

# SO WHAT DOES GUESSING THE RIGHT ANSWER OUT OF FOUR HAVE TO DO WITH COMPETENCE ANYWAY?

by Geoff Norman, Ph.D.

**E**verybody hates multiple-choice questions. They are viewed as an academic version of Trivial Pursuit, with little relevance to actual performance. In my field, medical education, there is a strong preference for performance-based assessment, based on actual work situations or elaborate “high-fidelity” simulations. But even in written testing situations, short-answer and essay tests are uniformly viewed as more valid and more valuable than multiple-choice questions.

I suspect that it isn’t all that different in law. In a recent issue of *THE BAR EXAMINER*,<sup>1</sup> Susan Case said, “Some people believe that essay and performance tests assess the most important aspects of readiness to practice law.” The presence of essay tests on virtually every U.S. bar exam implies that more than a few people feel that such tests have an essential role in assessing competence in law.

Essays do not marshal the same support in medicine, probably because doctors rarely have to make carefully reasoned written arguments (except, perhaps, when they find themselves involved in policymaking or lawsuits). But performance, in the ward, clinic, or operating room, is an essential component of professional medical practice. And in the medical field we have good methods of assessing performance, primarily an assessment tool called the Standardized Patient Examination or “SP exam,” in which the candidate goes from room to room in a simulated clinical setting, listening to a heart,

examining a knee, taking a history of a “patient” who presents with a cough, or counseling another patient about smoking cessation. The SP exam is now so ubiquitous that it is an essential component of the medical licensing examination in both Canada and the U.S., and about 15,000 applicants per year go through the exercise. The fact that licensing bodies in both countries are prepared to invest the resources to mount such an examination might be viewed as *prima facie* evidence that multiple-choice examinations are insufficient for assessment in this field.

However, if we move the issue from one of belief to one of evidence, and ask ourselves the question “What is the *evidence* that performance-based tests are more valid measures of clinical competence and better predictors of eventual outcomes such as malpractice?” then some quite counterintuitive results emerge. It is not a large body of literature—not very many people are prepared to hang around for a decade or two to see what happens to their graduates. But an interesting trend has emerged over the past 20 years.

In 1989, Paul Ramsey<sup>2</sup> looked at the relationship between scores on the American Board of Internal Medicine examination and performance in practice. The test is a full-day multiple-choice exam taken by physicians who wish to be certified in the specialty of internal medicine. Ramsey identified a sample of 259 internists who had been in practice an average of

7 to 10 years and had them evaluated by their peers using a standard peer-review rating form. Ramsey then correlated these evaluations with the internists' scores on the multiple-choice test taken 7 to 10 years earlier. The correlation was about 0.59, which is astonishingly high. (The correlation between LSAT scores and first-year grades in law school is 0.38; the correlation between undergraduate GPA and first-year law school grades is 0.24.)<sup>3</sup>

A more recent study produced even more dramatic results, because it looked at death due to heart attack. John Norcini<sup>4</sup> linked the records of 10,619 patients who had suffered heart attacks in the state of Pennsylvania to the records of the 2,078 physicians who cared for them in the coronary care units. Physicians who had passed their specialty board exams—multiple-choice tests in either internal medicine or cardiology—had, on average, a 19 percent lower mortality rate with these patients than did practicing physicians who had failed the specialty board exams.

Of course, there was a possibility that if Norcini had used results from a performance-based examination taken at the time of the board exams, the correlation would have been even higher. This comparison was made in another study. Paul Ram<sup>5</sup> took results from a 250-question multiple-choice test and from an eight-patient SP exam, and compared these results with videotaped patient consultations in real practice. The correlations were no higher for the SP exam than for the multiple-choice test—the correla-

tion with the multiple-choice test was 0.32–0.41 and the correlation with the SP exam was 0.23–0.41.

More recently, Robyn Tamblyn<sup>6</sup> looked at the relationship between performance on the Canadian medical licensing examination and complaints to the licensing body (some of which would end up as malpractice suits). The licensing exam is a two-part examination with multiple-choice and short-answer sections and, a year later, a performance-based SP exam in which candidates interact with simulated patients. Test score results were related to complaints to the provincial regulatory bodies in Ontario and Quebec. A total of 3,424 physicians were involved in the study; there were 696 complaints, most of which were for poor communication or poor quality of care.

The best predictor of complaints was the short-answer component of the written exam. The SP performance exam's measure of communication skills was nearly as good. However, the multiple-choice component of the written exam was a close third. In short, the written test was as good as the performance test in predicting subsequent malpractice-related complaints.

Why is it that multiple-choice tests appear to consistently outperform performance tests in terms of measured validity? One reason is technical, as described in detail in Case's article. Multiple-choice tests can test far more knowledge per unit of time than any other testing method. And the more observations you can make, the more confidence you

can have in the average score. Just as CONSUMER REPORTS' car ratings are useful because they represent the accumulated experience of hundreds or thousands of owners of a particular vehicle instead of the testimonial of your neighbor who ended up with a lemon, performance on a multiple-choice test of several hundred items provides more information on each candidate than can be gleaned from a few essay questions. It's the same principle used on the many election polls that flood the airways daily in an election year. Pollsters know that by sampling several thousand individuals, they can be reasonably confident that their reported percentages are quite accurate—that's what "This poll is accurate to +/-2 percent 95 percent of the time" means.

Nevertheless, if multiple-choice questions were measuring something that was unrelated to competence, it would be irrelevant how many questions were in the test. So there must be something of value associated with all those little black marks on the answer sheet. In short, is expertise primarily a matter of how much you know about a domain? Surprisingly, the answer, based on three decades of research, appears to be "Yes." Back in the 1960s the prevailing view was that expert doctors possessed "clinical problem-solving skills" that students lacked, and that medical school was really all about acquiring these skills. Studies were initiated at McMaster University<sup>7</sup> and Michigan State University<sup>8</sup> to explore exactly this premise. Expert physicians and students were videotaped as they

interviewed and examined standardized patients, and then the videos were reviewed in what was called "stimulated recall," in which the subjects were encouraged to reflect on their thinking at each stage as they watched the video. All this was coded in infinite detail and analyzed, in the hope of elucidating the clinical problem-solving process.

STUDIES OF EXPERTISE IN MANY OTHER DOMAINS—CHESS, COMPUTER PROGRAMMING, PHYSICS—SHOW SUBSTANTIALLY THE SAME THING. EXPERTS BECOME EXPERTS BY AMASSING A HUGE BODY OF BOTH FORMAL AND EXPERIENTIAL KNOWLEDGE. THAT'S WHAT LAW SCHOOL AND MEDICAL SCHOOL ARE ALL ABOUT. AND IT'S THE ACQUIRED SPECIALIZED KNOWLEDGE THAT DISTINGUISHES LAWYERS FROM PHYSICIANS OR SOCIAL WORKERS, NOT INTERPERSONAL OR GENERAL PROBLEM-SOLVING SKILLS.

But when the dust settled, an entirely different picture emerged. All the analysis showed that in terms of process, first-year students looked just like practicing physicians. Everyone was generating hunches (hypotheses) within a minute or two of seeing the patients. The difference was simply that doctors generated better hypotheses. And that ability led back to knowledge. Studies of expertise in many other domains—chess, computer programming, physics—show

substantially the same thing. Experts become experts by amassing a huge body of both formal and experiential knowledge. That's what law school and medical school are all about. And it's the acquired specialized knowledge that distinguishes lawyers from physicians or social workers, not interpersonal or general problem-solving skills.

Of course, that's not the whole story. A law student who aspires to become a litigator but gets stage fright every time he stands to speak in public, one who lacks the interpersonal skills to interact effectively with clients and colleagues, or one whose written language skills are so poor that she

cannot communicate effectively in writing with the precision required of legal documents will be unlikely to achieve any success in the legal profession. This being the case, it appears a straightforward implication that the bar exam should, almost axiomatically, contain both written and oral components.


Maybe, but then again, maybe not. Oral examinations have as long a history in medicine as they do in law; in fact, one mainstay in medical licensure even has the Latin name *viva voce*, and consists of a panel of three or four physicians who will interrogate the examinee for up to three hours. The examination has almost disappeared in the past couple of decades, as a consequence of evidence showing that, while judges in one session may agree that the candidate's performance was good or bad, there will be little relation between this judgment and the ratings assigned in the next session.<sup>9</sup> Again, it's an issue of sampling. The sample of knowledge assessed by any one panel is both small and biased, in that examiners tend to ask their pet questions, and so bears little relation to the sample assessed by the next panel of examiners.

As for the essay, things are even more complicated. Not only is there a sampling problem—you just can't sample enough knowledge in an hour or two of an essay—but study after study has shown that it is almost impossible to get judges to agree on scores for essay answers. Norcini<sup>10</sup> tried unsuccessfully to get inter-rater agreement in the American Board of Internal Medicine examination, but even after seven hours of training, agreement was only marginal. The Medical College Admissions Test, the equivalent of the LSAT for medical school applicants, has a writing sample component, administered at enormous expense, which consistently shows no correlation with subsequent performance in medical school and

is therefore ignored by many schools. At my university for many years we had a writing competency test to be taken by all undergraduates. Surprisingly, it was multiple-choice. But it was multiple-choice for a reason—a review of the literature showed clearly that writing competency could be better assessed by a well-designed multiple-choice test than by an essay exam.

Why are essay questions such weak assessment tools? Undoubtedly it is in part because of the inherent limits on sampling. But it also appears well-nigh impossible to even get score agreement between raters. This is pure speculation, but I think the process of grading essays is just too complex. One rater may be angered by illegible writing, another by deficient grammar or spelling, another by poor sentence structure, and a fourth by poor arguments and inadequate knowledge.

Which leaves us back where we began. It may be worth assessing legal skills more broadly than simply focusing on knowledge with a multiple-choice test. But the other test components, whatever they may be, should be additions to, not replacements for, the multiple-choice component and should be introduced with due attention to issues of reliability and validity.

Nobel laureate Herb Simon said, "The essence of intelligence is less a matter of general problem solving and more a matter of knowing a lot about the world." The same could be said for legal and medical competence. And that's what should be tested in the licensing examination. 

## ENDNOTES

1. Case, S.M., *The Testing Column: Best Practices with Weighting Examination Components*, 77 BAR EXAMINER 1:43 (February 2008).

2. Ramsey, P.G., J.D. Carline, T.S. Inui, E.B. Larson, J.P. LoGerfo & M.D. Wenrich, *Predictive Validity of Certification by the American Board of Internal Medicine*, 110 ANNALS OF INTERNAL MEDICINE 719 (1989).
3. Linn, R.L. & C.W. Hastings, *A Meta-Analysis of the Validity of Predictors of Performance in Law School*, 21 JOURNAL OF EDUCATIONAL MEASUREMENT 245 (1984).
4. Norcini, J.J., R.S. Lipner & H.R. Kimball, *Certifying Examination Performance and Patient Outcomes Following Acute Myocardial Infarction*, 36(9) MEDICAL EDUCATION 853 (2002).
5. Ram, P., C. van der Vleuten, J. Rethans, B. Schouten, S. Hobma & R. Grol, *Assessment in General Practice: The Predictive Value of Written-Knowledge Tests and a Multiple-Station Examination for Actual Medical Performance in Daily Practice*, 33(3) MEDICAL EDUCATION 197 (1999).
6. Tamblyn, R., M. Abrahamowicz, D. Dauphinee, E. Wenghofer, A. Jacques, D. Klass, S. Smee, D. Blackmore, N. Winslade, N. Girard, R. Du Berger, I. Bartman, D.L. Buckeridge & J.A. Hanley, *Physician Scores on a National Clinical Skills Examination as Predictors of Complaints to Medical Regulatory Authorities*, 298(9) JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION 993 (Sept. 5, 2007).
7. Neufeld, V.R., G.R. Norman, J.W. Feightner & H.S. Barrows, *Clinical Problem-Solving by Medical Students: A Cross-Sectional and Longitudinal Analysis*, 15 MEDICAL EDUCATION 315 (1981).
8. Shulman, L.S., A.S. Elstein & S.A. Sprafka, *MEDICAL PROBLEM-SOLVING: AN ANALYSIS OF CLINICAL REASONING*. Cambridge, Mass.: Harvard University Press, 1978.
9. Turnbull, J., D. Danoff & G. Norman, *Content Specificity and Oral Certification Exams*, 30(1) MEDICAL EDUCATION 56 (1996).
10. Day, S.C., J.J. Norcini, D. Diserens, R.D. Cebul, J.S. Schwartz, L.H. Beck, G.D. Webster, T.G. Schnabel & A. Elstein, *The Validity of an Essay Test of Clinical Judgment*, 65 ACADEMIC MEDICINE S39 (1990).



GEOFF NORMAN, PH.D., is Professor of Clinical Epidemiology and Biostatistics, McMaster University. He holds a B.Sc. in physics from the University of Manitoba, a Ph.D. in nuclear physics from McMaster University, and an M.A. in educational psychology from Michigan State University. He is the author of 10 books in education and in measurement and statistics and over 200 journal articles.