What Does an OSCE Checklist Measure?

Heidi S. Chumley, MD

In this issue of *Family Medicine,* Drs Turner and Dankoski[1] raise questions about the reliability and validity of objective structured clinical examinations (OSCEs). They propose that many institutions may be using an OSCE that does not have adequate reliability, and their interpretation of the literature is that the validity of OSCEs has not been established. I agree with their conclusions.

The reader, however, may be left with the impression that the reliability and validity of OSCEs, and specifically of OSCEs using standardized patients, can be improved to acceptable levels in their own institutions simply with added resources and attention to details. I believe this is possible when standardized patients are used to test a narrow range of communication skills. I am less convinced about their use—even with the best of resources—as a comprehensive test of clinical competency.

I think of clinical competency as a concept that can be displayed in a Venn diagram, with overlapping circles of knowledge, clinical skills, and clinical reasoning. Knowledge can be measured with a well-written multiple-choice test, and it is thought that skills and reasoning can be measured with OSCEs. But, many institutions use standardized patient OSCEs as an overall test of clinical competency and score those OSCEs with a checklist. In my opinion, standardized patient OSCEs that are graded with a checklist probably do not effectively measure knowledge, clinical skill, or reasoning. We should keep standardized patients but abandon the checklist.

See related article on page 574.

(Fam Med 2008;40(8):589-91.)

From the Department of Family Medicine, Kansas University Medical Center.

The Problem With OSCE Checklists

Checklists combine content-specific and content non-specific items, creating a mixture of testing materials that together may not assess knowledge, skill, or reasoning. Most standardized patient checklists are developed in one of three ways: by a panel of experts, by case writers, or by experts reacting to checklists proposed by case writers.[2] Typically, checklists represent an agreed-on selection of critical items that a trainee should address in a specific encounter. However, for a particular encounter, even among experts there will be extreme variance in the specific items addressed. As a result, the checklist items include only those items on which everyone agrees.

Confusion Between Clinical Skills and Knowledge

The items on which everyone agrees include items like general characteristics of a symptom (eg, did trainee ask when symptoms started?) and some elements of the other parts of the medical history (eg, did trainee ask what medications were being taken?). No content-specific knowledge is needed to obtain credit for asking these questions. For other cases, the agreed-upon checklist may contain physical examination items that do require content knowledge if the trainee is to know what to examine. I would wager that many of those physical examination items have poor positive and negative likelihood ratios for diagnosing the conditions under consideration (but that is a different argument). The checklist may also contain content non-specific communication items and content-specific management items.

This combination of different types of items can give confusing results when evaluating a trainee's score. For example, does a trainee who scored poorly do so because of problems with knowledge, skill, or reasoning? Most often, it is impossible to tell. As such, the usefulness of a standardized patient OSCE to assess overall clinical performance is plagued by content specificity, and this may partly be because an OSCE measures a skill set for which the underlying knowledge is assumed.

For example, to competently evaluate clinical skills using a standardized patient presenting with an acute symptom, a student must have the knowledge of likely causes of that symptom. A case designed to test counseling about

tobacco use requires content knowledge of tobacco dependence and behavior change. We should not be surprised, therefore, when there is no correlation between students' performances on two counseling cases, one on tobacco use and one on weight loss. Some students will perform better on one case or the other based on their underlying knowledge of each subject. This leads to a low reliability of these two cases for assessing students' abilities to counsel for behavior change.

An interesting tactic would be to remove as much of the interference with case-specific knowledge as possible. If you told students that their skill at counseling for behavior change will be assessed on two patients, one who smokes and the other who desires weight loss, and if the student prepared for the assessment by reviewing the required knowledge, and if the checklist had only general behavior change items, then I would anticipate improved reliability on this two-station test of counseling skills. I am not sure how often this is done, however, even for formative assessments.

Similarly, I think it is important to consider if the element of surprise is critical to the assessment of specific skills. Norman's statement[3] that an OSCE is about as good at predicting performance as a multiple-choice test of relevant knowledge is only true because we have not controlled for knowledge differences among students undergoing skills assessment. Granted, at present, a student will not be allowed to review a content area before seeing each standardized patient on the USMLE-2CS. But, I would put forward that many faculty members review content before or during a patient encounter in their practices, yet we test students without permitting them to undertake such reviews.

Assessing Clinical Reasoning

In contrast to the problems involved in distinguishing clinical skills from knowledge, assessment of clinical reasoning, especially diagnostic reasoning, seems feasible with a standardized patient OSCE. But, can we use a checklist to do this? Consider a set of 15 standardized patient cases, each presenting with an undifferentiated problem, and how a checklist might be used to assess a student's clinical reasoning by determining whether the student uncovers the key features of the cases outlined on the checklist.

Checklists are ideal for assessing skills that require several steps that should be completed the same way every time, such as starting an intravenous line or preparing an airplane for take-off. There is one best way. But, no two students will evaluate the 15 standardized patients using the exact same history questions and physical examination items in the same sequence.

Consider, for example, two students evaluating a patient with a headache. The first asks about vision changes and nausea in the history of present illness because some information has prompted a consideration of migraine. The second student asks about vision changes and nausea during a review of systems that the student uses on every patient regardless of presentation. Both students will get credit for these items on a checklist, but I would argue that these two students are functioning at two very different levels. A trained observing physician can see the first student incorporate an understanding of common causes of headache into the history, but scores from a checklist do not discern this difference. This limits the ability of a checklist to separate students performing at different levels.

Reevaluating the OSCE

Despite the aforementioned limitations, I nonetheless recommend continued use of standardized patients in OSCEs. They offer an ideal structure for infusing deliberate practice into medical education. Deliberate practice uses specific well-defined tasks, immediate feedback, and opportunity for repetition. This type of practice is needed to advance from acceptable performance to an expert level. I do not know how often standardized patients are used in this way, but they could be.

I do, however, recommend abandoning checklists or at least rethinking our approach to creating checklists and supplementing checklists with other measures. Supplementation with global assessment by a physician has improved testing characteristics,[4] but we need to go further.

We need to incorporate what science has taught us about physical and cognitive skill development. There are three levels of cognitive skill development: novice, acceptable performance, and expert. We need a method to determine where each student lies on this continuum. Indeed, experts generally score low on standardized patient OSCE checklists because they are able to reach diagnostic and treatment decisions with fewer steps,[5] thus not completing all the items on the checklist. This divergence between what experts do and what we ask trainees to do on an OSCE creates a problem with trying to define a single set of history and physical examination items and a single management strategy that can be used in a checklist. There are measurable differences between novices, acceptable performers, and experts, but they are difficult to capture with a checklist.

Because experts gather information and organize knowledge structures differently than do novices and acceptable performers, we need to make these differences the basis for assessment. We need methods to analyze students' information-gathering patterns. Fifteen years ago, novice and expert patterns of diagnosis were separated by pattern-recognition software;[6] we need to advance that work and apply it to current-day trainees. We

need methods to assess knowledge structures. For example, key feature problems[7] and script concordance tests[8] could be coupled to standardized patients to provide additional information about knowledge structures.

Conclusions

It's not just OSCEs to which the concerns I've expressed apply. Similar concerns exist for many other commonly used assessment strategies. For example, what is the reliability and validity of evaluation by a supervising physician who does not directly observe the student?

So, while I do not recommend abandoning OSCEs (what would we do instead?), I echo Turner's and Dankoski's call for attention to planning, commitment to larger-scale testing, and judicious use of assessment data. The concept of scoring with a checklist that tries to measure content-specific and non-specific items should be rethought. I also recommend that educational researchers study the effect that removing knowledge interferences has on the reliability of an OSCE checklist to measure a specific skill across content areas.

I would advocate for OSCEs to be a critical part of true deliberate practice for specific clinical skills. Spend the resources to add direct observation or video review with debriefing done by trained physicians and study the effect on reliability and validity. Consider the information-gathering pattern. But most importantly, assess as often as possible, in as many ways as possible.

Correspondence: Address correspondence to Dr Chumley, Kansas University Medical Center, 3901 Rainbow Blvd, Mail Stop 4010, Kansas City, KS 66160. 913-588-1996. Fax: 913-588-0195. hchumley@kumc.edu.

References

1. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. Fam Med 2008;40(8):574-8.
2. Gorter S, Rethans JJ, Scherpbier A, et al. Developing case-specific checklists for standardized-patient-based assessments in internal medicine: a review of the literature. Acad Med 2000;75(11):1130-7.
3 Norman G. Inverting the pyramid. Advances in Health Sciences Education 2005;(10):85-8.
4. 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(9):993-7.
5. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med 1999;74(10):1129-34.
6. Stevens RH, Lopo AC. Artificial neural network comparison of expert and novice problem-solving strategies. Proc Annu Symp Comput Appl Med Care 1994:64-8.
7. Fischer MR, Kopp V, Holzer M, Ruderich F, Junger J. A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. Med Teach 2005;27(5):450-5.
8. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12(4):189-95.