Medical Education

Student assessment: Issues and dilemmas regarding objectivity

TEJINDER SINGH

I recently purchased a laptop. The manufacturer claimed that its battery time was over 8 hours. However, when I started using the laptop, the battery never lasted that long. I called the customer care helpline. They told me that the figure of 8 hours was arrived at by using a very advanced and standardized software, which estimates the battery time under 'standard' conditions (for the uninitiated, this means putting the machine on at its lowest brightness and then not using it, except for low-end applications such as word processing). Now that was a problem. I hate carrying the chargers in my handbag. How do I know how long the battery will last under actual work conditions? So I started using the laptop as I would normally do, i.e. for word processing, making slides, connecting to the Internet, listening to music and occasionally watching movies. After about a week, I thought 5 hours was a fair estimate. Just to be sure, I also requested my son to use it for a week (you guessed it, for gaming), and he also thought 4–5 hours was a good estimate. Now when I travel, I do not carry my charger along if I estimate my computer use to be less than 4 hours.

This incident got me thinking about the assessment of medical students. We are fond of objective and standardized tests, which are administered under standard test-taking conditions and in which the students are awarded certain grades. However, what happens when these doctors face a real-life situation? Are we incorrectly estimating the competences of our students in a controlled environment? Whether it is estimating the time of a laptop battery, the mileage of a new car or the competence of students, the issue seems to be the same—one-shot observation using standardized tools in artificial settings or long-term observation in real-life situations.

CHANGING PARADIGMS

We have come a long way with regard to the assessment of students. During the *gurukul* era, the guru had the prerogative of deciding if a student was fit enough to leave the *gurukul* and face the world. The sociocultural changes that accompanied the ensuing Mughal and British eras resulted in a drastic change in the role of the teacher.[1] Further, a need was felt to have external agencies show to society that the students had reached a certain level of educational attainment. There was a parallel change in the form of establishment of boards and universities to address the area of assessment of students and their certification as fit/unfit. What followed was an era of objective and standardized tests. The role of the teacher as a resource to assess students continuously and provide feedback on learning diminished progressively. We became so obsessed with objectivity that anything which could not be objectively measured was (is) not considered worthy of assessment. The wheel of time has turned a full circle and brought us back to a point where this deficiency is characterizing our system again. Globally, educational psychologists, assessment experts and teachers are going back[2] to emphasizing the role of expert opinion—even though subjective—on performance. However, Indian medical education seems to be ill prepared to accept this change.

Let us take the example of internal assessment (IA). Educationally, IA provides some of the best opportunities for assessing skills and competencies which cannot be assessed by traditional examinations.[3] It provides wonderful opportunities for giving the students feedback, and these have been shown to be the best input for improving their performance. We hoped that after its rechristening in 1997 by the Medical Council of India,[4] IA would make a meaningful difference to the way medicine is taught and learnt. Unfortunately, this has not happened. The failure of IA can be attributed partly to the lack of awareness among and training of teachers. However, the major reason for which teachers at large have failed to accept it appears to be the subjectivity involved in the process.[5] The myth that objective is reliable and subjective is unreliable seems to have invaded our psyche and is preventing us from making use of a simple, effective and useful intervention. Before further discussing the need to move away from this self-imposed restriction, we should be clear on the facts.

WHAT IS OBJECTIVE ASSESSMENT?

Objective assessment refers to assessment within a restricted domain through the use of methods such as multiple-choice questions (MCQs), objective structured clinical examinations (OSCEs) and patient management problems (PMPs) at a lower level of simulation.[6] Generally, objective assessments use a *norm-referenced* approach with no specified criteria (although some cut-off like 50% may be used). The advantage of these methods is that large domains of knowledge can be sampled within a small time-frame. Subjective assessment, on the other hand, could mean expert assessment in which performance is rated at a higher level of simulation, as in an extended period of supervised practice. The comparison of students' performance is generally with a predetermined set of criteria and, therefore, the approach is *criterion-referenced.*[6] To put it in an oversimplified manner, it can be said that objective assessment means nothing more than everyone marking students the same way.

Most objective tests of knowledge or performance make use of well-structured problems. This is in contrast to real-life scenarios, in which most problems are poorly structured.[7] In these circumstances, identifying the problem and generating hypotheses become as important as finding the solution to the problem. Variability is an inseparable part of the clinical process. The patient, the illness, the context and the student all contribute to variability.[8] However, objective tests are not suitable for helping the student deal with the variability of clinical practice.

WHAT IS RELIABILITY (AND WHAT IT IS NOT)?

It is commonly believed that objective assessments are the most reliable.[5] How much truth is there in this belief? Reliability is an

Department of Paediatrics, Christian Medical College, Ludhiana 141008, Punjab, India; *drtejinder22@gmail.com*

important attribute of assessment and, to me, it seems to have been the most misunderstood one. Traditionally, reliability has been viewed as *consistency of results* or *the same results under the same conditions*.[9] These views of reliability have their own flaws. First, they focus on *consistency of marking* and not *consistency of performance*. Second, it is not possible for a doctor to always encounter the same or a similar patient throughout her/his practice. The presentation, clinical findings or course of illness, even for the same disease, will always be variable. These definitions may therefore be appropriate for a biochemical test, but not for educational testing. Educational testing involves a certain degree of prediction—will the student who is performing well at this test or at this time be able to perform similarly on a different patient or at a different time? Reliability refers to the interpretations we make from assessment data and is not the inherent property of a tool.

Reliability is now considered a part of the validity argument, in which validity refers to the accuracy of measurement.[10] To be valid, an assessment has to be reliable, though in real life, there is often a trade-off between validity and reliability. In any given situation, the accuracy of measurement (validity) is more important than its precision (reliability). Kane's conceptualization[11] of validity reinforces the importance of adequate and representative sampling (generalization) as one of the four validity arguments. Kane has tried to put reliability in its proper perspective by identifying different dimensions, i.e. scoring (that is what we traditionally look for), extrapolation (representativeness of the test items) and generalizability (adequacy of the test items). This perspective shifts the focus of reliability from the individual test to the reliability of the assessment programme.[12] Reliability implies the degree of confidence that we can place in our results (try reading it as 'rely-ability').

Marker variability is *not* a major reason for low reliability, though it may appear to be so. The biggest threat to reliability comes from context-specificity.[13] Physicians are known to deal differently with different cases. For example, if a physician is able to deal satisfactorily with a disorder of the central nervous system, it is no guarantee that she/he will be able to do the same with a case of anaemia or pneumonia. It is this case-specificity which makes things difficult for us. Increasing the *size* and *representativeness* of the sample of tasks to be assessed seems the best approach to building (validity and) reliability in assessment.

The importance of an adequate and representative sample to ensure validity and reliability can never be overemphasized. For example, if the test paper for the final professional examination in medicine contained only five MCQs, the results would be highly objective but not a reliable (and valid) measure of the students' knowledge. Similarly, the results would again not be reliable if 80 of, say, 100 MCQs were from only one system.

OSCE AS AN EXAMPLE

OSCE makes use of checklists, which are marked by the observer(s). It is presumed that everyone will mark the checklists similarly, resulting in highly objective assessment. Let us consider this system from the point of view that any test of clinical competence should be able to assess the students' level of expertise and distinguish differences in the levels of expertise of different students. There is evidence that experts score low on OSCE checklists because they are able to make diagnostic and therapeutic decisions with fewer steps.[14] Experts gather information and organize knowledge differently from novices. This creates a divergence between what experts do and what students should be taught to do at an OSCE. There are measurable differences between levels of expertise that checklists fail to capture.[15] It would probably be better to have an expert observer passing global judgements on performance rather than to use checklists, howsoever elaborate these ratings may be. 'Subjective' expert assessment of performance through global rating scales has been reported to be highly reproducible.[16]

The reported figures of reliability tell some interesting tales (Table I).[17] The reported reliability of an hour-long OSCE is 0.47, compared with 0.60 for a long case of equal duration. Another tool, mini-CEX, has been reported to have a reliability of 0.73. Compared with OSCE, long case and mini-CEX are branded as highly subjective, but these end up being more reliable due to the comprehensiveness of the tasks assessed. What is more interesting is that the reliability of all tools increases with an increase in the testing time. The best inference one could draw from these data is that the testing time and sampling have a greater effect on reliability than the objectivity of a tool.

While high reliability (e.g. >0.85 for high-stake examinations) is important, it is difficult to attain it within a 2- or 3-hour test. Many of us are especially fond of using the United States Medical Licensure Examination (USMLE) as an example, but we forget to keep its logistics in mind. The USMLE step 1, for example, has an 8-hour test. Step 2 CK is 9 hours, step 2 CS is 8 hours, while step 3 is 16 hours,[18] making for a total of over 41 hours. Many medical colleges, in the name of objectivity, use 20–25 MCQs or 5–8 OSCE stations of 3–5 minutes each, making a mockery of the entire concept.

EXPERIENCE WITH SUBJECTIVE RATINGS

Subjective evaluation has the advantages of being flexible and involving much less cost, time and effort. It is also suitable for the assessment of domains not amenable to measurement by objective methods. Subjective global ratings for OSCE by experts have been shown to have reliability comparable to objective evaluation. The reliability of a 5-point rating scale for communication skills has been reported to be higher than a 17-item checklist.[19] Another study comparing the reliability of 5-point global ratings with a 25-item checklist reported that the reliability of the former was acceptable.[20] Checklists give the appearance of objectivity by measuring thoroughness, an element which is easy to measure but is a poor

TABLE I. Reliability as a function of testing time

Testing time (hours)	MCQ	Case-based essays	PMP	Oral examination	Long case	OSCE	Mini-CEX	Video assessment	*In cognito* SP
1	0.62	0.68	0.36	0.50	0.60	0.47	0.73	0.62	0.61
2	0.76	0.73	0.53	0.69	0.75	0.64	0.84	0.76	0.76
4	0.93	0.84	0.69	0.82	0.86	0.78	0.92	0.93	0.82
8	0.93	0.82	0.82	0.90	0.88	0.96	0.93	0.93	0.86

MCQ multiple choice questions PMP patient management problems OSCE objective structured clinical examination
mini-CEX mini-clinical evaluation exercise SP standardized patient *Source*: Van der Vleuten (2006)[17]

surrogate to many qualities that we want to develop in medical students. In experimental situations, clinicians have been shown to collect only about two-thirds of the data available in a case[21] without the accuracy of the diagnosis being compromised due to incomplete data collection.[22] Even untrained patients have been shown to provide reliable opinions on the clinical skills of students.[23]

The utility of subjective ratings has been demonstrated in other fields as well. For example, the review of manuscripts submitted to the *Journal of the American Medical Association (JAMA)* was rated on a 5-point scale and the assessment correlated very well with the reviewer's ability to report flaws in the manuscript.[24] Subjective expert assessment is considered superior for domains in which art and science are interwoven.[25]

Objective measures tend to break the skill into smaller and smaller components, the presumption being that the total of all the sub-parts will be equal to the whole. This presumption is as unfounded as the supposition that if everyone uses a similar amount of ingredients, they will bake a cake which tastes the same.

COMPLETELY OBJECTIVE?

There is no assessment that is 'completely' objective. All assessments are coloured by the values, thought processes, experiences and expectations of the assessor(s). If this were not so, all entrance examinations would have had similar sets of questions and all OSCEs would have had similar checklists. However, that does not happen. What really happens is that we decide on the test format and items subjectively and then try to measure them objectively. This process has been called 'objectification' and does not necessarily result in higher reliability than subjective assessments.[26] Rather, in the process, we tend to discard important aspects which are not objectively measurable and this takes a toll on the validity of the assessment. In India, communication, professionalism and ethics are not assessed, despite their perceived importance, simply because they cannot be measured objectively.

A recent paper[2] has cautioned against objectifying competency-based assessment. Clinical competence is inseparable from variability and heterogeneity. A rigid objective assessment stands to lose its validity in such a situation. Another commentary[27] has very aptly summarized the situation as 'the lack of clarity about the purpose of assessment at the implementation level, fuelled by an *incessant effort to objectify* assessment data and a *misconception that judgement-based assessments are inferior* in validity and reliability is, at least in part, responsible for what might be described variously as "reductionist", "deconstruc-tive", "tick-box", "mechanistic" or "instrumentalist" approaches to assessment' (emphasis added).

EDUCATIONAL IMPACT

One has to be careful to monitor the unintended 'side-effects' of assessment on learning, e.g. postgraduate entrance examinations have killed internship in India. The use of checklists in OSCE has been shown to promote undesirable learning. Students tend to memorize the checklists rather than learn the skill.[28] Similarly, students reported that they would focus more on comprehension when appearing for open-ended tests.[29]

We have argued that objectivity and reliability are not synonymous. Whereas objectivity is a *measurement issue*, reliability is a *decision-making issue*. Subjective judgements, especially those coming from experts and after a period of prolonged observation, are more than or at least as reliable as snapshot objective assessments.[30] Similarly, subjectivity and bias are not synonymous. I often give subjective ratings to my students without being biased for or against them. On the other hand, if I want to be biased, nothing will prevent me from putting a tick against each item of an OSCE checklist, irrespective of what the student does.

IMPLICATIONS

Over the past two decades or so, our understanding of assessment (and learning) has undergone a sea change. The process of learning is as important as learning. As knowledge gets dated very fast, it is all the more important that the process of learning should be sound. We have also learnt that one-shot observations are unlikely to tell us much about learning; instead, we need to look for evidence of learning.[31] The most important shift in our understanding has been with respect to the assessment of soft learning skills, which do not easily lend themselves to objective assessment.[32] They are important because of two reasons: (i) success in the workplace depends on them,[33] and (ii) when things go wrong in practice, it can often be attributed to the lack of these skills.[34] None of these skills can be assessed objectively, yet we can ill afford to ignore them in our assessments. We have to move out of the laboratory into real life to get a true picture of clinical learning. Norcini[35] observes: 'The venue for assessment has moved from the relatively controlled and homogeneous settings in education to the uncontrolled and heterogeneous world of work.'

Subjective ratings, though low on consistency on any given task, show higher consistency across tasks than do objective ratings. In fact, the low inter-item correlations of subjective ratings negate the common feeling that a halo effect can distort ratings.[6] Subjective ratings are easily usable and the cost involved is much less than that involved in objective assessment. No other discipline stands to gain as much from subjective expert judgements as medicine.

We know that the existing assessment tools have been developed with a lot of psychometric rigour, objectification and standardization. In comparison, the emerging tools are unstructured, heterogeneous and subjective by nature, as well as less standardized. The challenge is to build that rigour into subjective assessment. Cassidy[30] has rightly said that 'valid subjectivity' (based on long-term observation in a climate of trust and reciprocity) is as reliable as any objectified method.

This is not a plea for discarding objectivity, nor an effort to push subjectivity. Both have their own place. Objectivity may be more desirable for selection tests, but subjective expert judgements are more appropriate for formative purposes and that includes most examination situations we are involved in as teachers. Norman *et al.*[36] point out that '…it is clear that the choice of a test format—written or performance—cannot be made on the basis of an unconditional appeal to objectivity. Objectivity, in an empirical sense, does not necessarily result from the strategies of objectification, and the application of these strategies may have undesirable consequences. Decisions must be made as a result of careful consideration of other issues resulting from the purpose of the testing situation—practicality, educational impact, acceptability—rather than on the dogma that objective methods, like Orwell's four-legged animals, are inherently superior.'

Feldt and Brennan[37] have said that 'quantification of consistency in examinee *performance* (rather than scoring) constitutes the essence of reliability analysis'. This consistency in performance will come only with long-term observation of performance in multiple contexts, in which multiple tools are used and multiple examiners involved. Further, the climate should be one of trust and reciprocity. The onus lies on teachers who are teaching the subject. Term-end and summative examinations are simply unable to quantify consistency in performance.

A recent article[38] points out that human judgement is indispensable in any assessment programme. While human judgement may be less accurate than number-based decisions, what is surprising is how people still manage to do a good job when faced with ill-defined problems, insufficient information and situations which are less than ideal.

THE RIGHT PERSPECTIVE

Lest you start thinking otherwise, I must state that an OSCE which is designed and used properly can provide a wealth of information on the clinical competence of students.[39] If we use OSCE to widely sample and assess clinical skills, we gain a lot; however, if we use OSCE because it is objective and fashionable to do so, then we are missing the point. The thrust of the discussion above is not that OSCE is not useful. The aim was merely to make a comparison between *objectivity* and *considered subjectivity*.

It must also be kept in mind that reliability is not an inherent quality of OSCE. Rather, it depends on the inferences we make from the OSCE results. Tools are never good or bad; it is their use which is.

And yes, I am happy that though subjective, my assessment of the battery time of my laptop is more 'rely-able' than that of advanced standardized software!

REFERENCES

1 Kaul R. Whither equity? *Seminar* 2000;**494:**23–5.

2 Lurie SJ. History and practice of competency-based assessment. *Med Educ* 2012;**46:**49–57.

3 Singh T, Anshu. Internal assessment revisited. *Natl Med J India* 2009;**22:**82–4.

4 Medical Council of India. Regulations on graduate medical education, 1997. Available at *http://mciindia.org/RulesandRegulations/GraduateMedicalEducationRegulations1997.aspx* (accessed on 25 Dec 2011).

5 Tongia SK. MCI internal assessment system in undergraduate medical education. *Natl Med J India* 2010;**23:**46–7.

6 Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. *Med Educ* 1987;**21:** 477–81.

7 Epstein RM. Assessment in medical education. *N Engl J Med* 2007;**356:**387–96.

8 Cox K. No Oscar for OSCA. *Med Educ* 1990;**24:**540–5.

9 Anonymous. Reliability in assessment. Available at *http://www.education.com/reference/article/reliability-assessments/* (accessed on 20 Dec 2011).

10 Downing SM. Reliability: On the reproducibility of assessment data. *Med Educ* 2004;**38:**1006–12.

11 Kane M. Current concerns in validity theory. *J Educ Measurement* 2001;**38:** 319–42.

12 Schuwirth LW, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012;**46:**38–48.

13 Neufeld VR, Norman GR. *Assessing clinical competence. 1st ed.* New York: Springer; 1985.

14 Chumley HS. What does an OSCE checklist measure? *Fam Med* 2008;**40:** 589–91.

15 Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;**74:**1129–34.

16 Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. *Med Educ* 1987;**21:** 477–81.

17 van der Vleuten CPM. Life beyond OSCE. Available at *http://www.fdg.unimaas.nl/educ/cees/wba* (accessed on 30 May 2010).

18 USMLE. *Bulletin of information 2011*. Available at *http://usmle.org/pdfs/bulletin.pdf* (accessed on 26 Dec 2011).

19 Cohen R, Rothman AI, Poldre P, Ross J. Validity and generalizability of global ratings in an objective structured clinical examination. *Acad Med* 1991;**66:**545–8.

20 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:**1053–5.

21 Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: An analysis of clinical reasoning. In: Bender RJW (ed). *Teaching and assessing clinical competence.* Groningen: Boekverk Publishers; 1978.

22 Neufeld VR, Norman GR, Feightner JW, Barrows HS. Clinical problem-solving by medical students: A cross-sectional and longitudinal analysis. *Med Educ* 1981;**15:** 315–22.

23 Wilkinson TJ, Fontaine S. Patients' global ratings of student competence: Unreliable contamination or gold standard? *Med Educ* 2002;**36:**1117–21.

24 Callaham ML, Baxt WG, Waeckerle JF, Wears RL. Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *JAMA* 1998;**280:**229–31.

25 Eisner EW. *The educational imagination*. New York:Macmillan; 1979.

26 Van der Vleuten CP, Norman GR, de Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Med Educ* 1991;**25:**110–18.

27 Amin Z. Purposeful assessment. *Med Educ* 2012;**46:**4–7.

28 Rooney PJ, Allen SW, Dodd P, Norman GR, Powles ACP, Rosenfeld J, *et al.* Can objective assessment of clinical skills be contained within the small group setting? In: *Research in medical education, 1989: Proceedings of the twenty-eighth annual conference Association of American Medical Colleges.* Washington DC:Section for Student and Educational Programs Association of American Medical Colleges; 1989:1–273.

29 Stalenhoef-Halling BF, van der Vleuten CPM, Jaspers TAM, Fiolewt JFBM. The feasibility, acceptability and reliability of open-ended questions in problem-based learning curriculum. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwiestra RP (eds). *Teaching and assessing clinical competence.* Groningen:Boekwerk Publishers; 1990.

30 Cassidy S. Subjectivity and the valid assessment of pre-registration student nurse clinical learning outcomes: Implications for mentors. *Nurse Educ Today* 2009;**29:**33–9.

31 van der Vleuten CP, Schuwirth LW. Assessing professional competence: From methods to programmes. *Med Educ* 2005;**39:**309–17.

32 Norman G. Non-cognitive factors in health sciences education: From the clinic floor to the cutting room floor. *Adv Health Sci Educ Theory Pract* 2010;**15:**1–8.

33 Papadakis MA, Teherani A, Banach MA, Knettler TR, Rattner SL, Stern DT, *et al.* Disciplinary action by medical boards and prior behavior in medical school. *N Engl J Med* 2005;**353:**2673–82.

34 Papadakis MA, Hodgson CS, Teherani A, Kohatsu ND. Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. *Acad Med* 2004;**79:**244–9.

35 Norcini JJ. Current perspectives in assessment: The assessment of performance at work. *Med Educ* 2005;**39:**880–9.

36 Norman GR, Van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Med Educ* 1991;**25:**119–26.

37 Feldt LS, Brennan RL. Reliability. In: Linn RL (ed). *Educational measurement* (3rd ed). New York:Macmillan; 1989:105–46.

38 Schuwirth LWT, Van der Vleuten CPM. An overview of assessment in medical education: General overview of the theories used in assessment. *Med Teach* 2011;**33:**787–97.

39 Gupta P, Dewan P, Singh T. Objective structured clinical examination (OSCE) revisited. *Indian Pediatr* 2010;**47:**911–20.