Assessing Clinical Performance
Where Do We Stand and What Might We Expect?

During the past decade the field of clinical assessment has demonstrated a clear move away from dependence on pencil and paper tests and the multiple-choice format as measures of predicting clinical competence to assessing what physicians do in the real world of practice.¹ This change is reflected in the orientation of numerous conferences on newer approaches to clinical assessment.²

See also p 735.

Among the many reasons for this shift are the increasing feasibility of administering large-scale, multicenter performance tests and the documented improvement in the measurement qualities of performance-based tests. These tests include the Objective Structured Clinical Examination (OSCE), in which candidates proceed through multiple, timed, independent stations designed to assess predetermined clinical skills, and the use of standardized patients (SPs), that is,lay people who are trained to simulate a variety of medical problems in a consistent, reliable, and realistic manner. Another reason for the shift is the realization that variation in candidates’ differential results on many clinical assessment instruments may be due more to content or case specificity than to inherent methods’ biases or errors.⁶⁷ While the potential impact of focusing on performance has been well understood in the measurement and undergraduate medical education communities, only recently has it begun to influence graduate education and postresidency physician assessment.

Based on a review of currently evolving concepts and approaches, this Editorial suggests that the field of clinical assessment will continue to emphasize performance-based testing and the evaluation of what physicians “do” in the everyday reality of clinical practice and training. To demonstrate that such approaches meet the basic measurement standards of reliability and validity, we will need many more valid “biopsies” of many clinical situations over time. Similarly, data gathered through proxy measurements of performance from sources such as administrative databases and assessments by patients and colleagues will have to meet the same quality standards. If properly done, this broader and more balanced sampling of physician behavior will effectively guide physicians in training and in practice to address their individual educational needs.

Performance-Based Testing

What do we mean by performance-based testing and content (case) specificity? The concept of performance-based testing is not new. George Miller’s familiar hierarchical pyramid of physician behaviors differentiates between “knows” as a reflection of knowledge needed for appropriate physician behavior (base of the pyramid); “knows how” to act when presented with a particular circumstance on direct questioning or on a test (the classical way to imply competency in traditional testing formats) (second level); “shows how” when asked to deal with a particular clinical circumstance (performs) (third level); and “does” (how one acts in real life) (fourth level). The fundamental concept of performance-based testing is that a candidate must “show how” and/or be observed “doing” the behavior to be assessed.

In contrast, during the past 25 years, the emphasis in clinical testing has been on “knows” (eg, multiple-choice questions [MCQs]) and “knows how” (eg, patient management problems) as Professor McGuire has so ably recounted in this issue of THE JOURNAL.³ With the discontinuation of the clinical oral examination by many specialty boards in the 1970s, much of the evaluation of “does” in clinical assessment has occurred during in-training. Relatively little emphasis was placed on the assessment of performance, or “show how.” During the last few years, however, educators have become disillusioned with the “does” approach to the in-training assessment format, mainly because rating scales have been the basic evaluation method.

Rating scales can be fraught with difficulties from a measurement point of view, primarily because they have poor reliability when the judges are not well trained and the evaluation situations are not well defined. In many clinical settings
where the optimal conditions are lacking or are not feasible, the efficacy of the rating scale suffers. For example, when rating scales are part of a national evaluation system or in use across multiple institutions, quality control of the conditions and the evaluators' performance often falls short of the ideal. In those cases, the reliability of global rating scales is acceptably low. One recent review of a national evaluation system used in residency training programs pointed out that the reliability of currently used rating scales was essentially zero.

In contrast, technical developments and improvements of SPs and OSCEs have led to their increasing use as a way to "show how" on performance-based tests. In 1989, the Josiah P. Macy Foundation sponsored a national conference calling for the greater use of performance-based testing in medical schools; around the same time leaders in the field of assessment formed the Clinical Skills Assessment Alliance both to share information on the expanding field and to promote the use of performance-based testing. Groups such as the American Medical Association, Association of American Medical Colleges, Federation of State Medical Boards, Medical Council of Canada (MCC), American Board of Medical Specialties, National Board of Medical Examiners (NBME), and Educational Commission for Foreign Medical Graduates (ECFMG) were active in the Alliance.

The lessons of that initiative were that performance-based testing was possible, reliable, and had the advantages of face validity because candidates were actually observed. The result was that the ECFMG and the NBME in the United States and the Quebec licensing authority and the MCC in Canada embarked on pilot projects and development programs, using the OSCE format with SPs and other simulations. While none of these organizations are promoting performance-based testing as the only answer, each organization has either adopted these methods or reasserted its commitment to introduce performance-based testing as part of its assessment packages.

But there were problems. The cost of OSCEs using SPs was a concern, although the use of sequential testing may help minimize costs in the future. And there were concerns about scoring, namely, the use by medical examiners of discrete objectified checklists vs global ratings. These current methods also have technical and contextual limitations, such as the difficulty of not unmasking the SPs to the candidate (who is blind as to which patient is an SP) if more than one visit is required in an office setting or the inability to submit the SP to invasive tests needed for the workup of certain clinical problems. None of the groups are considering dropping the assessment of knowledge, which along with experience is a key determinant of the development of clinical expertise.

Content (Case) Specificity
The issue of content or case specificity also needs to be examined. Studies looking at a number of evaluation formats have repeatedly shown that performance on one problem or case does not necessarily predict performance on other problems or cases. These findings are in keeping with studies on clinical reasoning in which content (case) specificity has been a consistent finding. That content specificity is a critical factor in the development of expertise challenges the earlier concept that an individual can possess the generic trait of being a good problem solver across cases. To rephrase an adage, once you have seen one problem solved, you have seen one problem solved!

What practical implications do these observations have for clinical assessment? It is not being argued that content specificity is reason enough to shift to performance-based assessment. While content specificity is a reason to retain widely sampling multiple-item formats for knowledge-based assessment, such as MCQ examinations, it also has implications for planning clinical assessment sampling. For example, given limited resources and the reality of content specificity, the best strategy in a formal assessment of clinical competence is to sample more situations with one observer per problem than to assess fewer situations with two or more observers per problem. The latter strategy requires the ability to deal with the additional issues of numbers of observers and interrater reliability. The key notion is that it is better to sample many situations with one observer as long as the same observer is not always used. While questions of reliability are important, issues of interrater reliability should not discourage sampling widely with one observer per problem or case.

Defining Objective Assessment Methods
The oversimplified notion that performance-based testing is better because it appears to be more objective must be challenged. Evaluators must guard against the temptation to create another intuitively attractive but incorrect concept analogous to the notion of a "good problem solver." The pursuit of objectivity in relation to questions of reliability, validity, and efficiency has been examined by Norman, de Graaff, and van der Vleuten. These authors emphasize not confusing the pursuit of objectivity (the empirical demonstration through studies of reliability that a measure is free of subjective influences) with the act of objectification (the use of methods, such as behavioral checklists, that appear to be objective). The evidence they review suggests that more objective-appearing methods are not automatically more reliable than so-called subjective methods.

Therefore, when referring to an assessment method as being objective, it is important to understand that despite the intuitive comfort of our desire for objectification, all so-called objective methods should be chosen on the basis of their demonstrated superiority to other existing methods and their appropriateness to the specific testing situation. These concerns of case specificity and objective methods vs subjective methods are just as important to performance-based testing as they are to the assessment of knowledge or any other tests traditionally used to imply competency or "knows how."

Future Approaches to Performance Assessment
Where will the next decade lead us in the field of clinical assessment and physician performance? Clearly, technical advances, such as wonderful forms of simulation, will evolve. However, because of increasing emphasis on physician accountability (e.g., in managed care), clinical assessment will continue to focus on "shows" and "does." The notion of case specificity, combined with the continuing push for assessment of "shows" and "does," will refocus our attention on direct observation or direct proxies of those activities physicians undertake on a daily basis.

In-training evaluation will refocus on more frequent direct observation of trainees and thereby, one hopes, surmount the
problems currently caused by the poor measurement qual-
ties of rating scales used in that context. Improvement will
occur as many more assessments are made over limited pe-
riods, guided by a sampling plan. In short, the notion of
multiple samples of a physician's performance in many clinical
situations during shorter, specified periods will prevail over
the current ritual of filling in rating scales on a monthly or
quarterly basis. A major contribution of technology will be to
simplify the recording of these direct observations on a regu-
lar and immediate basis by use of on-line data systems, with
feedback or summative assessment immediately accessible.21
Furthermore, the documentation of what is done and how
well it is done will extend beyond residency programs to the
practicing community.20 The impact of managed care and
more resource-conscious systems of care is already driving
the assessment equation in that direction.

The focus on individual performance will extend to group
performance. Program directors will want to know if gradu-
ate programs collectively and systematically show a certain
tendency, if that tendency is good or bad, and what should be
done about it. More importantly, the feedback from these
observations of performance will become part of physicians'
daily self-assessment, thereby assisting them in determining
their educational needs.21,22

Profiling and current uses of "report cards" in assessing the
clinical outcomes of physicians' practices have given rise to
concerns.23-25 All new approaches must be validated and shown
to be reliable. In all likelihood, approaches found to meet such
standards will be combined with other proven assessment
methods to give a more balanced and accurate view of the
individual physician's needs. Eventually, these assessment
approaches for entry into practice will be linked with mea-
sures of performance in practice and will lead to studies of
predictive validity for licensure and certification examinations
and other assessment methods.26,27 Finally, this informa-
tion will likely be used to inform test developers and improve
the assessment process.

As noted at the outset, the view from 1995 is continued
emphasis on "shows" with the use of innovative technologies
to support and simulate assessment; emphasis on assessing
what physicians "do" in the everyday practice of medicine,
with many more samplings of actual clinical situations over
time; and the continuing shift of clinical assessment in prac-
tice and training toward the use of information from admin-
istrative databases (when valid) and from patient and peer
assessments. Finally, given the continuing trend in managed
care in the United States and regional decision making in
Canada, the individual physician will become more focused on
maintaining skills and knowledge and therefore on opportu-
nities to provide feedback to help focus self-directed learning.
Sound a little Orwellian? Perhaps, but only for those who do
not see maintenance of performance as a personal responsi-
bility. I hope most will or already have accepted that personal
goal as part of the responsibility of the physician.

W. Dale Dauphinee, MD, FRCP
c
18. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med. 1990;2:56-76.

JAMA, September 6, 1995—Vol 274, No. 9

Editorial 743

Downloaded From: http://jamanetwork.com/pdftocaccess.aspx?url=/data/journals/jama/9402/ by a University of Ottawa User on 02/28/2017