Assessment in the post-psychometric era: Learning to love the subjective and collective

BRIAN HODGES
Faculty of Medicine, University of Toronto, Canada

Abstract

Since the 1970s, assessment of competence in the health professions has been dominated by a discourse of psychometrics that emphasizes the conversion of human behaviors to numbers and prioritizes high-stakes, point-in-time sampling, and standardization. There are many advantages to this approach, including increased fairness to test takers; however, some limitations of overemphasis on this paradigm are evident. Further, two shifts are underway that have significant consequences for assessment. First, as clinical practice becomes more interprofessional and team-based, the locus of competence is shifting from individuals to teams. Second, expensive, high-stakes final examinations are not well suited for longitudinal assessment in workplaces. The result is a need to consider assessment methods that are subjective and collective.

The rise of psychometric discourse in assessment

In the last half of the twentieth century, medical education witnessed the rise of a new language, concepts, and practices of assessment which, taken together, constitute the discourse of psychometrics. Particular statements of truth about practices and theories, as well as roles that people and institutions should play, characterized this era. We are now seeing a decline of the dominance of psychometric discourse and a rise in discourses anchored in subjectivity and collectively.

Before discussing the future however, we first need to understand how a particular set of truth statements became accepted as unquestionable for decades. First articulated in 1922 (although not fully adopted into medical education until a few decades later), an exemplar is: "Knowledge of educational products and educational purposes must become quantitative, take the form of measurement" (Thorndike 1922, p.1). What arose from the legitimation of such truths was a huge discursive shift; what was deemed to be true about assessment in medical education, what was good or proper practice, and what was judged to be right or fair changed radically. The notion of converting human behaviors to numbers constituted a way of thinking that found its way into every corner of medical education.

Many concepts are central to psychometric discourse, although perhaps none is more important than reliability. It is difficult to overstate the degree to which concerns with reliability dominated assessment during the late twentieth century. No assessment guide or article was complete without the admonition that all tests must have a (Cronbach's alpha) reliability coefficient of at least 0.8. The origin of this imperative was psychology measurement textbooks; for example, Nunnally and Bernstein's (1994) textbook Psychometric Theory states, "if important decisions are made with respect to specific test scores, a reliability of 0.90 is the bare minimum." This concept was pulled strongly into the field of assessment in medical education.

The psychometric era brought not only the concept of reliability, but also other new concepts that gave credence to some practices and delegitimizing others. The most important discursive shift was the negative connotation taken on by the word subjective. Framed in opposition to objective, the use of subjective in conjunction with assessment came to mean biased and biased came to mean unfair. There was also a strong association forged between assessment that was objective and tools that were standardized.

I have documented elsewhere how the adoption of psychometric discourse came to shape the ways it was possible to think about and speak about assessment (Hodges 2009). A telling example is a quote from medical educator Howard Barrows, a man credited with the first use of what he called programmed patients to help improve the learning of clinical skills. In the late 1960s, when he was writing about his new technique, he made little reference to standardization or multiple sampling. But just two decades later he wrote, "The standardized-patient technique... has many of the same advantages of the multiple-choice question. It is a standardized item, can be given in multiples, and can be scored in reliable and valid ways" (Barrows 1993, p.448).

Practice points

- Psychometric approaches to assessment have yielded gains but also created challenges.
- Subjective judgment in assessment has value if multiple sampling is used.
- Competence assessment should focus on collective competence and not solely individual competence.
The rise of psychometric discourse meant that reliability was a desirable characteristic of all tests; reliability would be increased with a greater number of observations, sampled from a homogeneous set of test items, across standardized testing conditions. Seen through this lens, some forms of assessment, such as what were called “old orals exams,” were deemed unfair and unsuitable because of the one or two examiners and unstandardized questions. A pivotal event occurred in the late 1960s when the National Board of Medical Examiners in the United States discontinued the use of all oral examinations on the basis of a large-scale study of more than 10,000 final oral exams in which the average correlation between the two examiners was less than 0.25 (McGuire 1966). From the end of the 1960s, until the adoption of a multi-station, standardized clinical skills (CSA) assessment examination in 2005, the United States required only standardized written exams to assess the competence of graduating medical students.

The dominance of psychometric standards as criteria for assessment quality filtered down from large testing organizations to medical schools, and reliability and validity became markers of the acceptability of test instruments. Attention to reliability contributed to broader sampling that undoubtedly did make assessment fairer. Another positive effect was the rise in examiner training. No longer was it assumed that experience was all that was needed; examiners required orientation to testing methods and even calibration. Attention to psychometric validity helped focus test creators on a shared understanding about what exams were for, parsing various forms of variance related to scoring from what would be called the variable of test taking ability. Assessment research flourished with new attention given to the relationship of expertise and scoring; the psychometric properties of different kinds of assessment tools; and the generalizability of scores across domains, sites, and cultures.

Psychometric high-stakes testing and its discontents

Nevertheless, adverse effects of the dominance of psychometric discourse became apparent. This included the finer and finer atomization of competencies into sub-sub-domains that could be standardized and inserted into blueprints to boost the reliability of examinations. Standardization of examination content led to the homogenization of test materials and scenarios, while diagnostic, contextual, or inter-personal variables that might be part of the authentic variability of real practice settings were often removed to make tests equivalent for all test takers. Another problematic effect was the need to create large testing banks because of exam security. Often feedback from examinations disappeared so as to preserve the security and confidentiality of the test bank. Finally, the increased expense of standardized, multiple sampling examinations (such as OSCEs, MCQs) meant that exams were given infrequently, often at the end of a training block, year, or program, putting the test results out of range of students’ learning needs. Ultimately, there were calls to re-examine the dominance of the psychometric discourse itself (Schuwirth & van der Vleuten 2006).

What are the fundamental concepts on which psychometric discourse rests? First and foremost, it is a set of practices to convert human phenomena into numbers. Such conversion is not an exact process; data are lost during the conversion. The numbers generated during the process represent—form of existence of something—but they are not that entity, in and of themselves. Just as a measure of 3.5 represents a blood serum potassium level but it does not, in and of itself, constitute blood serum potassium; a score of 121 on a standardized intelligence test is not, in and of itself, a person’s intelligence. In the psychometric model, reliability is the accuracy of the transformation from real-world phenomenon to a number, and validity is the stability of that numerical representation. These assumptions are grounded on the ideas that phenomena are located within individuals; that there is a quantity or amount that can be measured; that this measure, or true score, is obscured by sources of statistical noise from extraneous factors that needs to be eliminated; and that the ability of tests to discriminate between individuals is something positive.

From a conceptual perspective there are several difficulties with these assumptions. Among them, in no particular order, are that competence is not a characteristic of individuals but is embedded in collectivities; competence is not a fixed, stable characteristic but one that varies in different contexts; tests have the power to shape the thoughts and behaviors of individuals; and finally, discriminating between individuals might be less helpful than some form of differentiation of abilities within individuals.

Turning from conceptual concerns to practices of assessment, psychometric discourse provided three key imperatives: to identify sub-components of competence; to standardize assessments and take multiple samples; and to aggregate sub-scores to reconstitute competence. The need to identify sub-components that make up competence has led to efforts around the world to articulate competence frameworks: for example, the CanMEDS roles, the ACGME competence framework, and Tomorrow’s Doctor in the UK. These are widely held to have helped better articulate the competencies required of graduates as the profession feels its way toward a competence-based model of training. Yet, the competence-based model rests on an assumption that medical competence can be subdivided into separate measurable, stable traits. Interestingly, in Dissecting the good doctor, Whitehead traces the evolution of medical education from a concern with a holistic notion of character to a focus on characteristics (Whitehead et al. 2012). She argues that while there are many advantages to identifying individual domains of competence, to place more and more reliance on measurements of smaller and smaller dimensions is to risk losing the art of judging character. This concern is reinforced in research on the visual structure of the evaluation forms in Canada, which, according to Zbrowski et al. (2009, p.745), leads residents to distort the CanMEDS construct from its original holistic philosophy. As one study participant commented, “there’s not always a true connect between what’s done on a daily basis and these little categories.”

While pattern recognition and the apperception of gestalt are at the heart of medical diagnostic competence, somehow
we have moved to thinking of supervisor judgments of trainees as being “riddled with bias.” Refocusing on the value of gestalt in assessment raises the possibility of capturing the wisdom in holistic supervisor judgments. Indeed, a new conversation has sprung up around this notion. Work by Ginsburg (2011) on respecting the expertise of clinician assessors, by Bogo et al. (2004) on the implicit criteria used by experienced field instructors, and by Gingerich et al. (2011) on rater-based assessments as social judgments illustrates ways in which notions of subjectivity and judgment in assessment are gaining new credibility.

The second psychometric imperative has been to standardize assessments and take multiple samples. Yet, as van der Vleuten and Schuwirth (2005) have shown, the major determinant of reliability is total testing time, not the standardization of the instrument used. Extrapolating from a number of studies, they showed that MCQs, essays, oral examinations, OSCEs, and other tools can all have reliability above 0.8 with sufficient testing time (about 4 hours). If the type of tool does not matter, the implication is that those tools that are more standardized (MCQ, OSCE) are not necessarily more reliable than those that are more subjective (essays, oral examinations). The caveat, of course, is that reliability is strongly tied to the number of examiners (Swanson 1987). The critical variables in attaining reliability, therefore, are testing time and multiple sampling, not standardization. We should not, therefore, be afraid of holistic judgment, although we should sample widely across sources of subjective influences (raters, examiners, patients).

The third psychometric imperative is that once assessed, component sub-scores should be recombined to determine competence. Imagine that we put together the world’s highest quality bicycle tire, a top quality tractor engine, and one wing from a state of the art airplane. In terms of transportation we will not have created anything of value. In the late twentieth century, there was a great focus on improving the technical aspects of various assessment tools, whereas today our need is to examine the utility of the assessment program as a whole (van der Vleuten & Schuwirth 2005). Adding 72% on a MCQ +80% on OSCE +4/5 on in-training evaluation + scores on case reports, mini-CEXs, and SP interviews gives us…what? Converting human phenomena into numbers is not an exact process, but recombining them compounds the problem. Reliability is very useful at the level of individual assessment tools, but it is not of much use when we combine very heterogeneous sources of information collected with different types of instruments.

To summarize, of the three imperatives given to us in psychometric discourse, we should have reservations about too much division into sub-competencies, near certainty that multiple sampling is helpful but also that a high degree of standardization is unnecessary, and strong reservations about arriving at a determination of competence by recombining a set of numerical scores derived from instruments with different purposes.

**Learning to love the collective**

Psychometric discourse is based on the idea that constructs of interest are located in individuals. Yet, the rise of team-based health care is shifting the focus from individuals to collaboration and collectivity (Lingard 2012). Ringsted et al. (2007, p. 2764) wrote, “In the assessment of physicians it must be acknowledged that physicians often work in teams and systems, rendering it impossible to attribute quality of practice to a single person.” For Lingard (2009, p. 626), “our individualist healthcare system and education culture [focuses] attention on the individual learner” nevertheless “competent individual professionals can—and do, with some regularity—combine to create an incompetent team.” And Kuper (2007, p. 1122) argues, “some aspects of medical education are better thought of as social constructs: instead of being considered as expressions of a single individuals abilities, they are conceived of as the products of interactions between two or more individuals or groups.”

Attention to the collective in competence began in earnest in the operating room in the late 1990s, at a time when comparisons were made between the scrutiny given to safety and risk management in aviation and the relative neglect of those same parameters in medicine. Over a few decades, evidence about the importance of team-based competence in health care and in particular links to patient outcomes propelled the adoption of a number of new practices, including the pre-operative briefing (“safe surgery checklist”). Many studies have since demonstrated improved patient outcomes when team-based training is employed (Haynes et al. 2009; Marr et al. 2012; Stevens et al. 2012). As research emerges that team training has a meaningful impact on patient outcomes, the notion that competence is something held by an individual surgeon, anesthetist, nurse, or individual anyone becomes more and more untenable.

**Learning to love the subjective**

In the late twentieth century, medical education tried to solve the problem of subjectivity when it became equated with unfairness, by turning to methods from psychology and psychometrics. Today there is a fundamental re-examination of the role of subjectivity and an effort to rehabilitate subjective judgment. Among the first to make this point was Epstein (2007) who called for caution regarding the unintended effects of testing and the assumption that qualitative data were more valuable than qualitative expert judgment. Eva and Hodges (2012) point out that, while the literature is replete with critiques of the dangers of subjectivity in assessment, there is a literature showing that many fallible judgments, summed together, create value. This is the key argument in James Surowiecki’s (2004) book, “The wisdom of crowds,” in which he writes that the value of subjectivity increases with the number of judgments, the independence of those judgments, and, interestingly, the diversity (not homogeneity) of perspectives.

One of the most powerful ways to rethink subjective judgment is to relate it to something of great value in medical practice—clinical diagnosis. In the clinical domain, pattern recognition is highly valued. Experienced clinicians rely on a gestalt impression of presenting features to engage in further questioning and investigation. What they do not do is use an exhaustive checklist of symptoms and signs before forming
preliminary diagnostic impressions. That is what novices do. With experience, expert clinicians become more rapid and more accurate in their recognition of patterns. There is no reason to believe that this process does not also operate in education. Experienced teachers also form rapid impressions of the competence of their trainees. Of course, just as clinicians must clarify first impressions with follow-up questions and investigations, so too, educators need to use specific assessment tools and observations to confirm their holistic impressions.

Try a thought experiment. Think about one of your current medical students, residents, or colleagues. Now ask yourself, would you send a family member to see this doctor? For most people the answer is easy, and automatic. This is a gestalt impression. On its own, your answer would be too unstable to be of much use. However, if multiple such impressions were collected and integrated—using something called a jury model—the collective judgment would have significant value. Further, if we went beyond just asking a yes or no question and had each rater describe why they would (or would not) refer a family member to this doctor, we would have a robust source of information for evaluation and feedback. Qualitative researchers call such a detailed narrative a thick description.

Imagine such a procedure was used to evaluate two individuals. For each of the two, 20 physicians, nurses, members of other health professional multidisciplinary team, peers, and patients are asked, "Would you send a family member to this doctor?" All raters are also asked to write down in detail why or why not. From a psychometric point of view the inclination would be to average these data and create a mean score. Yet, there is great value in the fact that different individuals, who encounter the learner in different contexts, see things in contrasting ways. In this model, the assessor’s role would be to interpret, not simply to apply transformation to numbers; interpretation would be both numerical and linguistic. If, for example, the impression of not wanting to send a family member to the doctor came systematically from patients, or from non-medical health professional colleagues, that would mean something. The assessor of the future must be an aggregator of such independent judgments. It may be that a team best performs this function. In the United States, the ACGME recommends using competence committees as part of residency education to perform this function. Just as we must learn to love the collective in gathering assessment data, a collective perspective may serve us well in the aggregation and interpretation of data, or put more simply, in deciding how metrics might inform our decisions, without becoming them” (Girdharadas 2009, p. 2).

Assessment in the post-psychometric era

If we are truly entering a post-psychometric era in assessment, what does the future hold? The twenty-first-century medical doctor will need the curiosity to tackle and sustain interest in problems of varying complexity, the capacity to engage cognitive and emotional resources, the ability and willingness to collaborate with others, and the desire to work in teams to create effective solutions. Our assessment programs must support all of this. To get there, we need to raise our game and focus on the overall impacts of assessment programs. As Rowntree (1987, p. 2) has written, “It is easy to find writers concerned with how to produce a better multiple choice question, how to handle test results statistically, or how to compensate for the fact that different examiners respond differently to a given piece of student work. It is much less easy to find writers questioning the purpose of assessment, asking what qualities it does or should identify, examining its effects on the relationship between teachers and learners, or attempting to relate it to such concepts as truth, fairness, trust, humanity or social justice.” The challenge before us then is to build rigor into our assessment programs, and to recognize that competence is contextual, constructed, and changeable and, at least in part, also subjective and collective.

Declaration of interest: The author reports no conflicts of interest. The author alone is responsible for the content and writing of the article.

Notes on contributor

BRIAN D. HODGES is a Professor in the Faculty of Medicine and Faculty of Education (OISE/UT) at the University of Toronto, the Richard and Elizabeth Gariepy Chair in Health Professions Education Research at the Wilson Centre for Research in Education and Vice President Education at the University Health Network.

References


Surowiecki J. 2004. The wisdom of crowds. Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. New York: Doubleday.


