Adv in Health Sci Educ (2010) 15:379–393
DOI 10.1007/s10459-009-9205-z

A new framework for designing programmes of assessment

J. Dijkstra · C. P. M. Van der Vleuten · L. W. T. Schuwirth

Received: 9 June 2009 / Accepted: 28 September 2009 / Published online: 10 October 2009

Abstract Research on assessment in medical education has strongly focused on individual measurement instruments and their psychometric quality. Without detracting from the value of this research, such an approach is not sufficient to high quality assessment of competence *as a whole*. A programmatic approach is advocated which presupposes criteria for designing comprehensive assessment programmes and for assuring their quality. The paucity of research with relevance to programmatic assessment, and especially its development, prompted us to embark on a research project to develop design principles for programmes of assessment. We conducted focus group interviews to explore the experiences and views of nine assessment experts concerning good practices and new ideas about theoretical and practical issues in programmes of assessment. The discussion was analysed, mapping all aspects relevant for design onto a framework, which was iteratively adjusted to fit the data until saturation was reached. The overarching framework for designing programmes of assessment consists of six assessment programme dimensions: *Goals*, *Programme in Action*, *Support*, *Documenting*, *Improving* and *Accounting*. The model described in this paper can help to frame programmes of assessment; it not only provides a common language, but also a comprehensive picture of the dimensions to be covered when formulating design principles. It helps identifying areas concerning assessment in which ample research and development has been done. But, more importantly, it also helps to detect underserved areas. A guiding principle in design of assessment programmes is *fitness for purpose*. High quality assessment can only be defined in terms of its goals.

Keywords Assessment · Design principles · Model · Programmatic approach · Theory development

J. Dijkstra (✉) · C. P. M. Van der Vleuten · L. W. T. Schuwirth
Department of Educational Development and Research, Maastricht University, Faculty of Health Medicine and Life Sciences, P.O. Box 616, Maastricht 6200 MD, The Netherlands
e-mail: joost.dijkstra@educ.unimaas.nl

Introduction

For long, research on assessment in medical education has strongly focused on individual measurement instruments and their psychometric quality. This is not illogical given the prevailing view of medical competence as consisting of separate elements—knowledge, skills, attitude, and problem solving—and the quest for the single best measurement instrument for each. Good examples of this approach are the established position of the Objective Structured Clinical Examination as the preferred instrument for skill measurement (Van der Vleuten and Swanson 1990) and key feature as approach of choice for problem solving skills (Page et al. 1995; Schuwirth 1998). Without detracting from the value of psychometric criteria and the focus on single instruments, which has provided valuable insights into the strengths and weaknesses of instruments as well as into the trade-offs that have to be made (Newble et al. 1994; Schuwirth and Van der Vleuten 2004; Van der Vleuten 1996), such an approach is not sufficient to high quality assessment of competence *as a whole*. From the point of view that medical competence is not the sum of separate entities but an integrated whole, it is only logical to conclude that no single instrument, however psychometrically sound, will ever be able to provide all the information for a comprehensive evaluation of competence in a domain as broad as medicine.

A currently popular model, Miller's pyramid (Miller 1990), frames assessment of "professional services by a successful physician" using a four-layered pyramid. While being a useful aid in selecting appropriate instruments for discrete elements of competence, Miller's pyramid does not describe the relationships between the layers or within combinations of instruments. Unfortunately, little is known about relations, compromises and trade-offs at this highly integrated level of assessment. Of course not just any mix of instruments will suffice: a purposeful arrangement of methods is required for measuring competence comprehensively. Similar to a test being more than a random sample of items, a programme of assessment should be more than a random selection of instruments. An optimal mix of instruments would be the best possible match between a programme of assessment and the goals of assessment (and/or the curriculum at large).

So a programmatic approach to assessment design is advocated (Lew et al. 2002; Schuwirth et al. 2002; Van der Vleuten and Schuwirth 2005). It is not easy to provide a single definition of such a "programme of assessment", but central to the concept is a design process that starts with a clear definition of the goals of the programme. Based on this; well-informed, literature-based, and rational decisions are made about the different assessment areas to be included, the specific assessment methods, the way results from various sources are combined, and the trade-offs that have to be made between strengths and weaknesses of the programme's components. In this way we see not just any set of assessment methods in a programme as the result of a programmatic approach to assessment, but reserve the term programmes of assessment for the result of the design approach as described above.

In this, design and development of assessment programmes must be underpinned by ideas and decisions on how to reconcile the strengths and weaknesses of individual instruments and how to complement and synthesise different kinds of information. Studying programmatic assessment can only be at the level of comprehensive competence, framing medicine as an integrated whole task. This in contradiction to the view of competence as split up into separate entities, or even as the sum of these entities. From a holistic perspective on assessment, a programmatic approach offers several theoretical advantages.

- It can help to create an overview of what is and what is not being measured. This promotes the balancing of content and other aspects of competence and counteracts the pitfall of overemphasising easy-to-measure elements, like unrelated factual knowledge.
- It allows for compensation for the deficiencies of some instruments by the strengths of other instruments, resulting in a diverse spectrum of complementary measurement instruments that can capture competence as a whole.
- Matching instruments can increase efficiency by reducing redundancy in information gathering. When data on a subject are already available from another test, test time and space is freed for other subjects.
- In high-stakes examinations, information from different sources (tests or instruments) can be combined to achieve well-informed and highly defensible decisions.

Of course, many existing examples of programmes of assessment are around already, much of which are based on extensive deliberation and good expertise and which are probably of high quality (Dannefer and Henson 2007). Unfortunately however, there is little research in this area that would help to support or improve their quality.

In our notion of a programmatic approach to assessment we presupposed that criteria for designing comprehensive assessment programmes and for assuring their quality would already be available in the literature, but when we searched the literature for guidelines for designing assessment programmes, the results were disappointingly scant. One of the early developments in this area, based on the notion that assessment drives learning, was the alignment of objectives, instruction, and assessment to achieve congruent student behaviour (Biggs 1996). Although in theory it might encompass an entire assessment programme, probably due to the complexity of educational environments, the application level of this alignment has rarely extended beyond the content of measurement (Webb 2007), i.e. blueprinting assessment based on curriculum objectives. Another approach focused on the application of psychometric criteria to combinations of methods (Harlen 2007), resulted in a framework for quality analysis which relied heavily on a “unified view of validity” (Birenbaum 2007) and research into high-stakes assessment programmes for certification of physicians aimed at high composite reliability (Burch et al. 2008; Knight 2000; Wass et al. 2001). Neither achieved a coherent programmatic approach to assessment, however.

Not only the search for single best instruments, but also the strong and almost unique reliance on psychometric quality in assessment can be challenged (Schuwirth and Van der Vleuten 2006) Undeniably, psychometric quality is important, but so are practical feasibility of instruments, educational goals, and context and environment of assessment. Baartman (2008) recently proposed adding education-based criteria, such as authenticity and meaningfulness. Her set of criteria for competence measurement was a valuable theoretical step with strong practical relevance, but the exclusive focus on competence (although cost and efficiency were considered too) disregarded the relationship of assessment programmes with their environment. Likewise, little attention was given to integrating or weighting criteria.

This paucity of research with relevance to programmatic assessment, and especially its development, prompted us to embark on a research project to develop design principles for programmes of assessment. Fearful of the pitfalls of a blunderbuss technique, we first set out to develop a model to frame programmes of assessment and determine which dimensions have to be covered in formulating design criteria, before we could—in a subsequent study—start defining the individual design criteria. Because of the absence of a common language for programmatic assessment and uncertainty about criteria, we used an exploratory, open, qualitative method to probe the views and ideas of experts in assessment (in medical education). From this resulted an overarching model for programmatic assessment, which we present in this paper.

Method

Study design

We conducted focus group interviews to explore the experiences and views of assessment experts concerning good practices and new ideas about theoretical and practical issues in programmes of assessment. The focus group approach was chosen because it allows participants to freely express ideas without having to reach consensus and leaves room for issues not previously considered in research (Hollis et al. 2002). Prior to data collection, the research team devised a rough and ready framework (list of topics) as a starting point for the discussions. The framework consisted of six elements of assessment relating to theoretical issues as well as practical suggestions for an assessment programme (see Fig. 1). The overall purpose of the assessment (*Goals*) and objectives of the curriculum, determine what needs to be tested (*Collecting information*) to gain data about medical competence of students. The data from different tests or sources needs to be merged (*Combining information*) into an overview which can be distributed among various stakeholders (*Reporting*). Based on the goals and data a further action needs to be taken (*Decision taking*). Finally in order to ensure high-standard assessment, a system of quality checks and measures should be in place (*Quality control*).

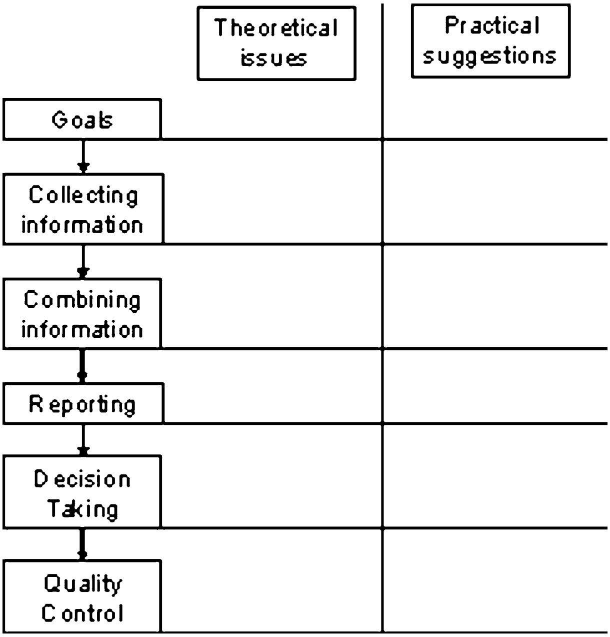

Fig. 1 Initial framework

Participants

An email giving details of the objectives and the topics of the focus groups invited 12 experts with extensive experience with difficulties and problems associated with programmes of medical assessment to participate in the study. A total of nine experts voluntarily took part in two focus groups. Three had to decline because of diary or health problems. The experts, five from North America and four from Europe, fulfil different (and some multiple) roles in their assessment practice i.e. Program Directors (5), National Committee Members (6). The experts represented different domains ranging from undergraduate and graduate education (4), to national licensing (5) and recertification (2) and had published extensively on assessment. Purposeful selection based on the experts' longstanding involvement in different assessment organisations ensured heterogeneity of the focus groups. To facilitate participation, we organised the sessions directly after the 2007 AMEE conference in Trondheim and paid all related expenses.

Procedure

The meeting was divided in four sessions on 1 day: a plenary introductory session in which the guiding (initial) framework was presented; two sessions split into groups, first on theoretical issues; and second on practical recommendations; and a plenary retrospective session summarising the discussions. It was explained to the participants that we were interested in variety of views and that there were no correct or incorrect answers. Dissent was encouraged. All sessions were semi-structured using the framework. Two of the researchers (LS & CvdV) moderated the sessions of one group each. A third researcher (JD) took field notes.

Data analysis

All sessions were audio recorded, transcribed, and read by the research team. One coder (JD) analysed the transcripts, starting with using the categories from the initial framework. Because this exploratory research requires an informed but open mind, the framework, including concepts and theories, was further developed in a continuous process of checking and refinement, without adhering to this pre-set framework. Furthermore the data was analysed by identifying and labelling new emerging themes and issues. When the research team met to evaluate the resulting themes and issues, they were forced to conclude that the first draft of the model (the framework guiding the discussions) was overly simplistic, causing ambiguities in coding and occasionally precluding coding altogether. The model was revised until the research team reached consensus that saturation of coding was reached and no new topics emerged. Finally the model was send to the participants to check if it reflected the discussion correctly and whether our interpretation of the discussion was accurate. No major revisions were suggested by the participants, just a minor suggestion as to the specific captions in English was made by a native English speaking participant.

Results

There is a risk the result section becomes more confusing in stead of clarifying as a result of the differences between the initial framework and the end result. Therefore some thoughts and explanation about the development from the initial framework to the final framework are provided first. Next the frameworks are compared on the top level, and similarities and differences are briefly described, before the dimensions of the final model are described in more detail and illustrated with quotes from the discussion to clarify some terminology. The selected quotes are accompanied by a (randomly assigned) number corresponding to a specific participant. This selection of quotes is no quantitative reflection of the participation during the focus group discussion as only the most clear and illustrative quotes are included. Some quotes are edited for reasons of clarity without changing the meaning and/or intention of the participant.

Coding the transcripts with the initial framework was complicated by the fact that this framework covered only a small proportion of the topics of assessment programmes that were discussed, and by the interrelatedness of the different elements, which had initially been conceived of as discrete. The distinction between theory and practice proved problematic as well, with theoretical issues often requiring adjustment due to practical considerations and practical suggestions requiring translation into general guiding principles, which could become increasingly theoretical. The alternative framework (see Fig. 2) is based on the refinement of the initial framework and new themes which emerged. It is more interrelated and comprehensive than our initial framework, but is less sequential in nature.

Comparing the frameworks the dimension *Goals* is a central in both. Next, the four elements from the initial framework—*Collecting*, *Combining*, *Reporting*, and *Decision Taking*—are closely related activities that are represented in one dimension in the new framework, named *Programme in Action*. With the exception of some changes in definition, the two frameworks are similar in this respect. In contrast, the analysis yielded a huge amount of information on *Quality Control*. It appeared that our first framework did not do justice to the diversity in activities related to quality and the importance the experts placed on this issue. Quality turned out to be multi-layered and integrated with *Goals* and the

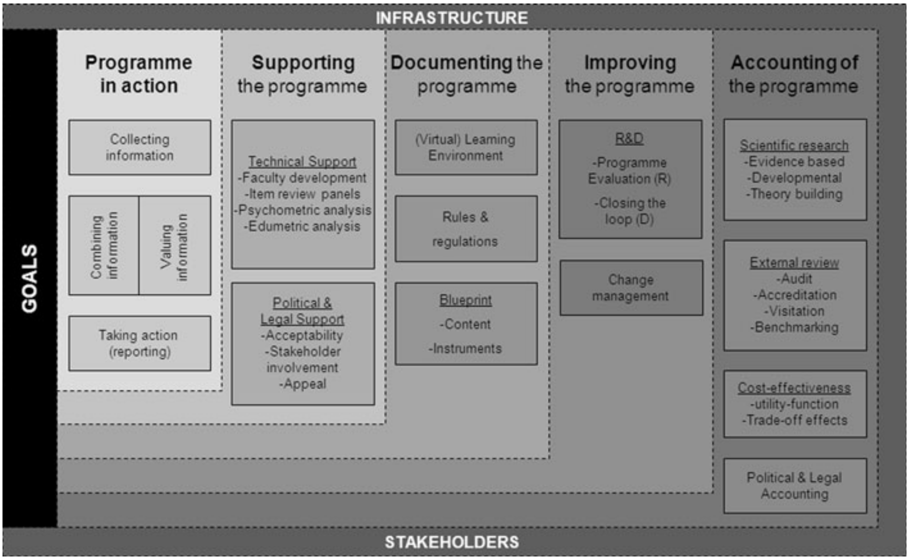

Fig. 2 New framework for programmes of assessment

Programme in Action in stead of a single element at the end of the process. In the final framework four layers (dimensions) were identified, which were placed on the same level as *goals* and *programme in action*. These are *supporting*, *documenting*, *improving*, and *accounting*.

Goals

Goals dominated the discussions, with experts typically linking ideas and suggestions to specific programme goals.

> I think another way to think about the goal at the top level is eh, that there should be a purpose statement to the assessment programme just as there should be a purpose statement to each of the components. [...] there should be a purpose of the assessment system that guides the whole of planning. (P8)

> ... did you meet your goals, there has to be some sort of relationship between the quality control and the purpose and the goals of what you are trying to do (P4)

Although *goals* were also part of our initial framework, we were struck by their unexpected centrality in almost every discussion on the other programme elements. Apparently, it was impossible to consider these elements in isolation from the goals of the assessment. The content of goals seemed to be of lesser importance, however.

> ... they are implied in goals which themselves will have a dynamic relationship to each other and to the context within it's being applied... (P6)

> ... cause the ones where they run into problems are where they're not agnostic where there is a religious devotion to a particular tool [and everything else has to fit in] and it is used for everything where it's not appropriate. (P2)

Regardless of educational concept (e.g. traditional education, problem-based learning) or the specific function of assessment (e.g. learning tool, licensing decisions), the quality of assessment programmes was framed in terms of *fitness for purpose*. This implies that clearly defined programme goals are prerequisite for high-quality programmes.

As *fitness for purpose* was regarded as the central premise of programme design, care should be taken to avoid a too normative view of design principles and quality criteria. Not all programmes are based on identical educational ideas. Today's popularity of competence-based programmes does not imply that a competence-based design should be the universal standard. Assessment aimed at selecting candidates uses different principles but that does not detract from their fitness-for-purpose.

Programme in action

The focus group discussions focused predominantly on *Programme in Action* or—in other words—on all the activities minimally required to *have* a running assessment programme. These activities encompass activities ranging from collecting information to taking action based on that information.

Emerging themes that were similar to elements of the initial framework were *collecting information*, *combining information*, *reporting*, and *decision making*, which were regarded as core activities of virtually any assessment programme. *Collecting information* was understood as referring to all activities for gathering the various kinds of information about assessees' abilities, including e.g. numeric (quantitative) data as well as descriptive (qualitative) data. Topics of consideration could be assessment content, selection of test formats, use of instruments, scoring systems, and scheduling of assessment.

With regard to *combining information*, an interesting distinction was made between technical and meaningful aspects. Technical aspects relate to combining data from multiple sources and combining different kinds of data. Combining data often seems a lot like comparing apples and oranges. For example, many programmes of assessment employ a compensatory test model (compensation of results on different items of the test or OSCE-stations) and a conjunctive model disallowing compensation between tests, (e.g. between an OSCE and an MCQ test on the same subject).

Using multiple instruments often results in a large amount data from different sources. In order to take an action based on a versatile and rich data set, interpretation of the data is needed to add value to the information collected. Meaningful aspects refer to the use of combined information, including interpretation, valuing, and selecting data. Although closely linked to—and sometimes intertwined with—combining data, *valuing* data was regarded as a separate element. So, in the new framework, *valuing information* is presented alongside *combining* information.

> Another common problem is that lots of sources of information are gathered but the system is not set up so that they are all considered […] they're not integrating and considering all of the material that is gathered… (P2)

> … the problem is how you can make it, so that you can get it in one place and that you can relate it to each and that you can understand the importance of different things and you can come to a judgment […] Don't inappropriately combine things which shouldn't be combined to force them together when they shouldn't be. (P6)

According to the experts, valuing information involved not only setting a pass-fail score, but also determining candidates' strengths and weaknesses or prioritising which learning goals to distil from the information provided by the assessment.

With regard to fitness for purpose, our initial definitions of *reporting* and *decision making* were too restrictively tied to common (summative) purposes of assessment, which—although general—are not necessarily universally applicable.

> But … there is an issue … about considering which stakeholders need to have this information or appropriate to have this information, so it is not a way of never giving it out. (P1)

> … but I don't agree either with the idea that every test provides feedback to every stakeholder, that to me, no… [Mod: It's depending on the goals]… the nature of the test will be greatly influenced by the feedback that will be given. (P2)

Based on these views, reporting and decision making were merged into a more generic element in the new model, *taking action*, which includes all activities resulting from the collected, combined and valued information relating to assessments. Without taking action, information from previous activities was considered pointless. Taking action implies closing the loop, and may vary from go/no-go decisions to feedback or even remediation. Taking action attaches consequences to assessments.

As *Programme in Action* focuses on core activities that have practical consequences and are essential to determine students' abilities, it deserves extensive attention. *In Action* signifies that conducting the activities is indispensable for any assessment. In summary, the four core activities of *Programme in Action* are: *Collecting Information*, *Combining Information*, *Valuing Information* and *Taking Action.*

Supporting the programme

Although the elements of *Programme in Action* suffice to establish a programme of assessment, they cannot guarantee a high standard. The activities contributing to the quality of the programme of assessment were more often than not related to, if not interwoven with, activities categorised under programme in action. In other words, a major part of the activities classified as relating to quality control in the initial framework appear to be qualified more appropriately as activities in support of the programme in action (activities).

For an activity to support the programme in action and contribute to overall programme quality it should be directed at the goals of the assessment programme. Supporting activities must ensure that the programme in action is of sufficient quality to contribute optimally to the purpose of the assessment programme.

Two support-related themes matched the concept of quality as fitness for purpose. One is *technical support*, contributing to the quality of assessment materials. A distinction was made between proactive activities before an assessment is conducted (e.g. item review panels, faculty development) and monitoring after the assessment (e.g. psychometric and other analyses). Test quality depends on *review*, which determines whether test items or elements meet the required characteristics. *Psychometric* and other analyses serve to determine the quality of an assessment and whether steps are needed to make improvements. As the success of an assessment depends largely on its users, *faculty development* is important to promote the quality of assessment programmes. The term *technical* also captures the knowledge, skills, and attitudes necessary for designing and conducting an educationally sound assessment system.

It was also pointed out that even a technically sound design of an assessment programme does not preclude the risk of failure due to resistance from stakeholders.

> you have to establish providence… do you have the right to do what you are doing […] you need to identify the people that are involved within that and then they need to go through a process by which there is agreement within those people and that could be stakeholders (P5)

The second support-related theme concerned *political and legal support*, targeted at increasing the acceptability of the assessment by early involvement of stakeholders and by putting in place an appeal procedure to avoid unfair conduct. Without acceptability, support will likely be insufficient to achieve high quality. Stakeholder involvement in the design of assessment programmes not only promotes input of creative ideas, but also ensures a certain fitness for practice. It can give stakeholders a sense of ownership of the programme, thereby gaining their support, without which goals can remain elusive. Issues related to (inter)national or local legal considerations need to be considered too and can influence the degrees of freedom in programme design.

> in court when you stand up and you go through this whole due process business it's whether or not every body was treated in equal manner, did everybody have an opportunity to demonstrate their abilities…(P5)

> … well the government has just passed a law that says every doctor will have a 360 degree appraisal every 5 years whether you need it or not. (P6)

Support-related actions have an immediate effect on the currently running assessment practice. Together with programme in action, *supporting the programme* forms a cyclic process aimed at optimising the internal assessment system.

Documenting the programme

Documenting assessment serves two purposes. Firstly, documentation will facilitate learning of the organisation by allowing the cyclic system of optimising the programme in action to function properly. Secondly, it enhances the clarity and transparency of the programme.

> That is an important point. Disclosure … about exactly what the procedures are going to be like and exactly how scores are going to be combined in psychometric characteristics I don't know whether that goes on reporting or something else… (P4)

Thus all the elements of programme in action and supporting the programme, including responsibilities, rights, obligations, rules, and regulations, must be recorded to ensure that the assessment process is unambiguous and defensible. Three elements deserve special attention in this respect.

Because assessment programmes do not function in a vacuum, it is of vital importance to address the first element, the *(virtual) learning environment and context* of a programme, which must be linked to the purpose of the assessment programme.

> I was thinking about the importance … eh, the purpose and the setting and the context in which this is occurring to a range of stakeholders who might very well have a view about how important it was, […] I think eh, in different circumstances of acceptability to quite a wide range of stakeholders as well. (P1)

The context and applicability of an assessment programme have to be clearly described. Stakeholders must be able to determine for themselves if and how the programme affects them.

Secondly, *rules* and *regulations*, establishes a reference for stakeholders to review the purpose of the assessment and the rights and duties of all stakeholders in relation to programme in action and supporting the programme. Often the conditions under which the assessment is to be conducted and specific demands on stakeholders can be captured in rules. Regulations describe the consequences and actions to be taken in specific (standard) situations. Responsibilities can be clearly defined and allocated on all levels of the programme, so that the proper person is approached in cases of errors or mistakes. Clear documentation of regulations can prevent shirking of responsibilities.

Obviously, in assessment design on any level content is part of the equation. Although there can be no assessment without content, the specific content does not influence the general design process. Because content is strongly related to assessment goals, it should however be recorded for future reference. So the third element, *blueprinting*, is a tool to map content to the programme and the instruments to be used in the programme. In this respect, it is strongly tied to the design principles relating to *information collecting*. Blueprinting can also be regarded as a tool to sample the domain efficiently.

To summarise, documenting the programme is about recording information that can help to establish a defensible programme of assessment and support improvement.

Improving the programme

Two different types of quality activities can be distinguished. We have described activities aimed at optimising the programme in the dimensions *supporting* and *documenting*. But, another type is aimed at *improving the programme* in response to critical appraisal from a more distant perspective. Activities in this dimension generally have no immediate effect on the currently running programme, but take only effect as they become apparent in the (re)design of (parts of) the programme, usually at a later date.

Most improvement activities involve *research and development* aimed at careful evaluation of the programme to ascertain problematic aspects. It is imperative, however, that the evaluation loop should not stop at data gathering: it must be closed by the actual implementation of measures to address diagnosed problems.

> … the goals change because the professional needs change and if it's frozen in time …, that's not good; so it means … some concept of periodically revisiting the effectiveness of the whole system somehow (P2)

> Is there something also about closing the loop, I mean there is no point in evaluating side-effects if you never have some mechanisms in place for putting it right. (P7)

Apart from measures to solve problems in a programme, political change or new scientific insights can also trigger improvement. A concept that cropped up in relation to improvement was *change management*, comprising procedures for change and activities to cope with potential resistance to change. (Political) acceptance of changes refers to changes in (parts of) the programme.

> we haven't had the concept, yet… but it is so important in assessment systems is this idea of change management and how you, you know, move from one approach to another if it's starting the evidence is starting show a good idea eh who says what when and how and the impact. (P6)

> … eh implementation is part of change management to me, take something from nothing and you implement it but they actually test the administration (P5)

Improvement is driven by the purpose of the assessment programme, which determines whether a change is an improvement or not. What may be an improvement for a licensing institute may be a change for the worse in an educational programme and vice versa.

Accounting for the programme

While the previous dimensions of the framework related to internal aspects of the institution or organisation responsible for the assessment programme, *Accounting for the programme* relates to the increasing demand for public accountability. The purpose of activities in this dimension is to defend the current practices of the programme in action and demonstrate that goals are met in light of the overarching programme goals. Accounting for the programme deals with the rationale of the programme.

Four major groups of accounting activities can be distinguished. The experts identified a need for *scientific research*, frequently attributing uncertainty about assessment activities to a lack of research findings and calling for research to support practices with sound evidence, which is in line with the prominence in medicine of the drive for evidence-based practice.

> well we said everything had to be evidence-based I mean if you don't have some sort of research programme or you don't have some sort of reporting mechanism then I'll never be able to prove to you that was right so I agree […] things should be either proven or being in a research mode or some research and development. (P5)

The influence of scientific research is also manifest in the application of new scientific insights to assessment programmes.

Accountability also requires *external review* of programmes of assessment. A common method is external review by outside experts, who judge information on the programme and in some cases visit an institution to verify information and hear the views of local stakeholders. External review is generally conducted for accreditation and benchmarking purposes.

> Actually that is a good principle from time to time, the processes put in place, should be reviewed by an outside body or somebody who is less associated with… (P5)

Assessment programmes are also shaped by the needs and wishes of external stakeholders. As assessment programmes do not exist within a vacuum, *political and legal* requirements often determine how (part of) the programme of assessment has to be (re)designed and accounted for.

In every institution or organisation, resources—including those for assessment programmes—are limited. *Cost-effectiveness* is regarded as a desirable goal. Although fitness for purpose featured prominently in the discussions, the experts thought more attention should have been paid to accountability and especially to costs, which can be a formidable obstacle to new ideas. The success of assessment programmes often hinges on the availability of resources. Obviously, greater efficiency is desirable but there is a cost-benefit trade-off. In other words, the quality of a programme is also defined in terms of the extent to which it enables the attainment of the goals, despite the boundaries of available resources.

Discussion

The main purpose of this study was to produce a framework for programmes of assessment with appropriate dimensions for design. The model that resulted from the focus group discussions with experts was far more comprehensive and integrated than the model used to guide the discussions. The quality of assessment in particular turned out to be a much broader dimension than we had envisaged. During the focus group meetings it became clear that—even though there was general agreement on topics with relevance to programmes of assessment—a shared frame of reference for programmatic assessment was glaringly absent. As a consequence, while some elements of assessment received a lot of attention, others remained underexposed.

We believe the model described in this paper can help to frame programmes of assessment, because it not only provides a common language (shared mental model) for programme developers and users but also a more comprehensive picture of the dimensions to be covered when formulating design principles. However this makes it hard to relate our findings to previous research. Where research is done on design criteria with respect to assessment it, focuses on specific, isolated elements, and where research is done at the level of assessment programmes is does not focus on design, but for example on quality in terms of content, validity, reliability, or alignment with education (Biggs 1996; Harlen 2007; Baartman 2008). This is not to say that all elements of the model we propose are completely new. There is for example good research on the combinations of information from various assessment methods; not only at the level of conjunctive versus compensatory combinations but also about how scores correlate between tests with identical content than between tests with identical format (Van der Vleuten et al. 1989) Yet most assessment programmes still allow for full compensation between format-similar elements (the separate stations in an OSCE) and not between format dissimilar elements (e.g. combining scores on an OSCE station with scores on a content-similar written test). Such a paradox cannot be resolved when one designs an assessment programmes starting from the individual methods, only a programmatic design perspective may be useful here.

A central concept was that high quality assessment and the activities needed to achieve it can only be defined in terms of the goals of an assessment programme. Goals underpin the guiding principle of programme design: *fitness for purpose*. Quality is inextricably interwoven with goals, which are closely tied to all activities related to assessment. Achieving appropriate interrelatedness of goals and activities requires design principles that are prescriptive, but take into account context and/or specific goals. Thus normative statements can only be included in design principles with explicit reference to specific purposes.

To explain and support this argument further we come back to our most important and maybe most obvious finding that quality of an assessment programme can only be judged in light of its purpose. The purpose of an assessment programme is often not included in research on relations between separate elements of an assessment system. In studying these relations the outcome measure should be what is the optimal configuration to contribute to our goals.

Initially we took a same isolated approach when drawing up our initial model to guide the focus groups, in which we defined discrete and sequential steps. The new model values interrelatedness and complexity of assessment, while undeniably, an intuitively logical sequence retains. For example within the *programme in action* (first collect, then combine and value, and finally take action), but this sequence can also be reversed, especially from the design point of view. Key is the interrelatedness of the elements within the framework for the design of assessment programmes that resulted from this study.

Remarkably, the prime focus of the discussions was the programme in action and, within this dimension, collecting information. This is not surprising since this dimension deals with the core activities of assessment and the visible aspects of the assessment process. The experts disapproved of what they regarded as an obsession with assessment tools in the assessment literature, whereas elements like accreditation standards tended to be neglected. We think that our model can attenuate this obsession by raising awareness that programmatic assessment consists above all of variegated components which are integrated and interconnected and bear no resemblance whatsoever to an assessment toolkit with different instruments suited to specific tasks.

When we looked at the literature from the perspective of the new model, a similar picture presented itself. It seems that in terms of our model the topics of the literature on assessment can largely be categorised as *collecting information* and as the major elements of programme in action and supporting the programme. Regrettably, the interrelatedness of these elements is largely ignored, which is only to be expected as they are generally considered in isolation, an approach that has also characterised the search for the one superior instrument for each type of test to which we referred earlier.

The focus group approach fitted the purpose of this study, which was to explore experts' experiences and ideas on the largely uncharted topic of programmatic assessment. The experts agreed that so far little work had been done on programmatic approaches to assessment, also by themselves, and that the discussions had been enlightening. However, the focus groups had limitations as well. The selection of experts was biased by our social network and field of educational expertise (medical education), and the group was small. Although we are convinced that the experts were open minded, their long-standing experience and fields of interest may have given rise to some blind spots. Although they had been instructed to think outside the box, during the wrap-up evaluation the experts expressed concern that the discussion had been heavily dominated by what they were most comfortable with or where their experience was. Their fear was that the discussion had resulted in more traditional ideas than intended. Yet the data gave rise to many new insights and ideas, reinforcing our resolve to move this research forward. Experts are only one source of information, so we will have to triangulate the results by tapping into other sources of information, such as the opinions of teachers and medical students as end-users of assessment programmes.

Although the new model is comprehensive, it is possible that relevant issues were overlooked in the discussions leaving gaps in our model that need to be filled by further research. The question is how. It was suggested that incorporating ideas from other cultures and practices could generate fresh ideas, admittedly with a concomitant risk of reduced generalisability as was illustrated during the discussions. These were sometimes less general than intended due to cultural differences between educational settings (undergraduate, postgraduate and continuing education) and countries of origin of the experts. So this note of caution on generalisability applies equally to our model because the experts' experiences and views were inevitably contextual. Although we strove to keep the model general and applicable to different contexts, it would be interesting to investigate its applicability (robustness) in different cultural contexts. A further concern about the application of criteria in different contexts led to the recommendation to look to a wider context (for example society at large) as a possible framework to make the general criteria transferable to different contexts.

Numerous ideas worth pursuing were produced by our study, pointing the way to topics of further research. One obvious next step would be to apply this model to an existing assessment programme and determine whether all the dimensions and elements are identifiable and relevant. Further steps could also include producing concrete design criteria and validating them by application to existing programmes of assessment.

Acknowledgments The authors would like to thank all experts for their participation in the discussion: Dr. David Blackmore, Dr. Susan Case, Dr. Bob Galbraith, Dr. Brian Hodges, Prof. Pauline McAvoy, Prof. Peter McCrorie, Dr. John Norcini, Prof. Dame Lesley Southgate, and Prof. Val Wass. They would also like to thank Kristof van Royen for his help with the transcription of the audio recordings and Mereke Gorsira for the useful comments and editing of this paper.

References

Baartman, L. K. (2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes.* Dissertation, Universiteit Utrecht.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364.

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33*(1), 29–49.

Burch, V., Norman, G., Schmidt, H., & van der Vleuten, C. (2008). Are specialist certification examinations a reliable measure of physician competence? *Advances in Health Sciences Education, 13*(4), 521–533.

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

Harlen, W. (2007). Criteria for evaluating systems for student assessment. *Studies in Educational Evaluation, 33*(1), 15–28.

Hollis, V., Openshaw, S., & Goble, R. (2002). Conducting focus groups: Purpose and practicalities. *British Journal of Occupational Therapy, 65*, 2–8.

Knight, P. T. (2000). The value of a programme-wide approach to assessment. *Assessment & Evaluation in Higher Education, 25*(3), 237–251.

Lew, S. R., Page, G. G., Schuwirth, L. W. T., Baron-Maldonado, M., Lescop, J. M. J., Paget, N. S., et al. (2002). Procedures for establishing defensible programmes for assessing practice performance. *Medical Education, 36*(10), 936–941.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63–S67.

Newble, D., Dawson, B., Dauphinee, W., Page, G., MacDonald, M., Swanson, D., et al. (1994). Guidelines for assessing clinical competence. *Teaching and Learning in Medicine, 6*(3), 213–220.

Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine, 70*(3), 194–201.

Schuwirth, L. W. T. (1998). *An approach to the assessment of medical problem solving: computerised case-based testing*. Dissertation, Maastricht University.

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

Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education, 38*(9), 974–979.

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

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

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

Van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58–76.

Van der Vleuten, C. P. M., Van Luyk, S. J., & Beckers, H. J. M. (1989). A written test as an alternative to performance testing. *Medical Education, 23*(1), 97–107.

Wass, V., McGibbon, D., & Van der Vleuten, C. (2001). Composite undergraduate clinical examinations: How should the components be combined to maximize reliability? *Medical Education, 35*(4), 326–330.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7–25.