A model for programmatic assessment fit for purpose

C.P.M. van der Vleuten[1], L.W.T. Schuwirth[1], E. Driessen[1], J. Dijkstra[1], D. Tigelaer[2], L.K.J. Baartman[3], J van Tartwijk[3])

[1]Department of Educational Development and Research, Maastricht University, The Netherlands

[2] Eindhoven School of Education of Eindhoven University of Technology

[3]Faculty of Social and Behavioural Sciences, Utrecht University

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
Email: c.vandervleuten@educ.unimaas.nl

Cees van der Vleuten 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 of Maastricht University

Lambert Schuwirth is a Professor of Education, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences of Maastricht University

Erik Driessen is Senior Lecturer of Education, Faculty of Health, Medicine and Life Sciences of Maastricht University

Joost Dijkstra is an Assistant Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences of Maastricht University

Dineke Tigelaar is Assistant Professor, ICLON-Leiden University Graduate School of Teaching

Liesbeth Baartman is postdoctoral researcher, Eindhoven School of Education of Eindhoven University of Technology

Jan van Tartwijk is Professor of Education, Faculty of Social and Behavioural Sciences, Utrecht University

Key words: assessment, programmatic, model, systems approach

Abstract

A model for programmatic assessment in action is proposed that optimizes assessment for learning as well as decision making on learner progress. It is based on a set of assessment principles that are interpreted from empirical research. The model specifies cycles of training, assessment and learner support activities that are completed by intermediate and final moments of evaluation on aggregated data-points. Essential is that individual data-points are maximized for their learning and feedback value, whereas high stake decisions are based on the aggregation of many data-points. Expert judgment plays an important role in the program. Fundamental is the notion of sampling and bias reduction for dealing with subjectivity. Bias reduction is sought in procedural assessment strategies that are derived from qualitative research criteria. A number of challenges and opportunities are discussed around the proposed model. One of the virtues would be to move beyond the dominating psychometric discourse around individual instruments towards a systems approach of assessment design based on empirically grounded theory.

Introduction

In 2005 we published a plea for thinking about assessment in a programmatic approach (C. P. M. Van der Vleuten & Schuwirth, 2005). We described a program of assessment as a planned arrangement of methods of assessment in such a way that it optimizes its fitness for purpose. Fitness for purpose is a functional definition of quality, in which quality can only be judged in terms of contribution to achieving the purpose of the assessment programme. Fitness for purpose is an inclusive approach towards quality as other definitions of quality (e.g. zero defects) can be regarded as a purpose (Harvey & Green, 1993). With overall quality in mind we advocated that an assessment program is deliberately constructed, that its elements are accounted for, that it is governed in its implementation and execution and that it is regularly evaluated and adapted. Much like it is generally accepted that a good test is more than a random set of good quality items, a good program of assessment is more than a randomly selected set of good instruments. The problem of programmatic assessment goes even beyond this analogy. Where there might be all good items there are never ideal instruments. In 1996 we already described that any individual assessment requires a compromise on quality criteria (C. P. M. Van der Vleuten, 1996). The decision on the exact compromise is dependent on which quality element needs to be optimized and is then determined by the specific assessment context. In a program of assessment the combination of assessment activities will alleviate the compromises of the individual methods, rendering to total more than the sum of its parts.

Since the introduction of the notion of programmatic assessment, further work has been done to define and assess quality criteria for assessment programs (L. K.J. Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006), (L.K.J. Baartman, Prins, Kirschner, & Van der Vleuten, 2007). On a different strand, design guidelines are being formulated with a first publication on a framework structuring these guidelines (Dijkstra, Van der Vleuten, & Schuwirth, 2009) and a next study in which concrete guidelines are formulated [Dijkstra, under editorial review]. Although these theoretical developments are important it may still be hard to visualize how such recommendations could lead to an assessment program in action, with a reference to its theoretical underpinning. What still lacks is a theoretically funded framework or generic model that provides concrete recommendations on how to structure an assessment program in action (according to Dijkstra's et al. model) in order to maximize its fitness for purpose. The purpose of this paper is to provide such a model.

The proposed model is limited to programmatic assessment in the educational context, thereby excluding licensing assessment programs. The model is generic to type of learning program. Learning programs can either be 'school-based', i.e. classroom teaching, or 'work-based', i.e. a postgraduate specialty training program. We do however assume the learning program to be learner centered, with holistic approaches to learning (as opposed to atomistic mastery-oriented learning) and a focus on deep learning strategies. An assessment model for mastery-oriented learning program would probably be different from our model. This doesn't mean that some tasks in a learner centered program are mastery-oriented and should learned and assessed that way. We define three fundamental purposes that we wish to unite in an assessment program: a program that maximally facilitates learning (assessment *for* learning), a program that maximizes the robustness of high stakes decisions (i.e. on promotion/selection decisions of learners), and a program that provides

information for improving the instruction or the curriculum. For the moment we will park the latter for the discussion and will focus on optimizing the first two purposes. The aim here is to provide such a theory based model. In order to motivate choices in this model, we will first shortly describe a number of theoretical principles of assessment that are based on empirical research, but clearly represent our interpretation of that research. The account is deliberately brief; a fuller account of most of these principles is given elsewhere (C. P. Van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges).

Principles of assessment

1. Any single assessment data-point is flawed

Single shot assessments, a single administration of an assessment method at any level of Miller's pyramid (Miller, 1990), a point-measurement, have their limitations. For content specificity reasons (K. W. Eva, 2003), performance of individuals is highly context dependent, requiring large sample of test items (in the broadest sense of the term) and long testing time for producing minimally reliable results (C. P. M. Van der Vleuten & Schuwirth, 2005). Furthermore, any single method can only assess a part of Miller's pyramid and there is no magical method than can do it all. In our view, this provides the legitimization for thinking about programs of assessment.

2. Standardized assessment can have 'built-in' validity in the instrument

For all methods that can be standardized (first three levels of Miller assessing knows, knowing, how, and shows how) validity can be built into the test by careful construction of the content, the scoring and administration procedures. Quality control procedures around test construction have a dramatic effect on test material quality (Verhoeven, Verwijnen, Scherpbier, Schuwirth, & Van der Vleuten, 1999), (Jozefowicz et al., 2002). If applicable, assessors can be trained, scoring lists can be objectified, simulated patients can be standardized, 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-standardized assessment lies more in the users than in the instrument

A complete assessment program will often also have to employ unstandardized methods. Particularly if we need to assess in real practice, the top of Miller's pyramid (the 'does' level), we cannot always standardize. The real world is unstandardized and haphazard, and if we try to standardize here, we quickly trivialize the process (Norman, Van der Vleuten, & De Graaff, 1991). The assessment literature is currently developing its 'technologies' for assessing this level of performance, for example in the field of work-based assessment (J. J. Norcini, 2003), (J. Norcini & Burch, 2007). However, assessment in daily educational settings (e.g., in the classroom, tutorial, or practical) fall under the same category of assessing habitual performance (e.g., assessment of a presentation or assessment of professional behavior). Typically, in such situations 'standardized forms' do not determine the validity of the assessment. The users, assessors, learners, patients, are more important than the instrument itself. Their expertise in using the instrument, the extent to which they take it seriously, the time they can spend on using it all determine whether the assessment will be performed well or not. No extensive training is needed for someone who hands out a multiple choice test to the learners, extensive teacher training on the other hand is considered essential for all

those who are involved in unstandardized observational assessment. The way in which this last group takes the assessment task seriously (i.e. by taking time to give feedback or to complete a narrative on a form) really defines the utility of these methods. Creating understanding of their role requires training, facilitation, feedback, expertise development, etc. Since a program cannot do without unstandardized methods we need to develop a 'technology' that can help these users to function appropriately in their assessment role. In doing this, we need to realize that learners are learners, even if they are assessors, teachers, or supervisors. They all learn in the same way, preferably by training, doing and feedback. Simply providing information or the instruments will not suffice. If the users do not understand the meaning and purpose of the assessment, the assessment will trivialize.

4. Stakes of the assessment is a continuum and proportionally related to the number of data-points

From the conceptual framework of programmatic assessment, the formative-summative distinction is not very useful, as all assessments in the framework are both formative and summative but in varying degrees. A distinction in stakes of the assessment seen as a continuum from low to high stakes is more useful. In low-stakes assessment the results of the assessment have limited consequences for the learner in terms of promotion, selection or certification, whereas in high-stakes assessment they can be dramatic. In a program of assessment single data-points of assessment should only lead to low stakes decisions, whereas high stake decisions should always be based on many data-points. The role of the teacher as a helper can be compromised by a high stake assessment. Being a helper and a judge (for high stake decisions) are conflicting roles. The conflict often leads to inflation of judgments (Dudek, Marks, & Regehr, 2005),(Govaerts, Van der Vleuten, Schuwirth, & Muijtjens, 2007). The risk is trivialization of the assessment process through the stakes that are involved. If high stake decisions are to be taken on the basis of many data-points it would be foolish to ignore information of the rich material derived from single data-points. The combined low stake information feeds into high stake information. Low stake as an individual data-point may be, it is not without any stake.

5. Assessment drives learning

This is a widely shared opinion in the assessment literature, but it is poorly understood. Most assessment probably drives learners in negative way not in being in line with curriculum objectives, particularly in purely information-poor summative systems. We need more theoretical clarification on why and how assessment drives learning and research on this is emerging (Cilliers, Schuwirth, Adendorff, Herman, & van der Vleuten). The objective is to drive learning in a desirable way, fostering deep-learning approaches to learning (and mastery-learning where appropriate). There is a wealth of evidence that formative feedback can foster learning (Kluger & DeNisi, 1996), (Hattie & Timperley, 2007), (Shute, 2008). We note that meaningfulness of the assessment information is imperative for driving learning. That means that the assessment information should be as rich as possible. Richness of information can be achieved through many different ways, both quantitatively as qualitatively. We note that assessment is often associated to grades (only), but grades are one of the poorest form of feedback (Shute, 2008). Other 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 feedback and contributes strongly to the meaningfulness of the information (Sargeant et al., 2010). We finally note that

feedback seeking and giving are skills (Sluijsmans, 2003) that need to be developed, which ties us back to our previous point of investing in the users of the assessment.

The absence of meaningfulness leads to trivialization. Assessment often risks trivialization. If learners are required to memorize checklists for passing the OSCE but have no connection with the patients, their performance becomes trivial; if an assessor completes a professional behavior rating form by one strike of the pen from top to bottom, the assessment is not meaningful and becomes trivial. If the assessment information has meaning, learning will be enhanced in a meaningful way. We argue that low stake individual data-points should be as meaningful as possible, fostering learning, and we argue that high stake decision-making should be based on many data-points. With the aggregation of meaningful data-points a meaningful high stake decision can be taken. In all elements of the assessment program trivialization is prevented.

There is one exception that individual data-points can be high stake. That 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 where they occur in the program. The proposed model should accommodate this exception. This doesn't imply that mastery-tasks can do without feedback.

6. Expert judgment 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 in competencies or competency domains (Ref CanMeds, ACGME), interpreting results of assessment always requires human judgment. By providing support, e.g. in scoring rubrics, training, performance standards, we can reduce the subjectivity in the judgment (Malini Reddy & Andrade, 2010), but if we try to objectify it completely, we will trivialize the assessment process (see also the examples described above in principle 5). We therefore need to rely on the expert judgment of knowledgeable others at various points in the assessment process. We also need expert judgment to combine information across individual data-points. Often, we use quantitative strategies for aggregating information sources (e.g. by averaging scores, or by counting the number of passes). When individual data-points are information rich (e.g., multisource-feedback or mini-CEX), particularly also when containing qualitative information, simple quantitative aggregation is impossible and expert judgment is required. From a vast amount of literature in the decision making literature we know that the human mind is quite fallible if compared to actuarial decisions (Shanteau, 1992). We argue that random bias in judgment can be overcome by sampling strategies and systematic bias by procedural measures. The sampling perspective has been effectively proven in many types of assessment situations (C. P. M. Van der Vleuten, Norman, & De Graaff, 1991), (Williams, Klamen, & McGaghie, 2003), (K. W. Eva, Rosenfeld, Reiter, & Norman, 2004): simply by using many judgments we can produce reliable information. Actually the sample needed for assessment methods that heavily rely on judgment is considerably smaller than in most objectified methods (C. P. Van der Vleuten et al.). Bias is difficult to prevent. We argue that through procedural measures around the decision making biases can be reduced. For example, a decision on a borderline candidate will require much scrutiny of information gathering, perhaps even more data-gathering and deliberation of that information. In a recent paper we proposed methodologies from qualitative research to serve as inspiration for developing procedural measures in assessment (C. P. Van der Vleuten et al.). The example just given stems from the triangulation criterion. Member checking, another criterion, would suggest to incorporate the learner's view in the assessment

procedure. Table 1 provides an overview of such procedural strategies. Depending on the carefulness of these procedures, biases will be reduced and the resulting decision will be more trustworthy or defensible. We think these strategies can handle subjective information (in combination with objective) and fortify the robustness of the resulting decisions. It prevents having to objectify every part of the assessment program and risking reductionism and trivialization of the learning process.

Model of programmatic assessment in action

With the above principles we will propose a model that is optimized for fitness of purpose. The program's purpose is to maximize assessment for learning while at the same time is able to arrive at robust decisions over learners. Figure 1 provides a graphical representation of the model. We will describe its elements systematically and will provide arguments for its coherence.

We distinguish between training activities, assessment activities and learner support activities as a function of the time of the ongoing curriculum.

Learning activities

We start with a first period of training activities consisting of *learning tasks* denoted by small circles (after the 4-CID model (J.J.G. Van Merriënboer, 1997)). The learning task can be anything that leads to learning: a lecture, a practical, a patient encounter, an operation, a PBL tutorial, a project, a learning assignment or self-study. When done appropriately these learning tasks themselves provide a coherent program or curriulum and are developed according to principles of instructional design (Harden, Sowden, & Dunn, 1984), (J.J.G. Van Merriënboer & Kirschner, 2007). Some learning tasks may yield *artifacts of learning,* denoted by the larger circles. These artifacts can be outcome related, for example a report on a project, or can be process oriented such as for example a list of operations done in the surgical theatre.

Assessment activities

The assessment activities in period 1 are denoted by small pyramids. Each one represents a *single data-point of assessment*. The symbolic shape is purposefully chosen because each single data-point can be of any method of any layer of Miller's pyramid. It could be a written test, an OSCE, an observation of a clinical encounter (i.e. Mini-CEX), it could be a peer evaluation in a PBL tutorial assessment, etcetera. Some of these assessments are evaluations of the artifacts or learning tasks. An example is the assessment of a patient information folder that has been written by a learner or the evaluation of a presentation that has been given on a research report. The arrangement of these assessment activities maximally supports the ongoing learning of the learner thereby adhering to our principle number 3 (assessment driving learning). Therefore all assessment is maximally meaningful to learning. It should provide information-rich feedback on the performance of the learner, either quantitatively or qualitatively. The information is documented, that means physically or electronically traceable. Each data-point is low stakes (principle 5). Naturally the performance feedback provides information in relation to some kind of performance standard, but we warn against passing or failing someone as in a mastery test. Each data-point is but one element in a longitudinal array of data-points (principle 1). Each data-point is of low stakes. It doesn't mean it cannot be used later on for making a decision about progress. The assessor's task is to provide

feedback as much as possible, not just to declare someone competent on a competency. The assessor is protected in his role as a teacher or facilitator, not as a judge (principle 5). Both roles are disentangled as much as possible (naturally in the realization that any assessor will judge whether the performance is done well or not). There is one exception and this is represented by the black pyramid. Some tasks are mastery oriented and require a 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 be certified on laparoscopic operation skills on the simulator before being allowed to conduct similar procedures on a patient. But most assessment tasks will not be mastery oriented but developmental in terms of reaching some proficiency in a competency. We similarly warn against grades if that is the only feedback given. Grades are poor feedback carriers and tend to have all kinds of undesirable educational side effects (learners hunting for grades but ignoring what and how they have learned; teachers being happy of the supposed objectivity of grades and excusing them to provide performance feedback). We advocate applying all assessment technology as was linked to our assessment principles 2 and 3 above. We 'sharpen' the instruments and/or people as much as possible. We are agnostic to any preference of assessment method since any assessment approach may have utility depending on its function within the program. We do not avoid subjective information or judgments from experts (principle 6). Experts are defined flexibly and refer to any knowledgeable person. Depending on the context this may be the teacher, the tutor, the supervisor, the peer, the patient, or, not to be forgotten, the self. Naturally self-assessment should never stand alone (K. W. Eva & Regehr, 2005), but in many cases the person self is a knowledgeable source of expertise. In all, the assessment activities in a given period of the training program are meaningful and traceable data-points of learner performance maximally connected to the learning program reinforcing desirable learning behavior.

Supporting activities

The supporting activities in that same period are twofold. First, the learner will reflect on the information derived from the learning and assessment activities (principle 4 and 6 combined). This is denoted as underscored connected small circles. Perhaps at the start and at the end there is more of *reflective activity* but it is an ongoing *self-directed learning activity*. The feedback is interpreted and used for planning new learning tasks or learning goals (J. G. Van Merriënboer & Sluijsmans, 2009). From the literature we know how hard it is to get people reflect and self-direct (E. Driessen, van Tartwijk, van der Vleuten, & Wass, 2007) , (Korthagen, Kessels, Koster, Lagerwerf, & Wubbels, 2001), (Mansvelder-Longayroux, Beijaard, & Verloop, 2007). One of the paradoxes of self-directed learning is that it requires a lot of external direction and scaffolding (E. W. Driessen, Overeem, & Van Tartwijk, 2010). Therefore we propose to scaffold this self-directed learning with some sort of social interaction. In the model this is the bottom rectangular with oppositely connected circles. The most prominent one is coaching or mentoring (supervision activities), but alternatively this could also be done with senior learners or with peers (intervision activities). Dedicated instruments can also be used to facilitate this process, in which reflective activity is structured (in time, content and social interaction) and documented (Embo, Driessen, Valcke, & Van der Vleuten, in press). In general we would encourage some documentation of this reflective process. At the same time, we shouldn't exaggerate these documented reflective activities, for they need to be 'lean and mean' and have direct meaningful learning value. If they don't have intrinsic learning value, they become bureaucratic tigers of thick ritualistically produced paper darts trivializing the learning activity. The social interaction is a requirement for providing meaningfulness to this process.

Intermediate evaluation

At the end of this period all artifacts, assessment information and (selected) information from the supporting activities are going to be assessed in an intermediate evaluation of progress. The aggregate information across all data-points is held against a performance standard by an independent and authoritive group of assessors, i.e. a committee of examiners. We think a committee is appropriate because expert judgment 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 do this when appropriate and possible. In one of our own programs (Maastricht) we for example use an online performance-data base for progress testing that can flexibly aggregate across an infinite number of comparisons and it can predict future performance based on past performance. However, some data-points are narrative and qualitative. This needs a human interpretation of the information (just like the patient chart) (principle 6). Aggregation of data-points preferably happens across meaningful entities. Traditionally we aggregate with methods (or layers of Miller's pyramid) as entities. Other, more meaningful aggregation categories are possible as well, for example in themes representing the training program or in terms of a competency-framework. We naturally advocate measures to make this evaluation robust. The committee consists of experts, knowledgeable in terms of what they have to assess. They are trained, perhaps certified, and use supporting tools such as rubrics and performance standards. They learn with accumulated experience and make changes to the procedures and supporting tools. The committee size matters as well as the extent of deliberation. For most learners the assessment process will be fast and efficient depending on the consistency and level of the information from the original data-points. For some learners there will be much debate, deliberation and argumentation. Their decision is informative in relation to the performance standard, but also informative in its diagnostic, therapeutic and prognostic value. They provide information on areas of strength and of improvement (diagnosis), they may suggest remediation for achieving desirable performance objectives (therapy), they may predict certain performance outcomes later in the training program (prognosis). Very importantly, the assessment done here is remediation oriented. This is very different from conventional types of assessment which are typically mastery-oriented: if mastery is not achieved, the course is simply repeated. Our approach is quite developmental: we propose an information rich recommendation for further learning, tailored to the individual and contingent to the diagnostic information. The assessment done by the committee is of intermediate stake. The assessment information doesn't have dramatic consequences for survival in the learning program, but it is not to be neglected information for further planning of learning.

The intermediate evaluation leads to a *firewall dilemma* that may have multiple ways of resolving. The dilemma is the input from the actors in the support system. According to the criterion of prolonged engagement information from a coach, mentor or learner will provide the richest information. At the same time by vesting the power of decision making in the actors of the support system their relationship will be compromised. One way to resolve this is to completely firewall both activities of support and decision making. The consequence is that the committee remains oblivious of valuable information, probably also leading to more work for the examiners, potentially more bias and more cost. 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 give a recommendation on the performance decision that is amended by the learner. There is no single right strategy and

compromises are in order depending on the resources, argumentation, sentiments, culture, and the stakes involved (Van Tartwijk & Driessen, 2009).

The presented cycle of training, assessment and supporting activities in the first cycle may be repeated. The fact that the model provides 3 cycles is not of any significance. It will depend on the exact nature of the training program and resources available. The 3 cycles depicted in the model might reflect a first year of a medical school. Actually, each period might have multiple courses. Important is a logical longitudinal development of the learner trough learning tasks, appropriate feedback and (supported) self-direction. This is quite opposite to a pure mastery-oriented approach where passing an exam is being declared competent for life. Important is also that sufficient data-points and remediation moments have occurred before a final high stake decision occurs.

Final evaluation

After sufficient cycles a final evaluation takes place at a moment where a learner-progress-decision is in order. This is a high stakes decision with imposing consequences for the learners. The decision is taken by the same committee of examiners as from the intermediate evaluation (prolonged engagement) with even more stringent procedural safeguards as are feasible. Examples are procedures of appeal, procedures of learner and coach input (firewall dilemma), training and benchmarking of examiners, committee size, deliberation and documentation, performance standards and/or rubrics, quality improvement measures on the evaluation procedure as a whole and, finally and very importantly, through the inclusion of all data points from the preceding period including the intermediate evaluations (principle 5).

In the ideal assessment situation the decision is motivated through a justification. The decision may not only be a pass or fail, but may also indicate distinctive high performance. One should note that more performance classifications provide more subtlety but also more classification errors and judgmental headache. If the system works well outcome decisions will not surprise the learner (or coach). In a minority of cases it will, and the fact that it occurs more or less 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 resulting decision is robust and based on rich information and lots of 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 court.

The model in Figure 1 depicts a certain learning period and ends with a natural moment of decision making over learner promotion. It does not reflect a total curriculum. Depending on the curriculum the learning period from the model may be repeated in as many cycles as are appropriate to complete the curriculum. Each cycle doesn't have to be of equal length, all depending on the nature of the curriculum and the natural decision moments therein.

Discussion

We think the model proposed is optimally fit for purpose. It optimizes the learning value consistently across the assessment program. No compromises are made on the meaningfulness of data in the assessment program. At the same time, high stake decision making is robust and credible, providing

an internal and external (societal) account for the quality of the graduating learners. A third purpose of an assessment program could be to evaluate the curriculum. Information from the supporting actors such as mentors/coaches and information from the actors in the intermediate and final evaluation are excellent data-points for curriculum evaluation, both in terms of the process of training as well as in terms of outcomes of training. We formulated the model in generic terms as much as possible. Some may conclude that we are describing portfolio learning and assessment. We deliberately avoided suggesting specific assessment methods or show any preference of methods. The purpose was to theorize beyond a single assessment method approach. Our model is informed through extensive previous research in assessment and brings together strategies from various theoretical strands crossing the boundaries of the quantitative and qualitative discourse. It also reinstates the value of expert judgment as an indispensable source of information. We will finish with describing some challenges and opportunities of the model presented.

Challenges

An obvious first challenge of the suggested programmatic approach is *cost* and the required *resources* to run such a program. Our first remark would be that in reducing costs it is wise to do fewer things well than to do a many things poorly (the 'less is more' principle). There is no point having many data-points which provide little information; it is a waste of time and effort. A second remark is that the boundaries between assessment and learning activities are vanishing. The ongoing assessment activities are very much part of the learning program, actually fully embedded within it (Wilson & Sloane, 2000). Thirdly, economic compromises can and should be made. Some of the assessment activities, particularly when they are low stakes, can also be cheap. For example, to assess a certain area of knowledge an online bank of questions could be used for self-assessment. The sharing of test material across schools is a smart strategy that was mentioned earlier. Assessing certain professional qualities as professionalism or communicative behavior can be assessed through the use of peers (Falchikov & Goldfinch, 2000). Choices could also be made about compromising on certain elements of the model in certain periods of time in the curriculum, all depending on the balance between stakes and resources. For example, mentoring or coaching is only done in certain parts of the curriculum and not in others. And finally, a quote attributed to McIntyre and Bok seems quite appropriate here: "If you think education is expensive, try ignorance".

A second huge challenge is *bureaucracy, trivialization and reductionism*. The word trivialization has occurred frequently in this paper. Indeed, trivialization lurks everywhere. As soon as an assessment instrument, an assessment strategy or procedure becomes more important than the original goal it was set out for, trivialization occurs. We see it happening all the time. Learners performing tricks to pass the exam, teachers completing forms with the stroke of the pen (administrative requirement completed but totally useless activity), procedures we simply follow because we doing them for ages ("we want grades because they are objective and accountable to society"). As soon as we see the exchange of test materials on black markets or new internet resources with infinite ready-made reflections we have trivialized the process. All actors in the programmatic assessment should understand why they are doing what, otherwise they will lose sight of their function and will start to rely on bureaucratic procedures and artifacts. This is probably the most difficult task of all to realize programmatic assessment programs such as proposed. To prevent bureaucracy we need support systems that facilitate the entire process and computer technology is naturally an important

facilitator (Bird, 1990). We have only begun to explore these technologies, but they are a promise and may reduce the workload and provide intelligent solutions to some of the problems.

A third challenge factor are *legal restrictions*. Curricular programs are restricted by university or national legislative rules. These rules are usually very conservative using very much a mastery-oriented approach to learning with courses, grades and credits.

This then turns into a final challenge factor: the *novelty* and the *unknown*. The proposed model of programmatic assessment is quite different from a classical summative assessment program as probably the majority of current assessment programs are. When confronted with this new approach stakeholders think we have turned soft. Particularly the role of subjective information, the reliance on judgment is seen as soft. We wholeheartedly disagree and we hope we demonstrated that the decision-making procedures are quite tough, but require a lot of actors who need to understand why they are doing what for which purpose. Not an easy task to complete.

Opportunities

The opportunities are manifold. We hope to have demonstrated, at least theoretically, that it is possible to assess for learning while being able to take robust decisions. Naturally, the proof of the pudding is in the eating. Some good practices do exist. For example, the Cleveland Clinical Lerner College of Medicine of Case Western Reserve University has such a program of assessment in operation (Dannefer & Henson, 2007). They use portfolios for documenting learning in a very feedback-intense assessment program followed by a very strict and firewalled system of appraisal of portfolios. From personal communication we know it is functioning quite well. We know of several other implementations, including our own that require further research and publication. Naturally we need more research and documentation, but the model is not an unreachable cloud in the theoretical sky.

We also hope that we move beyond the exclusively psychometrically driven discourse of individual assessment instruments (Hodges, 2006). This is not to say that this discourse is unimportant, nor that individual methods should be valid. We think however it is incomplete. Moving towards assessment programs and to more theory-based design of these programs is an extension that we hope advances our assessment knowledge, indeed very much similar to the scientifically underpinned approaches to instructional design.

A third opportunity is the infinite number of research possibilities. Any attempt to summarize them is bound to fail and we mention just a few. Quite interesting (and challenging) would be to develop formal models for decision making. When can we trust the information we have if we aggregate across multiple sources and when is enough enough (Schuwirth et al., 2002)? Are Bayesian or similar approaches possible to support the decision making progress? Can we demonstrate empirically that we can reduce bias through procedural measures? Can we describe the process of decision-making in expertise judgments? What are underlying mechanisms? Can we use and optimize judgments by applying theory and empirical outcomes from other disciplines such cognitive theories on decision making decision making (Dijksterhuis & Nordgren, 2006), (Marewski, Gaissmaier, & Gigerenzer, 2010), the psychology of judgment and decision-making (Morera & Dawes, 2006), (Karelia & Hogarth, 2008), (Weber & Johnson, 2009), cognitive expertise theories (K.W. Eva, 2004), naturalistic decision-making (Klein, 2008)? Can we train making judgments? How, why and when is learning facilitated by

assessment information? We could easily continue with many other important questions, but we pause for the moment.

Conclusion

The model proposed for programmatic assessment for the curriculum in action may serve as an aid to actually design such assessment programs. We belief it is a coherent structure that in its synergy of elements is fit 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 program in action, not to the other elements (program support, documentation, improvement, justification) of the framework for programmatic assessment (Dijkstra et al., 2009). Design guidelines on all these elements are important for programmatic assessment to come to live. These in turn may also be used for evaluative or even accreditation purposes for overall fitness for purpose.

References

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2006). The wheel of competency assessment. Presenting quality criteria for competency assessment programmes. *Studies in Educational Evaluations, 32*(2), 153-170.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determining the quality of assessment programs: a self-evaluation procedure. *Studies in Educational Evaluation, 33*, 258-281.

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

Cilliers, F. J., Schuwirth, L. W., Adendorff, H. J., Herman, N., & van der Vleuten, C. P. The mechanism of impact of summative assessment on medical students' learning. *Adv Health Sci Educ Theory Pract*.

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*(5), 493-502.

Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science, 1*(2), 95-109.

Dijkstra, J., Van der Vleuten, C. P., & Schuwirth, L. W. (2009). A new framework for designing programmes of assessment. *Adv Health Sci Educ Theory Pract*.

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*(12), 1224-1233.

Driessen, E. W., Overeem, K., & Van Tartwijk, J. (2010). Learning from practice: mentoring, feedback and portfolios. . In T. Dornan, K. Mann, A. Scherpbier & J. Spencer (Eds.), *Learning Medicine*. Dordrecht: Elsevier.

Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to fail: the perspectives of clinical supervisors. *Acad Med, 80*(10 Suppl), S84-87.

Embo, M., Driessen, E. W., Valcke, M., & Van der Vleuten, C. P. M. (in press). An instrument to integrate feedback and assessment to support self-directed learning in clinical practice: a qualitative study of students' perceptions. . *Medical Teacher*.

Eva, K. W. (2003). On the generality of specificity. *Med Educ, 37*(7), 587-588.

Eva, K. W. (2004). What every teacher needs to know about clinical reasoning. *Medical Education, 39*, 98-106.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: a reformulation and research agenda. *Acad Med, 80*(10 Suppl), S46-54.

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: the multiple mini-interview. *Med Educ, 38*(3), 314-326.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Govaerts, M. J., Van der Vleuten, C. P., Schuwirth, L. W., & Muijtjens, A. M. (2007). Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Adv Health Sci Educ Theory Pract, 12*, 239-260.

Harden, R. M., Sowden, S., & Dunn, W. R. (1984). Educational strategies in curriculum development: the SPICES model. *Medical Teacher, 18*(4), 284-289.

Harvey, L., & Green, D. (1993). Defining quality. *Assessment and Evaluation in Higher Education, 18*(1), 9-34.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hodges, B. (2006). Medical education and the maintenance of incompetence. *Med Teach, 28*(8), 690-696.

Jozefowicz, R. F., Koeppen, B. M., Case, S. M., Galbraith, R., Swanson, D. B., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77*(2), 156-161.

Karelia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A met-analysis of Lens Model studies. *Psychological Bulletin, 134*(3), 404-426.

Klein, G. (2008). Naturalistic decion making. *Human factors, 50*, 456-460.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. . *Psychological Bulletin, 119*, 254-284.

Korthagen, F. A. J., Kessels, J., Koster, B., Lagerwerf, B., & Wubbels, T. (2001). *Linking theory and practice: The pedagogy of realistic teacher education.* . Mahwah, NY: Lawrence Erlbaum Associates.

Malini Reddy, Y., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment and Evaluation in Higher Education, 35*(4), 435-448.

Mansvelder-Longayroux, D. D., Beijaard, D., & Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. . *Teaching and Teacher Education, 23*(1), 47-62.

Marewski, J. N., Gaissmaier, W., & Gigerenzer, G. (2010). Good judgments do not require complex cognition. *Cognitive processing, 11*(2), 1612-4782.

Miller, G. E. (1990). The Assessment of Clinical Skills/Competence/Performance. *Academic Medicine, 65*(9), S63 - 67.

Morera, O. F., & Dawes, R. M. (2006). Clinical and statistical prediction after 50 years: a dedication to Paul Meehl. *Journal of Behavioral Decision Making, 19*, 409-412.

Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Med Teach, 29*(9), 855-871.

Norcini, J. J. (2003). Work based assessment. *Bmj, 326*(7392), 753-755.

Norman, G. R., Van der Vleuten, C. P. M., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education, 25*, 119-126.

Sargeant, J., Armson, H., Chesluk, B., Dornan, T., Eva, K., Holmboe, E., et al. (2010). The processes and dimensions of informed self-assessment: a conceptual model. *Acad Med, 85*(7), 1212-1220.

Schuwirth, L. W., Southgate, L., Page, G. G., Paget, N. S., Lescop, J. M., Lew, S. R., et al. (2002). When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ, 36*(10), 925-930.

Shanteau, J. (1992). The psychology of experts: an alternative view. In G. Wright & F. Bolger (Eds.), *Expertise and decision support.* (pp. 11-23). New York: Plenum Press.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research, 78*, 153-189.

Van der Vleuten, C. P., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol*.

Van der Vleuten, C. P. M. (1996). The assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Science Education, 1*(1), 41-67.
Van der Vleuten, C. P. M., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25*, 110-118.
Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessment of professional competence: from methods to programmes. *Medical Education, 39*, 309-317.
Van Merriënboer, J. G., & Sluijsmans, M. A. (2009). Toward a synthesis of cognitive load theory, four-component instructional design, and self-directed Learning. *Educational Psychology Review, 21*, 55-66.
Van Merriënboer, J. J. G. (1997). *Training complex cognitive skills.* New Jersey: Englewood Cliffs Educational Technology Publications.
Van Merriënboer, J. J. G., & Kirschner, P. A. (2007). *Ten steps to complex learning: A systematic approach to four-component instructional design*. Mahwah, New Jersey Lawrence Erlbaum Associates.
Van Tartwijk, J., & Driessen, E. W. (2009). Portfolios for assessment and learning: AMEE Guide no. 45. *Med Teach, 31*(9), 790-801.
Verhoeven, B. H., Verwijnen, G. M., Scherpbier, A. J. J. A., Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (1999). Quality assurance in test construction: the approach of a multidiciplinary central test committee. *Education for Health, 12*(1), 49-60.
Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision Making. *Annual Review of Psychology, 60*, 53-85.
Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270-292.
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.

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	Training of assessors. The persons who know the student the best (a coach, peers) provide information for the assessment. Incorporate in the procedure intermittent feedback cycles.
	Triangulation	Many assessors should be involved and different credible groups should be included. Use multiple sources of assessment within or across methods. Organize a sequential judgment procedure in which conflicting information necessitates the gathering of more information.
	Peer examination (sometimes called Peer debriefing)	Organize discussion between assessors (before and intermediate) for benchmarking and discussion of the process and the results. Separate multiple roles of the assessors by removing the summative assessment decisions from the coaching role.
	Member checking	Incorporate the learner's point of view in the assessment procedure. Incorporate in the procedure intermittent feedback cycles.
	Structural coherence	Organize assessment committee to discuss inconsistencies in the assessment data.
Transferability	Time sampling	Sample broadly over different contexts and patients.
	Thick description (or Dense description)	Incorporate in the assessment instruments possibilities to give 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). Organize quality assessment procedures with external auditor. Give learners the possibility to appeal to the assessment decision.

Figure1: Model for programmatic assessment in action fit for the purpose of assessment for learning and making robust decisions on learner's achievements, selection and promotion.