2012; 34: 205–214 MEDICAL TEACHER

A model for programmatic assessment fit for purpose

C. P. M. VAN DER VLEUTEN[1], L. W. T. SCHUWIRTH[2], E. W. DRIESSEN[1], J. DIJKSTRA[1], D. TIGELAAR[3], L. K. J. BAARTMAN[4] & J. VAN TARTWIJK[5]

[1]Maastricht University, The Netherlands, [2]Flinders Medical School, Australia, [3]Leiden University Graduate School of Teaching, The Netherlands, [4]Utrecht University of Applied Sciences, The Netherlands, [5]Utrecht University, The Netherlands

Abstract

We propose a model for programmatic assessment in action, which simultaneously optimises assessment for learning and assessment for decision making about learner progress. This model is based on a set of assessment principles that are interpreted from empirical research. It specifies cycles of training, assessment and learner support activities that are complemented by intermediate and final moments of evaluation on aggregated assessment data points. A key principle is that individual data points are maximised for learning and feedback value, whereas high-stake decisions are based on the aggregation of many data points. Expert judgement plays an important role in the programme. Fundamental is the notion of sampling and bias reduction to deal with the inevitable subjectivity of this type of judgement. Bias reduction is further sought in procedural assessment strategies derived from criteria for qualitative research. We discuss a number of challenges and opportunities around the proposed model. One of its prime virtues is that it enables assessment to move, beyond the dominant psychometric discourse with its focus on individual instruments, towards a systems approach to assessment design underpinned by empirically grounded theory.

Introduction

In 2005, we made a plea for adopting a programmatic approach in thinking about assessment in education (Van der Vleuten & Schuwirth 2005). We described a programme of assessment as an arrangement of assessment methods planned to optimise its fitness for purpose. Fitness for purpose is a functional definition of quality, the essence of which is the notion of contributing to the achievement of the purposes of the assessment programme. Fitness for purpose is thus an inclusive notion of quality, encompassing other quality definitions (e.g. zero defects) which are interpreted as purpose (Harvey & Green 1993). With overall quality in mind, we advocated that an assessment programme should be constructed deliberately, its elements should be accounted for, it should be centrally governed in its implementation and execution and it should be regularly evaluated and adapted. Analogous to the now generally accepted view that a good test is more than a random set of good quality items, a good programme of assessment is more than a random set of good instruments (Schuwirth & Van der Vleuten 2011). The problem of programmatic assessment extends even beyond this analogy. For, whereas good quality items are achievable, there is no such thing as an ideal instrument. As early as 1996, we contended that any single assessment implies a compromise on quality criteria (Van der Vleuten 1996). The choice on which criterion(s) to compromise should be based on a well-considered decision as to which quality element is to be optimised on the specific assessment context. A programme of assessment, combining different assessments, can alleviate the compromises on individual methods, thereby rendering the total more than the sum of its parts.

> **Practice points**
>
> - Good assessment requires a programmatic approach in a deliberate and arranged set of longitudinal assessment activities.
> - A model of programmatic assessment is possible that optimises the learning and certification function of assessment.
> - Individual data points in the assessment programme are maximally informative to the learning.
> - Aggregated data points are used for higher stake pass/fail and remediation decisions; the higher the stakes in the assessment decision the more data points are needed.
> - Expert professional judgement in assessment is imperative and requires new approaches to deal with biases.

Since the first introduction of the notion of programmatic assessment, further work has been done to define and assess the quality criteria for assessment programmes (Baartman et al. 2006, 2007). On a different strand, work is going on in the area of designing guidelines. Recently, this has resulted in a published framework for structuring such guidelines (Dijkstra et al. 2010) followed by a study in which concrete guidelines are formulated (Dijkstra et al. Under editorial review). Notwithstanding the importance of these theoretical developments, it remains hard to imagine how such recommendations

Correspondence: C. van der Vleuten, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel: +31433885725; fax: +31433885779; email: c.vandervleuten@maastrichtuniversity.nl

ISSN 0142–159X print/ISSN 1466–187X online/12/030205–10 © 2012 Informa UK Ltd.
DOI: 10.3109/0142159X.2012.652239

could be translated into a concrete assessment programme in action that is in alignment with defensible theoretical underpinnings. The link that is still missing today is a theory-based framework or generic model that offers concrete recommendations for structuring an assessment programme in line with Dijkstra's model so as to maximise its fitness for purpose. The purpose of this article is to present the outlines of such a model.

The proposed model is limited to programmatic assessment in the educational context, and consequently licensing assessment programmes are not considered. The model is generic with respect to types of learning programmes, which may be 'school based', emphasising classroom teaching, or 'work based', such as postgraduate specialty training programmes. We do assume, however, that the learning programme is learner centred, favouring holistic approaches to learning (as opposed to atomistic mastery-oriented learning) and deep learning strategies. An assessment model for a predominantly mastery-oriented learning programme would probably differ from our model, although this does not preclude the inclusion in our model of tasks requiring mastery-oriented learning and assessment. We define three fundamental purposes that should be united within an assessment programme that fits our model: a programme that maximally facilitates learning (assessment *for* learning); a programme that maximises the robustness of high-stake decisions (on promotion/selection of learners); a programme that provides information for improving instruction and the curriculum. For the moment, we will park the third purpose to return to it briefly in the discussion. Our main focus for now is a theory-based model (Schuwirth et al. 2011) designed to achieve optimisation of the first two purposes. In order to motivate the choices we have made in creating this model, we first present some theoretical principles of assessment based on empirical research or, more accurately, on our interpretation of that research. We deliberately keep this account short, as a fuller account of most of these principles can be found elsewhere (Van der Vleuten et al. 2010).

Principles of assessment

(1) Any single assessment data point is flawed

Single-shot assessments, such as a single administration of an assessment method at any one level of Miller's (1990) pyramid, in other words, all point measurements are intrinsically limited. Due to content specificity (Eva 2003), the performance of individuals is highly context dependent, requiring large samples of test items (in the broadest sense of the term) and long testing times to produce minimally reliable results (Van der Vleuten & Schuwirth 2005). Profile scores are inherently less reliable. However, there are more characteristics to optimise than reliability. One single method can only assess a part of Miller's pyramid and there is no magic bullet that can do it all in one go. A one-off measure will also not be able to establish change or growth. This limitation of single data points of assessment drives, legitimises and informs our thinking about programmes of assessment.

(2) Standardised assessment can have validity 'built-in' the instrument

All methods that can be standardised (the first three levels of Miller's pyramid, assessing knows, knows how and shows how) can have validity built into the test instrument by careful construction of content and scoring and administration procedures. Quality control procedures around test construction can have a dramatic effect on the quality of the test material (Verhoeven et al. 1999; Jozefowicz et al. 2002). If applicable, assessors can be trained, scoring lists objectified, simulated patients standardised, etc. Through careful preparation, the validity of the instrument can be optimally enhanced. For virtually all assessment methods, best practice technology is available.

(3) Validity of non-standardised assessment resides in the users and not so much in the instruments

A complete assessment programme will inevitably also have to employ non-standardised methods. Particularly, if we wish to assess in real practice, i.e. at the top of Miller's pyramid (the 'does' level), standardisation is out of reach. The real world is non-standardised and haphazard, and, more importantly, any attempt at standardisation will only trivialise the assessment (Norman et al. 1991). In the assessment literature, we are currently seeing the development of 'technologies' for assessing the 'does' level of performance, for example in the field of work-based assessment (Norcini 2003; Norcini & Burch 2007). However, assessment in regular educational settings (e.g. classroom, tutorials and laboratory) also comes under the same category of assessment of habitual performance. Examples are assessment of a presentation or assessment of professional behaviour. It is typically not 'standardised forms' that determine the validity of the assessment in such situations (Hodges et al. 2011). The users, i.e. the assessors, learners and patients, are more important than the instrument. Their expertise in using the instrument, the extent to which they take the assessment seriously and the time they can spend on it, these aspects together determine whether or not the assessment is performed well. While extensive training is not required for someone handing out multiple choice test booklets to students, with non-standardised observational assessment it is of crucial importance that all those involved in the assessment process should receive extensive training. The extent to which the users take their assessment task seriously, as reflected in their taking time to give feedback or record a narrative on a form, ultimately determines the utility of these methods. Ensuring that the users have a proper understanding of their roles requires training, facilitation, feedback, expertise development, etc (Govaerts et al. 2007). Since an assessment programme without non-standardised methods is unthinkable, we need to develop a 'technology' to help users to function appropriately in their assessment role. In doing so, we need to realise that someone who learns is a learner, even if most of the time they are assessors, teachers or supervisors. All people learn in the same way, preferably by training, practice and feedback. It will not suffice to simply provide assessors with information or instruments. If the users, assessors and assesses do not fully understand the meaning and purpose of the assessment, the assessment is doomed to be trivialised.

(4) The stakes of the assessment should be seen as a continuum with a proportional relationship between increases in stakes and number of data points involved

From the perspective of a conceptual framework of programmatic assessment, the formative–summative distinction is not a very useful one, considering that the framework predicates that any assessment should be both formative and summative, only to varying degrees. Therefore, conceptualising the stakes of the assessment as a continuum from low to high stakes seems more useful. In low-stake assessment the results have limited consequences for the learner in terms of promotion, selection or certification, whereas high-stake assessment can have far-reaching and dramatic consequences. In a programme of assessment, only low-stake decisions can be based on single data points, whereas all high-stake decisions require input from many. With higher stake assessment, the role of the teacher as helper is more easily compromised. Combining the roles of helper and judge (in high-stake decisions) confronts teachers with a conflict of interest (Cavalcanti & Detsky 2011). A conflict that is aggravated as the stakes increase, and which can easily lead to inflation of judgement (Dudek et al. 2005; Govaerts et al. 2007), with the concomitant risk of trivialisation of the assessment process. However, when high-stake decision making is informed by many data points, it would be foolish to ignore the information from the rich material derived from all the single data points. Information from combined low-stake assessments should therefore feed into high-stake information. However low stake an individual data point may be, it is never zero stake.

(5) Assessment drives learning

This is a generally accepted concept in the assessment literature, but at the same time it remains poorly understood. In all likelihood, many assessments drive undesirable learning strategies because the assessment is not at all or ill aligned with curriculum objectives. This situation is particularly common in poor information, purely summative systems (Al Kadri et al. 2009). We need more theoretical clarification as to why and how assessment drives learning, and research on this is emerging (Cilliers et al. 2010, 2011). The objective is to have assessment drive learning in a desirable direction and foster deep-learning approaches (but mastery-learning too wherever appropriate). There is a wealth of evidence that formative feedback can enhance learning (Kluger & DeNisi 1996; Hattie & Timperley 2007; Shute 2008). We note that, if assessment is to drive learning, it is imperative that it should produce meaningful information to the learner. In other words, assessment information should be as rich as possible. Information can be rich in many different ways, both quantitatively and qualitatively. At this point, we should note that assessment is often associated with grades (only), and that grades are one of the poorest forms of feedback (Shute 2008). Different types of quantitative information are needed, such as profile scores and reference performance information. However, we also note the importance of qualitative information. Narrative information is a powerful tool for qualitative feedback and can contribute substantially to the meaningfulness of the information (Sargeant et al. 2010). We finally note that feedback seeking and giving are skills (Sluijsmans et al. 2003) that need to be developed, a notion that is in agreement with our previous point emphasising the need to invest in the users of assessment.

Lack of meaningfulness leads to trivialisation, a serious and frequent hazard in assessment. If learners are required to memorise checklists for passing the objective structured clinical examination (OSCE) but have no connection with patients, their performance is trivial; if an assessor completes all items on a professional behaviour rating form by one strike of the pen, the assessment loses all meaning and is trivialised. However, if the assessment information is meaningful, learning will be enhanced in a meaningful way. We argue that low-stake individual data points should be as meaningful as possible to foster learning, and we also argue that high-stake decisions should be based on many individual data points. Aggregation of meaningful data points can result in a meaningful high-stake decision. In all elements of the assessment programme we should be on our guard against trivialisation.

There is one exception where individual data points can be high stake. This is when the learning task is a mastery task (i.e. the tables of multiplication for children, resuscitation for medical students). Mastery tasks need to be certified as and when they occur in the programme. The proposed model should accommodate this exception. This does not imply, however, that mastery tasks do not require feedback.

(6) Expert judgement is imperative

Competence is a complex phenomenon. Regardless of whether it is defined in terms of traits (knowledge, skills, problem-solving skills and attitudes) or competencies or competency domains (Frank and Danoff 2007; Accreditation Council for Graduate Medical Education [ACGME] 2009), interpreting assessment results always requires human judgement. By providing support, e.g. scoring rubrics, training and performance standards, we can reduce the subjectivity in judgements (Malini Reddy & Andrade 2010), but if we try to achieve complete objectification, we will only trivialise the assessment process (see the examples of principle 5). We have no choice but to rely on the expert judgements of knowledgeable individuals at various points in the assessment process. We also need expert judgement to combine information across individual data points. Often, we use quantitative strategies to aggregate information sources (averaging scores and counting the number of passes), but when individual data points are information-rich, and particularly when they contain qualitative information, simple quantitative aggregation is out of the question and we have to resort to expert judgement. From a vast amount of literature on decision making, we know that the human mind is nothing if not fallible, compared to actuarial decision making (Shanteau 1992). We argue, however, that random bias in judgement can be overcome by smart sampling strategies and systematic bias by procedural measures. The sampling perspective has been proven to be effective in many types of assessment situations (Van der Vleuten et al. 1991; Williams et al. 2003; Eva et al. 2004): we can produce reliable information simply by using

Table 1. Illustrations of potential assessment strategies related to qualitative research methodologies for making robust assessment decisions.

Strategies to establish trustworthiness	Criteria	Potential assessment strategy
Credibility	Prolonged engagement	Train assessors
		People who know that the learner best (coach, peers) provides information for assessment
		Incorporate intermittent feedback cycles in the procedure
	Triangulation	Involve many assessors and different credible groups
		Use multiple sources of assessment within or across methods
		Organise a sequential judgement procedure where conflicting information necessitates the gathering of more information
	Peer examination (sometimes called peer debriefing)	Assessors talk about benchmarking, the assessment process and results before and halfway an activity
		Separate assessors' multiple roles by removing summative assessment decisions from the coaching role
	Member checking	Incorporate the learner's point of view in the assessment procedure
		Incorporate intermittent feedback cycles
	Structural coherence	Assessment committee discusses inconsistencies in the assessment data
Transferability	Time sampling	Sample broadly over different contexts and patients
	Thick description (or Dense description)	Assessment instruments facilitate inclusion of qualitative, narrative information
		Give narrative information a lot of weight in the assessment procedure
Dependability	Stepwise replication	Sample broadly over different assessors
Dependability/ confirmability	Audit	Document the different steps in the assessment process (a formal assessment plan approved by an examination board, overviews of the results per phase)
		Quality assessment procedures with external auditor
		Learners can appeal the assessment decision

many judgements. In fact, assessment methods that rely heavily on judgement require considerably smaller samples than are required for most objectified and standardised methods (Van der Vleuten et al. 2010). Bias is difficult to prevent, but we argue that systematic biases can be ameliorated by putting in place appropriate procedural measures around decision making. A decision on a borderline candidate, for example, will require much scrutiny of the information gathering process, and perhaps even more data gathering and more deliberation on the additional information. In a recent paper, we proposed that methodologies from qualitative research could serve as inspiration for the development of procedural measures in assessment (Van der Vleuten et al. 2010). The example we just gave stems from the triangulation criterion. Another criterion, member checking, would suggest incorporating the learner's view in the assessment procedure. Table 1 provides an overview of such procedural strategies. Depending on the care taken in creating and conducting these procedures, biases can be reduced and the resulting decisions will be more trustworthy and defensible. We think these strategies can handle subjective information (combined with objective information) and fortify the robustness of the resulting decisions. This obviates the need to objectify every part of the assessment programme, which, as we have noted earlier, will only lead us to reductionism and trivialisation of both assessment and learning.

Model of programmatic assessment in action

Based on the above principles, we propose a model that is optimised for fitness of purpose. The purpose of an assessment programme is to maximise assessment for learning while at the same time arriving at robust decisions about learners' progress. Figure 1 provides a graphical representation of the model. We will describe its elements systematically and provide arguments for its coherence. In the model, we make a distinction between training activities, assessment activities and learner support activities as a function of the time in the ongoing curriculum.

Learning activities

We start with a first period of training activities consisting of *learning tasks* denoted by small circles (after the 4C-ID model (Van Merriënboer 1997)). A learning task can be anything that leads to learning: a lecture, a practical, a patient encounter, an operation in the hospital operating theatre, a problem-based learning (PBL) tutorial, a project, a learning assignment or self-study. When arranged appropriately, these learning tasks in themselves provide a coherent programme or curriculum constructed in accordance with the principles of instructional design (Harden et al. 1984; Van Merriënboer & Kirschner 2007). Some learning tasks may yield *artefacts of learning,* as denoted by the larger circles. These artefacts can be outcome related, such as a project report, or they can be process oriented, such as a list of surgical procedures performed in the operating theatre.

Assessment activities

The assessment activities in period 1 are shown as small pyramids, each representing a *single data point of assessment.* This symbolic shape is deliberately chosen, because each

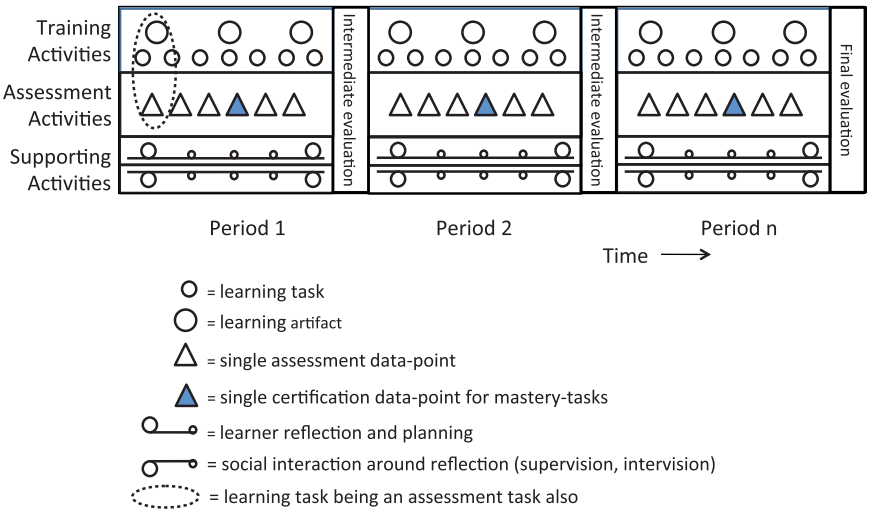

Figure 1. Model for programmatic assessment in action fit for the purpose of assessment for learning and robust decision making on learners' achievements, selection and promotion.

single data point can relate to any method at any layer of Miller's pyramid, be it a written test, an OSCE, an observation of a clinical encounter (i.e. Mini-CEX), a peer evaluation in a PBL tutorial assessment, etc. Some of these assessments are evaluations of artefacts resulting from learning tasks. Examples are the assessment of a patient information leaflet produced by a learner or the evaluation of a presentation on a research report (denoted by the dashed ellipse). All assessment activities should be arranged so as to maximally support the learner's ongoing learning to ensure adherence to principle 3 (assessment drives learning). This principle requires that all assessment be maximally meaningful to learning and provide feedback on the learner's performance that is information-rich, whether quantitatively or qualitatively. The information is documented, i.e. physically or electronically traceable. Each single data point is low stake (principle 5). Although performance feedback obviously provides information in relation to some kind of performance standard, we strongly caution against passing or failing a learner based on one assessment point, as can be done in a mastery test. Each data point is but one element in a longitudinal array of data points (principle 1). Although single data points are low stake, this does not preclude their use for progress decisions at a later point in the curriculum. With each single assessment, the principal task of the assessor is to provide the learner with as rich and extensive feedback as possible. It is not useful to simply declare whether or not someone has achieved a certain standard. Assessors are protected in their role as teacher or facilitator, but not in their role as judge (principle 5). Both roles are disentangled as much as possible, although, obviously, any assessor will judge whether or not the learner did well. There is one exception, which is represented by the black pyramid. Some tasks are mastery oriented and require demonstration of mastery. For example, resuscitation is a skill that needs to be drilled until mastery is achieved. In the same way, a postgraduate trainee may have to be certified on laparoscopic surgical skill performance on the simulator before being allowed to perform a procedure on a patient. Nevertheless, most assessment tasks are not mastery oriented but developmental in terms of working towards proficiency in a competency. We similarly warn against grades as the only feedback that is given. Grades are poor feedback carriers and tend to have all kinds of adverse educational side effects (learners hunting for grades but ignoring what and how they have learned; teachers being content to use the supposed objectivity of grades as an excuse for not giving performance feedback). We advocate applying all assessment technology in accordance with our assessment principles 2 and 3. We should 'sharpen' the instruments and/or people as much as possible. We are agnostic with respect to any preference for specific assessment methods, since any assessment approach may have utility depending on its function within the programme. We explicitly do not exclude subjective information or judgements from experts (principle 6). The designation 'expert' is defined flexibly and can apply to any knowledgeable individual. Depending on the context, this may be the teacher, the tutor, the supervisor, the peer, the patient and, last but not least, the learner him or herself. Granted that self-assessment should never stand alone (Eva & Regehr 2005), in many cases, the learner can be a knowledgeable source of expertise. In summary, all activities in the assessment programme conducted during a given period of the training programme should present meaningful and traceable data points of learner performance which are maximally connected to the learning programme and reinforce desirable learning behaviours.

Supporting activities

The supporting activities in the same period are twofold. First, the learner reflects on the information obtained from the learning and assessment activities (principles 4 and 6 combined). This is shown as underscored connected small circles. There may be more *reflective activity* at the start and at the end, but *self-directed learning activity* is continuous. Feedback is interpreted and used to plan new learning tasks or goals (Van Merriënboer & Sluijsmans 2009). From the

literature, we know how hard it mostly is to get people to reflect and self-direct (Korthagen et al. 2001; Driessen et al. 2007; Mansvelder-Longayroux et al. 2007). One of the paradoxes of self-directed learning is that it takes considerable external direction and scaffolding to make it useful (Sargeant et al. 2008; Driessen et al. 2010). We therefore propose scaffolding of self-directed learning with some sort of social interaction. In the model this is the bottom rectangle with circles connected to it at the opposite ends. The principal form of support for self-directed learning is coaching or mentoring (supervision activities), but alternatively, support can be provided by more senior learners or peers ('intervision' activities). This process can also be facilitated by dedicated instruments in which reflective activity is structured (with respect to time, content and social interaction) and documented (Embo et al. 2010). In general, we encourage documentation of the reflective process, but warn against overdoing it. Documented reflective activities will only work if they are 'lean and mean' and have direct meaningful learning value (Driessen et al. 2007). Otherwise, they are just bureaucratic chores, producing reams of paper for the rubbish bin. This type of trivialisation can be avoided if we keep firmly in mind that social interaction is prerequisite to lend meaningfulness to reflective activities.

Intermediate evaluation

At the end of the period, all artefacts, assessment information and (selected) information from the supporting activities are assessed in an intermediate evaluation of progress. The aggregate information across all data points is held against a performance standard by an independent and authoritative group of assessors, i.e. a committee of examiners. We think a committee is appropriate because expert judgement is imperative for aggregating information across all data points (principle 6). We do not wish to downplay the virtues of numerical aggregation of information and we should use it whenever appropriate and possible. In one of our programmes at Maastricht, for example, we use an online performance database of progress testing, which can flexibly aggregate across an infinite number of comparisons and predict future performance based on past performance (Muijtjens et al. 2010). However, some data points are narrative and qualitative, necessitating human interpretation of information (like a patient chart! principle 6). Data points should preferably be aggregated across meaningful entities. Traditionally, these entities have been methods (or layers of Miller's pyramid), but other, more meaningful aggregation categories are thinkable, such as the themes of the training programme or a competency framework (Schuwirth & Van der Vleuten 2011). We are obviously in favour of measures that enhance the robustness of this evaluation. The committee consists of experts, knowledgeable in terms of what they have to assess. They are trained, perhaps even certified, and use supporting tools such as rubrics and performance standards. They learn as their experience accumulates and can change the procedures and supporting tools. The committee's size matters as well as the extent of its deliberations. For most learners, the assessment process will be fast and efficient depending on the consistency and level of the information from the single data points. For some learners, however, the committee will have to engage in substantial debate, deliberation and argumentation. Their decision is informative in relation to the performance standard, but also informative in its diagnostic, therapeutic and prognostic value. The experts provide information on areas of strength and improvement (diagnosis), and they may suggest remediation to help the learner achieve desirable performance objectives (therapy) and predict certain performance outcomes later in the training programme (prognosis). Very importantly, this intermediate assessment is remediation oriented. This is very different from conventional types of assessment, which are typically mastery-oriented: if mastery is not achieved, the learner simply has to re-do the course and be re-assessed. Our approach is first and foremost developmental: we propose an information-rich recommendation for further learning, tailored to the individual learner and contingent on the diagnostic information. The committee's assessment can be qualified as intermediate stake. Although the assessment information has no dramatic consequences for the learner's survival in the learning programme, the information it provides is not to be ignored and the learner should use it to plan further learning activities.

The intermediate evaluation poses a *firewall dilemma,* which can be resolved in multiple ways. The dilemma is posed by the actors' input into the support system. According to the criterion of prolonged engagement (Table 1), a coach, mentor or learner provides the richest information. At the same time by vesting the power of decision making in the actors of the support system, the relationship between helper and learner can be compromised (Cavalcanti & Detsky 2011). One rigorous way of resolving this is to erect an impenetrable firewall between activities of support and activities of decision making. However, this would mean that the committee remains oblivious of valuable information, it would likely lead to more work for the examiners and potentially more bias and higher costs. Intermediate solutions are equally possible. One protective approach is to require the coach to authenticate the information from the learner: a declaration that the information provides a valid picture of the learner. One step further: the coach may be asked to make a recommendation on the performance decision, which can be amended by the learner. To sum up, there is no single best strategy to resolve the firewall dilemma and compromises are in order depending on the available resources, argumentation, sentiments, culture and the stakes involved (Van Tartwijk & Driessen 2009).

We have presented a first cycle consisting of training, assessment and supporting activities. This cycle can be repeated indefinitely. The number of cycles depends on the nature of the training programme and the availability of resources. The fact that the model shows three cycles is of no significance. The three cycles could represent the first year of a medical school. Each period could actually comprise multiple courses. The logical longitudinal development of the learner through learning tasks, appropriate feedback and (supported) self-direction is of key importance. This is entirely the opposite of a purely mastery-oriented approach where passing an exam means being declared competent for life. It is also important

that sufficient data points and remediation moments should have occurred before a final high-stake decision is made.

Final evaluation

After an appropriate number of cycles, a final evaluation takes place at a moment when a decision on the learner's progress is in order. This is a high-stake decision with major consequences for the learner. The decision is taken by the same committee of examiners that conducted the intermediate evaluation (prolonged engagement) but with even more stringent procedural safeguards in so far as these are feasible. Examples are procedures of appeal, procedures of learner and coach input (firewall dilemma), training and benchmarking of examiners, committee size, extent of deliberation and documentation, performance standards and/or rubrics, quality improvement measures for the evaluation procedure as a whole and, last but by no means least, the inclusion of all data points from the preceding period including the intermediate evaluations (principle 5).

Ideally, the decision is motivated by a justification. The decision may not be limited to a mere pass or fail, but also indicate distinctive excellence of performance. One should note here that more performance classifications (i.e. grades) do not only augment the subtlety of judgement but also the risk of classification error and judgemental headache. If the system works well, outcome decisions will come as no surprise to the learner (or coach). In a minority of cases, the decision will belie the learner's expectations and their frequency of this occurrence validates the existence of the committee. Depending on the nature of the progress decision, the committee may provide recommendations for further training or remediation. Overall, the final decision is robust and based on rich information and numerous data points (principle 6). The robustness lies in the trustworthiness of the decision. If the decision is challenged, it should be accountable and defensible, even in a court of law.

The model in Figure 1 depicts a certain learning period, ending with a natural moment of decision making over learner promotion. It does not represent a curriculum in its entirety. Depending on the curriculum, the learning period in the model can be repeated in as many cycles as are appropriate to complete the curriculum. The cycles do not have to be of equal length: the number and length of the cycles depend on the nature of the curriculum and the natural decision moments therein.

Discussion

We think our proposed model is optimally fit for purpose. It consistently optimises learning value across the assessment programme. No compromises are made on the meaningfulness of the data in the assessment programme. At the same time, high-stake decision making is robust and credible, providing internal and external (societal) accountability for the quality of graduating learners. As we said in the introduction, the third purpose of an assessment programme is to evaluate the curriculum. Information from the supporting actors, such as mentors/coaches, and information from the actors in the intermediate and final evaluation offer excellent data points for curriculum evaluation in terms of both the process and the outcomes of education and training.

We have taken care to formulate the model in the most generic terms possible. Some may conclude that what we describe is portfolio learning and portfolio assessment. We have, however, deliberately avoided making any suggestions for specific assessment methods or showing any preference for specific methods. Our purpose here was to theorise beyond a single assessment method approach. Our model is informed by extensive previous research in assessment and brings together strategies from various theoretical strands crossing the boundaries of the quantitative and qualitative discourse (Hodges 2006; Hodges et al. 2011). It also reinstates the value of expert professional judgement as an irreplaceable and valuable source of information (Coles 2002). We will finish with describing some challenges and opportunities of the model we have presented.

Challenges

An obvious first challenge of the suggested programmatic approach is the *cost* and *resources* needed for running such a programme. Our first remark here is that, in keeping costs down, it is wiser to do fewer things well than to do many things badly (the 'less is more' principle). There is no point in gathering a vast amount of data that provides little information; it would only be a waste of time, effort and money. A second remark is that, in our programmatic approach, the boundaries between assessment and learning activities are blurred. The ongoing assessment activities are very much part and parcel of the learning programme, indeed they are inextricably embedded in it (Wilson & Sloane 2000). Third, economic compromises can and must be made. Some of the assessment activities, particularly low-stake ones, can be done well at low cost. For example, an online item bank would enable students to self-assess their knowledge in a certain domain. Furthermore, the sharing of test materials across schools is a smart strategy, as we have pointed out earlier (Van der Vleuten et al. 2004). Certain professional qualities, like professionalism or communication, lend themselves very well to peer assessment (Falchikov & Goldfinch 2000). It is also thinkable that compromises are made on certain elements of the model or in certain periods in the curriculum, depending on the balance between stakes and resources. For example, mentoring or coaching could be done in certain parts of the curriculum but not in others. And finally, a quote attributed to McIntyre and Bok seems appropriate here: 'If you think education is expensive, try ignorance'.

A second huge challenge that must be faced squarely is *bureaucracy, trivialisation and reductionism.* The word trivialisation has cropped up time and again in this article. Our frequent usage of it is intentional, for trivialisation lurks everywhere. As soon as an assessment instrument, an assessment strategy or an assessment procedure becomes more important than the original goal it was intended to accomplish, trivialisation rears its ugly head. We see it happening all the time. Learners perform tricks to pass exams, teachers complete forms with one stroke of the pen (administrative requirement

completed but judgement meaningless), we stick to procedures for no other reason than that we have always done it this way (we want grades because they are objective and accountable to society) or because of institutional policy. As soon as we notice the exchange of test materials on the black market or new internet resources peddling rafts of ready-made reflections, we can be sure that we have trivialised the assessment process. All actors in programmatic assessment should understand what they are doing, why they are doing it and why they are doing it this way. Otherwise they are in danger of losing sight of the true purpose of assessment and will fall back on bureaucratic procedures and meaningless artefacts. Steering clear of trivialisation is probably the hardest yet most urgent task we have to tackle if we are to realise programmatic assessment as advocated here. To prevent bureaucracy, we need support systems to facilitate the entire process. Computer technology seems an obvious candidate for an important role as facilitator (Bird 1990; Dannefer & Henson 2007). We have only begun to explore these technologies, but they show great promise to reduce workload and provide intelligent solutions to some of the problems.

A third challenge is *legal restrictions*. Curricula have to comply with university regulations or national legislation. These are usually very conservative and tend to favour a mastery-oriented approach to learning with courses, grades and credits.

This brings us to the final challenge: the *novelty* and the *unknown*. The proposed model of programmatic assessment is vastly different from the classical summative assessment programme familiar to most of us from personal experience as learner and teacher. When confronted with our new model, many stakeholders are likely to tell us we have turned soft on assessment. Our willingness to rely on subjective information and judgement, in particular, is seen by many as a soft option. We fervently disagree and we hope to have demonstrated that the decision-making procedures we propose can actually be extremely tough, provided they are put in the hands of a large body of actors who really understand why they are doing and for which purpose. A daunting task indeed, but the one we support wholeheartedly.

Opportunities

The opportunities are manifold. We hope to have demonstrated, at least theoretically, that it can be worthwhile and feasible to assess for learning and at the same time take robust decisions. Naturally, the proof of the pudding is in the eating. In fact, a number of good practices already exist, some of them are reported in this issue of this journal. We clearly need more research and documentation, but we feel quite confident that the model is not an unreachable star in the theoretical sky.

We also hope that, with this model, we can move beyond the exclusively psychometrically driven discourse of individual assessment instruments (Hodges 2006). This is not to claim that the psychometrical discourse is irrelevant or that individual methods cannot have validity. All we are saying is that the psychometric discourse is incomplete. It does not capture the full picture. Moving towards programmes of assessment and towards a more theory-based systems design of these programmes is an extension of the discourse, which we hope will advance not only the assessment of learning but learning in all its facets.

A third exciting opportunity is the infinite number of research possibilities. Any attempt to summarise them can only be futile but we will mention just a few. It would be quite interesting (and challenging) to develop formal models of decision making. How can we be confident that our information is trustworthy when we aggregate across multiple sources? And when is enough (Schuwirth et al. 2002)? Are Bayesian or similar approaches useful to support the decision making process? Can we show empirical proof that we can successfully reduce bias through procedural measures? Can we describe the process of decision making in expert judgements as a constructive process (Govaerts et al. 2011)? What are the underlying mechanisms? Can we use and optimise judgements by applying theory and empirical outcomes from other disciplines, like cognitive theories on decision making (Dijksterhuis & Nordgren 2006; Marewski et al. 2010), the psychology of judgement and decision making (Morera & Dawes 2006; Karelia & Hogarth 2008; Weber & Johnson 2009), cognitive expertise theories (Eva 2004) and naturalistic decision making (Klein 2008)? Can we train the judges? How, why and when is learning facilitated by assessment information?

Conclusion

The model of programmatic assessment for the curriculum in action that we propose here can serve as an aid in the actual design of such assessment programmes. We believe its coherent structure and synergy of elements ensure its fitness for purpose. Fit for purpose in its learning orientation and in its robustness of decision making. We think it is well grounded in theoretical notions around assessment, which in turn are based on sound empirical research. We note that the model is limited for the programme in action, but not for the other elements (programme support, documentation, improvement and justification) of the framework for programmatic assessment (Dijkstra et al. 2010). Design guidelines for all these elements are important to make programmatic assessment really come to life. These guidelines can also be used for evaluative or even accreditation purposes to truly achieve overall fitness for purpose.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on contributors

C. P. M. VAN DER VLEUTEN, PhD, is a Professor of Education, Chair of the Department of Educational Development and Research and Scientific Director of the School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands, Honorary Professor at King Saud University (Riyadh, Saudi Arabia), Copenhagen University (Copenhagen, Denmark) and Radboud University (Nijmegen, The Netherlands).

L. W. T. SCHUWIRTH MD, PhD, is a Professor of Medical Education, Health Professions Education, Flinders Medical School, Adelaide, South Australia.

E. W. DRIESSEN is an Associate Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

J. DIJKSTRA, MA, is an Assistant Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

D. TIGELAAR, PhD, is an Assistant Professor, ICLON – Leiden University Graduate School of Teaching, Leiden, the Netherlands.

L. K. J. BAARTMAN, PhD, is a Senior Researcher and Lecturer, Faculty of Education, Research Group Vocational Education, Utrecht University of Applied Sciences, the Netherlands.

J. VAN TARTWIJK, PhD, is a Professor of Education, Faculty of Social and Behavioural Sciences, Utrecht University, the Netherlands.

References

ACGME. 2009. Accreditation Council for Graduate Medical Education. Common program requirements: IV.A.5. General Competencies 2007 [Internet]. Available from: http://www.acgme.org/acWebsite/dutyHours/dh_dutyhoursCommonPR07012007.pdf

Al Kadri HM, Al-Moamary MS, van der Vleuten C. 2009. Students' and teachers' perceptions of clinical assessment program: A qualitative study in a PBL curriculum. BMC Res Notes 2:263.

Baartman LKJ, Bastiaens TJ, Kirschner PA, Van der Vleuten CPM. 2006. The wheel of competency assessment. Presenting quality criteria for competency assessment programmes. Stud Educ Eval 32:153–170.

Baartman LKJ, Prins FJ, Kirschner PA, Van der Vleuten CPM. 2007. Determining the quality of assessment programs: A self-evaluation procedure. Stud Educ Eval 33:258–281.

Bird T. 1990. The schoolteacher's portfolio: An essay on possibilities. In: Millman J, Darling-Hammond L, editors. The new handbook of teacher evaluation: Assessing elementary and secondary school teachers. Newbury Park, CA: Corwin Press. pp 241–256.

Cavalcanti RB, Detsky AS. 2011. The education and training of future physicians: Why coaches can't be judges. JAMA 306:993–994.

Cilliers FJ, Schuwirth LW, Adendorff HJ, Herman N, van der Vleuten CP. 2010. The mechanism of impact of summative assessment on medical students' learning. Adv Health Sci Educ Theory Pract 15:695–715.

Cilliers FJ, Schuwirth LW, Herman N, Adendorff HJ, van der Vleuten CP. 2011. A model of the pre-assessment learning effects of summative assessment in medical education. Adv Health Sci Educ Theory Pract, DOI: 10.1007/s10459-011-9292-5.

Coles C. 2002. Developing professional judgment. J Contin Educ Health Prof 22:3–10.

Dannefer EF, Henson LC. 2007. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. Acad Med 82:493–502.

Dijksterhuis A, Nordgren LF. 2006. A theory of unconscious thought. Perspect Psychol Sci 1:95–109.

Dijkstra J, Galbraith R, Hodges B, McAvoy P, McCrorie P, Southgate L, Van der Vleuten C, Wass V, Schuwirth L. under editorial review. Fit-for-purpose guidelines for designing programmes of assessment.

Dijkstra J, Van der Vleuten CP, Schuwirth LW. 2010. A new framework for designing programmes of assessment. Adv Health Sci Educ Theory Pract 15:379–393.

Driessen E, van Tartwijk J, van der Vleuten C, Wass V. 2007. Portfolios in medical education: Why do they meet with mixed success? A systematic review. Med Educ 41:1224–1233.

Driessen E, Overeem K, Tartwijk van E. 2010. Learning from practice: mentoring, feedback, and portfolios. In: Dornan T, Mann K, Scherpbier A, Spencer J, editors. Medical Education, Theory and Practice. pp 211–227.

Dudek NL, Marks MB, Regehr G. 2005. Failure to fail: The perspectives of clinical supervisors. Acad Med 80(Suppl. 10):S84–S87.

Embo MP, Driessen EW, Valcke M, Van der Vleuten CP. 2010. Assessment and feedback to facilitate self-directed learning in clinical practice of Midwifery students. Med Teach 32:e263–e269.

Eva KW. 2003. On the generality of specificity. Med Educ 37:587–588.

Eva KW. 2004. What every teacher needs to know about clinical reasoning. Med Educ 39:98–106.

Eva KW, Regehr G. 2005. Self-assessment in the health professions: A reformulation and research agenda. Acad Med 80:S46–S54.

Eva KW, Rosenfeld J, Reiter HI, Norman GR. 2004. An admissions OSCE: The multiple mini-interview. Med Educ 38:314–326.

Falchikov N, Goldfinch J. 2000. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. Rev Educ Res 70:287–322.

Frank JR, Danoff D. 2007. The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies. Med Teach 29:642–7.

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

Govaerts MJ, Van der Vleuten CP, Schuwirth LW, Muijtjens AM. 2007. Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. Adv Health Sci Educ Theory Pract 12:239–260.

Harden RM, Sowden S, Dunn WR. 1984. Educational strategies in curriculum development: The SPICES model. Med Teach 18:284–289.

Harvey L, Green D. 1993. Defining quality. Assess Eval High Educ 18:9–34.

Hattie J, Timperley H. 2007. The power of feedback. Rev Educ Res 77:81–112.

Hodges B. 2006. Medical education and the maintenance of incompetence. Med Teach 28:690–696.

Hodges BD, Ginsburg S, Cruess R, Cruess S, Delport R, Hafferty F, Ho MJ, Holmboe E, Holtman M, Ohbu S, et al. 2011. Assessment of professionalism: Recommendations from the Ottawa 2010 Conference. Med Teach, 33(5):354–363.

Jozefowicz RF, Koeppen BM, Case SM, Galbraith R, Swanson DB, Glew RH. 2002. The quality of in-house medical school examinations. Acad Med 77:156–161.

Karelia N, Hogarth RM. 2008. Determinants of linear judgment: A meta-analysis of Lens Model studies. Psychol Bull 134:404–426.

Klein G. 2008. Naturalistic decision making. Human factors 50:456–460.

Kluger AN, DeNisi A. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 119:254–284.

Korthagen FAJ, Kessels J, Koster B, Lagerwerf B, Wubbels T. 2001. Linking theory and practice: The pedagogy of realistic teacher education. Mahwah, NJ: Lawrence Erlbaum Associates.

Malini Reddy Y, Andrade H. 2010. A review of rubric use in higher education. Assess Eval High Educ 35:435–448.

Mansvelder-Longayroux DD, Beijaard D, Verloop N. 2007. The portfolio as a tool for stimulating reflection by student teachers. Teach Teach Educ 23:47–62.

Marewski JN, Gaissmaier W, Gigerenzer G. 2010. Good judgments do not require complex cognition. Cogn Process 11:103–121.

Miller GE. 1990. The assessment of clinical skills/competence/performance. Acad Med 65:S63–S67.

Morera OF, Dawes RM. 2006. Clinical and statistical prediction after 50 years: A dedication to Paul Meehl. J Behav Dec Making 19:409–412.

Muijtjens AM, Timmermans I, Donkers J, Peperkamp R, Medema H, Cohen-Schotanus J, Thoben A, Wenink AC, van der Vleuten CP. 2010. Flexible electronic feedback using the virtues of progress testing. Med Teach 32:491–495.

Norcini JJ. 2003. Work based assessment. BMJ (Clin Res Ed) 326:753–755.

Norcini J, Burch V. 2007. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach 29:855–871.

Norman GR, Van der Vleuten CPM, De Graaff E. 1991. Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. Med Educ 25:119–126.

Sargeant J, Armson H, Chesluk B, Dornan T, Eva K, Holmboe E, Lockyer J, Loney E, Mann K, van der Vleuten C. 2010. The processes and dimensions of informed self-assessment: A conceptual model. Acad Med 85:1212–1220.

Sargeant J, Mann K, van der Vleuten C, Metsemakers J. 2008. "Directed" self-assessment: Practice and feedback within a social context. J Contin Educ Health Prof 28:47–54.

Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, Pangaro L, Ringsted C, Swanson D, Van Der Vleuten C, et al. 2011. Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 33(3):224–233.

Schuwirth LW, Southgate L, Page GG, Paget NS, Lescop JM, Lew SR, Wade WB, Baron-Maldonado M. 2002. When enough is enough: A conceptual basis for fair and defensible practice performance assessment. Med Educ 36(10):925–930.

Schuwirth LW, Van der Vleuten CP. 2011. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach 33:478–485.

Shanteau J. 1992. The psychology of experts: an alternative view. In: Wright G, Bolger F, editors. Expertise and decision support. New York, NY: Plenum Press. pp 11–23.

Shute VJ. 2008. Focus on formative feedback. Rev Educ Res 78:153–189.

Sluijsmans DMA, Brand-Gruwel S, van, Merriënboer J, Bastiaens TR. 2003. The training of peer assessment skills to promote the development of reflection skills in teacher education. Stud Educ Eval 29:23–42.

Van der Vleuten CPM. 1996. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1:41–67.

Van der Vleuten CPM, Norman GR, De Graaff E. 1991. Pitfalls in the pursuit of objectivity: Issues of reliability. Med Educ 25:110–118.

Van der Vleuten CPM, Schuwirth LWT. 2005. Assessment of professional competence: From methods to programmes. Med Educ 39:309–317.

Van der Vleuten CP, Schuwirth LW, Muijtjens AM, Thoben AJ, Cohen-Schotanus J, van Boven CP. 2004. Cross institutional collaboration in assessment: A case on progress testing. Med Teach 26:719–725.

Van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B. 2010. The assessment of professional competence: Building blocks for theory development. Best Pract Res Clin Obstet Gynaecol 24:703–719.

Van Merriënboer JJG. 1997. Training complex cognitive skills. Englewood Cliffs, NJ: Educational Technology Publications.

Van Merriënboer JJG, Kirschner PA. 2007. Ten steps to complex learning: A systematic approach to four-component instructional design. Mahwah, NJ: Lawrence Erlbaum Associates.

Van Merriënboer JG, Sluijsmans MA. 2009. Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed Learning. Educ Psychol Rev 21:55–66.

Van Tartwijk J, Driessen EW. 2009. Portfolios for assessment and learning: AMEE Guide no. 45. Med Teach 31:790–801.

Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. 1999. Quality assurance in test construction: The approach of a multidisciplinary central test committee. Educ Health 12:49–60.

Weber EU, Johnson EJ. 2009. Mindful judgment and decision making. Annu Rev Psychol 60:53–85.

Williams RG, Klamen DA, McGaghie WC. 2003. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 15:270–292.

Wilson M, Sloane K. 2000. From principles to practice: An embedded assessment system. Appl Meas Educ 13:181–208.