

Exploring Clinical Reasoning Strategies and Test-Taking Behaviors During Clinical Vignette Style Multiple-Choice Examinations: A Mixed Methods Study

BRIAN SANJAY HEIST, MD, MSc
JED DAVID GONZALO, MD, MSc
STEVEN DURNING, MD, PhD
DARIO TORRE, MD, MPH, PhD
DAVID MICHAEL ELNICKI, MD

Abstract

Background Clinical vignette multiple-choice questions (MCQs) are widely used in medical education, but clinical reasoning (CR) strategies employed when approaching these questions have not been well described.

Objectives The aims of the study were (1) to identify CR strategies and test-taking (TT) behaviors of physician trainees while solving clinical vignette MCQs; and (2) to examine the relationships between CR strategies and behaviors, and performance on a high-stakes clinical vignette MCQ examination.

Methods Thirteen postgraduate year-1 level trainees completed 6 clinical vignette MCQs using a think-aloud protocol. Thematic analysis employing elements of grounded theory was performed on data transcriptions to identify CR strategies and TT behaviors. Participants' CR strategies and TT behaviors were then compared

with their US Medical Licensing Examination Step 2 Clinical Knowledge scores.

Results Twelve CR strategies and TT behaviors were identified. Individuals with low performance on Step 2 Clinical Knowledge demonstrated increased premature closure and increased faulty knowledge, and showed comparatively less ruling out of alternatives or admission of knowledge deficits. High performers on Step 2 Clinical Knowledge demonstrated increased ruling out of alternatives and admission of knowledge deficits, and less premature closure, faulty knowledge, or closure prior to reading the alternatives.

Conclusions Different patterns of CR strategies and TT behaviors may be used by high and low performers during high-stakes clinical vignette MCQ examinations.

Editor's Note: The online version of this article contains tables showing descriptions and examples of themes representing clinical reasoning strategies and test-taking behaviors.

Introduction

Since the 1970s, research has explored how physicians think when approaching clinical scenarios in an attempt to

enhance our understanding of clinical reasoning.¹ Assessment of trainees' clinical reasoning has evolved, including the use of testing modalities, such as patient management problems, key features examinations, script-concordance testing, and think-aloud protocols. The utility of these methods is limited by their resource requirements and consequent inefficiency in testing. Multiple-choice question (MCQ) examinations are relatively efficient to administer and have excellent psychometric performance.² Over the past 2 decades, clinical vignette MCQ examinations have become the primary standardized method to evaluate trainees' clinical reasoning (CR) skills.³

Despite the widespread use of clinical vignette MCQs, strategies used by trainees when approaching these questions have not been well described. A 1994 study qualitatively explored this topic.⁴ Employing "think-aloud" methodology, where participants verbalize thoughts while answering questions, Skakun et al⁴ characterized medical students' item-related activities and application of successful problem-solving activities on MCQs, and observed both backward CR (starting with a hypothesis and working backward) and forward CR (moving forward from observations to generate hypotheses).⁴

Brian Sanjay Heist, MD, MSc, is Assistant Professor of Medicine, Division of General Internal Medicine, University of Pittsburgh Medical Center (UPMC) Shadyside Hospital; **Jed David Gonzalo, MD, MSc**, is Assistant Professor of Medicine, Pennsylvania State University College of Medicine; **Steven Durning, MD, PhD**, is Professor of Medicine, General Medicine Division, Uniformed Services University of the Health Sciences; **Dario Torre, MD, MPH, PhD**, is Associate Professor of Medicine, Department of Medicine, Drexel University College of Medicine; and **David Michael Elnicki, MD**, is Professor of Medicine, Division of General Internal Medicine, University of Pittsburgh School of Medicine.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

Corresponding author: Brian Sanjay Heist, MD, MSc, UPMC Shadyside Hospital, 5230 Centre Avenue, SON #313, Pittsburgh, PA 15232, 412.623.3963, heistb@upmc.edu

Received March 1, 2014; revisions received June 18, 2014, and July 13, 2014; accepted July 14, 2014.

DOI: <http://dx.doi.org/10.4300/JGME-D-14-00176.1>

Subsequent research has compared self-chosen and instructed diagnostic reasoning strategies on specific clinical vignette MCQs to success on those same items,^{3,5–7} and evaluated the influence of test question format on predominant reasoning strategy.⁸ However, although recent years have seen extensive evaluation and modification of clinical vignette MCQs,⁹ no current studies have evaluated what trainees do when approaching such questions; nor have they assessed test-taking (TT) behaviors, or “test-wiseness.” Commonly defined as “a subject’s capacity to utilize the characteristics of the test and/or test-taking situation to receive a high score,”¹⁰ test-wiseness can influence performance on standardized tests.¹¹ Principles of test-wiseness relevant to clinical vignette MCQs include elimination of answers known to be incorrect and interpretation of answers in view of the test purpose.

Exploring how trainees answer clinical vignette MCQs may help improve our understanding of CR assessment. More specifically, exploring the relationships between CR strategies and TT behaviors may provide insight into why some physician trainees perform better than others. For example, might premature closure or overconfidence contribute to examination performance? Recent research has identified premature closure and overconfidence as major contributors to diagnostic error in clinical practice,^{12,13} but their effect on clinical vignette MCQ performance has not been assessed.

In this exploratory study we (1) describe the CR strategies and TT behaviors of postgraduate year (PGY)–1 level trainees while solving clinical vignette MCQs; and (2) examine relationships between the use of these strategies and behaviors, and performance on high-stakes examinations.

Methods

Study Approach

We performed a thematic analysis integrating principles of grounded theory to develop a conceptual understanding of CR strategies and TT behaviors used while solving clinical vignette MCQs. Consistent with Glaser,¹⁴ to limit bias in our data interpretation, a review of the literature on CR during examinations and TT strategies was delayed until after initial coding.

Selection of Participants

From March through July 2011, we invited all 28 PGY-1 level trainees rotating in internal medicine at the University of Pittsburgh Medical Center Shadyside Hospital. A total of 11 trainees participated during that period (1 participant was subsequently excluded, and data were analyzed for 10 participants). In September 2012, 2 PGY-2 categoric

What was known

Clinical vignette multiple-choice questions (MCQs) are widely used, but little is known about the clinical reasoning strategies trainees use when taking these examinations.

What is new

Low-performing individuals demonstrated higher rates of premature closure, higher rates of faulty knowledge, less ruling out of alternatives, and less effective self-monitoring.

Limitations

Small sample of first-year residents limits generalizability; test-taking performance may not be relevant to performance in a clinical context.

Bottom line

Different clinical reasoning strategies and test-taking behaviors appear to be used by high and low performers taking clinical vignette MCQ examinations.

residents were recruited. Both PGY-2 residents had taken leave for 3 months during their PGY-1, and, therefore, were at the same level of training as the other participants. Participants received \$20 gift cards for their time.

Selection of Test Items

Two investigators selected 6 clinical vignette format test items, each representing a different subspecialty area, from the American College of Physicians’ Medical Knowledge Self-Assessment Program 15 digital question databank.¹⁵ Each item probed comprehension and/or application according to Bloom’s taxonomy,¹⁶ eliciting CR suitable for analysis. Two items asked for the most likely diagnosis, and 4 items asked for the most appropriate next step in evaluation/management (TABLE 1).

The Institutional Review Board at the University of Pittsburgh approved the study.

Data Collection

All participants agreed to be individually recorded “think-aloud” while completing the test items. Think-aloud methodology is recognized as an optimal methodology to capture thought processes,¹⁷ and recent evaluation employing functional neuroimaging supports its validity.¹⁸ The start of each session contained specific instruction to “Say everything that goes through your mind as you try to solve each question.” Next, participants were given the question packet and audio recording was initiated. All participants completed the questions within 25 minutes. A research assistant transcribed the deidentified audio recordings.

Data Analysis

After the first 11 think-aloud sessions, transcripts were analyzed independently by 2 investigators, with data

TABLE 1 TOPIC AND QUESTION-TYPE OF CLINICAL VIGNETTE-STYLE MULTIPLE-CHOICE QUESTIONS USED IN THINK-ALoud PROTOCOL

Topic Area	Question Type
1. Infectious disease (HIV complication)	Diagnosis
2. Pulmonary (asthma exacerbation)	Management (What's the next step?)
3. Hematology (deep vein thrombosis)	Management (What's the most appropriate management?)
4. Endocrine (thyroid nodule)	Management (What's the next step?)
5. Gastroenterology (intrahepatic cholestasis)	Management (What's the most appropriate management?)
6. Nephrology (renal tubular acidosis)	Diagnosis

management support with Atlas.ti 6.0 (Scientific Software). One participant was excluded because of a paucity of verbalized thinking, resulting in 10 transcripts suitable for analysis. We used process coding, memos, and the constant comparison method to identify and reach consensus on emerging themes. To evaluate thematic saturation, we analyzed the data for 2 additional participants at the same level of training (final N = 12). Subsequently, we performed a literature review and categorized themes into CR strategies and/or TT behaviors. Transcripts were again reviewed to confirm application/absence of the identified CR strategies and TT behaviors. Disagreements were resolved through discussion.

The US Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge (CK) was identified as the high-stakes examination similar in content and format to the sample questions. A research assistant compiled participants' examination scores from program application files and matched them to deidentified participant transcripts. Four participants had taken the Comprehensive Osteopathic Medical Licensing Examination of the United States (COMLEX-USA) Level 2. Using a published conversion formula,^{19,20} the research assistant converted COMLEX scores to the corresponding USMLE scores.

Results

Of the 12 participants in the study, 4 were men and 8 were women. The average age was 29 years (range, 25–35 years). The students graduated from 9 medical schools, including 2 osteopathic medical schools and 3 international medical schools.

CR Strategies and TT Behaviors

Despite variation in test question type and content, CR strategies were relatively consistent for each participant rather than for a given question. Twelve themes emerged from the coding of transcripts for the first 10 participants, and no new themes emerged from the 2 added participants.

Based on our literature review, these themes were divided into 3 categories: (1) clinical reasoning strategies (described in literature on CR, but not TT behaviors); (2) test-taking behaviors (described in literature on TT behaviors, but not CR); and (3) clinical reasoning and test-taking behaviors (described in literature on both TT behaviors and CR).

“Strategies” denote themes in the CR literature consistently using this term, and that have been shown to enhance physicians' diagnostic accuracy.^{3,21–23} “Behaviors” are themes described in the educational literature that may not enhance performance. Explanations and examples for the 3 categories are available as online supplemental material.

We identified “nonanalytic reasoning” and “reaching closure prematurely” among participants' CR strategies and TT behaviors. We defined nonanalytic as the generation of a hypothesis after verbalizing fragments of the vignette, in accordance with Eva's description²⁴ of the unconscious filtration of clinical features based on prior experiences to generate a hypothesis. We defined reaching closure prematurely as failure to consider the correct alternative in answering the test item, consistent with Berner and Graber's definition.¹³ We recognize the lack of uniformity in nomenclature within the CR literature,²⁵ and that other researchers may have interpreted or labeled the identified themes differently.

USMLE Step 2 CK Examination Performance and Relationships With CR Strategies and TT Behaviors

Participants' Step 2 CK scores ranged from 189 to 260. In consideration of the national Step 2 CK score mean and SD,²⁵ participants' scores were clustered into 3 categories: low (189, 191, 193, 200), middle (208, 211, 212, 213, 214, 228), and high (251, 260).

TABLE 2 summarizes the frequency of CR strategies and TT behaviors on the 6 test items by low, middle, and high Step 2 CK performer categories. The 4 participants with low Step 2 CK scores ruled out alternatives (17% of questions), admitted knowledge deficits (13% of questions)

TABLE 2 FREQUENCY OF HIGH, MODERATE, AND LOW STEP 2 CLINICAL KNOWLEDGE (CK) EXAMINATION PERFORMERS' DEMONSTRATION OF EACH CLINICAL REASONING (CR) STRATEGY AND TEST-TAKING (TT) BEHAVIOR ON SAMPLE MULTIPLE-CHOICE QUESTION TEST ITEMS			
CR Strategy and TT Behavior	Percentage of Test Items Demonstrating CR Strategy and TT Behavior in Each USMLE Step 2 CK Score Performance Category ^{a,b}		
	High Performer	Moderate Performer	Low Performer
CR strategy			
Reading vignette with restating and summarizing	17	72	88
Summarizing key features after reading alternatives	83	50	71
Using scheme-inductive reasoning	0	11	8
Using nonanalytic reasoning	67	69	67
TT behavior			
Reading question and alternatives before vignette	0	17	0
Querying test writer's objective	0	14	13
CR strategy and TT behavior			
Reaching closure before reviewing alternatives	0	8	8
Ruling out alternatives	92	69	17
Reaching closure prematurely	0	6	25
Reaching closure with difficulty or delay	8	22	4
Admitting knowledge deficits	58	22	13
Applying faulty knowledge	8	28	46

Abbreviation: USMLE, US Medical Licensing Examination.

^a USMLE Step 2 CK scores of the 12 study participants organized by performance based on national Step 2 CK score mean and SD: high performer = 260, 251; moderate performer = 228, 214, 213, 212, 211, 208; low performer = 200, 193, 191, 189. Comprehensive Osteopathic Medical Licensing Examination of the United States (COMLEX-USA) Level 2 score converted to USMLE Step 2 CK equivalent for 2 moderate and 2 low examination performers.

^b The number of test items completed by performer category varied because of differing numbers of participants in respective categories. The 2 high performers completed a total of 12 test items (ie, 2 participants × 6 sample test items). Similarly, the 6 moderate and 4 low performers completed a total of 36 and 24 questions, respectively.

least frequently, and demonstrated premature closure (25% of questions) and faulty knowledge (46% of questions) most frequently. The 2 participants with high Step 2 CK scores ruled out alternatives (92% of questions) and admitted knowledge deficits (58% of questions) most frequently, did not reach closure before reviewing alternatives (0% of questions), and demonstrated premature closure (0% of questions) and faulty knowledge (8% of questions) least frequently.

Discussion

We identified 12 CR strategies and TT behaviors demonstrated by medical residents during clinical vignette MCQs near the end of their first year of training. The findings are illuminating for 2 reasons. First, 2 TT behaviors demonstrated by trainees—reading question and alternatives before vignette and querying test writer’s objective—are clearly inauthentic to clinical practice. Second, unique

patterns of CR strategies and TT behaviors were identified across our 3 Step 2 CK performance categories. Overall, the likelihood of ruling out alternatives and admitting knowledge deficits increased with Step 2 CK performance, while the likelihood of reaching closure prematurely and applying faulty knowledge decreased with Step 2 CK performance. Additionally, *reaching closure before reviewing alternatives* was infrequently demonstrated, particularly by high examination performers.

To assess the implications of our findings for clinical vignette MCQ examinations as a measure of CR, we first considered participants’ knowledge, because knowledge is required for high Step 2 CK performance.²⁶ The most intuitive marker of knowledge we identified, applying faulty knowledge, decreased with higher Step 2 CK performance. Two other behaviors, *reaching closure prematurely* and *admitting knowledge deficits*, may appear to reflect knowledge, but a direct linkage is unlikely.

Reaching closure prematurely has been demonstrated previously as the most common form of faulty synthesis, distinct from faulty knowledge, and a far more common source of cognitive-based errors in clinical practice.¹² For admitting knowledge deficits, the increased likelihood of this behavior among high Step 2 CK performers is consistent with research demonstrating superior self-monitoring by high-performing medical student test-takers.²⁷ During analysis of the think-aloud transcripts, we found admitting knowledge deficits as an attitudinal difference among participants. Although not definitive from the results, it is plausible that premature closure and failure to admit knowledge deficits, and their increased likelihood with low Step 2 CK performance, may relate to overconfidence. Overconfidence has been proposed as a contributor to error in diagnosis, more specifically discerned as an underlying attitudinal influence for negative cognitive behaviors, including premature closure.¹³ Numerous studies have documented overconfidence in clinical decision-making,¹³ and studies comparing diagnostic accuracy and confidence in diagnosis observed no correlation and negative correlation, respectively.^{28,29}

The theme demonstrating the strongest positive relationship with Step 2 CK performance was ruling out alternatives, categorized under “clinical reasoning and test-taking behaviors.” Studies addressing CR during clinical vignette MCQs have described ruling out alternatives as an analytic CR method.^{4,5,8} However, because test-takers answering MCQs are not required to generate hypotheses, this behavior may be more reflective of TT than CR. For example, “thoughtful consideration of various answer choices” is demonstrated by college students on reading comprehension tests.³⁰ While we cannot exclude the contribution of analytic CR strategies to ruling out alternatives, we consider this behavior more as a TT strategy.

Collectively, our participants reached closure before reviewing alternatives on less than 10% of items, with no high examination performers demonstrating this behavior. Rather than self-generating answers, participants focused on selecting an answer from the provided alternatives. This behavior is consistent with Heemskerk’s hypothesis⁸ that “presenting a list of diagnostic options, of which one is definitely the correct answer, might facilitate a process of excluding all diagnoses except one.” The association between more alternatives on Step 2 CK examination items and increased difficulty and longer time required for completion suggests this practice is widespread³¹ and supports the impression that the number of alternatives on a test item affects CR strategies and TT behaviors.

Our results suggest clinical vignette MCQs elicit a mixture of CR strategies and TT behaviors, categorized as authentic or inauthentic to behaviors observed in clinical

practice. We observed 2 inauthentic behaviors: (1) reading questions and alternatives before vignette, and (2) querying the test writer’s objective, and have explained that a third behavior, ruling out alternatives, has concerning differences from clinical practice. Importantly, lack of authenticity does not connote lack of validity. Absence of benefit to querying the test writer’s objective or reading the question and alternatives before the vignette are clearly desired results. The benefit of ruling out alternatives also supports examination validity, albeit perhaps less obviously. While we described the negative influence of a list of alternatives on the authenticity of the test-taker’s CR process, we recognize that systematically considering self-generated differential diagnoses contributes to diagnostic accuracy in complex simulated cases,⁶ and also those likely in clinical practice. With regard to patterns of authentic CR strategies and behaviors, likelihood of faulty knowledge and the proposed markers of overconfidence decreased with higher Step 2 CK performance, supporting the examination’s validity.

This study has several limitations. It used a small number of study participants and examination items, which may be insufficient to achieving thematic saturation or elucidating relationships completely. We converted COMLEX-USA scores for 4 participants. Multiple formulas exist, opening the possibility for misclassification of USMLE performance.^{20,32} Although think-aloud methodology is recognized as a means to capture thought processes,¹⁷ it does not capture all thinking, particularly nonanalytic reasoning.²¹ Additionally, USMLE scores may disproportionately reflect quick nonanalytic reasoning,⁷ which was not encouraged by this study protocol. Participants’ behavior may also have been influenced by the Hawthorne effect (ie, impact of being studied on participant behavior). Lastly, our participants were in their first postgraduate year. CR strategies change with increasing training,⁵ suggesting our results cannot be extrapolated to more senior trainees.

Further research should include confirmation and elicitation of CR strategies and TT behaviors, include participants at different levels of training, and examine the relationship between high-stakes clinical vignette MCQ examination scores and performance on other measures of CR.

Conclusion

We qualitatively probed reasoning strategies and TT behaviors of physician trainees during MCQs and found different patterns of CR strategies and TT behaviors were used by high and low performers. Patterns found also may be more reflective of MCQ TT behaviors than skills activated by real life clinical scenarios.

References

- 1 Elstein AS, Kagan N, Shulman LS, Jason H, Loupe MJ. Methods and theory in the study of medical inquiry. *J Med Educ.* 1972;47(2):85–92.
- 2 Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* 2001;357(9260):945–949.
- 3 Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Med Educ.* 2004;4:23.
- 4 Skakun EN, Maguire TO, Cook DA. Strategy choices in multiple-choice items. *Acad Med.* 1994;69(suppl 10):7–9.
- 5 Coderre S, Mandin H, Harasym PH, Fick GH. Diagnostic reasoning strategies and diagnostic success. *Med Educ.* 2003;37(8):695–703.
- 6 Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Med Educ.* 2008;42(5):468–475.
- 7 Ilgen JS, Bowen JL, McIntyre LA, Banh KV, Barnes D, Coates WC, et al. Comparing diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought. *Acad Med.* 2013;88(10):1545–1551.
- 8 Heemskerk L, Norman G, Chou S, Mintz M, Mandin H, McLaughlin K. The effect of question format and task difficulty on reasoning strategies and diagnostic performance in internal medicine residents. *Adv Health Sci Educ Theory Pract.* 2008;13(4):453–462.
- 9 McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach.* 2004;26(8):709–712.
- 10 Millman J, Bishop CH, Ebel R. An analysis of test-wiseness. *Educational and Psychological Measurement.* 1965;25:707–726.
- 11 Rogers WT, Bateson DJ. Verification of a model of test-taking behavior of high school seniors. *J Exp Educ.* 1991;59(4):331–350.
- 12 Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med.* 2005;165(13):1493–1499.
- 13 Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med.* 2008;121(suppl 5):2–23.
- 14 Glaser BG. *Basics of Grounded Theory Analysis: Emergence vs. Forcing.* Mill Valley, CA: Sociology Press; 1992.
- 15 Alguire PC, ed. *MKSAP Digital 15.* Philadelphia, PA: American College of Physicians; 2010.
- 16 Bloom BS. *Taxonomy of Educational Objectives: The Classification of Educational Goals.* New York, NY: Longmans; 1956.
- 17 Ericsson KA, Charness N, Feltovich PJ, Hoffman RR. *The Cambridge Handbook of Expertise and Expert Performance.* Cambridge, MA: Cambridge University Press; 2006.
- 18 Durning SJ, Artino AR Jr, Beckman TJ, Graner J, van der Vleuten C, Holmboe E, et al. Does the think-aloud protocol reflect thinking?: exploring functional neuroimaging differences with thinking (answering multiple choice questions) versus thinking aloud. *Med Teach.* 2013;35(9):720–726.
- 19 Slocum PC, Louder JS. How to predict USMLE scores from COMLEX-USA scores: a guide for directors of ACGME-accredited residency programs. *J Am Osteopath Assoc.* 2006;106(9):568–569.
- 20 Chick DA, Friedman HP, Young VB, Solomon D. Relationship between COMLEX and USMLE scores among osteopathic medical students who take both examinations. *Teach Learn Med.* 2010;22(1):3–7.
- 21 Eva KW, Brooks L, Norman GR. Forward reasoning as a hallmark of expertise in medicine: logical, psychological, and phenomenological inconsistencies. *Adv Psychol Res.* 2002;8:41–69.
- 22 Norman GR, Brooks LR. The non-analytical basis of clinical reasoning. *Adv Health Sci Educ Theory Pract.* 1997;2(2):173–184.
- 23 Patel VL, G.J. Knowledge-based solution strategies in medical reasoning. *Cognitive Sci.* 1986;10:91–116.
- 24 Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ.* 2005;39(1):98–106.
- 25 US Medical Licensing Examination. Score and score reporting. <http://www.usmle.org/bulletin/scores/>. Accessed September 2, 2014.
- 26 O'Donnell MJ, Obenshain SS, Erdmann JB. Background essential to the proper use of results of step 1 and step 2 of the USMLE. *Acad Med.* 1993;68(10):734–739.
- 27 McConnell MM, Regehr G, Wood TJ, Eva KW. Self-monitoring and its relationship to medical knowledge. *Adv Health Sci Educ Theory Pract.* 2012;17(3):311–323.
- 28 Landefeld CS, Chren MM, Myers A, Geller R, Robbins S, Goldman L. Diagnostic yield of the autopsy in a university hospital and a community hospital. *N Engl J Med.* 1988;318(19):1249–1254.
- 29 Potchen EJ. Measuring observer performance in chest radiology: some experiences. *J Am Coll Radiol.* 2006;3(6):423–432.
- 30 Farr R, Pritchard R, Smitten B. A description of what happens when an examinee takes a multiple-choice reading comprehension test. *J Educ Meas.* 1990;27(3):209–226.
- 31 Swanson DB, Holtzman KZ, Allbee K. Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Acad Med.* 2008;83(suppl 10):21–24.
- 32 Lee AS, Chang L, Feng E, Helf S. Reliability and validity of conversion formulas between comprehensive osteopathic medical licensing examination of the United States level 1 and United States medical licensing examination step 1. *J Grad Med Educ.* 2014;6(2):280–283.