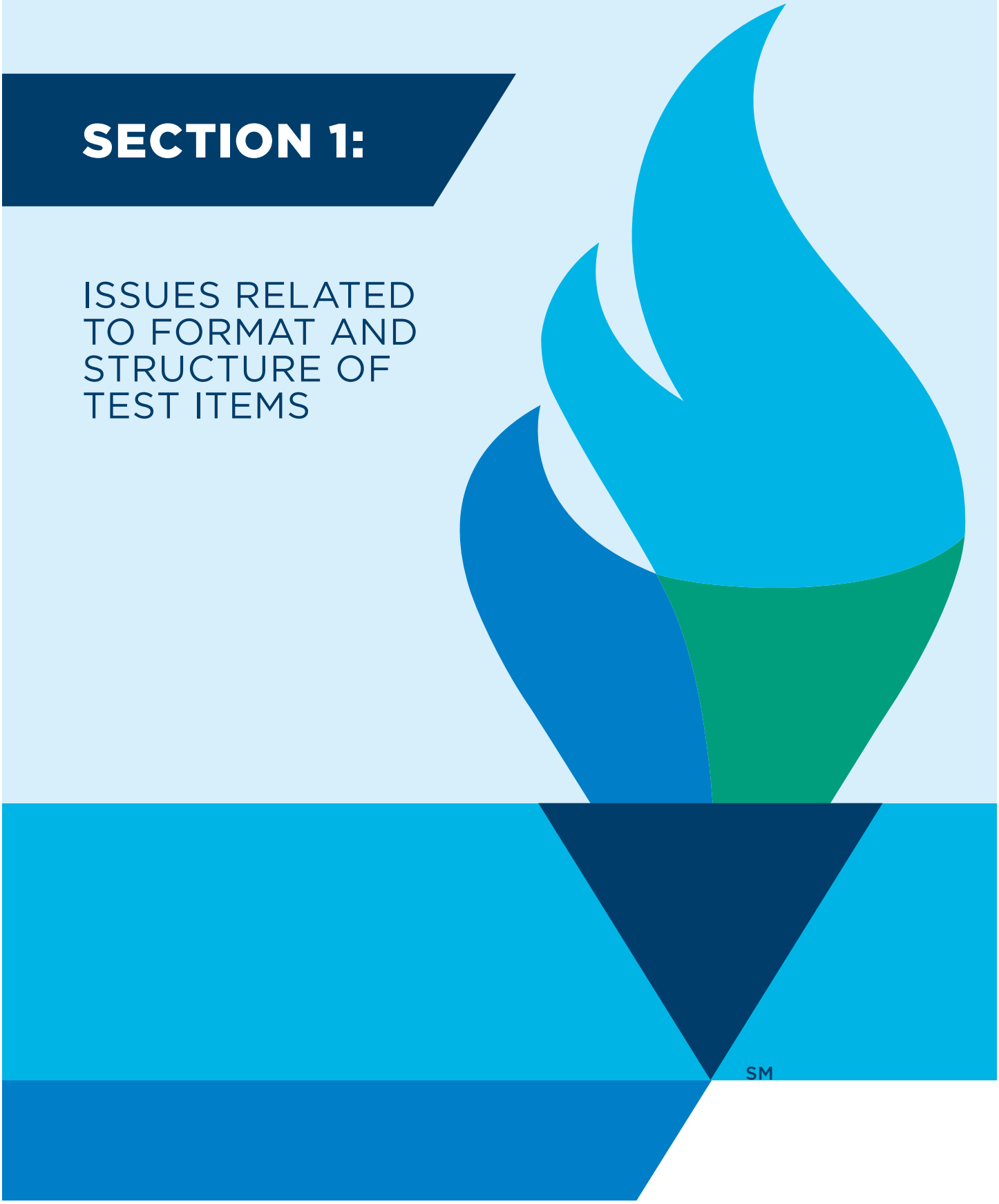
NBME® ITEM-WRITING GUIDE

Constructing Written Test Questions for the Health Sciences

FEBRUARY 2021

SECTION 1:

ISSUES RELATED TO FORMAT AND STRUCTURE OF TEST ITEMS



CHAPTER 1: INTRODUCTION

ASSESSMENT: AN IMPORTANT COMPONENT OF INSTRUCTION

Assessment, also known as testing, is a critical component of health professions education. When properly used, it can aid in determining the learners' knowledge and skills, based on criteria related to the stated educational goals. A primary purpose of testing is to communicate what you, as the instructor or preceptor and item writer, view as important. Tests are a powerful motivator, and your test-takers or students will learn the educational concepts they believe you value. Assessment also helps to fill instructional gaps by motivating students to seek educational resources and opportunities beyond course work. This outcome of testing is especially important in clinical learning environments, in which the experienced curriculum may vary from student to student, depending on factors such as the setting and the flow of patients. This outcome may also be important in some basic (foundational) science settings, in which the educational experiences may also vary from student to student. As students progress toward competence or even excellence, they are aided by ongoing feedback from their instructors or preceptors. Tests are often an important and useful component of that feedback in activating further learning.

Because tests have such a powerful influence on student learning, it is important to develop tests that properly align with educational goals and objectives. This manual focuses on the process of writing high-quality multiple-choice questions (MCQs), or items, that can be used to assess a wide variety of clinical knowledge and skills within the framework of the basic and clinical sciences.

Two issues that are of concern when developing and constructing quality MCQ-based exams are content sampling and psychometric performance.

Issues of Content Sampling

The purpose of any assessment is to permit inferences to be drawn concerning the ability (knowledge, skills) of test-takers or examinees. Inferences are defined here as decisions, judgments, or conclusions that extend beyond the set of items included in the exam into the larger domain from which the items were sampled. Performance on the test provides a basis for estimating achievement in the broader domain of interest, and this broader domain should be made explicit with statements about the inferences to be made from the test.

The first decision to be made involves the content to be sampled on the test; content decisions will drive the number and topic areas of the MCQs to be developed. The amount of attention given to evaluating a content area should reflect its relative importance, and it is often impractical to cover all testing topics in equal lengths. Tests are point-in-time measurements that take a specific and limited amount of time; if one content area contains many items, there is less testing time for other content areas. The nature of the test determines the extent to which the estimate of achievement is reproducible (aka reliable or generalizable) and accurate (provides sufficient validity evidence to support the intent). If the test questions are not adequately representative of the broader domain of interest (eg, if a test of competence in general medical practice includes only cardiovascular-related content), the test results may be biased and may not provide a good basis for estimating achievement in the full domain of interest. If the overall test length is too short, the scores may not be sufficiently precise or reliable to ensure they are a good representation of true proficiency. In order to generate a reproducible score, the item writer needs to sample content broadly (ie, typically 100 or more MCQs for written assessments).

Issues of Psychometric Performance

The process of evaluating the psychometric characteristics of an assessment and weighting their relative importance is determined by the purpose of the test and the decisions that will be made based on the results. For tests with higher stakes, such as those used for promotion or graduation decisions, those used for course grades, or those used in isolation for decisions, the scores must be reasonably reproducible (as demonstrated by high reliability) and evidence should be presented to demonstrate the accuracy of the test (eg, showing how content outlines for the test match the inferences to be made). For tests with lower stakes, such as those for which the score is but one element of the decision-making process, the amount of required psychometric evidence is less, but attention should be paid to evidence of test reliability and validity of score use nonetheless (see Figure 1 in Chapter 6 for more information).

PURPOSES OF TESTING

- Inform students about material that is important
- Motivate students to study
- Identify areas of deficiency, in need of remediation, or further learning
- Determine final grades or make promotion decisions
- Identify areas in which instruction can be improved

WHAT MATERIAL SHOULD BE TESTED?

- Exam content should align with course or clinical experience objectives
- Predetermined important topics should be weighted more heavily than less important topics
- The testing time devoted to each topic should reflect the relative (predetermined) importance of the topic
- The breadth of items should be representative of the instructional goals and objectives (curricular alignment)

NOTES

CHAPTER 2: MULTIPLE-CHOICE ITEM FORMATS

One of the most crucial aspects of a multiple-choice test item or question (MCQ) is its type or structure. Different item types can be used for different topic areas, and each item type carries with it advantages and disadvantages. A critical aspect to consider when choosing an item type is the inclusion of potential flaws that might benefit the savvy test-taker or introduce irrelevant difficulty. This chapter covers the basics of several multiple-choice item formats and introduces some potential flaws that are common to specific formats, while Chapter 3 will discuss specific item flaws in more detail.

ONE-BEST-ANSWER VS. TRUE-FALSE ITEMS

The universe of multiple-choice items can be divided into two families:

- Items that require test-takers to indicate a single, most accurate response (one-best-answer)
- Items that require test-takers to indicate all responses that are appropriate (true-false)

NBME has used multiple item formats within each family in the past, listed below by designating letter.

One-best-answer item formats that require test-takers to select the single best response:

- A-type (4 or more options, single items or sets)
- F-type (2 to 3 items grouped into a set around specific content or case scenario where test-takers **cannot** return to previously seen items in the set)
- G-type (2 or 3 items grouped into a set around specific content where test-takers **can** return to previously seen items in the set)

True-false item formats that require test-takers select some set of options that are true:

- C-type (A/B/Both/Neither response items)
- K-type (complex true-false items)
- X-type (simple true-false items)

The letters used to label the item formats hold no intrinsic meaning; letters were assigned more or less sequentially to new item formats as they were developed. For an extended list of item types formerly used by NBME, ordered by their designated letters, see Appendix C: NBME Retired Item Formats.

The True-False Family

True-false items require test-takers to select all the options that are "true," which could be anywhere from one to all of the listed options. In answering these items, the test-taker must decide where to make the cutoff and determine to what extent a response must be "true" in order to be keyed as "true." While this task requires additional judgment beyond what is required to select the true answer(s), that additional judgment may be unrelated to clinical expertise or knowledge. Too often, test-takers have to guess what the item writer had in mind because the options are not either completely true or completely false.

Sample of Acceptable True-False Item

Which of the following are X-linked recessive conditions?

1. Cystic fibrosis
2. Duchenne muscular dystrophy
3. Hemophilia A (classic hemophilia)
4. Tay-Sachs disease

This item is an example of a reasonably acceptable true-false item from a structural perspective. Note that the stem is clear, and the options are absolutely true or false with no ambiguity. Following tradition, for true-false items, the options are numbered. Options should be homogeneous (all are conditions), clearly worded, and of similar length, and the question should be closed and focused.

The options can be diagrammed as follows.

1	2
4	3
Totally Incorrect	Totally Correct

Sample of Flawed True-False Item

True statements about cystic fibrosis (CF) include:

1. CF is an autosomal recessive disease
2. Patients with CF usually live into adulthood
3. Males with CF are sterile
4. The incidence of CF is 1:2000

This item demonstrates a commonly seen flaw for true-false items that often occurs when options are not homogeneous and vaguely worded. Options 2, 3, and 4 cannot be judged as absolutely true or false, because a group of content experts would not necessarily agree on the answers. For example, for option 4, experts would demand more information to determine incidence: Is this in the United States? Is this among all ethnic groups? Similar issues arise with options 2 and 3, whereas option 1 is clear. Revision of this item would most likely include editing options 2, 3, and 4, to be statements of fact like option 1, and revising the question itself to be closed.

Sample of Flawed True-False Item

In children, ventricular septal defects are associated with:

1. cyanosis
2. pulmonary hypertension
3. systolic murmur
4. tetralogy of Fallot

The problems with this true-false item are more subtle. The difficulty is that the test-taker has to make assumptions about the severity of the disease, the age of the patient, and whether or not the disease has been treated. This is due in part to the vagueness in the question itself ("associated with"). Different assumptions lead to different answers, even among experts. Revising this question would require adding additional text, perhaps a lot of it, in order to allow the test-taker to judge the options as wholly true or wholly false.

General Rules for True-False Items

Because test-takers are required to select all the options that are "true," true-false items must satisfy the following rules:

- Item and option text must be clear and unambiguous. Avoid imprecise phrases such as "is associated with" or "is useful for" or "is important"; words that provide cueing such as "may" or "could be"; and vague terms such as "usually" or "frequently."
- The lead-in should be closed and focused.
- Options must be absolutely true or false; no shades of gray are permissible.
- Options should be homogeneous so that they can be judged as entirely true or entirely false on a single dimension.

Challenges with Using True-False Items

We recommend avoiding true-false questions if possible. Although many item writers believe true-false items are easier to write than one-best-answer items, this type can often be more problematic. The writer may have something particular in mind when writing the item, but careful review subsequently reveals subtle difficulties that were not apparent to the item author. Often the distinction between "true" and "false" is not clear, and it is not uncommon for subsequent reviewers to alter the answer key. As a result, reviewers tend to rewrite or discard true-false items far more frequently than items written in other formats. Some ambiguities can be easily clarified, but others cannot. In addition, to avoid ambiguity true-false questions often test on recall of an isolated fact, which we recommend avoiding.

ONE-BEST-ANSWER QUESTIONS ARE BETTER AT ASSESSING TEST TAKERS' JUDGMENT, SYNTHESIS, AND APPLICATION OF KNOWLEDGE.

CHAPTER 3: TECHNICAL ITEM FLAWS

Good content and good structure contribute to the quality of an item. However, quality can be impacted negatively by the inclusion of technical item flaws. There are two kinds of technical item flaws:

1. A flaw that adds irrelevant difficulty to the item can confuse all test-takers. These flaws make the item challenging for reasons unrelated to the testing objective/point of the item and can add construct-irrelevant variance to the final test score.
2. A flaw that cues the more savvy and confident test-takers (aka the "testwise") and aids them in guessing the right answer. These flaws related to "testwiseness" make it easier for some students to answer the item correctly based on their test-taking skills alone, without necessarily knowing the content.

The item writer's goal is to develop and structure items to eliminate both types of flaws as much as possible, in order to create a test that ensures a level playing field for all test-takers. A test-taker's probability of answering an item correctly should be determined by his or her amount of expertise on the topic being assessed; ideally, that probability will not decrease due to a suboptimally written item and will not increase due to test-taking strategies.

FLAWS RELATED TO IRRELEVANT DIFFICULTY

Long or Complex Options

The item below has several flaws. The vignette contains extraneous information, and in fact, the vignette is not needed to answer the item. More importantly, the options themselves are overly long and complicated. The number of words in each option increases the reading load, which can shift the construct that is being measured from content knowledge to reading speed. Please note that this flaw relates only to options. There are many well-constructed test items that include a long vignette, and decisions about vignette length should be made in accordance with the testing point of the item. If the purpose of the item is to assess whether or not the student can interpret and synthesize information to determine, for example, the most likely diagnosis for a patient, then it is appropriate for the vignette to include a fairly complete description of the situation.

Example of Item with Long, Complex Options

Peer review committees in HMOs may move to take action against a physician's credentials to care for participants of the HMO. There is an associated requirement to ensure that the physician receives due process in the course of these activities. Due process must include which of the following?

A. Notice, an impartial forum, counsel, and a chance to hear and confront evidence
B. Proper notice, a tribunal empowered to make the decision, a chance to confront witnesses, and a chance to present evidence in defense
C. Reasonable and timely notice, an impartial panel empowered to make a decision, a chance to hear evidence and to confront witnesses, and the ability to present evidence in defense

Numeric Data Presented Inconsistently

When numeric options are used, the options should be listed in numeric order and in a single format (ie, as terms or ranges). Confusion can occur when formats are mixed or when options are listed in an illogical order. In this example, options A, B, and C are expressed as ranges, whereas options D and E are specific percentages. All options should be expressed as ranges or as specific percentages; mixing them is ill-advised. In addition, the range for option C includes options D and E, which almost certainly rules out options D and E as correct answers for the testwise examinee.

Example of Item with Inconsistent Numeric Data

After a second episode of infection, which of the following is the likelihood that a woman is infertile?

A. Less than 20%
B. 20 to 30%
C. Greater than 50%
D. 75%
E. 90%

Vague Terms

Vague frequency terms in the options (such as “often” or “usually”) are not consistently defined or interpreted by the readers, and sometimes not even by experts. Different interpretations of these terms can lead to multiple correct answers or a set of options that cannot be rank ordered in terms of correctness.

Example of Item with Vague Terms

Severe obesity in early adolescence:

A. has a 75% chance of clearing spontaneously
B. often is related to endocrine disorders
C. shows a poor prognosis
D. usually responds dramatically to dietary regimens
E. usually responds to pharmacotherapy and intensive psychotherapy

“None of the Above”

The phrase “None of the above” is problematic in items for which judgment is involved and the options are not absolutely true or false. If the correct response is intended to be one of the other listed options, knowledgeable students are faced with a dilemma because they have to decide between the option that the item writer has intended as correct and an option that encompasses everything not listed in the option set. Test-takers can often intuit an option that is more correct than the item writer intended to be correct, which would lead them to use the more expansive option. Use of “None of the above” essentially turns the item into a true-false item; each option has to be evaluated as more or less true than the universe of unlisted options. It is often possible to fix such items by replacing “None of the above” with an option that is more specific. In this example, which asks a test-taker to specify the most appropriate pharmacotherapy, option E, “None of the above” should be replaced by “No pharmacotherapy is indicated at this time,” to eliminate any ambiguity while still requiring the test-taker to commit to a management decision.

Example of Item with “None of the Above”

A 3-day-old male newborn is brought to the office by his parents because his crying has increased during the past night compared with his first 2 days of life. The parents have been unable to calm the newborn during the past 2 hours. The newborn also has had mild shaking of his hands and legs during the past 4 hours. He was delivered at 38 weeks’ gestation via uncomplicated spontaneous vaginal delivery. His mother, gravida 2, para 2, is age 19 years. She has a history of dysthymia for which she took escitalopram during pregnancy. The newborn takes no medications. He is at the 50th percentile for length, weight, and head circumference. Temperature is 37.2°C (98.9°F), pulse is 155/min, respirations are 35/min, and blood pressure is 84/50 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 100%. The newborn has a high-pitched cry and is inconsolable with swaddling. He has tremors of his hands and feet with crying. Moro reflex is present. Which of the following is the most appropriate pharmacotherapy?

A. Citalopram
B. Lorazepam
C. Morphine
D. Naloxone
E. None of the above*

Nonparallel Options

The next item illustrates a common flaw in which the options are not only too long but the structure of each option is different, both of which add to the reading time. Generally, this flaw can be corrected by careful editing to ensure that the options all have the same format and the same structure. In this particular item, the lead-in can be changed to “Which of the following is the most likely reason no conclusion can be drawn from these results?” Each option can then be edited to fit a logical and parallel answer to the lead-in.

Example of Item with Nonparallel Options

In a vaccine trial, 200 two-year-old boys were given a vaccine against a certain disease and then monitored for 5 years for occurrence of the disease. Of this group, 85% never contracted the disease. Which of the following statements concerning these results is correct?

A. The number of cases (ie, 30 cases over 5 years) is too small for statistically meaningful conclusions
B. Vaccine efficacy (%) is calculated as 85-15/100
C. No conclusions can be drawn because the trial involved only boys
D. No conclusion can be drawn since no follow-up was done with nonvaccinated children

Complicated Stems

This item, as written, requires that the test-taker (a) understands the concepts of genetics that are represented and (b) is able to rank order Roman numerals (the second of which is an irrelevant and unnecessarily difficult addition to the goal of the item). This item should be rewritten to focus on a single karyotype, such as the greatest risk, with the karyotypes arranged in the options themselves, so that the test-taker who understands the order of risk of occurrence can more easily identify the correct answer.

Example of Item with Complicated Stem

Arrange the parents of the following children with Down syndrome in order of highest to lowest risk of recurrence. Assume that the maternal age in all cases is 22 years and that a subsequent pregnancy occurs within 5 years. The karyotypes of the daughters are:

I: 46,XX,–14,+T(14q21q)pat

II: 46,XX,–14,+T(14q21q)de novo

III: 46,XX,–14,+T(14q21q)mat

IV: 46,XX,–21,+T(14q21q)pat

V: 47,XX,–21,+T(21q21q) (parents not karyotyped)

A. III, IV, I, V, II
B. IV, III, V, I, II
C. III, I, IV, V, II
D. IV, III, I, V, II
E. III, IV, I, II, V

Suggested Revision

Five couples come to the office for counseling prior to conception. Each couple has one child with Down syndrome. The karyotypes of each of the children are shown. The parents of the child with which of the following karyotypes have the greatest risk for recurrence of Down syndrome in their next pregnancy?

A. 46,XX,–14,+T(14q21q)pat
B. 46,XX,–14,+T(14q21q)de novo
C. 46,XX,–14,+T(14q21q)mat
D. 46,XX,–21,+T(21q21q)pat
E. 47,XX,–21,+T(21q21q) (parents not karyotyped)

Negatively Phrased Lead-ins

A negative phrasing in the lead-in asks the test-taker to find the least accurate option, with the rest being accurate, rather than to find the most accurate option. If most of the items on a test are positively phrased, the inclusion of a negatively phrased item carries the risk that the test-taker will miss the word "except," even when it is set in bold and/or capitalized.

Example of Item With Negatively Phrased Lead-in

Each of the following statements about cholesterol is true EXCEPT:

A. cholesterol contains numerous fatty acids
B. cholesterol is not present in any foods of plant origin
C. cholesterol is required in many complex bodily functions
D. endogenous cholesterol is produced within the body

FLAWS THAT CUE THE TESTWISE EXAMINEE

Grammatical Cues

This flaw exists when an option does not follow grammatically from the lead-in. In this example, testwise students can eliminate B, C, D, and E as possible correct answers because they do not grammatically or logically follow the lead-in. This flaw can happen when an item writer focuses more attention on writing the correct answer than on the distractors, leading to the potential for grammatical errors. To avoid this flaw, read each option immediately following the stem to ensure that the language is a good fit. Another way to avoid the flaw is to always use closed lead-ins.

Example of Item with Grammatical Cue

A 12-year-old girl is brought to the office because of chest pain. She has recently experienced an upper respiratory infection with frequent coughing. Temperature is 37.2°C (99.0°F), pulse is 120/min, respirations are 22/min, and blood pressure is 95/65 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 99%. Physical examination shows tenderness to palpation over her costochondral joints on the left. Auscultation of the lungs discloses diffuse end-expiratory wheezes bilaterally. Her diagnosis is most likely to be an:

A. asthma attack*
B. costochondritis
C. pleurisy
D. rib fracture secondary to coughing
E. viral pneumonitis

Grouped or Collectively Exhaustive Options

This flaw exists when a savvy student can identify a subset of options that cover all possible outcomes (are collectively exhaustive) and rule out the options not in that subset. In this item, options A, B, and D are exhaustive—urine potassium can only increase, decrease, or not change—and thus one of these three options must be the correct answer. A less testwise student might spend time considering C and E. Often, item writers add options like C and E only because they want to have a total of five options, but it is not an improvement of the item to add options that have no merit. The item writer should be able to rank order each option on the same dimension, and no subset of options should include all possible outcomes.

Example of Item With Collectively Exhaustive Options

Administration of furosemide results in:

A. a decrease in urine potassium
B. an increase in urine potassium
C. improved glucose control in patients with type 2 diabetes mellitus
D. no change in urine potassium
E. requires decreasing the dose with renal failure

Absolute Terms

In this item, options A, B, and E contain terms that are less absolute than those in options C and D. The testwise student will eliminate options C and D as possibilities because they are less likely to be true than something stated less absolutely, and so this item is flawed with the inclusion of those terms. This flaw tends to arise when verbs are included in the options rather than in the lead-in. Focusing the lead-in, placing the verb in the lead-in, and shortening the options are possible ways to correct this flaw.

Example of Item With Absolute Terms

In patients with advanced dementia, Alzheimer type, the memory defect:

A. can be treated adequately with phosphatidylcholine (lecithin)
B. could be a sequela of early parkinsonism
C. is never seen in patients with neurofibrillary tangles at autopsy
D. is never severe
E. possibly involves the cholinergic system

Correct Option Stands Out

In this item, the correct answer, option A, is longer than the other options, and is the only "double" option, containing two components. This flaw is another potential outcome when item writers pay more attention to constructing the correct answer than the distractors. This results when item writers likely create the correct answer first and then write the incorrect distractors. In addition, item writers are often teachers and they will construct long correct answers that include additional instructional material, parenthetical information, caveats, and so on. This flaw can be avoided by reviewing the entire option set for length, ensuring the level of detail is consistent across options, and removing language that is purely for instructional purposes only.

Example of Item with Correct Option that Stands Out

Secondary gain is:

A. a complication of a variety of illnesses and tends to prolong many (>3) of them*
B. a frequent problem in obsessive-compulsive disorder
C. never seen in organic brain damage
D. synonymous with malingering

Word Repetition ("Clang Clues")

This flaw arises when language used in the stem is repeated in the correct answer. Here, the word "unreal" in the vignette can clue test-takers to the fact that the correct answer, "derealization," is the only option that also includes the word "real." The same flaw can appear even if a word is repeated only in an etymological sense, such as when a stem mentions bone pain and the correct answer begins with the prefix "osteo-." Item writers should scan the options and item stem to check for this word or phrase repetition.

Example of Item with Word Repetition

A 58-year-old man with a history of heavy alcohol use and previous psychiatric hospitalization is confused and agitated. He speaks of experiencing the world as unreal. Which of the following best describes this symptom?

A. Depersonalization
B. Derailment
C. Derealization*
D. Focal memory deficit
E. Signal anxiety

Convergence

This item flaw might be less obvious than the others, but it occurs frequently and is worth noting. The underlying flaw is that the correct answer is the option that has the most in common with the other options, and thus the testwise test-taker can converge on the right answer just by counting the number of times certain terms appear. In this example, the testwise test-taker would eliminate "anionic form" as unlikely because "anionic form" appears only once; that test-taker would also exclude "outside the nerve membrane" because "outside" appears less frequently than "inside." The test-taker would then have narrowed the options to B and D. Since three of the five options involve a charge, the testwise test-taker would then select option B, which is in fact the correct answer. This flaw can also occur without being directly reflected in the language; for example, if an item is asking which pharmacotherapy is most effective, and three of the five options are in one class of drugs, the savvy test-taker may rule out the other two as less likely. This flaw occurs when item writers start with the correct answer and write the distractors as permutations of the correct answer. The correct answer will then be more likely to have elements in common with the rest of the options, and the incorrect answers are more likely to be outliers. A useful check is to review all options and see if words or terms are repeated across options.

Example of Item with Convergence

Local anesthetics are most effective in the:

A. anionic form, acting from inside the nerve membrane
B. cationic form, acting from inside the nerve membrane*
C. cationic form, acting from outside the nerve membrane
D. uncharged form, acting from inside the nerve membrane
E. uncharged form, acting from outside the nerve membrane

NOTES

SUMMARY OF TECHNICAL ITEM FLAWS

Issues Related to Irrelevant Difficulty	
FLAWS	**SOLUTIONS**
Long, complex options	▶ Put common text in stem. ▶ Use parallel construction in options. ▶ Shorten options.
Tricky, unnecessarily complicated stems	▶ Include content that is necessary to answer the question or to make distractors attractive. ▶ Avoid teaching statements.
Inconsistent use of numeric data	▶ Avoid overlapping options. ▶ Ask for minimum or maximum value to avoid multiple correct answers.
Vague terms	▶ Avoid frequency terms, like usually and often. Such terms are interpreted differently by different people.
"None of the above" option	▶ Replace "None of the above" with specific action (eg, No intervention needed).
Nonparallel options	▶ Edit options to be parallel in grammatical form and structure.
Negatively structured stem (eg, "Each of the following EXCEPT")	▶ Revise lead-in to have a positive structure. ▶ If possible, use correct options to create a scenario.

Cues to the Testwise Examinee	
FLAWS	**SOLUTIONS**
Collectively exhaustive options (subset of options cover all possibilities)	▶ Replace at least one option in subset. ▶ When revising, avoid creating option pair.
Absolute terms ("always," "never") in options	▶ Eliminate absolute terms. ▶ Use focused lead-in and short homogeneous options.
Grammatical clues	▶ Make all options singular or all options plural. ▶ Use closed lead-ins.
Correct answer stands out	▶ Revise options to equal length. Remove language used for teaching points and rationales.
Word repeats (clang clue)	▶ Replace repeated word in either stem or option. **OR** ▶ Use repeated word in all options.
Convergence	▶ Revise options to balance use of terms.