Beyond standard checklist assessment: Question sequence may impact student performance

Jeff LaRochelle · Steven J. Durning · John R. Boulet · Cees van der Vleuten · Jeroen van Merrienboer · Jeroen Donkers

Abstract

Introduction Clinical encounters are often assessed using a checklist. However, without direct faculty observation, the timing and sequence of questions are not captured. We theorized that the sequence of questions can be captured and measured using coherence scores that may distinguish between low and high performing candidates.

Methods A logical sequence of key features was determined using the standard case checklist for an observed structured clinical exam (OSCE). An independent clinician educator reviewed each encounter to provide a global rating. Coherence scores were calculated based on question sequence. These scores were compared with global ratings and checklist scores.

Results Coherence scores were positively correlated to checklist scores and to global ratings, and these correlations increased as global ratings improved. Coherence scores explained more of the variance in student performance as global ratings improved.

Discussion Logically structured question sequences may indicate a higher performing student, and this information is often lost when using only overall checklist scores.

Conclusions The sequence test takers ask questions can be accurately recorded, and is correlated to checklist scores and to global ratings. The sequence of questions during a clinical encounter is not captured by traditional checklist scoring, and may represent an important dimension of performance.

Keywords Clinical skills · Medical education · Assessment

Essentials

- Recording question sequence is an important addition to checklist scores.
- Global ratings of student clinical skills are correlated with question sequence.
- Question sequence may better discriminate between low and high performing students.

Introduction

Learning how to conduct a successful patient encounter is a major goal in the pre-clerkship period of medical school and beyond. A successful patient encounter requires a complex integration of medical knowledge, communication and physical exam skills, in addition to clinical reasoning, which drives the entire process. The assessment of this complex interaction, whether on an objective structured clinical examination (OSCE) or individual patient encounters with real or standardized patients, often employs the use of a checklist [1]. These checklists typically reduce any particular encounter into the basic scenario-specific

J. LaRochelle (✉) · S. J. Durning
Department of Medicine, Uniformed Services University of the Health Sciences
Bethesda, USA
e-mail: jeffrey.la-rochelle@usuhs.edu

J. R. Boulet
Department of Research, FAIMER
Philadelphia, USA

C. van der Vleuten · J. van Merrienboer · J. Donkers
Department of Educational Development and Research,
Maastricht University
Maastricht, The Netherlands

historical and physical exam components that are ideally representative of data gathering tasks required to identify a particular diagnosis or condition. If properly constructed, these checklists represent a scoring method that should ideally capture the essential elements of the patient encounter [2]. In this way, standardized patients, faculty, or others can then use these checklists, or analytic tools, to assess candidates, and with appropriate training, demonstrate a reasonably high degree of inter-rater reliability in their assessments of a learner's data gathering skills [3].

The most common use of a checklist for these clinical encounters is to simply add up the number of items accomplished correctly for each encounter, which then becomes the overall score for that encounter. However, there are several potential problems with this approach, and one key issue relates to the particular importance of each individual checklist item. For example, in a patient with unexplained syncope, a detailed family history of any sudden death would be quite important whereas in a patient with a sinus infection, a detailed review of their medication allergies would rise to greater importance. One of the more common solutions to address this issue is to weight (assign higher priority to) individual checklist items based on how essential the item is believed to be for establishing the diagnosis and/or therapy. The difficulty associated with this solution is twofold. First, this could further complicate the checklist development process, and add another layer of complexity regarding which items should be given more weight. Second, the amount of weight given to these items may appear to be arbitrary, and may not have the intended impact on student assessments or performance [4–6]. A different approach to giving an additional weight to certain checklist items is to identify critical items to be analyzed separately from the overall checklist, or to design a checklist based solely on a key features approach for a particular patient encounter [7–9]. However, the identification of critical items must still be based on current practice guidelines using a transparent and standard approach; otherwise, from a validity perspective, placing a high degree of importance on these items could be questioned [10].

Another potential problem with the use of checklists that simply record the number of correct responses is that there is no accounting for how long student or resident trainees take to ask critical questions, or the sequence in which they ask these questions. A candidate who can simply ask a standard set of rapid-fire questions during a patient encounter may not be distinguishable from students who take an approach based upon the patient presentation. In other words, the logical sequence of questions is not taken into account with a scoring system that simply adds up correct responses, regardless of item weighting or key features checklists. We would expect an end of pre-clerkship student to accomplish a focused history and physical directed by a chief complaint, and be able to justify a reasonable differential diagnosis with data gathered using appropriate medical terminology. A lower performing examinee may simply stumble upon the critical items or key features within a patient encounter without having had a logical method or identifiable thought process applied to the clinical scenario. This method of assessment may inadvertently promote an approach to patient encounters that subverts the tenets of situated cognition theory, where clinical reasoning is nested in specifics of the situation, in this case a patient encounter, emphasizing both the environment and the participants (patient and candidate) in addition to their interactions which emerge based on these factors [11]. Therefore, an over-reliance on checklist assessments may encourage candidates to ask generic, perhaps inappropriate, sets of questions on every patient regardless of the situation, and thus devalue the concept that a patient encounter is actually an interactive and evolving process between the (candidate) physician and the patient where the sequence of questions, or their timing, would be expected to be important based on this theory.

Clinician educators can use global ratings in an attempt to counter these potential flaws in the way checklist-based assessments are scored. Global ratings have been shown to be better correlated with learner judgment and skills; therefore, they may provide performance measures (ability estimates) that can distinguish between students who may have received similar checklist scores, but who differ in their approach to the patient encounter, or who have missed any key features in the scenario [12]. Therefore, global ratings may be able to identify unique aspects of encounters with patients not readily captured by the traditional scoring of checklists.

Global ratings can suffer from poor inter-rater reliability, although generally studies have found both reliability and validity evidence with these assessments with appropriate experts, training, and task anchors [13, 14]. Given limitations in resources and/or available expert faculty, the ability to identify and deconstruct various components of a global rating, which could then potentially be related back to checklist items, may provide the same benefit of a global rating combined with the accuracy of a checklist scoring method. One such component of the global rating not typically identified in current checklists may be the sequencing of questions that candidates ask of a patient during a clinical encounter. In this study, we sought to determine if the sequence of checklist items on a single end of second year OSCE case differentiates performance between low and high performing students as determined by a global rating. We hypothesized that the sequence of questions asked would be able to discriminate between students, and that this sequencing component will be correlated with faculty global ratings and with overall checklist ratings of students.

Table 1 Sequentially ordered key factors scored during the patient encounters

1 Polydipsia	2 Duration
3 How much drinking	4 Urinary frequency
5 Nocturia	6 Urine volume
7 UTI symptoms	8 Weight changes
9 Polyphagia	10 Blurred vision
11 Paresthesias	12 Sexual dysfunction
13 Past medical history	14 CAGE (alcohol use screening)
15 Alcohol use	16 Dyspnoea on exertion
17 Short of breath	18 Family history of DM
19 Family history of CAD	20 Previous diagnosis of DM

CAD coronary artery disease; *CAGE* cutback, angry, guilty, eye-opener; *DM* diabetes mellitus; *UTI* urinary tract infection

Methods

The F. E. Hebert School of Medicine, within the Uniformed Services University of the Health Sciences (USU), is the only federal medical school in the US. The school matriculates 160–170 medical students annually, and students are exposed to both medicine and military in their curriculum. At the time of this study there was a traditional curricular model, whereby students complete two years of classroom work followed by two years of clerkship activities. The former is comprised of one year of basic sciences and one year of pre-clinical (or transition to clerkship courses), with the pre-clinical year being taught largely in a small-group setting. This study involved students at the end of their second or 'pre-clerkship' year at our medical school.

We were able to obtain the video recordings for 145 of the 168 students who completed the end of the second year objective structured clinical exam (OSCE). All students are required to complete a six station OSCE as part of their clinical skills and clinical reasoning courses, and we selected a single station from that OSCE for this study. Each OSCE station focuses on the domains of communication skills, history taking skills, physical exam skills, and clinical reasoning. Communication skills are assessed by standardized patients using a modified version of the Essential Elements of Communication (EEC) from the Kalamazoo conference [15]. Both history taking skills and physical exam skills are assessed by standardized patients using faculty developed checklists for each station, and the clinical reasoning domain is assessed using a combination of multiple choice questions and free-text response items.

For the standardized patient encounter under investigation, there were 20 possible key features on the standard checklist originally scored by trained standardized patients. These key features were sequentially numbered from 1 to 20 (Table 1) based upon the most appropriate logical sequence for the encounter by consensus of two study authors. Using a recorded videotape of each encounter, the sequence of questions asked was recorded independently by an outside physician with experience as a clinician educator, who then provided each encounter with a global rating based upon her expert opinion of what represented logical sequencing of questions using a scale from 0 (no logical sequence) to 5 (highly logical sequence) with 0.5 point increments. In this study we did not include the exact time a student asked about or elicited a key feature, and only took into account the sequencing of the key features. We also recorded the overall checklist score generated from the standardized patient and based on a traditional method of adding up all the unweighted correctly accomplished checklist items.

Since it is not possible to simply compare the student sequences with a 'best' ordering using a standard edit-distance measure such as the Levenshtein distance [16] as there might be multiple equally good 'best' sequences, we first grouped the encounters together based on their assigned global rating to determine whether the global rating for a given patient encounter correlated to the actual sequence of questions (key factors) asked by the student. The computational technique used in this paper enables us to measure how much the ordering of key features in a given encounter resembles all of the orderings in a set of encounters. We start with some assumptions and definitions. We will use the more general term 'item' instead of key factor and 'sequence' to indicate the order of key factors in an encounter.

First, we define the set of items involved (in this paper, the key features): $(K) = \{k_1, ..., k_n\}$, where n is the number of items. Each key factor occurs once in an encounter, and not all key factors are necessarily asked in each encounter. The key concept in our method is how often a given item occurs before or after another item in a set of sequences. Each encounter between a learner and a standardized patient produces a specific sequence of checklist items, and for each encounter a global rating (based upon the overall performance of the student) is independently generated by a clinician educator (Fig. 1, top left corner). Applying this concept to all the possible orderings of items for this set of sequences, we generate an average ordering matrix for a given global rating, where each cell represents the proba-
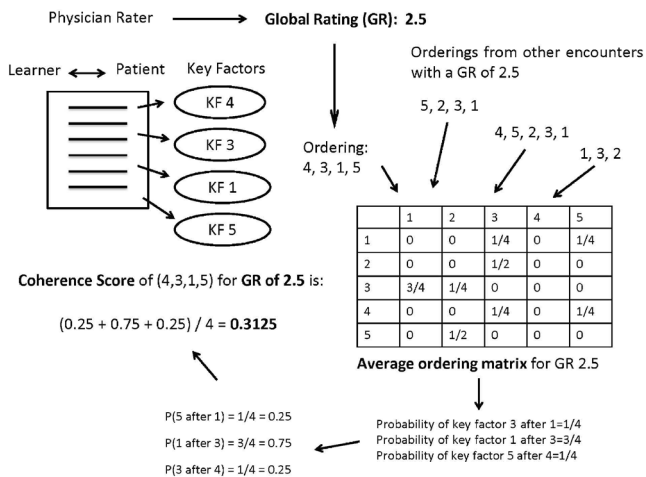
Fig. 1 Demonstrates a visual representation of the method to compute a coherence score. First, a checklist sequence is generated from a student/patient encounter, and a clinician educator (physician) independently generates a global rating score for the overall student performance. All of the sequences for a given global rating (2.5 in this example) are put into an average ordering matrix, and the probabilities of any key factor occurring after another are calculated. Finally, a coherence score for each individual checklist sequence associated with a specific global rating is calculated. In this example, the coherence score of (4, 3, 1, 5) is 0.3125 for a global rating of 2.5

bility that a given item occurs later than another given item within any of the sequences (Fig. 1, bottom right corner).

We then use the average ordering matrix to compute a measure for how closely a sequence ordering resembles those in a set of sequences with the same global rating, which we refer to as the coherence score. The matrix computes the average probability of all ordered pairs for a specific sequence within a given set of sequences, and the coherence score reflects the average probability of all these ordered pairs of items (Fig. 1, lower left corner). Therefore, a higher coherence score indicates a sequence that has a high probability of existing within the set of sequences, and closely resembles the ideal ordering, while lower coherence scores reflect increasingly random sequences within a particular set of sequences. We constructed an average ordering matrix for each set of encounters with the same global rating, which represents the average ordering of the key factors in this set of encounters. In total, we generated 11 of these average ordering matrices based on the number of different global ratings given. This method is similar to generating a matrix of transition probabilities used in a Markov chain analysis, which describes the probabilities of specific transitions in a sequence of random variables which have a serial dependence only between adjacent events [17]. This process is preferable to preference rank statistics, as we were not concerned with which key factor occurred first or last, but rather in the overall ordering of sequences.

Graphically, cells are populated in the upper right corner of the matrix as the items in a given set of sequences are similar to the ideal ordering (i. e. k_i is before k_j), and these cells become darker as the probability for this ordering increases within the set of sequences. For example, if all the sequences in a particular set are exactly equal to the ideal ordering (k_1, ..., k_n), then all the cells will be concentrated in the upper right corner of the matrix with the darkest shading to indicate a high probability. If sequences are more randomly ordered in a particular set of sequences, then graphically, this will be represented by a more even distribution of cells in both the upper right and lower left corners across the matrix with lighter shading to indicate lower probability values. In other words, more cells in the upper right corner indicate orderings that more closely resemble the ideal sequence, and more cells in the left lower corner indicate orderings that differ from the ideal sequence, while darker shading of cells indicates an increased probability of finding that specific sequence within the set of all sequences (Fig. 2).

We computed the coherence scores of all encounters at the level of the student and created an average ordering matrix for each global rating score. Additionally, we analysed the correlations between these coherence scores with the overall checklist scores in addition to the experts' global ratings that were assigned to the encounters using a Pearson's correlation coefficient and a standard linear regression.

This study was approved by the Uniformed Services University Institutional Review Board.

Results

We found several notable findings related to the sequence of questions asked during a patient encounter. We found that higher average coherence scores were more likely to be associated with higher global ratings (Table 2). Additionally, coherence scores associated with lower global ratings were similar, but coherences scores associated with matrices of high global ratings demonstrated better discrimination of students. In other words, there was a broader range of coherence scores for students receiving higher global ratings, indicating that using the coherence scores as an additional metric would further distinguish between students who may have received the same global rating. Fig. 3 provides a graphical representation of this finding where the coherence scores were distributed with increasing discrimination among global ratings with matrices for the higher global ratings. We also found positive correlations between coherence scores and global ratings that increased with average score matrices for higher global ratings, and higher global ratings explained more of the variance in coherence scores (Table 2) providing further evidence that the sequence in which a student obtains key factors during a patient encounter is not only correlated to, but predictive of expert

Fig. 2 Compact graphical representation of the average-ordering matrices for each of the ratings from 0 to 5. Each cell (i, j) in matrices is the probability that key factor j is asked later than key factor i in a consult that received that rating. Darker cells indicate a higher probability of i preceding j. The 20 key factors are ordered in the matrices in the most logical order for the encounter (see Table 1), hence the right upper triangle of the matrices are expected to be darker in the matrices with higher ratings. The matrix for grade 0.5 is missing since no encounter received this rating

Table 2 Pearson's correlation coefficients and linear regression analysis between global rating scores and coherence scores for each of the average-ordering matrices. Matrix '0' stands for the average ordering matrix computed for students with global rating 0

Matrix	Correlation	Intercept	Slope	$P_{intercept}$	P_{slope}	R^2
0	0.132	1.64	0.069	< 0.001	0.112	0.017
1	0.115	1.46	0.030	< 0.001	0.169	0.013
1.5	0.002	2.01	0.001	< 0.001	0.983	< 0.001
2	0.243	1.79	0.064	< 0.001	0.003	0.058
2.5	0.399	1.75	0.116	< 0.001	< 0.001	0.159
3	0.478	1.95	0.183	< 0.001	< 0.001	0.228
3.5	0.501	1.79	0.201	< 0.001	< 0.001	0.250
4	0.618	1.88	0.262	< 0.001	< 0.001	0.382
4.5	0.490	2.47	0.299	< 0.001	< 0.001	0.240
5	0.519	2.24	0.308	< 0.001	< 0.001	0.270

global ratings for that encounter. We also found that coherence scores were positively correlated with overall checklist scores, and that this correlation also increased with increasing global rating (Table 3).

Discussion

In our study of a single case on an end of second year OSCE, we found that we were able to differentiate student performance based upon the sequence of questions asked. This added dimension of assessment obtained from a standard checklist was not only correlated with both overall checklist scores and expert global ratings, but was able to explain an increasing percentage of the variance in global ratings as scores increased. The sequence in which a student obtains key features during a patient encounter may represent a unique aspect of student performance, which can now be obtained from a standard checklist instead of relying on expert global ratings. The nuance of the interchange of individual questions, or even sets of questions that are important to patient care, is effectively captured by an expert observing the encounter to provide a global rating. Given the complexity of a patient encounter, there is not likely to be a single ideal sequence, but rather a clustering of related sequences. In fact, as the global ratings for an encounter increase, a clustering of related sequences emerged, resulting in a broader range of coherence scores.
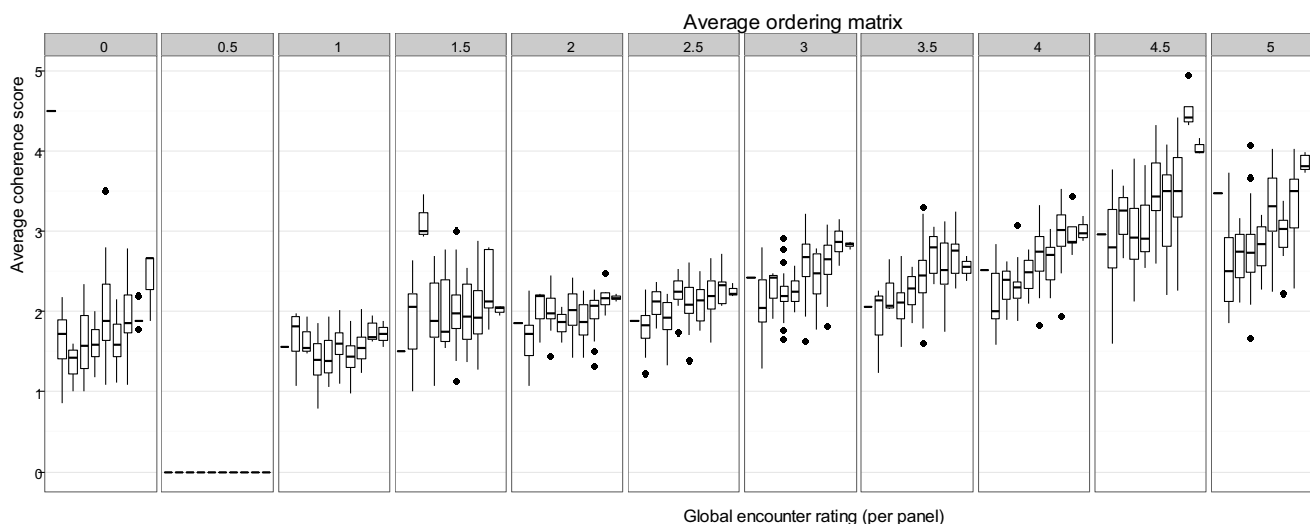
Fig. 3 Represents a distribution of the coherence scores within each global rating score (consult grade) as a function of average ordering matrix score (matrix grade). Within a panel, a distribution of boxplots is produced for the coherence scores for that average ordering matrix per global rating score

Table 3 Pearson's correlation coefficients analysis between overall checklist scores and coherence scores for each of the average-ordering matrices. Matrix '0' stands for the average ordering matrix computed for students with global rating 0

Matrix	Correlation
0	0.304
1	0.293
1.5	–0.081
2	0.135
2.5	0.313
3	0.484
3.5	0.398
4	0.520
4.5	0.489
5	0.488

Assuming that the ordering sequence will cluster around a more ideal sequence with higher global ratings, the individual coherence scores for learners in this group will be more differentiating than in students with lower global ratings where the coherence scores are low and range restricted. In other words, a student who receives a high global rating, but has an outlying sequence, will be easily identified through a lower coherence score. For that reason, our analysis using coherence scores as related to global rating scores adds more power to the use of coherence scores as a higher fidelity marker for logical sequences of questions even if there are slight differences between the sequences.

A potential benefit to the use of sequencing as a separate construct assessed during patient encounters is that it re-emphasizes the importance of engaging in an active communication process with the patient as opposed to rote memorization of questions to be asked. As stated

previously, situated cognition theory would argue that the sequence of questions is essential to high-quality patient care since the interaction between the physician (student in this case), the patient, and the environment is a critical component for effective reasoning. In other words, recording the sequence of questions asked as another component of the overall assessment will potentially discourage students from employing rapid-fire questions in a haphazard manner as a means of obtaining the appropriate information during a patient encounter, and encourage students to optimize a clinical encounter through a logical sequence of questions based on patient feedback and responses. In addition, a logical approach to history taking based on patient and physician interaction implies that the timing of questions will also be efficient and appropriate. In theory, a higher performing student may actually obtain the critical features of a particular encounter earlier based upon a more logical approach to their questioning. This may be another discriminating factor worthy of research that is not currently captured by the traditional scoring of checklist assessments.

Arguably, global rating rubrics using experienced and well-trained clinicians can be an effective alternative to this approach, and there are benefits to using clinicians as raters of medical students during patient encounters. However, given potential limitations in resources and variable inter-rater reliabilities, it could be particularly advantageous to deconstruct the overall global rating into readily identifiable components correlated to the sequence of checklist items. In this case, we were successfully able to dissect out the order in which key features were obtained during a patient encounter, and demonstrate a correlation to global ratings of sequence given by clinician educators. In fact, more logical sequences were able to better discriminate between

low and high performing students based on these global ratings. Using modified scoring of checklists (e. g. coherence scores) to capture not just historical information or physical exam items accomplished, but also capture the sequencing of questions during a patient encounter, could be a potentially powerful addition to the overall assessment of students' performance, informing or potentially replacing the information obtained on global rating scores. Although we recorded the sequence of questions by videotape after the completion of an encounter, this same information can be obtained using an electronic time stamp in real time as a standardized patient is recording data from a standard checklist.

There are several limitations to our study. First, this was a single institution study in a single class of medical students, and the information gathered may not be readily generalizable to the broader population of medical students at other medical schools. Second, we studied a single case on an end of second year medical student OSCE. The content of a particular case has an important impact on both clinical skills and reasoning, and we would need to evaluate the use of timing and sequence across a spectrum of clinical content, and different situations. Comparative generalizability studies could be done, comparing the added signal or universe score variance of sequencing with more standard uses of checklist scoring. Additionally, further studies on trainees at different levels of medical education (e. g., clerkship students or interns) could add insight into the development of clinical skills and clinical reasoning. This is particularly important, as our study was conducted on pre-clerkship students using standardized patients adhering to a specific script, and an organized approach with logical sequencing to an encounter on a real patient may be even more critical to realizing a successful outcome.

Conclusion

A successful patient encounter represents the integration of a variety of skills and abilities, and the logical timing and sequencing of questions appears to be an important marker of ability, and likely represents evidence of a hierarchical approach to problems commonly seen as trainees develop expertise. The use of scoring methods to incorporate sequence and even the timing of questions could represent an important advance in the assessment of novice medical students during patient encounters.

Disclaimer The views expressed in this manuscript are those of the authors and do not reflect the official policy or position of the Uniformed Services University of the Health Sciences, Department of Defence or the US Government.

References

1. Newble D. Techniques for measuring clinical competence: Observed structured clinical examinations. Med Educ. 2004;38:199–203.
2. Boulet JR, van Zanen M, de Champlain A, Hawkins RE, Peitzman SJ. Checklist content on a standardized patient assessment: An ex post facto review. Adv Health Sci Educ Theory Pract. 2008;13:59–69.
3. Heine N, Garman K, Wallace P, Bartos R, Richards A. An analysis of standardised patient checklist errors and their effect on student scores. Med Educ. 2003;37:99–104.
4. Clauser BE, Margolis MJ, Swanson DB. An examination of the contribution of the computer-based case simulations to the use of the USMLE Step 3 Examination. Acad Med. 2002;77(10 suppl):S80–S82.
5. Kane M, Case SM. The reliability and validity of weighted composite scores. Appl Meas Educ. 2004;17:221–240.
6. Kahraman N, Clauser BE, Margolis MJ. A comparison of alternative item weighting strategies on the data gathering component of a clinical skills performance assessment. Acad Med. 2008;83(10 suppl):S72–S75.
7. Bordage G, Brailovsky C, Carretier H, Page G. Content validation of key features on a national examination of clinical decision-making skills. Acad Med. 1995;70:276–281.
8. Hatala R, Norman G. Adapting the key features examination for a clinical clerkship. Med Educ. 2002;36:160–165.
9. Payne NJ, Bradley EB, Heald EB, et al. Sharpening the eye of the OSCE with critical action analysis. Acad Med. 2008;83:900–905.
10. Gorter S, Rethans JJ, Scherpbier A, Heijde D, Houben H, Vleuten C. Developing case-specific checklists for standardized-patient-based assessments in internal medicine: A review of the literature. Acad Med. 2000;75:1130–1137.
11. Durning S, Artino AR, Pangaro L, van der Vleuten C, Schuwirth L. Context and clinical reasoning: Understanding the perspective of the expert's voice. Med Educ. 2011;45:927–938.
12. Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Acad Med. 2001;76:1053–1055.
13. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med. 1998;73:993–997.
14. Spillane L, Hayden E, Fernandez R, et al. The assessment of individual cognitive expertise and clinical competency: A research agenda. Acad Emerg Med. 2008;15:1071–1078.
15. Joyce BL, Steenbergh T, Scher E. Use of the Kalamazoo essential elements of communication checklist (adapted) in an institutional interpersonal and communication skills curriculum. J Grad Med Educ. 2010;2:165–169.
16. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Doklady. 1996;10:707–710.

17. Kemeny JG, Snell JL. Finite Markov Chains. Princeton, NJ: Van Nostrand; 1960.

Jeff LaRochelle is an associate professor, Department of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD.

Steven J. Durning is a professor of medicine, Department of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD.

John R. Boulet is the vice president, Research and Data Resources, Foundation for Advancement of International Medical Education and Research, Philadelphia, USA.

Cees van der Vleuten is a professor of education, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

Jeroen van Merrienboer is a full professor of Learning and Instruction, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

Jeroen Donkers is an assistant professor of education, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.