Adv in Health Sci Educ (2013) 18:547–550
DOI 10.1007/s10459-013-9476-2

EDITORIAL

On objective: based education, objectivity, and rater cognition

Geoff Norman

The world is full of paradoxes. We have less pollution than we've had for centuries, yet we are preoccupied with the effect of the environment on health. Crime rates have been dropping in the western world for two decades, yet we are fearful of personal loss from street crime and are building more prisons (at least in Canada). Some of us are afraid to get on an airplane, yet we are at far greater risk driving to the airport than we are once airborne. We dreadfully fear the health effects of nuclear power, and ignore the thousands of miners killed in coal mines every year. And in health professions education, we praise objectivity with endless discourse on the value of learning objectives, the superiority of the objective structured clinical examination, the central role of objective tests, and now, in its latest incarnation, the apparent advantages of objective outcomes- or competency-based curricula. Yet in assessment, the single ubiquitous method used on every school at every level on every continent is the subjective, global rating scale, with a bunch of seven-point scales, completed by a supervisor, preceptor, or tutor after some extended period of instruction.

The present issue has quite a bit to say about both competency-based education and subjective ratings. Of course the first thing to note is that the present issue has quite a bit to say about a lot of things, mainly because it has grown, now containing more than 20 articles instead of our usual 12 or so. The same will be true of the last issue of 2013. While I would be delighted to report that this is a permanent change, reflecting the huge success of the journal, regrettably it is no more than a temporary aberration resulting from a policy at Springer to reduce the "Online First" queue to a manageable size. At last glance, the queue stood at 72, which means a delay of over a year from acceptance to publication, so something had to be done. However this does, in part, represent an increase in popularity of the journal. The number of submissions in 2012 was about 450, an increase of 100 % over 5 years earlier, and a trend that is continuing. To some degree this has been counterbalanced by a trend to lower acceptance rates, which in 2013 is down to about 12 %. The other good news about the journal is impact factor, which rose in 2011 from around 1.4–1.5 to over 2.0, and maintained this increase in 2012.

G. Norman (✉)
McMaster University, Hamilton, ON, Canada
e-mail: norman@mcmaster.ca

But let's get back to the paradoxes. The current issue contains a review article by Morcke et al. (2013), which is perhaps the most extensive comment on outcome-based education I have seen. Its origins are traced back to the behavioural objectives movement of the 1960s, which were themselves rooted in behavioural psychology, the dominant paradigm of psychology in those days. Morcke et al. document the rise and fall of the behavioural objective movement, and its rise again in our current love affair with competency- or outcome-based curricula. They correctly (in my eyes) view this renaissance as "a more or less unbroken line of inheritance" despite some proclamations that the new approach is distinct from behavioural objectives (Harden 2002). They also point out that outcome-based education has continued to rest more on advocacy and appeal to common sense than on empirical evidence of effectiveness "In the last decade, OBE has been advocated and implemented … but this has not been followed by substantial research on the impact of learning outcomes on teaching and learning in medicine."

Another paradox emerges from the paper. One of the strongest claims of competency statements like the CanMEDS roles is that they draw attention to those aspects of competence which are less easily learned and assessed like advocacy, communication, collaboration, professionalism. On the other hand, as Morcke et al., point out, the requirement that outcomes "be stated clearly and unambiguously" and the emphasis "on assessment systems that call for learners to demonstrate proficiency in the intended learning outcomes" mitigates against meaningful recognition of precisely those dimensions which are purported to be the strength of competency-based education.

This hypothesis finds support in another paper in AHSE currently in Online First. Sherbino et al. (2013) analyzed 1,800 workplace encounter ratings that used the CanMEDS competencies as items, and found that "all items loaded on a single factor … accounting for 87 % of the variance." Moreover, raters were required to assess "medical expert" and any 2 additional roles. Sherbino found that the expanded roles—scholar, collaborator, manager, advocate—were assessed on only 10–22 % of the encounters. Precisely what Morcke predicted.

Perhaps the fact that so much of medical expertise lies in these "soft" areas, where objective tests are of limited value, goes some towards explaining the survival of global ratings in the face of the inexorable march toward objectivity. We have known for nearly 40 years that these one-page ratings do not achieve even minimal levels of inter-rater reliability; a minimum criterion for application (Streiner 1985). Some (Schuwirth and Vleuten 2006) have argued that this represents a weakness in the psychometric paradigm—the assumption of a single underlying construct called "competence," not of the assessment method. Instead they argue, among other things, that each assessment may well have unique information, and its deviation from an average is to be identified and used in enhancing learning, not simply averaged with other ratings.

That may be so, but one is still left with no clear direction about how to proceed in the face of conflicting evidence from different assessments. To assume that all variation represents some kind of signal, with no possible error, amounts to a heroic assumption. Yet this is what results from the elevation of the particular over the average. If we accept that each assessment brings a unique perspective on the learner, then we must somehow dig deep to try to separate critical elements of the learner from other factors—rater bias, context, case difficulty, and so forth.

One response to these concerns may well be that my use of the term "rater bias" betrays my psychometric roots, and I have fallen victim to the old way of thinking. But when you carefully examine ratings from different raters and situations, some raters *are* tougher than others, and case difficulty *is* typically the largest source of variation in ratings. To presume that these can be easily disentangled in examining individual ratings is simplistic.

However, there is another way, as emerges from the several papers on rater cognition in this issue and other recent issues of AHSE. The articles cover the full gamut: two are more or less standard psychometric analyses (Paget et al. 2013; McGill et al. 2013), two involve qualitative analysis of think-aloud protocols (Berendonk et al. 2013; Govaerts et al. 2013); and two are review articles analyzing rater assessments from a cognitive psychology perspective (Tavares and Eva 2013; Wood 2013).

Looking first at the psychometric analyses, McGill et al. (2013) found, as is often the case, that the global workplace assessments had very high internal consistenct, with an alpha of 0.94. Moreover, while they extracted 3 factors, the first accounted for 54 % of the variance, the second 8 % and the third, 6 %. Paget et al. (2013) also did a psychometric analysis, but looked at rater variables, not student variables. They found, as might be anticipated, that most students were above average, with 66 % above expected level. However two rater variables modified the observation: the longer the assessor took to fill out the form (in days), and the number of evaluations the assessor completed. They provided a detailed explanation of these findings in terms of cognitive biases of raters, drawing on literature on rater cognition.

The approach to performance rating as a cognitive task is elaborated in the two review articles by Tavares and Eva (2013) (which appeared in AHSE 18:2) and Wood (2013) (which is available in Online First). Tavares and Eva explicitly introduce cognitive heuristics as a rater strategy to reduce cognitive load. Wood's central theme is also mental workload, but he deals with it by introducing a second theoretical framework from cognitive psychology, dual processing theory. He identifies literature showing that "thin slice" judgments, first impressions, can be as reliable and valid as more extensive judgments.

A different approach was taken by the two qualitative papers. Both Tavares and Wood, in exploring the cognitive factors underlying rater judgment, implicitly assume that raters are only partially aware of the basis of their judgments. By contrast, the Berendonk et al. (2013) and Govaerts (2013) papers use rater insights derived from interviews as their data source. However there is some convergence. For example, Govaerts suggests that "most raters started to observe person schemas the moment they began to observe trainee performance", which is exactly the kind of "thin-slice" judgment described by Wood. They also identified that rater idiosyncracy was substantial, consistent with the psychometric findings. Berendonk et al. (2013) begins with the "cognitive bias" perspective, but goes on to explore the assessor's perceptions of their role, the task, and the assessment context.

It appears to me that these studies are really addressing a previously neglected area in health sciences education. We have traditionally relied heavily on subjective judgments of performance despite their inherent limitations. As we move to competency-based curriculum, we must rely on these methods to a far greater degree than before, particularly to assess those areas which fall outside the traditional "medical expert" role. Whatever the purported advantages of competency-based learning may be, ultimately the project will stand or fall on the adequacy of assessment methods. This new research direction represents a critical way forward in improving the effectiveness and efficiency of these methods.

References

Berendonk, C., Stalemijer, R., & Schuwirth, L. W. T. (2013). Expertise in performance assessment: Assessor's perspectives. *Advances in Health Sciences Education*. doi:10.1007/s10459-012-9392-x.

Govaerts, M. J. B., Van de Wiel, M. W. J., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education, 18*(3), 375–396. doi:10.1007/s10459-012-9376-x.

Harden, R. M. (2002). Learning outcomes and instructional objectives: Is there a difference? *Medical Teacher, 24*, 151–155.

McGill, D. A., van der Vleuten, C. P. M., & Clarke, M. J. (2013). A critical evaluation of the validity and the reliability of global competency constructs for supervisor assessment of junior medical trainees. *Advances in Health Sciences Education*. doi:10.1007/s10459-012-9410-z.

Morcke, A. M., Dornan, T., & Eika, B. (2013). Outcome (competency) based education: An exploration of its origins, theoretical basis and empirical evidence. *Advances in Health Sciences Education*. doi:10.1007/s10459-012-9405-9.

Paget, M., Wu, C., McIlwrick, J., Woluschuk, W., Wright, B., & McLaughlin, K. (2013). Rater variables associated with ITER ratings. *Advances in Health Sciences Education*. doi:10.1007/s10459-012-9391-y.

Schuwirth, L. W., & Vleuten, C. P. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40*(4), 296–300.

Sherbino, J., Kulasegaram, M., Worster, A., & Norman, G. (2013). The reliability of encounter cards to assess the CanMEDS roles. *Advances in Health Sciences Education*. doi:10.1007/s10459-012-9440-6.

Streiner, D. L. (1985). Global rating scales. In V. R. Neufeld & G. R. Norman (Eds.), *Assessing clinical competence*. New York: Springer.

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education, 18*(2), 291–303. doi:10.1007/s10459-012-9370-3.

Wood, T. (2013). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education*. doi:10.1007/s10459-013-9453-9.