From Structured, Standardized Assessment to Unstructured Assessment in the Workplace

Lambert Schuwirth, MD, PhD

Real-time patient evaluation of physicians' communication and professional skills is a wonderful idea. Wasn't it Aristotle who said that the guests will be a better judge of the feast than the cook? But if such evaluations are so logical, why aren't patient evaluations an integral part of all assessment programs?

While reading the article by Dine et al[1] in the current issue of the *Journal of Graduate Medical Education*, I was reminded of a silly question I asked during my younger years in medical education: "Why do we insist on structuring and standardizing our assessments, and training our faculty to assess the competence of students to perform in unstandardized contexts for untrained bosses and with untrained patients?" My colleagues replied that tests must be fair and equitable, and therefore need to be structured and standardized to produce reproducible results. Of course I agreed with them because how can one disagree with the notion that tests must be fair?

Dine et al,[1] who study patient assessment of internal medicine resident skills, acknowledge that is it not possible to obtain standardization or structuring in all facets of an assessment program, yet these very aspects are important. As often occurs when patient assessments are used, no attempts were made to train the patients providing the assessments; instead their "raw" information was collected and used. There is good value in this approach; after all, it is patients who decide whether they trust and respect their physician, and it is patients who decide on physicians' communication skills and whether they are treated professionally and with respect.

If an assessment cannot be standardized or structured, how can we ensure that the assessment is of high quality and fair? This question is critical because, generally, our approach to assessment quality is based on the notion of a construct. In this context, a construct is a human characteristic that cannot be observed directly, but has to be inferred from things we can observe. A typical example of a construct in medicine is blood pressure. It cannot be observed directly but rather is inferred from reading the sphygmomanometer while lowering the cuff pressure and listening to the Korotkow tones with a stethoscope. In medical education assessment, we use constructs such as *knowledge*, *skills*, and *professionalism*.

In their seminal 1955 article, Cronbach and Meehl[2] shaped our thinking on construct validation. Part of their view is that an individual item of a test is not valid, per se; it is the total scores emanating from the aggregation of the performances on all items that make the test valid. With multiple-choice questions this concept is easy to understand: a single item does not tell us much about a candidate's competence. However, it is plausible that a larger set of item responses will provide more accurate information. This is why standardized tests typically contain large numbers of items. This principle is adhered to in many tests, and it is why in an objective structured clinical examination (OSCE) we add the performance on a chest examination station to those on resuscitation and communication stations to form a total score for "skills," despite the counterintuitive nature of such an approach.

Yet, collecting assessment information from patients, as performed by Dine et al,[1] is fundamentally different from the assessment modalities described above. One may question whether a construct-based approach is the most appropriate in this context or whether we need different approaches to determining the quality of patient-based assessments.

The first point to consider when comparing patient-based assessment to the more established assessments described above concerns a fundamental assumption in the construct approach. During a multiple-choice test, and even during an OSCE, it is a reasonable and necessary assumption that the object of measurement (ie, the student or resident) does not change. This is an important assumption because it allows us to assume that any variance between items or observations is most likely due to either systematic item (station) effects or due to interaction effects. But in the course of a long learning period it is not a reasonable assumption that the trainee, the object of measurement, remains unchanged. In fact, that is exactly what we are after: The student learns, gains competence, and is doing better in the next attempt than

Lambert Schuwirth, MD, PhD, is Professor of Medical Education, Flinders Innovation in Clinical Education, Health Professions Education, School of Medicine, Flinders University, Adelaide, South Australia.

Corresponding author: Lambert Schuwirth, MD, PhD, Flinders University, Sturt Road, Bedford Park, South Australia 5042, GPO Box 2100, Adelaide SA 5001, +61-8-82047174, lambert.schuwirth@flinders.edu.au

DOI: http://dx.doi.org/10.4300/JGME-D-13-00416.1

the previous one. Without this stability in the object of measurement, however, it is more difficult to disentangle all the sources of variance. If trainees perform better on a next occasion is that because they improved or because the case was easier?

In a slightly different way, the question of stability is relevant to the interns in the article from Dine and colleagues.[1] The interactions between an intern and a patient may all have occurred at different points during the intern's learning periods. Even if these were taken into account, which is statistically possible, the researcher would still have to assume that learning takes place exactly in the same way in all trainees, and that is not a very plausible assumption.

The second point relates to where we draw the line in accepting aggregation of information. It reminds me of the joke about the 3 statisticians on a bear hunt. The first aims and shoots 1 meter too far to the right, the second shoots 1 meter too far to the left, and the third averages the 2 shots, and shouts, "We've got him!" The lesson of this joke is that aggregating or averaging data from different events does not always make sense. The fundamental difference between the multiple-choice test and OSCE is that it is somewhat defensible to argue that poor performance on 1 item or station can be compensated for with good performance on the other. But in the assessment of constructs, such as empathy and professionalism, this is not realistic. Suppose a trainee says to a patient: "Please don't worry about my poor communication and empathy with you. I will do an excellent job with the next patient." Of course that would not reassure the patient as the communication and empathy shown during his or her consultation is the only relevant event; it cannot be compensated for by good communication with other patients.

Thus the question that needs to be answered is whether it is defensible or sufficiently meaningful to take an average of all patient opinions to form a pass-fail decision? Large numbers of data points cannot solve every assessment problem.

This leads to the third point, which is related to the nature of the construct of professionalism or communication. The exact way patient consultations will unfold is unpredictable beforehand; every utterance or action from the patient or physician leads to a reaction from the other. So it is nearly impossible to predict who will say what at any given time in the consultation or even which pathway the communication will follow. Good communication and professionalism skills require not only professional values but also the ability to translate those values into good behavior—because this behavior is what the patient judges. To be a good communicator, the intern has to have a broad range of communication modes or behaviors to effectively express his or her professional values and intentions. In addition, the intern has to have the flexibility to choose and change between those behaviors to optimally cater to every patient's individual expectations. This is an aspect of the granularity of the measurement—in this case the granularity needs to be very fine—and it could be argued that a single measurement event with a single instrument will likely be invalid and not generalizable.

With these considerations in mind, I suggest that the lack of reliability and validity of the instrument examined in the study by Dine and colleagues[1] is not an innate feature of the instrument itself but rather a reflection of the way it was used. In this case, the instrument was used as a single testing instrument with the main intent of distinguishing between competent and incompetent interns—and it may not be suited for this purpose. If, on the other hand, the instrument was used frequently and longitudinally with feedback to trainees it could be more useful. The summative aspect of the tool would then be how this feedback is used to define concrete learning goals and the extent to which these were then attained by each intern. If applied in this way, the instrument could lift the professionalism and communication skills of all interns while at the same time identifying residents who fail to respond to efforts to remediate their communication, interpersonal skills, or professionalism problems. This would constitute a more valid and generalizable use of the instrument.

These suggestions are based on current views of workplace-based assessment in which there is growing agreement that the user is more important than the instrument.[3,4] In other words, for observation in the workplace, the users—in this case, patients, supervisors, and interns—are the key elements in validity and generalizability. They should be the target of quality assurance and improvement rather than the psychometric properties of the instrument.

References

1 Dine CJ, Ruffolo S, Lapin J, Shea JA, Kogan JR. Feasibility and validation of real-time patient evaluations of internal medicine interns' communication and professionalism skills. *J Grad Med Educ.* 2014;6(1):71–77.

2 Cronbach L, Meehl P. Construct validity in psychological tests. *Psychol Bull.* 1955;52(4):281–302.

3 Govaerts MJ, Schuwirth LW, Van der Vleuten CP, Muijtjens AM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract.* 2011;16(2):151–165.

4 Govaerts MJ, Van de Wiel MW, Schuwirth LW, Van der Vleuten CP, Muijtjens AM. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ Theory Pract.* 2013;18(3):375–396.