LAMBERT SCHUWIRTH, HELENA WARD
AND SYLVIA HEENEMAN

15. ASSESSMENT FOR LEARNING

THE SETTING

Most students enter medical school with the intent to become a physician and work in patient care; and rightfully so because good patient care is important for society. But there is also a need in society for physicians or MDs with special abilities in conducting biomedical and clinical research. This special need was the reason why at the University of Maastricht a graduate-entry medical program was begun with an annual intake of 30 students. But it was not the only reason. Educational innovation and experimentation have become increasingly difficult due to the rapidly increasing enrolment in many medical programs. Our new program was therefore also started with the intent of enabling experimentation and innovation at the curriculum level.

The curriculum of this program (called AK-O which is the Dutch abbreviation for Physician-Clinical Investigator) is new in that it is eclectic in its educational philosophy. This implies that it does not see any educational approach as inherently superior. Instead, it seeks to build a medical education program of which the building blocks are based on what is known to work best from the literature on education, learning and development of expertise. This may not be entirely new, as there are probably many other educational programs that are optimised in this way. What is quite unique though is its radical approach to assessment. Instead of adopting a traditional view of assessment, where assessment is used mainly to determine whether students can make the cut or not and to motivate students for learning, assessment is used in the AK-O as an integral element of the learning process. This may seem obvious and not entirely new – many programs have some sort of formative assessment – but we argue here that this is radically different from the traditional approaches. Moreover, we think it is a successful example of turning an assessment culture around from a test-driven to an evaluation/improvement-driven one.

THE FOCUS

Though the radical implementation may be quite unique, the notion of this type of assessment is not. In the literature it is often called assessment for learning, as opposed to assessment of learning (Shepard, 2009). In assessment for learning the focus of the assessment is on using all possible information in an optimal way to steer, foster and motivate students at the individual level in their learning

J. Higgs et al. (Eds.), Realising Exemplary Practice-Based Education, 143–150.

processes. Assessment for learning, therefore, incorporates several fundamental principles. The first principle concerns the purpose of the assessment. The central purpose of assessment of learning is to determine almost exclusively whether student A is better than student B or if they are better than a cut-off score (for a virtual borderline student). Assessment for learning seeks to answer the question whether both students A and B are today optimally better or more competent than they were yesterday, and whether they will be optimally better tomorrow than they were today, in order in our case to stimulate every student to become the best doctor s/he can be. For this, assessment has to give directions as to what educational intervention would be best for each particular student to become optimally better tomorrow. And finally, such assessment is fully useful if it determines also whether students are sufficiently on track to becoming competent professionals. In summary, this "assessment for learning" program, in analogy to patient care, seeks to answer three questions:

1. Do we have enough information about the progress of this student or is additional assessment necessary (the "diagnostic" question: is the diagnostic work-up on this student complete)?
2. Which educational intervention is most useful for this student at this time (the "therapeutic" question: which therapeutic educational intervention is most indicated)?
3. Is this student on the right track to becoming a competent professional (the "prognostic" question: what is the prognosis for this student at this moment)?

In the design of the assessment program three major issues became clear right from the start. The first was that in such an assessment approach no single instrument can be seen as a panacea; no single instrument can do it all. Therefore a program of assessment is needed (Van der Vleuten & Schuwirth, 2005). This might not seem novel, but in fact it is. Our literature has been and still is full of papers trying to demonstrate the superiority of one assessment method over others, whereas from a programmatic view the question "where would the instrument fit best in an assessment program?" is much more relevant. In programmatic views on assessment there are no good or bad instruments per se; each has its indications, side-effects and contra-indications, and should be fitted into the program where it serves its best purpose. Programmatic assessment also requires considerations at the level of combining the results of different instruments, beyond the often arbitrary decisions (examination A counts for 80% and examination B counts for 20%) made in many assessment programs, in which no clear consideration is given to how those examinations contribute to the overarching goal of the assessment program.

A second issue was that for assessment to really inform at the level of integrated competencies, a 1:1 relationship would not work. By a 1:1 relationship we mean an assessment program in which for each competency there is one assessment method and each assessment method is used for only one competency; for example, a critically appraised topic for the role of scholar, a multiple-choice examination for the role of medical expert, and a presentation for the role of health advocate, etc. In

truly programmatic assessment the results of several assessments can be used to inform each competency domain and each assessment can inform several competencies. Suppose a student performed poorly on an objective structured clinical examination (OSCE) station on communication with a patient with abdominal complaints. These results could be used to judge the student's communication ability, understanding of pathophysiology, or preventive counselling skills (roles as communicator, medical expert and/or health advocate). In reverse, the results of various tests can be used to inform about a student's progress as scholar. In its extreme form this would imply breaking up assessments into parts, each of which can inform a different competency. This is not as strange as it might seem. There is good evidence in the literature that the content of the assessment determines much more what the assessment assesses than the format. An OSCE station on abdominal examination, for example, correlates better with a set of multiple-choice items on abdominal anatomy than with a station on neurological examination (Norman, Tugwell, Feightner, Muzzin, & Jacoby, 1985). Yet we still add up abdominal examination results with neurological examination results because they are of the same format, and we add up abdominal anatomy questions with neurology questions for the same reason. In patient care this would be similar to compensating for a low sodium level with a high blood glucose level and claiming that this patient is healthy because the average of his/her sodium level and blood glucose is above an arbitrarily set standard.

A final issue was that the assessment must be meaningful and taken seriously, but must also be fair and rigorous. This is not easy to achieve; many of the notions about rigour and fairness of assessment originate from research into assessment of learning, and so new ideas had to be found.

THE STRATEGY

Central to the assessment program of the AK-O is the portfolio; it is the backbone of the program and actually all study credits are assigned to it. All other assessments serve to produce information for the portfolio. Our portfolio is therefore not an instrument to assess reflection (that would be a 1:1 again), but it is used as analogous to a patient chart in patient care. The strength-weakness analyses (or the reflective analyses) with their learning goals are like the doctor's notes.

During each educational module, whether it is more theoretically oriented or more practical in nature, many assessment moments are built in, the results of which are all incorporated into the portfolio, need to be addressed in the strength-weakness analyses, and lead to concrete learning goals. The assessment program has longitudinal elements as well, including progress testing, continuous assessment of professional behaviour and a series of cumulative assessments concerning the topic of patient-doctor and society. The results of these assessments are of course part of the portfolio as well, as is all the informal feedback and evaluation a student receives.

Several times per year each student meets with a specifically assigned mentor to discuss progress. The student updates the strengths-weaknesses analysis and

reflects back on the completion of previous learning goals. The student and the mentor then discuss the updated portfolio addressing the three questions (diagnostic, therapeutic and prognostic). At the end of the year the mentor provides written advice about whether the student should be allowed into the next phase of the study or not, which is reviewed in a committee meeting of independent mentors in the presence of the student's mentor – with only an advisory role – after which a decision is reached.

This may seem a vulnerable process, but numerous measures have been put in place to make it rigorous. The portfolio needs to contain full information of all formal and informal assessments. To this end, the mentor has a list of all assessment moments and types of information to expect, so s/he can check at any time (and does so) whether the dossier in the portfolio is complete or whether the student has strategically omitted information. The portfolio is therefore not "student-owned," but is a shared document between student and faculty.

Minutes of each meeting are taken down on paper and these minutes have to be approved by both the student and the mentor as an accurate record of what has been discussed and agreed. These are a required element of the portfolio content.

Once each year, a second, independent mentor sits in on the meeting to provide a fresh pair of eyes on the process. At the end of each meeting the mentor provides in writing a prognosis about the student. Any negative decision at the end of the year, therefore, cannot come as a surprise to the student. The mentor is extensively trained for the role. In the training sessions, special care is taken to avoid common pitfalls such as the halo effect, primacy effects, cognitive dissonance, investment traps, etc. That is an extra reason why note-taking of the whole process is considered so important: it not only makes the process fully transparent and accountable but also counteracts some of the biases.

The mentor's advice is exactly what it says: advice. The decision is made by a committee of independent mentors. The procedure is as follows: one of the independent mentors reads the whole portfolio, prepares a summary of the most important information and prepares a decision. This is then presented to the group of mentors, who critically appraise this information (often consulting the portfolio specifically). When needed, the student's mentor may add his/her view or extra information. This whole procedure not only ensures greater "objectivity" in decision making; it also serves as an extra learning experience for all mentors.

This learning experience is on top of frequent peer-feedback mentor meetings. During these, mentors discuss difficult cases or situations – anonymously – and seek advice from their peers. They share strategies, pitfalls (e.g. judgement biases) and other experiences. Thus, ongoing improvement in expertise occurs.

All formal assessment moments may be highly informative but they are not completely formative nor are they completely without stakes, because they all serve to contribute to a summative decision at the end. The main difference is that an assessment moment is not automatically a decision moment; quite the contrary, most of them are not. Of course, all assessment is subject to regular quality control procedures such as item review processes and item analyses.

THE CHALLENGES FACED

This not to say that implementing this program has been easy – quite the contrary. The central idea may be intuitive but with the implementation all kinds of implications and consequences arose.

A first issue we had to deal with was the legal framework. Educational laws in the Netherlands prescribe a strong relationship between an educational module, its assessment and study credit points. Such laws are typically designed to cater for an assessment of learning approach and not assessment for learning. In planning the whole program we had considerable problems convincing the legal department of the university of the legality of it all. Actually, we were told "no" repeatedly until we managed to convince the Vice-Chancellor – who is a professor at law – of the program and he decided that it was in our university's remit to innovate and that this involves seeking the boundaries of the law. Once his consent was obtained we could go ahead and implement the program in full. In short, the permission enabled us to assign all study credit points to the portfolio and to give to students at the end of a phase and not during a phase. So far we have had no legal challenge, probably because of the transparency, carefulness and credibility of the program.

More difficult was changing the culture. The students who enrolled in our program were clearly all high achievers. They had completed a previous biomedical bachelor degree and had all obtained high grades. So they were used to succeeding in an assessment of learning environment and they were now asked to adapt to an assessment for learning environment. Because the former rewards performance orientation and the latter rewards a learning orientation this was not easy for most of them. Many struggled with this and just wanted grades (high grades, that is to say). And although feedback on all formal assessment included information with respect not only to what their strengths and weakness were but also to how well they had done overall, they still had considerable problems getting used to this assessment culture. On the rebound, once they had adopted the new culture it drove them into an extreme learning-orientated mode and they experienced problems deciding when enough was enough. Fortunately the mentors were able to help them with this. In their regular meetings with students mentors had the opportunity to help them better understand the culture and to slow students down if burnout threatened. Furthermore, regular class meetings were organised to explain and discuss the curriculum and the assessment with students. Finally, after the program had been running for some years the informal information-sharing between more senior and more junior students proved a useful means of acclimatising junior students to the new culture.

Not only students had to become accustomed to the new assessment culture; teachers had their problems too. Their most typical concern in this highly integrated approach to education and assessment was that their subject would not be covered with sufficient detail. It was therefore important to help them design assessments that would require integration of the essentials of their own topics with those of related disciplines; to produce genuinely integrated assessments rather than a stack of individual topic examinations. For this, it was necessary for teachers

to convene in groups and jointly produce assessment cases. Typically, such cases could only be solved and the related questions could only be answered correctly if the relevant knowledge of several topics and disciplines was combined successfully. We think it is obvious that this was not an easy task, but as experience grew it became easier, and we were able to tease out some templates and good examples to help in subsequent item-writing sessions.

Another aspect to which teachers needed to become accustomed was the provision of feedback. Most teachers were not really well experienced in this, and made only general comments such as "well done." Or they provided unworkable suggestions, such as "try to be a bit more assertive." So teacher training was focused on providing concrete feedback (what was good and why and what was bad and why?) and on behaviour rather than personality traits (if you are not assertive and this leads to you being given tasks you don't actually want, why not try out different strategies for saying "no").

By now you may have thought about the enormous costs associated with such an assessment program, and that seems to be one of its challenges. But actually the most challenging task is to show that the benefits outweigh the costs. The mentoring system seems costly, for example, but it also saves money. First year students see their mentor six times per year for half an hour. The mentor needs preparation time which is budgeted at an hour per meeting. For a starting mentor this is perhaps insufficient but as the mentoring experience grows the time needed to read and understand a portfolio decreases considerably. Also, mentors get to know their individual students and acquire good knowledge of what was in their portfolio; therefore they need considerably less than an hour. In all, mentoring requires the equivalent of 9 hours per student per year, so for 30 students this equals 270 hours or .16 FTE, costing roughly 20,000–25,000 Euros per year. If mentoring can prevent only one student from leaving the course or being delayed by a year, breakeven is already achieved. Unfortunately it is difficult to demonstrate that the mentoring system does prevent such attrition, and therein lies the challenge.

CRITICAL REFLECTIONS

Thus far everybody is enthusiastic about the program, and unfavourable reactions concern more minor implementation-related and organisational issues than the overarching concept of assessment for learning. But there is certainly no room for complacency, as much of what we do is experience-based and lacks rigorous scientific underpinning. At Maastricht University educational research is considered important, and especially the combination of fundamental and applied research is seen as essential to the support of any educational action.

Therefore this has set the research agenda in the following directions:

The AK-O program has adopted a programmatic approach to assessment and therefore it is important to understand better what determines quality of assessment at the level of a program. There is a shared opinion that the combination of reliability, validity, educational impact, cost-efficiency and acceptability are

elements of the quality of individual instruments (Van der Vleuten, 1996), but little is known about quality of programs. Therefore, a PhD project has begun on this topic, which has led to the development and early validation of a model for quality of assessment programs (Dijkstra, Van der Vleuten, & Schuwirth, 2010).

An old mantra in assessment is that summative and formative functions should not be mixed. In assessment for learning, however, these two functions have to be mixed, and a separation between assessment moments and decision moments is considered more useful. This puts extra pressure on teachers, because they have to combine formative and summative roles continuously. Thus we need to understand better what enables teachers to combine those two roles and – more importantly – which elements would constitute barriers to this combination. Research into the mental processes of teachers/assessors, how the summative and formative combination influences their feedback, and which organisational elements hamper this process has started and is part of another PhD project.

In combining information from various parts of assessment, both quantitative and qualitative, into meaningful conclusions about progress in competency domains, human judgement is indispensable. This might seem a difficult judgement task, but it is something humans do on a regular basis. Most doctors, for example, can easily combine complaints of thirst, poorly-healing wounds and fatigue with physical diagnostic findings of peripheral small artery dysfunction and a glucose level of 35 mmols/l into a possible diabetes mellitus, combining acoustic, visual, tactile and numerical information. The main reason they are able to do so is their extensive training and expertise development and their good understanding of the context as in illness and instance scripts. This leads them not only to understand the importance of each of the contributing features but also to be less susceptible to bias in evaluating the information. That is why we have started research into the nature of assessor expertise and the extent to which it is comparable to diagnostic expertise (Govaerts, 2011).

Another important implication is that standard psychometric methods cannot be applied to all decisions made during the assessment. Most standard approaches make firm assumptions about the nature of the aspect being assessed. Reproducibility, for example, assumes that the object of measurement is stable during the measurements. But what if repeated measurements (as in mini-CEX assessment) take place over a longer period of time and if considerable learning or development takes place between the observations? The student has then changed, and consequently scores on subsequent measurements will differ. Under the assumption that the object of measurement does not change these differences would be seen as error, whereas at least part of the change would have to be attributed to development and learning. Qualitative comments, feedback, and holistic decisions are all-important elements in a program of assessment but they do not fit well in a standard psychometric framework. Yet their quality needs to be demonstrated. Research has therefore been started to explore various methods of demonstrating assessment quality and to understand better which method is best for which element of the program. This is in its initial phase and needs further development.

FINALLY

We have presented a case of implementation of an assessment for learning program in one institute. A similar program has been established elsewhere (Dannefer & Henson, 2007). We have used this example for two reasons. First, we wanted to show a real implementation of what in the literature is often largely a merely philosophical notion, and to show that it can be done in a normal educational (though certainly not a worst-case) situation. Second, and more important, we wanted to show that just putting a good idea into practice does not suffice. One must realise the accompanying practical implications. Also one must understand that new ideas bring new questions, and that medical education development can only be taken seriously if it is supported by a critical research program.

REFERENCES

Dannefer, E., & Henson, L. (2007). The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine, 82*(5), 493-502.

Dijkstra, J., Van der Vleuten, C., & Schuwirth, L. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education, 15*, 379-393. doi: 10.1007/s10459-009-9205-z

Govaerts M. J. B. (2011). *Climbing the pyramid: Towards understanding performance assessment*. PhD dissertation, Maastricht University, Maastricht.

Norman, G., Tugwell, P., Feightner, J., Muzzin, L., & Jacoby, L. (1985). Knowledge and clinical problem-solving. *Medical Education, 19*, 344-356.

Shepard, L. (2009). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Science Education, 1*(1), 41-67.

Van der Vleuten, C. P. M., & Schuwirth, L. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317.

Lambert Schuwirth MD, PhD
Professor of Medical Education
Flinders University, Australia
Professor for Innovative Assessment
Maastricht University, The Netherlands

Helena Ward PhD
Heaslip Fellow for Medical Education
Flinders University, Australia

Sylvia Heeneman PhD
Program Director AK-O
Maastricht University, Maastricht, The Netherlands