Chapter

39

Section 5: Assessment

Programmatic assessment

C. P. M. van der Vleuten, S. Heeneman, L. W. T. Schuwirth

Trends

- A holistic view on assessment is emerging in which formative and summative assessment strategies are combined.
- Competency-based education is spreading both at the undergraduate and postgraduate level of training, but is hampered by an inappropriate assessment approach.
- Programmatic assessment has been proposed as a means to make assessment aligned with a constructivist view on teaching and learning.

Introduction

Programmatic assessment is an alternative way of arranging assessment in a training programme and a different take on how assessment may function to promote learning. Some have called programmatic assessment a paradigm shift in our thinking around assessment. Its origins lie in research and observations in practice. Any individual assessment involves a compromise on quality criteria (van der Vleuten, 1996); it cannot have perfect reliability, validity, educational impact, acceptability and low cost. The choice of compromise depends on the context and purpose of the assessment.

Any individual assessment is but one data point with limited utility. Consider a data point to be similar to a pixel of an image.

The question is really when to optimize what. The basic tenet of programmatic assessment is to capitalize on the complementarity of different assessment methods by seeking to achieve their most fruitful constellation, rather than pursuing perfection – in terms of fulfilment of all possible quality criteria – in each of these methods individually. When such constellation is reached, its constituent judgements of assessment together will bolster the strength of the assessment as a whole. Such approach requires us to make different choices. In making these choices educational arguments for maximizing learning by learners weigh heavily. In this chapter we will explain in which respects the traditional approach to assessment can gain from a shift to programmatic assessment. We will then proceed to explaining programmatic assessment itself, illustrated with an example of an existing programme. We will end with some reflections and issues that are found in recent implementations of programmatic assessment.

The traditional approach

The most dominant approach to assessment is to have an end-of-course assessment one has to pass. The performance of the learner is compared with a minimum standard. If a learner fails, then usually he or she has to resit the entire exam. In the event of multiple failures, the learner usually has to redo the course and repeat the assessment. This is how learners navigate through a full training programme. Many training programmes also include a comprehensive assessment at the end: a final exam. When all exams have been passed, the learner is considered qualified to pursue a more advanced programme in the field or to enter professional practice. This classical approach to learning is old and has served us well, but we argue here that there is scope for improvement.

The traditional approach is modular. It assumes that when learners pass, they have mastered the whole domain they were learning, often referred to as 'mastery learning'. Although the learning is triggered by the assessment occasion, the assumption is that learners will have 'mastered' the domain for the remainder of their lives. In most cases, this will likely not hold. Forgetting is quite normal, and the forgetting curves from psychology show that 50% of the learned subject matter is already forgotten after a few weeks. One of the most fundamental problems in education is the issue of transfer. Having the knowledge in no way guarantees that the learner will be able to apply this knowledge when appropriate, i.e. when needed to manage a professional task. Hence, mastery at one moment in time bears little relationship to its use at a later moment. Therefore, in many domains pure mastery learning is an outdated model of learning.

Modern learning programmes are based on 'constructivist' notions: learning occurs more effectively and efficiently if the learner 'constructs' the information or knowledge. Learning in this sense means 'processing information' rather than 'consuming' it. Internalizing information and being able to understand and use it is what makes learning productive. Hence, teaching is not only about transmitting information, but also about enabling learners to make maximum sense of it by allowing them to continuously practise transfer on the basis of professionally authentic tasks. Knowledge, skills and attitudes are integrated by using "whole tasks" (Vandewaetere et al., 2015). Problem-based learning, team-based learning, competency-based learning and outcome-based learning are illustrations of modern education approaches that are founded on this constructivist notion of learning. These approaches enjoy wide currency in undergraduate and postgraduate education in medicine.

A traditional approach to assessment may induce poor learning behaviour as passing the assessment becomes learners' main incentive to learn (Cilliers et al., 2012).

> *"Better utilization of assessment to influence learning has long been a goal in higher education (HE), though not one that has been met with great success."*
>
> *Cilliers et al., 2012, p. 40*

Learners will wish to maximize their success of passing the assessments and will do anything to pass. In their view, the assessment is what constitutes the curriculum. As a result, learning will be as good as the assessment requires. Many education practices, however, promote poor learning styles. Examples include the rewarding of rote memorization strategies, minimal preparation strategies due to minimal standards or competing exams, or procrastination due to the abundancy of resit opportunities.

Typical of modern programmes is a move beyond the knowledge domain. Many countries in the world have developed competency frameworks. These frameworks have been developed with substantial stakeholder input. What is striking is the commonality across these frameworks: although their descriptions differ, they all emphasize skills such as communication, collaboration and leadership skills, professionalism, reflective abilities, et cetera. Given the overlap across frameworks, there seems to be consensus on what we wish our professionals to be capable of and what skills are needed to improve healthcare. These skills are indeed important because they define success and failure in the labour market. Some therefore call them twenty-first-century skills or 'soft' skills. We choose the term 'domain-independent skills' because they are relevant to any domain of learning, also outside medicine. Embracing these skills in education has major consequences. First, they cannot be taught and tested in a single course. That is, one cannot have a 4-week course on 'communication', administer a test (e.g. an OSCE) and conclude the learner is a good communicator. These skills are learned longitudinally over longer periods of time. They have to be demonstrated in (daily or habitual) performance and are shaped through on-going feedback. Domain-independent skills therefore have to rely heavily on non-standardized assessment using methods from the top of Miller's pyramid (Miller, 1990). Modern training programmes embracing these competency frameworks typically interweave these competencies as continuous 'learning lines' throughout the curriculum. The longitudinal nature of this approach is not easily reconciled with a traditional mastery-oriented assessment approach.

Traditional assessment systems often lack feedback. For economic reasons many assessment practices do not disclose the content of the assessment (the items, for example) to learners. If anything more than pass or fail is communicated this is usually done in the form of grades. Grades represent a very poor form of feedback (Shute, 2008). When assessing domain-independent skills they are more or less useless because they do not provide information on how to improve. By capturing complex skills in a metric or in so-called objective lists of performance, we often trivialize what is being measured, and induce poor learning strategies as a result. Nothing stimulates learning more than high-quality feedback. Our education practices are frequently feedback deprived.

> *"The main goal of formative feedback—whether delivered by a teacher or computer, in the classroom or elsewhere—is to enhance learning, performance, or both, engendering the formation of accurate, targeted conceptualizations and skills."*
>
> *Shute 2008, p. 175*

Finally, traditional approaches to assessment do not reward self-directed learning, which is important for lifelong learning. In a traditional mastery-learning assessment approach there is not much to self-direct: everything is fixed and standardized through the stationary set of assessments.

To recap, traditional approaches to assessment can be characterized as rather reductionist. There is little information in the system, whereas promotion is completely based on discrete and stacked performance decisions (on *minimal* performance) often leading to unwanted educational side effects. This is worrisome because whenever there is a friction between the education goals and assessment goals, the latter tend to prevail. Hence, modern education programmes need a different approach to assessment.

Programmatic assessment

Programmatic assessment is based on a set of assessment principles that are derived from the research on assessment (van der Vleuten et al., 2010). Table 39.1 provides a summary.

A curriculum metaphor may help explain the concept of programmatic assessment. Where in the past a curriculum was the amalgamation of teachers' individual contributions, in modern programmes this is no longer the case. Nowadays, the curriculum is governed by a plan, which is implemented and evaluated to be finally rearranged. One makes a deliberate choice to address a topic over here to return to the topic over there. The sum is more than the whole of its parts. Likewise, programmatic assessment is based on an integral plan of assessment. Deliberate choices are made in terms of methods used and when to use them. The result is a variegated mix of methods that have been selected in accordance with their educational purpose at a particular moment in time and in relation to the assessment programme as a whole. Some will require verbalization, others synthesizing information, writing, reporting, performing, etc. The choices are deliberate because of the purposes they serve in the total programme. Like a curriculum, the assessment programme is evaluated and changes are made when needed. Fundamental to programmatic assessment is the elimination of pass/fail decisions from each individual assessment moment to be reintroduced only when sufficient information is gathered.

Remove pass-fail decision making from individual data points.

Based on the principles of assessment (Table 39.1) and on earlier research on what makes a good assessment programme (Dijkstra et al., 2012), we have formulated the following pillars in programmatic assessment (van der Vleuten et al., 2012):

Each assessment represents but one data point. Recognizing the notion that any single assessment involves a compromise and that no perfect assessment exists that is able to optimize all quality elements of assessment, we regard each assessment as a single data point only.

Each data point is optimized for learning. For a single assessment, no compromise is made on the educational consequences. A single data point is optimized for learning. It is feedback-oriented, rich in nature – either in (profile) scores or words – meaningfully informing the learner, and authentic for the learning task. Sometimes the learning task itself is also assessed. The intent is to support and promote good learning and good learning strategies.

Maximize meaningful feedback to the learner in individual data points.

The choice for a particular method depends entirely on the educational justification of this method in terms of the purpose it serves at a certain moment in time. There are no 'bad' methods. In the past, some methods were removed from the toolkit because they were considered to be too subjective (e.g. the oral exam or the long case). Subjectivity is not a problem if our prime purpose is to give feedback rather than make decisions (and multiple subjective judgements can be robust also; see principle 2, Table 39.1). When more complex skills need to be assessed, professional judgement is indispensable. Such judgements may also come from peers or patients. Agency and authenticity are essential elements in assessment programmes (Harrison et al., 2016). Any method – old or new – that accommodates these elements and that works educationally and meaningfully in the particular education context can be appropriate.

We advocate both course-related assessment *and* longitudinal or continuous assessment. Traditionally, course-related assessments dominate, but when a training programme relies on competency frameworks, attaching weight to the cultivation of domain-independent skills and personal development, more longitudinal assessment is required. Even knowledge can be assessed longitudinally, such as with progress testing (Wrigley et al., 2012).

Assessment consequences represent a continuum of stakes. In programmatic assessment the terms 'formative' and 'summative' are replaced by a continuum of stakes. This continuum ranges from low-stakes to high-stakes. A pass/fail decision is removed from a single data point, making the assessment low-stakes, which is not to be confused with 'no stakes'. The information from a low-stakes data point may be used later in higher-stakes decisions.

Stakes and number of data points are related. The higher the stakes of the decision, the more robust the information that is giving rise to the decision must be. A distinction is made between intermediate and final decisions. Intermediate decisions are made during the training programme, for example once or twice during the year. They are based on a number of data points. A decision can be expressed as a pass or fail or in any other qualification terminology. What is more important than the qualification is that intermediate decisions are also diagnostic (How is the learner progressing?), therapeutic (What remedies are needed?) and prognostic (What might happen with the learner?). Intermediate decisions may be followed by remediation, which is fundamentally different from re-take or re-sit assessments. Remediation is personal, and the learner will have to demonstrate the remediation has been carried out and has been effective. Final decisions are in order when a progression decision is needed (or selection or graduation decision). Final decisions are high-stakes and therefore based on many data points.

Table 39.1 Principles of assessment derived from past research, categorized into standardized and non-standardized assessment (van der Vleuten et al., 2010).

Standardised Assessment. Assessing 'knows', 'knows how', 'shows how' from Miller's pyramid	**Description**	**Practical implications**
1. Competence is context-specific, not generic	Any performance on a test element (item, case, oral, station, patient) is not very predictive of performance on another element, regardless of the method. This has been coined the 'content-specificity problem'. It relates back to the issue of transfer.	• Broadly sample performance across content within each assessment • Combine information across assessment or across time • Avoid high-stakes decision on a single assessment
2. Objectivity is not the same as reliability	Given the content-specificity, problem sampling is the dominant strategy for achieving reproducible test information. Subjective measures can be reliable, objective measures can be unreliable, all depending on the sampling.	• Use holistic professional judgement when it is needed • Use many subjective judgements in combination
3. What is being measured is more determined by the stimulus format than by the response format	The task given to the learner (stimulus format) in a test, much more than the way the response is captured, determines what the test is measuring. Different formats may measure similar or different things all depending on the stimulus format.	• Any method may assess higher-order skills • Produce stimulus formats that are as authentic as possible for the learning task (e.g. scenarios, cases etc.) • Use learning tasks as assessment tasks
4. Validity can be 'built-in'	Quality assurance measures in developing test material have a profound impact on the quality of the test material. Quality assurance can be done prior (e.g. item-writing), during (e.g. good instructions) and post-test (e.g. item and test analysis).	• Organize quality assurance cycles in item and test development • Use peer review • Use psychometric information • Use student input
Non-standardised Assessment. Assessing 'Does'	**Description**	**Practical implications**
5. Bias is an inherent characteristic of professional judgement	Whenever a judgement is made some bias will be introduced. Bias is *not* a reason to not use holistic professional judgement. Professional judgement is indispensable to assessing complex skills. Strategies to reduce bias should be used.	• Use sampling to reduce systematic errors • Use procedural measures of due diligence to reduce unsystematic errors and add to the credibility of the judgement (e.g. committee decisions, multiple cycles of feedback, learner agency in the decision process, etc.)
6. Validity lies in the users of the instruments, more than in the instruments	The seriousness with which the assessment is conducted defines the value of the assessment. Giving and receiving feedback is a skill. The people are therefore important.	• Prepare and train assessors and learners for their role in the assessment • Create working conditions that embed assessment possibilities
7. Qualitative, narrative information carries a lot of weight	For many assessment situations 'words' tell more than '"scores'. This is particularly true for complex skills such as the domain-independent skills.	• Use words for assessing complex skills • Be aware of unwanted side-effects of quantified information
8. Feedback use requires scaffolding	Feedback is often ignored, particularly in summative settings. Feedback use is promoted by the quality of the feedback, the credibility of the source, by reflection and by follow-up.	• Create feedback dialogues • Create feedback follow-up • Create meaningful relations between teacher and learners

Table 39.1 continued

Overall Assessment	Description	Practical implications
9. No single method is perfect	No single assessment method is able to cover all elements of Miller's pyramid. Any individual method will always involve a compromise.	• Vary in use of assessment methods • Combine information from multiple assessment sources
10. Assessment drives learning	The assessment dictates what and how the learner will learn. Any learner will optimize strategies for maximum success in the assessment.	• Verify the effect of assessment on learning • Use the effect strategically to promote desired learning effects

Data points can be compared to pixels in a photograph. Where an individual pixel will not tell you much, a combination of pixels will start to give an image. Sometimes a few pixels suffice to see the image clearly, while in other cases many more pixels are needed to arrive at the image. Information gathering and decision making are purposeful. Information is triangulated. Potential patterns emerge from the information. The stronger the pattern, the clearer the image becomes. Depending on the information, saturation will occur.

High-stakes decisions should be based on many data points; consider that multiple pixels will provide a clearer image.

Learners are guided in feedback use. Feedback use and self-directed learning are promoted by creating a dialogue through social interaction. One powerful way to achieve this is through mentoring. Creating a trusting relationship with a respected person/teacher with whom all assessment and feedback information is shared and discussed is a very effective strategy for using feedback. Learners prepare meetings. They self-direct their learning through analysis of assessment and feedback data and in some cases set learning objectives to follow-up in subsequent assessment of feedback moments. When done well, learners will actually describe the image emerging from the pixels. Mentors ask questions, probe, stimulate deep reflection, discuss remediation activities, and support the learner in any other way possible. Yet, they are not psychotherapists. The focus is on learning and wellbeing, but there are limits to the personal support given, as mentoring is a resource-intensive activity. Alternative social interaction might be possible as well, such as peer groups and buddy systems (e.g. pairing senior with junior learners). The longer the personal relationship lasts the more effective the mentoring will be.

Feedback use and self-directed learning require educational scaffolding through the creation of a dialogue between learner and entrusted teacher (mentor).

Aggregation of assessment through meaningful entities. Taking intermediate and final decisions implies that information must be aggregated. The traditional approach is to aggregate by method. For example, we aggregate information across stations of an OSCE. Yet, a meaningful relationship may not exist between these stations (e.g. history-taking and resuscitation), having us combine apples and oranges. Programmatic assessment, by contrast, aims for aggregation *across* methods to meaningful entities. For example, communication-related information is aggregated not only from the OSCE, but also from a multisource feedback (MSF) assessment and set of mini-CEXs. Having meaningful entities requires some overarching framework. This is often found in outcome systems and competency frameworks. They not only help us structure the curriculum, but also lend meaning to the assessment framework. It is important that instruments are structured according to the overarching framework. If not, meaningful aggregation will be complicated.

Due diligence procedural measures add to the trustworthiness of decision making. High-stakes decisions must be made with full confidence. The decision making on all pooled information cannot be a simple automated process. Often, both quantitative and qualitative information are available and simply averaging is not possible. Drawing inferences based on the 'picture' requires another professional judgement. To make this judgement robust or 'trustworthy' several measures can be taken. The first is to invest an appointed committee with decision-making responsibility, its members being sufficiently independent of learner and mentor. The committee weighs the information and deliberates to arrive at a well-founded decision. To be resource-efficient this appraisal process should be efficiently organized. The lion's share of the learners will not require much time or deliberation, but some cases will. Members of the committee may each prepare some of these cases. Basically, the level of expertise and number of assessors is tailored to the available information. When things are clear, little time investment is required, but when things are not, more investment is needed. The level of independence of the mentor will

bolster the trustworthiness of decision making but also create a firewall dilemma as the mentor is the one who knows the learner best. Yet, if the mentor also decides on progress, his or her relationship with the learner may be jeopardized. A compromise could be to have mentors submit to the decision-making committee a recommendation that has been annotated by the learner. This will increase the agency of the learner and enrich the decision-making process with more information. Yet another alternative is that the mentor does not give a judgement at all, but merely confirms that the evidence presented is authentic for the learner.

Many other procedural measures can be taken to improve the trustworthiness of the decisions. The size of the committee and audit trail of the processes and assessment information will matter, as will the degree to which decisions can be justified. Also, the preceding intermediate decisions will play a role because they reduce the unpredictability of the final decision. External factors are appeal procedures and the use of standards or milestones available. Finally, calibration of the assessors in the committee by training sessions or by discussing exceptional cases afterwards is pertinent. This is only a dip into the wealth of possible measures for due diligence around the decision making. Together they will bolster confidence in essentially what constitutes a human judgement.

> *"In making high-stake decisions based on aggregated information, protection could be provided by installing procedures that surpass the 'power' of the individual assessor."*
>
> *van der Vleuten et al., 2010*

The forenamed seven pillars support programmatic assessment with its purpose of optimizing both the learning and the decision-making aspects of assessment. Learning is promoted by focusing on feedback, learner guidance, growth and development, and self-directed learning. The decision making is robust in that it is based on a multitude of data points, significant expertise of the decision makers and richness in the data. Such information is stronger than a single examination, however big. Finally, programmatic assessment optimizes curriculum evaluation including the assessment system. Mentors will have a thorough impression of what is happening where in the curriculum. As such, they are an excellent source of information for curriculum improvement. In this chapter we have deliberately used the term 'optimization' to refer specifically to the optimization of the whole assessment process, as it can hardly be achieved in a single assessment alone.

An example

To demonstrate how programmatic assessment finds expression, in the following we will offer an example from practice. We will not discuss one of the first programmes, the Cleveland Clinic Lerner College of Medicine programme (Dannefer & Henson, 2007), or the many other practices that have emerged since, both at the undergraduate and postgraduate level and in and outside medicine. Instead, we will describe a programme we have experienced at first hand: the graduate entry programme in medicine at Maastricht University.

This graduate entry programme is a 4-year training programme that leads not only to an MD but also, through a pronounced emphasis on research skills, to a second degree of Master of Science in clinical research. So it is, in fact, a 4-year double-degree programme. The curriculum is structured according to the CanMEDS competencies. Didactically, the first year offers classical PBL, the second year offers a form of PBL based on real patients, the third year contains clinical rotations, and the final year consists of a long-term research participation in the context of a healthcare participation of choice. Students are selected by an MMI-type selection procedure and have at least a bachelor's degree in one of the biomedical sciences. Expectations are high, and students know they have to work hard to be successful.

The assessment programme consists of module-related assessments in the first 2 years and a cross-modular longitudinal assessment. Module-related assessments employ a multitude of different methods: MCQs, open-ended questions, assignments, projects, (mini) OSCEs et cetera. Some modules have a series of mini-tests. In the second year, learners write reports on patient cases and mostly take oral exams that are tailored to their individual experiences. Assessment in the last 2 years is comprised of an elaborate system of work-based assessment using Mini-CEX, OSATS, field notes, and MSF instruments. The longitudinal assessment includes a system of progress testing, which is a kind of final examination in the cognitive domain. It is a written test (200 MCQs) and includes questions from all disciplines and all organ system categories. The test is administered four times per year to all the students in the curriculum. Of course, junior-year students will not be able to answer as many questions as their more seasoned counterparts will, but we do not expect them to. Every 3 months a new test is constructed with new items. For students it is almost impossible to strategically revise for the test, nor do we want them to because anything could be asked. Instead, regular study activity is rewarded because it typically leads to better scores and less stress. For the assessment of the domain-independent CanMEDS competencies, reliance is placed on periodic peer and tutor assessment. The work-based assessment in years 3 and 4 is both module-related and longitudinal. Information is transmitted from one rotation to the other, as in a continuity-of-care approach. Professional behaviour is assessed against end-of-graduation performance.

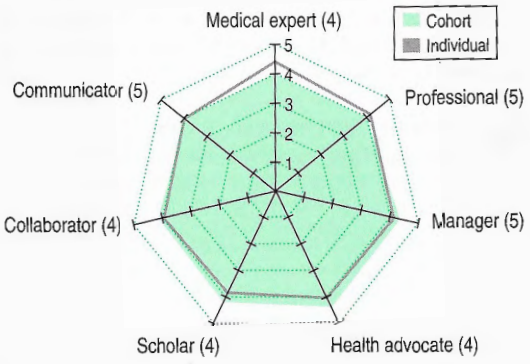

Fig. 39.1 A spider diagram presenting aggregate information from multiple data points on the discrete CanMEDS competencies. Individual performance is related to cohort performance. The number in brackets represents the number of observations.

All assessments are low-stakes; no credit points (European Credit Transfer Accumulation System [ECTS] points in Europe) are awarded to individual assessments. Yet, they are informative, with considerable attention being paid to feedback that may take the form of scores and/or words. For example, learners may review their own performance on an individual progress test or their progress made over the years, as can be gleaned from a series of consecutive tests, online. In doing so, they can select any type of result (e.g. regarding a particular discipline or organ system category). Individual performance is also set against year class performance to enable learners to make thorough analyses of their knowledge base as they progress through the curriculum. The spider diagram in Fig. 39.1 illustrates how feedback is given in years 3 and 4 as part of work-based assessment. In this case, the feedback is presented as a summary of overall performance, while in fact various kinds of graphical overviews can be generated. A student can easily access the underlying individual data points in these graphs and review the original assessment form that reports all quantitative and qualitative information. Finally, summary overviews of narrative information can be generated. All assessment information and all other relevant material are stored in an e-portfolio by the learner or are generated automatically through the e-portfolio assessment services (e.g. in managing an MSF round).

Every learner is assigned a mentor for the entire 4-year study period. This mentor is a regular faculty member overseeing a group of 5 to 10 learners. The mentor has full access to the e-portfolio and has regular meetings with the learners. To prepare these meetings, students write reports reflecting on their progress, and formulate and follow up on study plans or learning objectives based on the evidence included in the portfolio. At the end of the year, the mentor writes a recommendation for promotion to the next year, which is supplemented with a second recommendation put forward by all mentors collectively. In this latter recommendation, however, the learner's personal mentor has no say. The final decision on promotion is taken by an independent portfolio assessment committee. In the event a positive judgement is passed, the student is awarded all credit points (in our case, 60 ECTS per year).

This programmatic approach to assessment has served the programme well. Students have become real feedback seekers and are self-regulating their learning. Results on the progress tests are impressive compared to those of other medical training programmes (Heeneman et al., 2016). Teachers enjoy working with these students. Mentorship is considered a most rewarding role. Many learners publish their research work and some 50% of graduates pursue a PhD. The graduate entry programme has a yearly student intake of 50. We have currently implemented a similar programmatic approach in the much larger undergraduate medical training programme that accepts 340 students per year.

Implementing programmatic assessment

Though programmatic assessment is grounded in research, research has yet to demonstrate if and why it is successful. This research is ongoing. A number of findings have emerged from this research and our experiences so far.

- The transition from a traditional system to one that is based on programmatic assessment is a major change comparable to moving from a traditional curriculum to a PBL curriculum. It requires a different mindset of teachers and learners. The prerogative of the individual teacher to fail a learner in the course where the responsibility rests with the teacher is a deeply rooted tradition. Many universities have university-wide assessment regulations and grading systems in place that can often undermine change. Like in any other major education change, good change management is foundational. More specifically, effective change requires appropriate top-down and bottom-up strategies, which, in turn, calls for effective leadership. Faculty training is equally important. Training faculty on the job and just in time can help them make this change effectively. Exposure to programmes that have successfully made such change or following a course on programmatic assessment may accelerate the process. Teachers learn in the same way as our learners do, so all our knowledge on how to facilitate learning equally applies to them. It will not work to just give them the information, as was the case with our learners. Just like PBL can have many manifestations, so might programmatic

assessment. Some elements that are easier to implement, or that address a certain need in a particular organization, might create effective hybrids.

Programmatic assessment is a major innovation that requires a change management policy.

Getting the quality of the feedback right is what poses a challenge in programmatic assessment. The provision of feedback requires time, which is always lacking. Moreover, it is a skill that can be learned and therefore it may help to train learners and teachers alike. Similarly, feedback on feedback may help, as may the application of modern tools (apps) and software that facilitate feedback.

A recurrent finding is that learners do not perceive low-stakes assessment as low-stakes as intended. Programmatic assessment also involves a culture change. Any measure that reduces the 'stick' effect of the assessment is appropriate. This starts with good communication with all stakeholders about the purpose of low-stakes assessment with respect to learners. Some schools that have introduced programmatic assessment include re-take assessment (we have some in our programme), which inherently raises the stakes of the assessment.

Programmatic assessment can be applied to any part of the medical training continuum, although it seems to fit in most naturally with education that stresses experiential learning. This fit can probably be explained by the emphasis that is placed in experiential learning on non-cognitive and complex skills.

As may be clear, we have not used the standard language that often dominates the assessment discourse. Programmatic assessment introduces a take on assessment that differs from the more conventional psychometric view. This does not mean we are antagonistic to psychometric analysis of the assessments being used. As principle 4 in Table 39.1 has demonstrated, quality assurance is integral to good assessment. Psychometricians sometimes play an important role in serving that purpose, but they consider one perspective only. There are at least two other perspectives to programmatic assessment. One concerns education. The purpose is to train qualified professionals in an educationally optimal way. What becomes essential, then, is to foster engagement in learners, which can be achieved by giving them challenging tasks, autonomy, gratifying social relations and personal guidance. Consequently, learning will drive assessment, not the other way around, precisely how it should be. A second perspective is about qualitative inquiry (Govaerts & van der Vleuten, 2013). Qualitative inquiry has its own methodological conventions to deal with complexity. We have used many words that stem from this methodology. One could actually say that programmatic assessment is a mixed-method inquiry into a learner's achievements in a training programme using a multitude of quantitative and qualitative sources. It involves giving meaning to an individual person's development in an academic trajectory.

> *"We should aim for careful balancing of quantitative and qualitative approaches in our assessment programmes, justifying our choices on the basis of assessment purposes as well as conceptualisations of learning and performance/competence."*
>
> *Govaerts & van der Vleuten, 2013*

Summary

Programmatic assessment considers assessment as an optimization problem. For individual assessments, learning is optimized by selecting a method of assessment that is maximally aligned with the learning task and that provides meaningful feedback to the learner. Decision making is optimized by gathering rich information across many assessment moments in such a way that solid conclusions can be made about the progress of a learner. When implemented well, programmatic assessment may bring substantial benefits. The mushrooming of competency-based education makes it imperative that assessment be closely aligned with education, and we would argue that programmatic assessment can help achieve just that.

References

Cilliers, F.J., Schuwirth, L.W., Herman, N., et al., 2012. A model of the pre-assessment learning effects of summative assessment in medical education. Adv. Health Sci. Educ. Theory Pract. 17, 39–53.

Dannefer, E.F., Henson, L.C., 2007. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. Acad. Med. 82, 493–502.

Dijkstra, J., Galbraith, R., Hodges, B.D., et al., 2012. Expert validation of fit-for-purpose guidelines for designing programmes of assessment. BMC Med. Educ. 12, 20.

Govaerts, M.J.B., van der Vleuten, C.P.M., 2013. Validity in work-based assessment: expanding our horizons. Med. Educ. 47, 1164–1174.

Harrison, C.J., Könings, K.D., Dannefer, E.F., et al., 2016. Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. Persp. on Med. Educ. 5, 276–284.

Heeneman, S., Schut, S., Donkers, J., et al., 2016. Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment. Med. Teach. 1–9.

Miller, G.E., 1990. The Assessment of clinical skills/competence/performance. Acad. Med. 65, S63–S67.

Shute, V.J., 2008. Focus on formative feedback. Rev. Educ. Res. 78, 153–189.

van der Vleuten, C.P., Schuwirth, L.W., Driessen, E.W., et al., 2012. A model for programmatic assessment fit for purpose. Med. Teach. 34, 205–214.

van der Vleuten, C.P., Schuwirth, L.W., Scheele, F., et al., 2010. The assessment of professional competence: building blocks for theory development. Best Pract. Res. Clin. Obstet. Gynaecol. 24, 703–719.

van der Vleuten, C.P.M., 1996. The assessment of professional competence: developments, research and practical implications. Adv. Health Sci. Educ. Theory Pract. 1, 41–67.

Vandewaetere, M., Manhaeve, D., Aertgeerts, B., et al., 2015. 4C/ID in medical education: How to design an educational program based on whole-task learning: AMEE Guide No. 93. Med. Teach. 37, 4–20.

Wrigley, W., van der Vleuten, C.P., Freeman, A., Muijtjens, A., 2012. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. Med. Teach. 34, 683–697.