SAFETY BY DESIGN

Use of assessment to reinforce patient safety as a habit

R M Galbraith, M C Holtman, S G Clyman

Qual Saf Health Care 2006;15(Suppl 1):i30-i33. doi: 10.1136/qshc.2005.015941

The US spends far more than any other nation on health care. Physicians undergo lengthy and comprehensive training that is carefully scrutinized, and are held to high standards in national examinations. At best the care delivered matches or exceeds that in any other country. And yet, often simple preventable medical errors occur at alarming and unacceptable rates. The public, corporate consumers of health care, large payors and malpractice insurance carriers are all becoming impatient with the pace of improvement. The medical profession recognizes that dealing with this problem is an urgent priority and is grappling to find the best approaches. This paper focuses on the constructive use of assessment to embed a pervasive and proactive culture of patient safety into practice, starting with the trainee and extending out into the practice years. This strategy is based on the adage that "assessment drives curriculum" and proposes a series of new assessment tools to be added to all phases of the training-practice continuum.

SCOPE OF TRADITIONAL HIGH STAKES TESTS

Medical education has traditionally built instruction and assessment around three principles: knowledge, skills, and behaviors or attitudes. (The traditional classification specifies attitudes rather than behaviors. In that they reflect internal values and beliefs, attitudes may not be reliably amenable to inspection, but their external manifestations made explicit in the form of behaviors should be observable and perhaps modifiable.) The multiple choice questions (MCQ) upon which high stakes examinations have relied test primarily knowledge, and to some extent reasoning and judgment. However, MCQ examinations pose some difficulties for assessment related to patient safety:

- Although patient safety clearly involves a specific knowledge base, much of what is required is predicated on skills and behaviors—for example, communications, teamwork, professionalism, leadership, cultural competency—that would not necessarily be amenable to measurement by MCQ.
- MCQ are relatively low fidelity simulations that tend to measure potential to perform, rather than actual work performance. Their predictive value in the reality of patient care is unclear.
- MCQ yield test scores that are high in reliability. From the viewpoint of high stakes testing, this has the benefit of providing scores that are highly defensible. However, this has led many to avoid assessment methodologies that test skills and behaviors, or to characterize them as "squishy" simply because reliability measures cannot compete with those obtained with MCQ.

Regardless of whether the aim is to predict work performance, assure current competence, or to promote professional development in relation to patient safety, broadening the base of assessment (for example, to skills and behaviors) will most certainly imply lower reliability. This may be acceptable if in return we achieve improved patient safety.

CULTURAL DISSONANCE

The traditional culture of medicine has tended to a belief that physicians who strive continuously to perfect their knowledge of disease and of healthcare practice should never make any mistakes. Equally, medical facilities that conform to the detailed structural specifications of accrediting bodies ought to provide perfect mistake-free healthcare environments. This culture is dangerous because it does not fully recognize the

It has long been an article of faith that assessment drives education and learning, especially when the stakes are high. Perhaps the best example of this is the extraordinary care and intensity with which medical students approach the US Medical Licensing Examination (USMLE). Passing steps 1 and 2 is mandatory for promotion and graduation in most schools. Selection processes for residency, and especially those in specialties for which the earnings potential is highest, may key off step 1 scores. USMLE also creates concern among educators. Performance of *their* students and residents is under close scrutiny by their institution and its other component departments, by state legislators, and by national accreditation bodies. Students demand explicit and specific guidance as to what they need to know and do, and how to assimilate it most efficiently. Beyond the principle of independent audit entailed in a national licensing examination, the existence of USMLE as an absolute barrier to obtaining licensure makes people pay better attention to what they are teaching and/or learning.

It would be theoretically attractive if high stakes testing could be applied to patient safety, and could drive the emergence of the right proactive culture in which everyone is continually questioning how to do better. Unfortunately there are several problematic issues that complicate this ideal.

See end of article for authors' affiliations

Correspondence to:
Mr R M Galbraith,
Co-Executive Director,
Center for Innovation,
National Board of Medical
Examiners, 3750 Market
Street, Philadelphia, PA
19104, USA; rgalbraith@
nbme.org

dynamic nature of errors, which often arise at the ragged edges where constantly evolving teams and organizations meet relentlessly changing operational environments.¹⁻³

The traditional medical culture is also ineffective in dealing with errors when inevitably they do occur, in part because it lacks a sufficiently rich vocabulary to address the system level processes that are increasingly implicated when high technology breaks down. Even if individual physicians were perfect, health care is increasingly a team affair; errors are not automatically attributable to a single individual. This culture has perhaps been best exploded by the airline industry, where the inevitability of errors has been accepted and the overriding concern is now to continuously and relentlessly reduce opportunities for error. Given their culture and training, this is a difficult pill for independently minded physicians to swallow.

It is worth pointing out that pilots were also once fiercely independent, as evidenced by their widespread resistance to the introduction of gyroscopic artificial horizons in the 1920s.⁴ Many died in spiral dive crashes because they placed inappropriate trust in their natural sense of balance. Resistance to artificial horizon technology was eventually overcome by evidence based persuasion (and perhaps by demographic turnover), and it is to be hoped that evidence based medicine will eventually win over physicians who place great stock in personal clinical judgment. Pilots continue to be at risk for ego driven accidents, which is why such accidents are a major target of training along the lines of the Crew Resource Management model.^{5,6} The point is that educational processes in aviation have driven a cultural shift that marginalizes the more dangerous behaviors associated with excessive individualism, and emphasizes the communication skills and habits that air crews need to combat it. The same shifts may be expected to take place more slowly in medicine, because the environment is much more varied and the incentives for improvement may be less clearly focused for operators at the “sharp end” of systems.⁷ The constant interdisciplinary collaboration and discussion vital to identifying and rectifying problems are also a difficult sell, and fragmented government oversight and a thicket of less than transparent entanglements between providers, regulators, researchers, and payors is hardly helpful.

STRATEGIC FOCUS FOR ASSESSMENT

In its ground breaking series of publications,^{8,9} the Institute of Medicine (IOM) laid out an important inventory of risk factors contributing to medical errors, together with some key conceptual definitions. It summarized the major risk factors as:

- fatigue;
- communications failures;
- human-machine interface issues;
- systems failures; and
- culture of perfection and invincibility.

A convincing case can also be made for the importance of two other related domains—namely, increasing complexity and difficulties in its management; and the increasing reliance of effective care upon highly functional teams and systems within which the individual must function harmoniously and supportively. Furthermore, these issues interact in powerful and unpredictable ways. For example, in relation to fatigue, useful progress is being made as a consequence of the recently imposed 80 hour working week for residents, but many healthcare workers remain chronically overworked and stressed, particularly as organizations and health systems try to squeeze out cost savings by cutting back on personnel and increasing efficiency. In part, this is a consequence of a

cultural norm that still values work over sleep, and that is consonant with the culture of perfection and invincibility mentioned above. In addition, the increased emphasis on delivery of medicine as an efficient industrial product highlights the fundamental trade-off between efficiency and safety that is one of the central themes of systems failures.^{1,3,10} Communications gaps and suboptimal team functioning are further potent catalysts.

Designing effective strategies for change in these domains will clearly be challenging. While assessment may not be the most effective primary strategy for some of the risk factors, appropriate new assessment initiatives supporting salient education could play an important role.

MOVING BEYOND A PUNITIVE CULTURE

Physicians are sometimes accused of ignoring or even burying their mistakes. In fact, there is a long history of review of problems in formats such as the Morbidity and Mortality Conference. However, in seeking to find out who did what wrong, these frequently become “name and blame” sessions in which mental agility and protection of ego and reputation are sometimes more in evidence than honest efforts to identify problems and address them. The commercial airline industry has devised a system in which voluntary self-reporting of “non metal-bending” incidents provides personal learning opportunities while protecting the pilot from blame, humiliation, and other forms of punishment. Since incident data are voluminous and generally available, they are a rich source of important learning through which pilots, organizations, manufacturers, and regulators can reap greater benefit from their collective experience with rare but dangerous incidents. This model challenges us to find ways to link assessment with outcomes analysis through individual and community processes of learning.

The Aviation Safety Reporting System, one of the central critical incident reporting systems in US aviation, was in fact designed by a physician, Charles Billings, who has pointed out that medicine is a far bigger and more complex system than aviation.¹¹ Furthermore, critical incident data are characterized by many of the same kinds of “squishiness” that complicate the development of non-MCQ assessment tools. They are of limited use in developing traditional statistical indicators such as accident rates, but they are immensely valuable for identifying new clusters of problems that can then be investigated in depth using more rigorous methods.

Like critical incident reporting systems, assessments that are introduced around patient safety themes will benefit from being clearly couched in improvement rather than punishment, with plentiful feedback of the type that is usually missing from high stakes tests. They will benefit even more from incorporating growing knowledge of the ways in which real patient safety problems continue to be thorny, complex, and resistant to simple solutions. Observers of “high reliability organizations” have argued that operating safely in dangerous systems depends upon cultivating “requisite variety” in cognitive and social structures to match the complexity of the work environment.^{12,13} In the same way, future assessments for safety will have to reflect the richness and complexity of real clinical work. This will be an increasing challenge as the pace of technological change in medicine accelerates.

CLIMBING MILLER’S PYRAMID

In designing assessment supportive of patient safety, it is instructive to consider the pyramid first proposed by Miller (fig 1).¹⁴ At the lowest level, learning and assessment are initially focused around “Knows” and progressively ascends through “Knows how” and “Shows how” to “Does”. In

Figure 1 Miller's pyramid.¹⁴

general, lower fidelity simulations (e.g. MCQ and oral tests) are focused on cognitive knowledge and reasoning and map to the lower levels ("Knows" and, to a lesser extent, "Knows how". In contrast, higher fidelity simulations such as standardized patients, mechanical and/or virtual reality simulations live at the higher level of "Shows how". At the highest level of "Does", direct observation including multi-source feedback (MSF) and process and outcomes measures are more relevant. Broadening the base of assessment means progressively scaling Miller's pyramid, and the message for patient safety is that assessment should ideally be targeted at every level where it could contribute to making a difference. As one example, for communication skills, "Knows" could be tested by MCQ tests of knowledge—for example, how do you respond when a patient asks a particular question? "Knows how" could be assessed by computer simulations that provide standardized cues requiring response, while "Shows how" could be approached through the use of standardized patients depicting clinical situations—for example, breaking bad news or dealing with a disruptive patient. Finally, "Does" could be directly observed by MSF of behaviors observed in a 360 degree fashion. Some possible assessment approaches are briefly summarized.

ASSESSMENT APPROACHES AVAILABLE

Assessment approaches can be considered across the range of stakes from low to high that are attached to examination results. They can also be classified based on whether the assessment occurs in a simulated setting (so called competency based assessment) or in the actual work setting (work based assessment). Some aspects of the IOM risk factors could certainly be assessed by competency based assessment—for example, by MCQ. A lower stakes more formative approach which emphasizes assessment that is supported by feedback is potentially useful, particularly if this enables the "safe" non-punitive recognition of potential areas of deficiency and leads to tailored learning opportunities, including CME, followed by further self-assessment. This approach could in part be self-directed, but should also be steered by professional societies and specialty boards. However, in that MCQ tend to measure the potential to perform (or competence) rather than actual work performance,^{15 16} this approach is helpful and perhaps necessary but not sufficient. Other assessment modalities should be considered—both high fidelity simulations and work based assessment.

HIGH FIDELITY SIMULATIONS

Recent years have seen a profusion of new mechanical and/or virtual reality simulators incorporating haptic feedback. These can provide high fidelity testing for a variety of procedural and psychomotor skills. In relation to patient safety, this could be helpful for illustrating the importance of

human-machine interface issues, for demonstrating the effects of fatigue, and for testing communication skills. A number of simulations have also been built around testing teamwork and communications skills at the level of the team. Standardized patients provide a useful means of testing communications skills under conditions that are more realistic, although still not entirely authentic. They can also be used to assess cultural sensitivity and related aspects of communications skills.

Simulations have three major benefits over learning with real patients: (1) they reduce the risk of injury to the patient and the risk of psychological trauma to the physician who is learning a difficult and risky procedure; (2) they allow the individual and/or team to gain experience with relatively rare situations much faster than is likely in actual practice; and (3) they enable, at least in principle, a better balance between standardized measurement and fidelity in simulations that provide a more realistic representation of medical practice as it actually occurs, and thus greater confidence that the skills assessed have direct relevance to safety in practice.

The challenge lies in developing medical simulation technologies that have the right kinds of realism for the competencies that we want to measure. The obvious route to high fidelity simulations is through high technology, but there is also a role for experimenting with lower technology extensions—for example, in structured educational programs using deliberate practice or standardized patients combined with mannequins.^{17 18} These may come closer to allowing the measurement of actual work performance rather than potential to perform. However, there is a large gap in our knowledge around scoring performance on complex tasks in high fidelity simulations, and also in our knowledge about how clinicians think "in the wild".^{3 19–21} The future development of simulation to its full potential will depend on closing those gaps.

WORK BASED ASSESSMENT

There is increasing pressure to develop assessments that relate to what physicians actually do with their patients (that is, work performance) in the particular team, system, and environment in which they function. This is as opposed to competence, or whether or not they know what to do in "observed" simulated settings of variable realism.^{15 16} Measurement of work performance—processes and outcomes—in a real setting has long been considered to be simply too difficult to achieve within the scope of current testing approaches. Relevant measures and the resulting data sets are by definition heterogeneous and likely to be specific to the individual physician's practice, and measurement is ideally repetitive or even continuous. This is very different from the standardized episodic approach used for high stakes testing, and the inferences that can be drawn are likely to be correspondingly less robust. Work based assessment has recently been mandated legislatively in the UK, and there is growing interest in developing such approaches in the US. It is important to emphasize that such assessment should speak to all of the risk factors identified above, and to others that become evident in the future. Although work based assessment is at present rudimentary and the extent to which this is possible must remain conjectural, the current state of play with some methodologies of potential relevance is summarized below.

Multi-source feedback

One promising approach is the use of multi-source feedback (MSF), often termed 360° surveys, with continuing observation, frequent feedback and, if appropriate, mentoring and remediation.^{22–24} Even though MSF is widely used in corporate America, this methodology has not yet been fully

embraced for assessing physicians in the US. It is certainly true that MSF is unlikely to attain the impressive levels of reliability typical of highly standardized MCQ examinations. On the other hand, contemporary thinking is shifting towards the possibly positive benefits of more formative use, with plentiful feedback and lower stakes. This approach could certainly be suitable for addressing communications and the related domains of teamwork and leadership, systems failures, and the culture of practice. MSF has strong potential to contribute to cultural change for two reasons. Firstly, it sends a signal that what is being assessed is an important focus for education and, secondly, it allows us to perform a non-invasive biopsy of the traditionally closed off domain of informal social control in clinical training. When MSF is conducted properly (ensuring appropriate protections for the participants), it can provide a non-threatening view into the relationship between individual behavior and larger cultural pathologies. For example, MSF can make explicit how the modeling of inappropriate behavior by senior clinicians contributes to the perpetuation of a "hidden curriculum"²⁵ in clinical training that supports the general diffusion of unsafe practices.

Process and outcome measures

There are many identifiable processes and outcomes that have a direct bearing on patient safety issues,²⁶ and many of these are amenable to measurement. These range from potentially complex measures (such as postoperative morbidity and mortality) to relatively straightforward measures (such as rate of intravenous line infections or wrong sided surgery). There has been great reluctance to release such data, based on fears about the accuracy of the data and/or the consequences of its publication. Such concerns often revolve around malpractice exposure or negative impact on pay-for-performance measures. However, if the intent is to gather data that inform quality improvement and fears of inappropriate punitive consequences can be adequately addressed, "showing the data" can be a powerful propellant for change as well as a key measure of its success. The experiences of the Northern New England Cardiovascular Group in relation to mortality rates of cardiovascular bypass grafting²⁷ provide an especially powerful example of this.

CONCLUSION

Medicine is beginning to change as a proactive culture of patient safety becomes more widely accepted as a central value in healthcare delivery, and this trend must be actively encouraged. Education and assessment together share an important role in emphasizing the importance of patient safety, and assessment in particular is an important potential lever. The traditional assessment approach involves a focus on high stakes competency assessment with highly reliable standardized tests of knowledge in which accountability is applied externally by regulatory bodies. The resulting culture offers an alluring but ultimately false assurance that we have weeded out the "bad apples", selected for perfection, and that patient safety is now built in. That a punitive regulatory approach is suboptimal can be seen in the suboptimal reporting to and use of, for example, the National Practitioner Data Bank.²⁸

On the other hand, there is now a growing emphasis on lower stakes more formative assessment in which the gathering of real work performance data in a non-punitive fashion could provide practice-specific data that are currently unavailable. Such work based assessment—including

measurement of skills and behaviors and practice process and outcome measures—could inform and enable self-directed improvement. It is also important to stress that the locus of accountability is more likely to be internal, and the indications are that this might be effective in promoting improved patient safety as a habit.

Authors' affiliations

R M Galbraith, M C Holtman, S G Clyman, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA, USA

Funding: none.

Competing interests: none declared.

REFERENCES

- 1 Rasmussen J, Pejtersen AM, Goodstein LP. *Cognitive systems engineering*. New York: John Wiley & Sons, 1994:146–8.
- 2 Snook SA. *Friendly fire: the accidental shootdown of US Blackhawks over Northern Iraq*. Princeton: Princeton University Press, 2000:179–201.
- 3 Perrow C. *Normal accidents*. New York: Basic Books, 1984:179–83, 318, 341.
- 4 Langewiesche W. The turn. *The Atlantic Monthly* 1993;272:115–22.
- 5 Helmreich RL, Meritt AC, Wilhelm JA. The evolution of crew resource management training in commercial aviation. *Int J Aviat Psychol* 1999;9:19–32.
- 6 Helmreich RL. On error management: lessons from aviation. *BMJ* 2000;320:781–5.
- 7 Reason J. Human error: models and management. *BMJ* 2000;320:768–70.
- 8 Kohn LT, Corrigan JM, Donaldson M. *To err is human: building a safer health system*. Washington, DC: Institute of Medicine, 1999:49–68.
- 9 Institute of Medicine. *Crossing the quality chasm: a new health system for the 21st century*. Washington: National Academy Press, 2001:39–60.
- 10 Rochlin GI, LaPorte TR, Roberts KH. The self-designing high-reliability organization: aircraft carrier flight operations at sea. *Naval War College Review* 1987;40:76–91.
- 11 Cook RI, Woods DD, Miller C. *A tale of two stories: contrasting views on patient safety*. Report from a Workshop on Assembling the Scientific Basis for Progress on Patient Safety. Chicago: National Health Care Safety Council of the National Patient Safety Foundation at the AMA, 1998:52–61.
- 12 Weick KE. Organization culture as a source of high reliability. *Calif Manage Rev* 1987;28:112–27.
- 13 Weick KE, Sutcliffe KM, Obstfeld D. Organizing for high reliability: processes of collective mindfulness. *Res Organ Behav* 1999;21:81–123.
- 14 Miller G. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65(Suppl):S63–7.
- 15 Rethans JJ, Norcini JJ, Baron-Maldonado M, et al. The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;36:901–9.
- 16 Norcini, JJ. Current perspectives in assessment: the assessment of performance at work. *Med Educ* 2005;39:880–9.
- 17 Kneebone R. Simulation in surgical training: educational issues and practical applications. *Med Educ* 2003;37:267–77.
- 18 Issenberg SB, McGaghie WC, Gordon DL, et al. Effectiveness of a cardiology review course for internal medicine residents using simulation technology and deliberate practice. *Teach Learn Med* 2002;14:223–8.
- 19 Hutchins E. How a cockpit remembers its speed. *Cognitive Sci* 1995;19:265–88.
- 20 Hutchins E. *Cognition in the wild*. Cambridge, Massachusetts: MIT Press, 1995, xvi–xviii.
- 21 Klein G. *Sources of power: how people make decisions*. Cambridge, Massachusetts: MIT Press, 1998:2–6.
- 22 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005;330:1251–3.
- 23 Lockyer J. Multisource feedback in the assessment of physician competencies. *J Cont Educ Health Prof* 2003;23:4–12.
- 24 Ramsey PG, Wenrich MD, Carline JD, et al. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269:1655–60.
- 25 Hafferty FW, Franks R. The hidden curriculum, ethics teaching, and the structure of medical education. *Acad Med* 1994;69:861–71.
- 26 Donabedian A. Evaluating the quality of medical care. 1966. *Milbank Q* 2005;83:691–729.
- 27 O'Connor GT, Plume SK, Olmstead EM, et al. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. *JAMA* 1996;275:841–6.
- 28 Waters TM, Warnecke RB, Parsons J, et al. The role of the National Practitioner Data Bank in the credentialing process. *Am J Med Qual* 2006;21:30–9.