Guidelines for designing programmes of assessment

Guidelines for designing programmes of assessment

Joost Dijkstra

The research reported here was carried out at

Maastricht University Leading in Learning!

in the

SHE School of Health Professions Education

ISBN 978 94 6159 344 3

Cover design: Datawyse
Production: Datawyse / Universitaire Pers Maastricht

Guidelines for designing programmes of assessment

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. dr. L.L.G. Soete,
volgens het besluit van het College van Decanen
in het openbaar te verdedigen
op woensdag 25 juni 2014 om 10.00 uur

door

Joost Dijkstra

UPM
UNIVERSITAIRE
PERS MAASTRICHT

Promotores
Prof. dr. C.P.M. van der Vleuten
Prof. dr. L.W.T. Schuwirth

Co-promotor
Dr. M.J.B. Govaerts

Beoordelingscommissie
Prof. dr. S. Heeneman (voorzitter)
Prof. dr. J. Cohen-Schotanus (UMC Groningen)
Prof. dr. T. De Goeij
Prof. dr. A.D.C. Jaarsma (AMC Amsterdam)
Prof. dr. M. oude Egbrink

Table of Contents

Chapter 1

Introduction

Quality and design of assessment

Coming from a long tradition, research on assessment in medical education has been mainly focussed on single assessment instruments. This focus fits the typical assessment approach of dividing competence into separate, individually measurable elements (e.g. knowledge, skills, attitude, problem solving). The aim of most of these studies was to achieve the best possible instrument to measure these separate elements. This measurement approach to assessment builds on research and theory in psychological testing and, consequently, approaches to quality of the assessment are consistent with the psychometric framework. Quantifiable measures of assessment quality, such as reliability, are key in the evaluation of assessment instruments. This wealth of research in past decades has provided valuable insights into the strengths and weaknesses of single assessment instruments used in medical education (Van der Vleuten et al., 2010). It has further resulted in an extensive toolbox with assessment instruments that can be used to assess many different competency areas or skills (e.g. the Objective Structured Clinical Examination for measuring technical/procedural skills (Van der Vleuten and Swanson, 1990); key feature testing for problem solving ability (Page et al., 1995; Schuwirth, 1998), and Mini-CEX for assessment of performance in real life (authentic) situations (Norcini et al., 1995).

With this assessment approach, decisions about student achievement are typically based on the collection and combination of the separate outcomes of each of the examinations without taking into account if and how these building blocks represent a complete and integrated picture of professional competence. Results on a written multiple choice test measuring knowledge are combined with results on an OSCE for skills and professional behaviour assessment for attitude. Simply adding up or lumping together individual and independent exams does not capture competence comprehensively. This division of one test per element leads to loss of integration, as the total is more than the sum of its parts. Competence in regarded as a whole task, which cannot be broken down into separate parts. Competence does not consist of one-dimensional traits, but is a complex integrated construct (Schuwirth and Van der Vleuten, 2006). Assessing only easily measured elements of competence in isolation might provide a skewed view on the qualities of an individual. 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. Furthermore, while acknowledging the importance of psychometrics, it is clear that exclusively focussing on psychometrics is an insufficient basis (Schuwirth and Van der Vleuten, 2006) for selecting assessment instruments. It is a well-accepted fact that decisions about selecting assessment instruments require trade-offs between various quality aspects. Not only should reliability and validity be taken into account, but educational impact, acceptation, and costs need to be considered too (Newble et al., 1994; Schuwirth and Van der Vleuten, 2004; Van der Vleuten, 1996).

More important, assessment in medical education entails more than just determining competence (assessment of learning). Multiple and divergent goals also need to be addressed by assessment, such as facilitating or influencing development (assessment for learning) as well as evaluating instruction (quality improvement). Any single instrument – each with its own specific strengths and weaknesses – only has certain (limited) value and therefore cannot meet all the assessment purposes completely, nor are they able to accomplish even a single purpose. So, assessment in medical education requires a carefully designed assessment programme, consisting of a purposeful mix of various assessment components that correspond with the goals of assessment (and/or the curriculum at large) in the best possible way. Similar to an exam being more than a random sample of items, a programme of assessment should be more than a random selection of separate instruments. In order to grasp the complete picture of the qualities of an individual, assessment should be approached from a broader holistic perspective (Vleuten and Schuwirth, 2005).

A programmatic approach to assessment design is advocated in order to help assessment developers in dealing with the complexity of the design process and combining multiple assessment purposes (Lew et al., 2002; Schuwirth et al., 2002; Van der Vleuten and Schuwirth, 2005). The result of a programmatic approach to assessment is not just an arbitrary set of assessment methods, but a well-designed programme of assessment. This design must take into account more than just the strengths and weaknesses of separate assessment components. It must also include how the results of these components should be combined and the context in which assessment has to take place. Inevitably, such an approach does not consider assessment merely a measurement problem, but more as an educational design problem in which trade-off decisions have to be made to deal with dilemmas.

Designing assessment programmes in medical education settings is a complex process influenced by a broad range of factors, such as scheduling of assessments, combining different assessment outcomes and exam regulations, that have to be taken into account in order to optimally achieve assessment purposes. Contextual factors such as organisational and/or political issues may have a strong influence on the design process. Similarly, available infrastructure and resources may necessitate even further compromises, which may limit the options for design even more. Other sources of influence can be the pressure to progress students or avoid attrition (which is often a political and financial issue), and the required justification to the public needs. Obviously, assessment should do more than just correspond to the context and support the instructional design of the curriculum (or the educational philosophies); conflicts of interest may lead to necessary decisions that negatively influence the educational quality. Serving multiple purposes and needs makes assessment of the whole picture (of competence) complex and challenging. Assessment programmes that are perceived to be of high quality in one particular context may not be suitable in other contexts. We need, therefore, guidelines that not only provide a framework for design of an integrated assessment of professional competence, but are also applicable (or easily adaptable) in a broad range of assessment and/or educational contexts.

In contrast to the amount of literature on the quality of single instruments, literature that addresses what constitutes the quality of programmes of assessment and how to balance various factors in programmatic approaches to assessment design is scarce. The available standards for educational and psychological testing (AERA et al., 1999) contain generic descriptions of quality in testing, covering a wide range of assessment instruments and activities that are relevant in the assessment of professional competence. However, guidance provided by these standards mainly focuses on the quality of individual tests (i.e. measurement instruments), with far less attention to embedding individual instruments in a specific assessment programme.

The lack of scientific evidence or guidance for programmes of assessment does not imply that existing programmes are not of high quality. There are various examples of good programmes of assessment which are based on extensive deliberation and which are designed by experts (Dannefer and Henson, 2007; Davies et al., 2009; Ricketts and Bligh, 2011). Unfortunately though, there is little research in this area that would help to support or improve their quality. Several recent papers describe programmes of assessment that illustrate the added value of a programmatic approach, e.g. triangulation of information from multiple assessments (Dannefer and Henson, 2007). However, these examples are hard to replicate in other settings because these programmes are designed for a specific local context (e. g. Dannefer and Henson, 2007; Ricketts and Bligh, 2011). Scientific evidence on quality of such programmes as a whole is currently limited and in need of theory formation and applicable research outcomes.

Criteria for designing integrated, purposeful, high-quality comprehensive assessment programmes that also assure this quality are not readily available in the literature. Little is known about key relations, compromises and trade-offs needed to design and implement truly integrated assessment programs that serve assessment purposes as intended. Early developments determining quality of programmes of assessment focused on alignment of objectives, instruction and assessment to achieve congruent student behaviour (e.g. Biggs, 1996). Basic principles underlying constructive alignment focus on blueprinting assessment based on curriculum objectives, and criteria within this approach therefore tend to be limited by their focus on assessment content (Webb, 2007).

Other studies on assessment quality that evaluate combinations of different assessment instruments often apply psychometric approaches only, e.g. combined reliability estimates (Moonen-van Loon et al., 2013; Wass et al., 2001; Verhoeven et al., 2000; Harlen, 2007). Research on high-stakes assessment programmes for the certification of physicians typically aims for high composite reliability (e.g. Burch et al., 2008; Knight, 2000;

Wass et al., 2001). In addition, their applicability is often limited to specific well-described educational philosophies or contexts, such as competency-based assessment programmes (e.g. Baartman, 2008). Baartman (2008) took competency-based education as a basis for quality, and proposed adding education-based criteria, such as authenticity and meaningfulness, to the established psychometric criteria.

General aim and research questions

The previous description of the available criteria and guidance for assessment design clarifies the scarcity of literature addressing guidance for programmatic assessment design. The overall aim of this research is to provide support for achieving high-quality assessment programmes. Preferably, it would include the development of guidance to support programmatic assessment design decisions, and deal with dilemmas and trade-off decisions in a wide variety of contexts independent of educational philosophies and suitable for all kinds of goals of assessment. Therefore, we take a utilitarian approach, whereby quality is defined as fitness-for-purpose (Harvey and Green, 1993). Quality is determined based on the extent to which a programme of assessment fulfils its purpose or its function. The advantage of this perspective is that it makes the quality framework more widely applicable and less reliant solely on the current ideas on education and assessment. In contrast to an ideological or a deontological perspective, a utilitarian approach does not prescribe what the criteria should be. Hence, from this perspective, quality *criteria* are not a goal in and of itself. From a utilitarian perspective, the term 'guidelines for assessment design' is more appropriate than 'criteria for quality of assessment'. For example a criterion from an ideological perspective can be: 'an assessment programme should have summative tests', whereas from a utilitarian approach a *guideline* could be formulated as: 'the need for summative tests should be considered regarding the purpose'. In different contexts assessment designers need to decide how important or relevant a guideline is, and use their own expertise to make decisions based on specific contextual circumstances. In this sense the application of guidelines is eclectic. From a fitness-for-purpose view, weaknesses of assessment components can be perfectly acceptable if the strengths contribute optimal or sufficiently to the purpose of assessment.

In order to create guidelines, the areas or elements that constitute a complex design of assessment must first be determined. This leads to the first research question (RQ).

RQ1: What areas or elements can be distinguished in the design of high-quality assessment programmes?

If these areas are determined, the next step is to develop guidelines based on a utilitarian approach, as described above.

RQ2: What guidelines can be formulated for design support based on the areas of assessment design?

The success of implementation and application of guidelines depends on the evidence that supports them (Basinski, 1995). Therefore our next RQ is:

RQ 3: What evidence can be provided to substantiate the validity of guidelines based on utilitarian principles in practice?

Clear validation procedures for guidelines for educational design are not readily available. However, the validation process of the description of areas and elements relevant to assessment design are similar to the development of theories or frameworks and evaluation of clinical guidelines. To that end, evidence to support guidelines that apply utilitarian principles is based on Basinski's (1995) work on evaluation of guidelines and on the validation process of Prochaska et al. (2008) who describes a validation process for theory development.

Based on Basinski (1995), the provision of evidence for a framework for assessment design and guidance is divided into three phases: (1) evaluation during the development; (2) evaluation of programmes in which guidelines play a role; and (3) scientific evaluation. Several criteria for theory building (Prochaska et al., 2008) are applicable to developing guidelines for assessment development and can be linked to these phases. In order to evaluate the guidelines we conducted several studies, which will be briefly introduced in the following section.

Overview of studies

Chapter 2 describes the development of a framework for assessment programmes and determining which areas and elements have to be covered when formulating design guidelines. Because of the absence of a common vocabulary for programmatic assessment, we used an exploratory, open, qualitative method to probe the views and ideas of experts in assessment in medical education. This resulted in an overarching framework for programmatic assessment, which defines the scope of what constitutes a assessment programme, and should be covered by our guidelines.

Chapter 3 describes how a set of **g**uide**l**ines for **a**ssessment **d**esign (GLAD) was derived from this framework. Because the aim of this study was to formulate guidelines that are general enough to be applicable in a variety of contexts, and yet at the same time meaningful and concrete enough to support assessment designers, we started by generating ideas for guidelines based on our framework for programmes of assessment using the input of international experts in the field of assessment in medical education. In this first phase of gathering validity evidence *during* the development of guidelines, we used a consensus procedure around expert evaluation, focussing on achieving *clarity*, *consistency*, and *parsimony* (Prochaska et al., 2008) of the guidelines. More specific, attention was given to creating explicit terminology and defining the guidelines carefully. The guidelines were grouped logically to avoid any contradiction with each other. Finally, complexity as well as redundancy of the guidelines was minimized. This led to a comprehensive set of guidelines.

Chapters 4 and 5 describe the next steps in the validation process. In Chapter 4 the evaluation of GLAD was done in a real life setting. An instrumental case study and a multiple qualitative inquiry two-step approach were used to evaluate the *practicality* and *explanatory power* of GLAD (Prochaska et al., 2008). The practicality of GLAD was investigated through document analysis and interviews with multiple stakeholders in the assessment process. More specific, we investigated if GLAD are found in actual practice and if they are taken into account during the process of design. Results yielded in-depth information about decisions and considerations made during the design process. Based on the results from the practicality evaluation, the explanatory power of GLAD was investigated through analysis of statements about quality of the assessment programme, as perceived by relevant stakeholders. Explanatory power is determined by the ability to describe and evaluate these statements in terms of GLAD and explained by the outcomes of the practicality analysis.

The second case study as described in Chapter 5 aims to investigate the *utility* and *productivity* of GLAD (Prochaska et al., 2008). The utility of GLAD in the evaluation of assessment programmes was investigated by comparing evaluation outcomes and processes to a well-researched and validated set of quality criteria for Competence Assessment Programmes (CAP). A competence based assessment programme was purposefully selected and was evaluated using a two-step procedure. Firstly, we evaluated the programme using GLAD by conducting interviews and document analysis. Secondly, the programme was evaluated by the CAP criteria using a self-evaluation tool followed by a group interview (see: Baartman et al., 2007). Both evaluations are an interpretation of an in-depth qualitative analysis of the assessment programme.

The *productivity* of GLAD is determined by investigating whether GLAD contributes to existing research. More specifically, the productivity in this study looks at whether GLAD adds to the established and validated CAP criteria. Conclusions are based on comparison in evaluation outcomes, and especially the scope of the evaluation outcomes and the areas of assessment programmes that are addressed.

Finally, in Chapter 6, the main findings are summarized and discussed, and implications of this work are explored. Suggestions are presented for future development and evaluation of support for designing programmes of assessment

This dissertation consists of related articles. Since, each chapter was written to be read on its own, repetition and overlap across chapters are inevitable.

References

AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing*. AERA, Washington.

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

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Determining the Quality of Competence Assessment Programs: A Self-Evaluation Procedure. *Studies in Educational Evaluation, 33*(3-4), 258-281.

Basinski, A. S. (1995). Evaluation of clinical practice guidelines. *Canadian Medical Association Journal, 153*(11), 1575-1581.

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

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.

Davies, H., Archer, J., Southgate, L., & Norcini, J. (2009). Initial evaluation of the first year of the Foundation Assessment Programme. *Medical Education, 43*(1), 74-81.

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

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

Knight, P. T. (2000). The Value of a Programme-wide Approach to Assessment. *Assessment & Evaluation in Higher Education, 25*(3), 237-251.

Moonen-van Loon, J. M. W., Overeem, K., Donkers, H. H. L. M., Vleuten, C. P. M., & Driessen, E. W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*, 1087–1102.

Newble, D., Dawson, B., Dauphinee, D., Page, G., Macdonald, M., Swanson, D. et al. (1994). Guidelines for assessing clinical competence. *Teaching and Learning in Medicine: An International Journal, 6*(3), 213 - 220.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The Mini-CEX (Clinical Evaluation Exercise): A Preliminary Investigation. *Annals of Internal Medicine, 123*(10), 795-799.

Page, G., Bordage, G., & Allen, T. (1995). Developing Key-feature Problems and Examiniations to Assess Clinical Decision-making Skills. *Academic Medicine, 70*(3), 194-201.

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model. *Applied Psychology: An International Review, 57*(4), 561-588.

Ricketts, C., & Bligh, J. (2011). Developing a Frequent Look and Rapid Remediation Assessment System for a New Medical School. *Academic Medicine, 86*(1), 67-71.

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

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 psychometric models in educational assessment. *Medical Education, 40*, 296-300.

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

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 practice & research. Clinical obstetrics & gynaecology, 24*(6), 703-719.

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., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine: An International Journal, 2*(2), 58 - 76.

Verhoeven, B. H., Hamers, J. G. H. C., Scherpbier, A. J. J. A., Hoogenboom, R. J. I., & Vleuten, C. P. M. v. d. (2000). The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. *Medical Education, 34*(7), 525-529.

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.

Chapter 2

A new framework for designing programmes of assessment

Published in Advances in Health Science Education
2010, Vol 15, 379-393

Joost Dijkstra,
Cees P.M. Van der Vleuten,
Lambert W.T. Schuwirth

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 chapter 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.

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.

1. 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.
2. 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.
3. 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.
4. 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 et al., 2004). 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 chapter.

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 Figure 2.1). 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*).

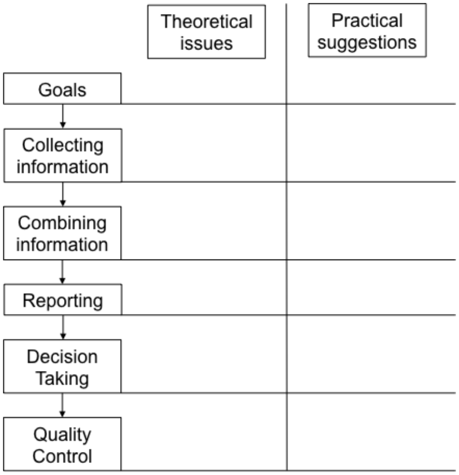

Figure 2.1: Initial framework

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 and 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 Figure 2.2) 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.

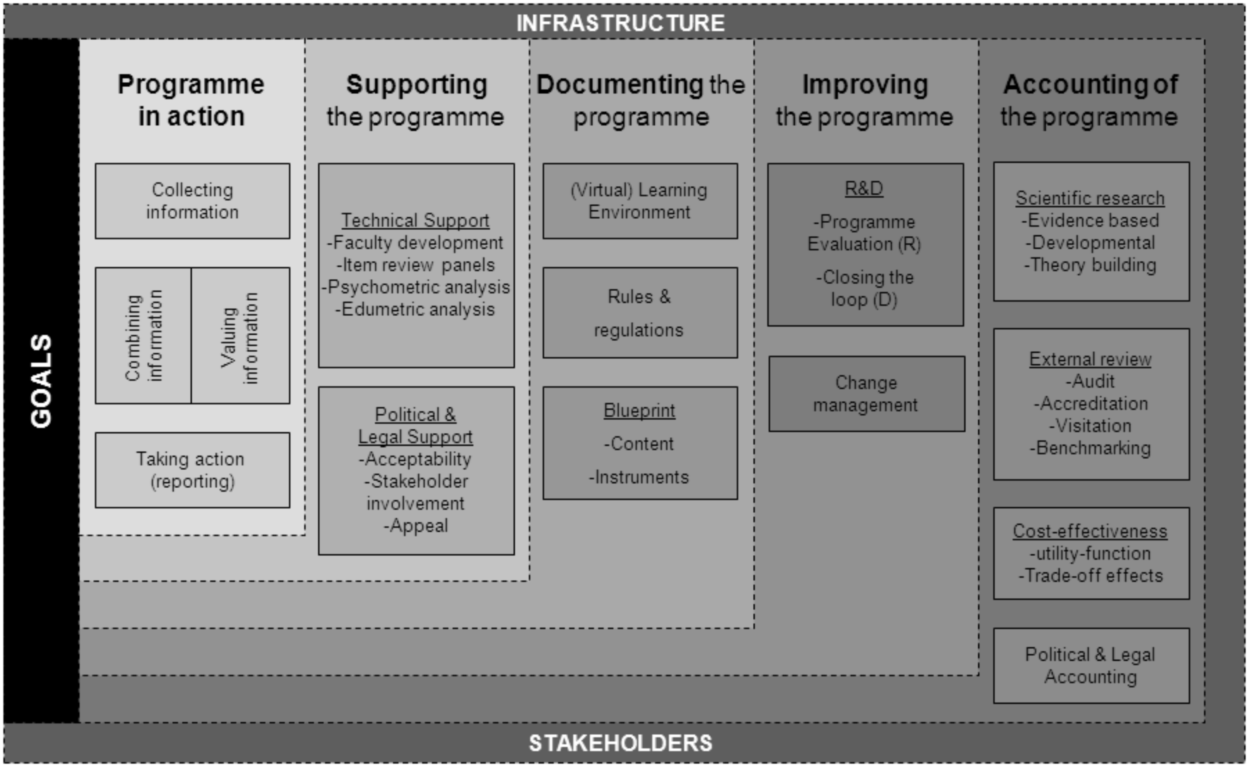

Figure 2.2: New framework for programmes of assessment
****Note: Figure as published in Advances in Health Science Education 2010, Vol 15, 379–393***
See Figure 3.1 (Page 41) for an updated version

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 *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 chapter 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., 1988) 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.

References

Baartman, L. K. (2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes*. Universiteit Utrecht, 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. *The 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), S63-67.

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. (1998). *An approach to the assessment of medical problem-solving: computerised case-based testing.* Maastricht: 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 psychometric models in educational assessment. *Medical Education, 40*, 296-300.

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

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

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: An International Journal, 2*(2), 58 - 76.

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.

Chapter 3

Expert validation of fit-for-purpose guidelines for designing programmes of assessment

Published in BMC Medical Education; 2012, 12:20

Joost Dijkstra, Robert Galbraith,
Brian D. Hodges, Pauline A. McAvoy,
Peter McCrorie, Lesley J. Southgate,
Cees P.M. Van der Vleuten, Val Wass,
Lambert W.T. Schuwirth

Abstract

An assessment programme, a purposeful mix of assessment activities, is necessary to achieve a complete picture of assessee competence. High quality assessment programmes exist, however, design requirements for such programmes are still unclear. We developed guidelines for design based on an earlier developed framework, which identified areas to be covered. A fitness-for-purpose approach defining quality was adopted to develop and validate guidelines.

First, in a brainstorm, ideas were generated, followed by structured interviews with 9 international assessment experts. Then, guidelines were fine-tuned through analysis of the interviews. Finally, validation was based on expert consensus via member checking.

In total 72 guidelines were developed and in this chapter the most salient guidelines are discussed. The guidelines are related and grouped per dimension of the framework. Some guidelines were so generic that these are applicable in any design consideration. These are: the principle of proportionality, rationales should underpin each decisions, and requirement of expertise. Logically, many guidelines focus on practical aspects of assessment. Some guidelines were found to be clear and concrete, others were less straightforward and were phrased more as issues for contemplation.

The set of guidelines is comprehensive and not bound to a specific context or educational approach. From the fitness-for-purpose principle, guidelines are eclectic, requiring expertise judgement to use them appropriately in different contexts. Further validation studies to test practicality are required.

Background

There is a growing shared vision that a *programme* of assessment is necessary to achieve a coherent and consistent picture of (assessee) competence (Lew et al., 2002; Schuwirth et al., 2002; Van der Vleuten and Schuwirth, 2005; Savage, 2006). A programme is more than a combination of separate tests. Just as a test is not simply a random sample of items; a programme of assessment is more than a random set of instruments. An optimal mix of instruments should match the purpose of assessment in the best possible way. However, there is less clarity about what is actually needed to achieve an integrated, high quality programme of assessment. Little is known about key relations, compromises, and trade-offs needed at the level of a highly integrated programme of assessment (Dijkstra et al., 2010). This does not imply that existing programmes of assessment are not of high quality, indeed there are numerous examples of good programmes of assessment which are based on extensive deliberation and which are designed by experts (Dannefer and Henson, 2007; Davies et al., 2009; Ricketts and Bligh, 2011).

However, scientific evidence on quality of such programmes in its entirety is currently limited, and certainly in need of theory formation and applicable research outcomes. The scant research that has been conducted into the quality of programmes of assessment, focuses on various aspects of assessment, with different aims and adopting diverse viewpoints on quality, and the results of the individual studies therefore are hard to compare. From a psychometric perspective quality has been almost exclusively defined as the reliability of combinations of decisions and a 'unified view of validity' (Birenbaum, 2007; Burch et al., 2008; Harlen, 2007; Knight, 2000; Wass et al., 2001).

From an educational perspective the focus has been on the alignment of objectives, instruction, and on using assessment to stimulate desirable learning behaviour (Biggs, 1996; Cilliers et al., 2010; Cilliers et al., 2011). In another study Baartman (2008) took competency-based education as a basis for quality, and proposed adding education-based criteria, such as authenticity and meaningfulness, to the established psychometric criteria. Most of this research determines assessment quality afterwards, when assessment has already taken place. Unfortunately, this does not provide assessment designers with much support when they intend to construct a high-quality programme. In our study we therefore investigate the possibility of enhancing quality of assessment programmes from a design perspective by providing guidelines for assessment design.

In various local contexts standards, criteria, and guidelines are used to support assessment development. However, the transferability of these to other contexts is fairly low as they are highly contextual and often based on local policy decisions. On the other hand guidance is available at a broader educational level, e.g. the Standards for educational and psychological testing (AERA et al., 1999). But these standards focus predominantly on single tests (i.e. the measuring instrument) instead of on programmes of assessment. And, despite the standards being open to expert judgement and acknowledging contextual differences (e.g. in regulations), they are still formulated from a specific testing framework and from the perspective of *assessment of learning* (Schuwirth and Van der Vleuten, 2011). This predetermines the goal of assessment and takes an ideological standpoint in the quality perspective and as a result, such standards are necessarily prescriptive. So, our aim in this study is to develop and validate more context-independent guidelines, applicable with different purposes in mind (including *assessment for learning*), and with a focus on programmes of assessment instead of single instruments. In addition we seek to develop and validate guidelines that support both assessment developers and decision makers. In this study we adopted the *fitness-for-purpose* principle (Dijkstra et al., 2010; Harvey and Green, 1993), in which quality is determined as the extent to which a programme of assessment fulfils its purpose or its function. The advantage of this is that it makes the quality framework more widely applicable and less reliant on contemporary ideas on education and assessment. From the fitness-for-purpose perspective defining *criteria* is avoided, and instead *design guidelines* are formulated. For example, a quality criterion would be: ‘An assessment programme should have summative tests’, whereas a guideline would be: ‘The need for summative tests should be considered in light of the purpose.’ Given the fitness-for-purpose principle the application of the guidelines are necessarily eclectic. In different contexts assessment designers need to decide how important or relevant a guideline is, and use their own expertise to make decisions based on specific contextual circumstances.

In this chapter we propose a set of design guidelines for programmes of assessment, based on a framework developed in our previous research (Dijkstra et al., 2010). This framework defines the scope of what constitutes a programme of assessment and should be covered by our guidelines (see Figure 3.1).

The framework is divided into several dimensions and is placed in the context of *stakeholders* and *infrastructure* (outer layer). The starting point is the *purpose of the programme* (key element in the framework). Around the purpose, 5 layers (dimensions) were distinguished. (1) *Programme in action* describes the core activities of a programme, i.e. collecting information, combining and valuing the information, and taking subsequent action. (2) *Supporting the programme* describes activities that are aimed at optimizing the current programme of assessment, such as improving test construction and faculty development, as well as gaining stakeholder acceptability and possibilities for appeal. (3) *Documenting the programme* describes the activities necessary to achieve a defensible programme and to capture organizational learning. Elements of this are: rules and regulations, learning environment, and domain mapping. (4) *Improving the programme* includes dimensions aimed at the re-design of the programme of assessment, after the programme is administered. Activities are R&D and change management. (5) The final dimension *justifying the programme* describes activities that are aimed at providing evidence that the purpose of the programme is achieved taking account of effectiveness, efficiency, and acceptability.

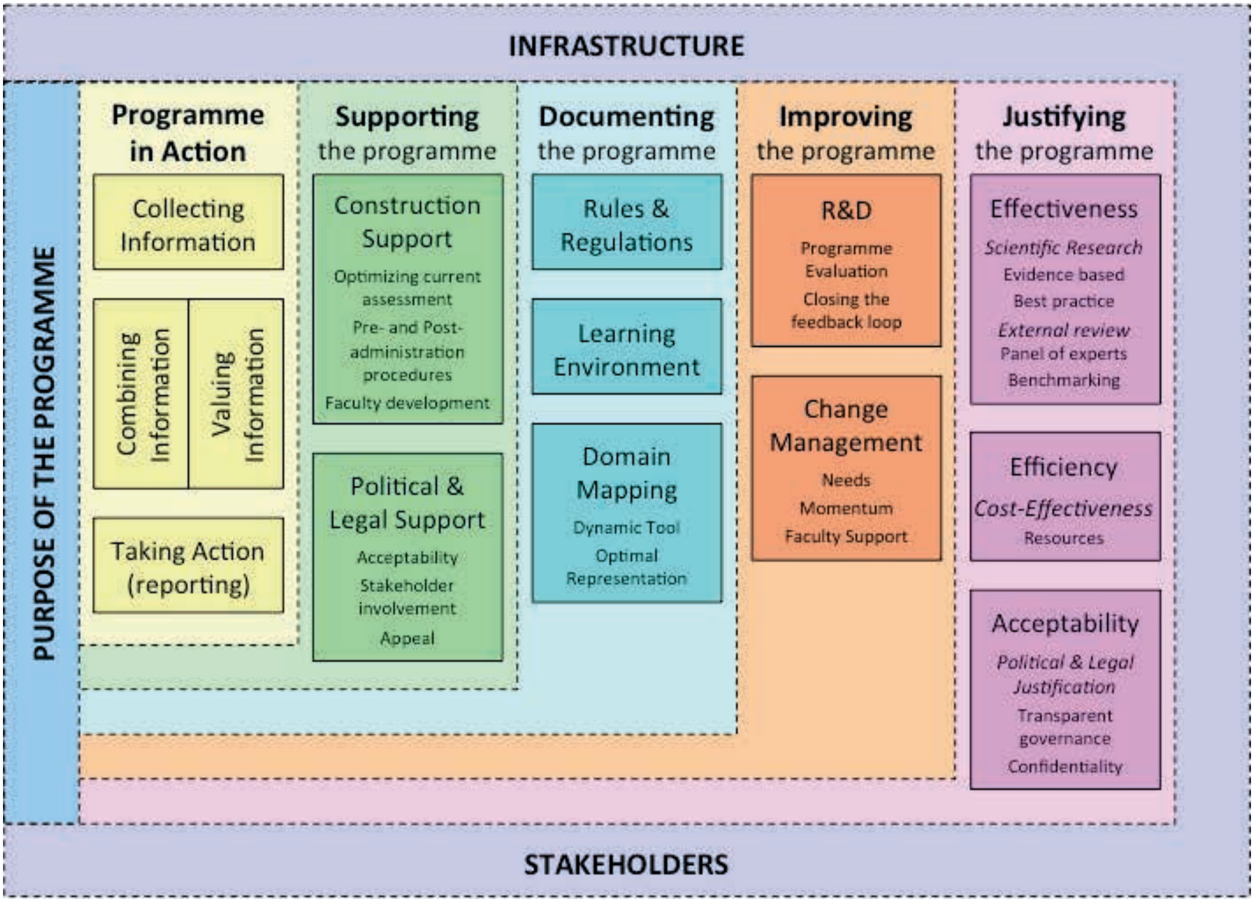

Figure 3.1: Framework for programmes of assessment

Because the aim of this study was to formulate guidelines that are general enough to be applicable to a variety of contexts, and yet at the same time meaningful and concrete enough to support assessment designers, we started by generating ideas for guidelines based on the above framework for programmes of assessment using the input of international experts in the field of assessment in medical education. In order to validate the guidelines we sought expert consensus. In this article we do not go into further detail about the framework; but kindly refer the reader to our previous publication (Dijkstra et al., 2010). In describing the results we will focus on the most important and salient findings (i.e. the guidelines). For the complete set of guidelines we refer to the Addendum.

Method

Study design

The development and validation of design guidelines was divided into four phases, starting with a brainstorm phase to generate ideas using a core group of experts (JD, CvdV and LWTS), followed by a series of discussions with a wider group of international experts to elaborate on this brainstorm. Next in a refinement phase, the design guidelines were fine-tuned based on the analysis of the discussions. Finally a member check phase was initiated to validate the guidelines based on expert consensus.

Participants

The participants were purposefully selected based on their experience with programmes of assessment. They all have published extensively on assessment. Given their backgrounds it was anticipated that these experts would provide the most valuable information. The nine participants of the focus group of the preceding study (Dijkstra et al., 2010) were invited by e-mail to participate in this follow-up study, explaining the goal and providing details about the method and procedures. One participant declined because of retirement, another declined because of other obligations, a third declined because of a change in field of work. With the addition of CvdV and LWTS a total of eight experts took part in this study. The experts (all co-authors) came from North America (2) and Europe (6). Within their institution, they fulfil different (and some multiple) roles in their assessment practice e.g. programme directors, national committee members, and other managerial roles. They represent different (educational) domains ranging from undergraduate and graduate education, to national licensing and recertification.

Procedure and data analysis

The brainstorm was done by the research team (JD, CvdV, LWTS) based on their experience and data from the preceding study (Dijkstra et al., 2010). This resulted in a first draft of the set of guidelines, which served as a starting point for the discussion phase. The discussion took place in multiple (Skype®) interviews with the participants. Individual interviews were held with each participant and led by one researcher (JD) with the support of a second member of the research team (either CvdV or LWTS). The interview addressed the first draft of guidelines and was structured around three open questions: 1. Is the formulation of the guidelines clear, concise, correct? 2. Do you agree with the guidelines? 3. Are any specific guidelines missing? The interviews were recorded and analysed by the research team to distil a consensus from the various opinions, suggestion, and recommendations. One researcher (JD) reformulated the guidelines and to avoid overly adherence to initial formulations the interview data (expert suggestions) were taken as starting point. The goal of the new formulation was to represent the opinions and ideas expressed by the experts as accurately as possible. Peer debriefing was done to check the reformulation by the research team (JD, CvdV, and LWTS) to reach initial consensus. After formulating a complete and comprehensive set of guidelines, a member-check procedure was conducted by e-mail. All participants were sent the complete set for final review and all responded. No content-related issues had to be resolved and some wording issues were resolved as a final consensus document was generated.

Results

A set of 72 guidelines was developed based on expert experience, and then validated based on expert consensus. Because of the length of this list we have decided not to provide exhaustive detail about all of them, but to limit ourselves to the most salient guidelines per dimension of the framework (the complete list is provided as an Addendum). For reasons of clarity, a few remarks on how to read this section and the addendum with the complete set of guidelines. Firstly, the guidelines are divided over the dimensions of the framework and grouped per element within each dimension. We advise the reader to regard the guidelines in groups rather than as separate guidelines. Also in application of the guidelines it is expected that it is not practical to apply guidelines in isolation. Secondly, there is no linear order in the guidelines presented. When reading the guidelines, you may not immediately come across those guidelines or important topics you would expect to be given priority. There is potentially more than one way of ordering the guidelines. For instance *costs* are important throughout the design process. However, because of the way this framework is constructed, *costs* are addressed near to the end. Thirdly, there is overlap in the guidelines. It appeared impractical and somewhat artificial to split every assessment activity into separate parts. The guidelines are highly related, and overlap and/or redundancy are almost inevitable. In the example of *costs*, which are primarily addressed as part of *cost-efficiency*, references to *costs* are actually made in several guidelines. Fourthly, the level of granularity is not equal for all guidelines. Determining the right level of detail is a difficult endeavour, variable granularity reflects the fact that some issues seem more important than others, and others may have been investigated in depth. Hence, the interrelatedness and the difficulty of determining the right level of granularity is also a reason to review the guidelines per group. The division of guidelines within elements of the dimensions was done based on key recommendations in the design process. However, in some situations this division might be arbitrary and of less relevance. Finally we have sought to find an overarching term that would cover all possible elements of the programme, such as assessments, tests, examinations, feedback, and dossiers. We wanted the guidelines to be broadly applicable, and so we have chosen the term assessment components. Similarly for outcomes of assessment components we have chosen assessment information (e.g. data about the assessees' competence or ability).

General

In addition to the fact that the number of guidelines exceeded our initial expectations, we found that most guidelines focused on the more practical dimensions of the framework (see Table 3.1). In particular, many of the guidelines deal with collecting information. This is not unexpected, since considerable research efforts are focused on specific assessment components for collecting information (measuring). On the other hand some guidelines (e.g. on combining information) are less explicit and straightforward and there is less consensus, resulting in less nuanced guidelines.

Table 3.1: Number of guidelines per dimension

Dimension	Number of guidelines
Purpose	3
Infrastructure	2
Stakeholder	2
Programme in Action	**21**
• Collecting information	• 13
• Combining information	• 3
• Valuing information	• 2
• Taking Action	• 3
Supporting the Programme	**12**
• Construction Support	• 5
• Political Support	• 7
Documenting the Programme	**12**
• Rules and Regulations (R&R)	• 6
• Learning Environment	• 2
• Domain Mapping	• 4
Improving the programme	**7**
• R&D	• 3
• Change Management	• 4
Justifying the Programme	**10**
• Scientific research	• 2
• External Review	• 2
• Efficiency	• 2
• Acceptability	• 4

Three major principles emerged and led to generic guidelines that are applicable in any design consideration are set out below. These are (1) the principle of proportionality, (2) the need to substantiate decisions applying the fitness-for-purpose principle, and (3) getting the right person for the right job. These were translated into the following general guidelines (I to III):

I) Decisions (and their consequences) should be proportionate to the quality of the information on which they are based.

This guideline has implications for all aspects of the assessment programme, both at the level of the design of the programme, and at the level of individual decisions about assessees' progress. The higher the stakes, the more robust the information needs to be.

In the dimension *Programme in Action* for instance, actions based on (collected) information should be proportionate to the quantity and quality of the information. The more high-stakes an action or decision, the more certainty (justification and accountability) is required, the more the information collection process has to comply with scientific criteria, and usually the more information that is required.

For example the decision that an assessee has to retake one exam, can be taken based on less information (e.g. the results of one single test) compared to a decision that the assessee has to retake an entire year of medical school, which clearly requires a series of assessments or maybe even a dossier.

II) Every decision in the design process should be underpinned preferably supported by scientific evidence or evidence of best practice. If evidence is unavailable to support the choices made when designing the programme of assessment, the decisions should be identified as high priority for research.

This implies that all choices made in the design process should be defensible and can be justified. Even if there is no available scientific evidence, a plausible or reasonable rationale should be proposed. Evidence can be sought through a survey of the existing literature, new research endeavours, collaborative research, or completely external research. We stress again that the fitness-for-purpose principle should guide design decisions. The evaluation of the contribution to achieving the purpose(s) should be part of the underpinning.

III) Specific expertise should be available (or sought) to perform the activities in the programme of assessment.

This guideline is more specifically aimed at the expertise needed for the assessment activities in the separate dimensions and elements within the assessment programme. A challenge in setting up a programme of assessment is to 'get the right person for the right job'. Expertise is often needed from different fields including specific domain knowledge, assessment expertise, and practical knowledge about the organisation. Some types of expertise, such as psychometric expertise for item analysis, and legal expertise for rules and regulations, are obvious. Others are less clear and more context specific. It is useful when designing an assessment programme to articulate the skill set and the body of knowledge necessary to address these issues.

Salient guidelines per dimensions in the framework

This section contains the more detailed and specific guidelines. We describe them in relation to the dimensions of our previously described model (see Figure 3.1), starting from the *purpose* towards the outer layers. In the Addendum all guidelines are described and grouped per element within each dimension.

Purpose, stakeholders, and infrastructure

From the fitness-for-purpose perspective, by definition the purpose of an assessment programme is an important key element. The authors all agreed that defining the purpose of the programme of assessment is essential and must be addressed at a very early stage of the (re)design. Although there was some initial debate on the level of detail and the number of purposes, it was generally acknowledged that, at least in theory, there should be one principal purpose.

A1 One principal purpose of the assessment programme should be formulated.

This principal purpose should contain the function of the assessment programme and the domains to be assessed. Other guidelines in this element address the need for multiple long and short term purposes and the definition of framework to ensure consistency and coherence of the assessment programme. The challenge in designing a programme of assessment will be to combine these different purposes in such a way that they are achieved in the optimal way with a clear hierarchy defined in terms of importance. This group of guidelines is aimed at supporting this combination.

Whereas in the original model *stakeholders* and *infrastructure* had been addressed last, they are now considered to be essential in many design decisions and are now considered at an early stage as well. Also, during the discussions, many guidelines led to questions about the organization and infrastructure, and the people needing to be involved. It was decided that it is imperative to establish parameters in relation to infrastructure, logistics, and staffing as soon as possible.

A4 Opportunities as well as restrictions for the assessment programme should be identified at an early stage and taken into account in the design process.

A7 The level at which various stakeholders participate in the design process should be based on the purpose of the programme as well as the needs of the stakeholders themselves.

Programme in action

Since the key assessment activities are within this dimension, it is no surprise that many of the guidelines relate to this aspect. Hence, most guidelines are about *collecting information*, especially the element that deals with selecting an assessment component. In line with general guideline (II), a rationale for the selection of instruments should be provided, preferably based on scientific research and/or best practice. The rationale should justify how components contribute to achieving the purpose of the assessment programme.

B1 When selecting an assessment component for the programme, the extent to which it contributes to the purpose(s) of the assessment programme should be the guiding principle.

During the interviews the experts agreed without much debate on the majority of guidelines about *collecting information* (B2 to B9). These should aid in demonstrating the underpinning of the selection choices. Different components have different strengths and weaknesses and these have to be weighed against each other in order to decide the optimal balance to contribute to the purpose of the assessment. The interrelatedness of the guidelines should be taken into account in the design, but feasibility (Infrastructure) and acceptability (Stakeholders) are also clearly important. This is not as obvious as it seems. Currently design is often focussed almost exclusively on the characteristics of individual assessment components and not on the way in which they contribute to the programme as a whole. Often there is a tendency to evaluate the properties of an assessment component per se and not as a building block in the whole programme.

Around the guidelines about *combining information* there was considerably more discussion, therefore we decided to formulate them more generically and provide more elaborate explanations. Important within this group of guidelines is an underpinning for combing information (general guideline II), whereas in practice data is often combined based in similarity in format. (e.g. the results a communication station and a resuscitation station in one OSCE).

B14 Combination of the information obtained by different assessment components should be justified based on meaningful entities either defined by purpose, content, or data patterns.

Guidelines on *valuing information* and on *taking action* both consider the consequences (e.g. side effects) of doing so. Also links with other elements are explicitly made in these groups of guidelines.

B21 Information should be provided optimally in relation to the purpose of the assessment to the relevant stakeholders.

Supporting the programme

In this dimension, we found extensive agreement among the authors. Within the guidelines on *construction support*, next to the definition of tasks and procedures for support, special attention was given to faculty development as a supporting task as part of the availability of expertise to perform a certain task (general guideline III).

C4 Support for constructing the assessment components requires domain expertise and assessment expertise.

Guidelines on *political and legal support* are strongly related to the proportionality principle (general guideline I) and address procedures surrounding assessment, such as possibilities for appeal. This relates to seeking acceptance for the programme and acceptance of change which forms a basis for and links with *improving the programme*.

C6 The higher the stakes, the more robust the procedures should be.

C8 Acceptance of the programme should be widely sought.

Documenting the programme

The fact that *rules and regulations* have to be documented did not raise much debate. These guidelines address the aspects that are relevant when considering the rules and regulations including the need for an organisational body, upholding the rules and regulations. The fact that the *context (e.g. learning environment)* in which the programme of assessment exists must be made explicit was self apparent.

A group of guidelines which received special attention in the discussions addressed *Domain Mapping*. The term blueprinting is deliberately not used here, because this term is often used to denote a specific tool using a matrix format to map the domain (content) to the programme and the instruments to be used in the programme. With Domain Mapping, a more generalised approach is implied. Not only should content match with components, but the focus should be on the assessment programme as a whole in relation to the overarching structure (e.g. the educational curriculum) and the purpose.

D9 A domain map should be the optimal representation of the domain in the programme of assessment.

Improving the programme

The wording in this dimension turned out to evoke different connotations. R&D in particular is defined differently in different assessment cultures. We therefore agreed to define r*esearch* in R&D as the systematic collection of all necessary information to establish a careful evaluation (critical appraisal) of the programme with the intent of revealing areas of strengths and areas for improvement. *Development* should then be interpreted as re-design. Once this shared terminology was reached, consensus on the guidelines came naturally.

E1 A regular and recurrent process of evaluation and improvement should be in place, closing the feedback loop.

Apart from measures to solve problems in a programme, political change or new scientific insights can also trigger improvement. *Change management* refers to activities to cope with potential resistance to change. (Political) acceptance of changes refers to changes in (parts of) the programme. Also these guidelines are related to the *political and legal support*.

E4 Momentum for change has to be seized or has to be created by providing the necessary priority or external pressure.

Justifying the programme

The guidelines in this dimension are more general, probably due to the fact that they are tightly related to the specific context in which a programme of assessment is embedded. Outcomes of good scientific research on assessment activities are needed to support assessment practices with trustworthy evidence, much like the drive for evidence-based medicine. Although this is a general principle which should guide the design of the programme as a whole, the guidelines about *effectiveness* become specifically important when one has to justify choices made in the programme.

F2 New initiatives (developments) should be accompanied by evaluation, preferably scientific research.

Guidelines on *cost-effectiveness* appear obvious as it is generally regarded as a desirable endeavour from a fit-for-purpose perspective. In every institution or organisation, resources - including those for assessment programmes - are limited. If the programme of assessment can be made more efficient, resources can be freed up for other activities. However, guidelines on this are rarely made explicit.

F6 A cost-benefit analysis should be made regularly in light of the purposes of the programme. In the long term, a proactive approach to search for more resource-efficient alternatives should be adopted.

The guidelines on *acceptability* are related to the issue of due practice. As an assessment programme does not exist within a vacuum, political and legal requirements often determine how the programme of assessment is designed and justified. An issue not often addressed during the design process is the use of outcomes by others, and related unintended consequences thereof.

F10 Confidentiality and security of information should be guaranteed at an appropriate level.

Discussion and conclusion

We developed a comprehensive set of guidelines for designing programmes of assessment. Our aim was to formulate guidelines that are general enough to be applicable to a variety of contexts. At the same time they should be sufficiently meaningful and concrete as to support assessment designers. Since we tried to keep away from specific contexts or educational approaches, it is likely that this set may be applicable beyond the domain of medical education. Although these guidelines are more general than existing sets of guidelines, criteria or standards, we cannot dismiss that our backgrounds (i.e. medical education) might have resulted in too restrictive formulations of guidelines. This stresses the need for further replication of our study and on application of these guidelines in a range of contexts.

Although establishing guidelines is an ongoing process, it is remarkable that in a short time such a good consensus was reached among the experts. Most of the debate actually focused around a few specific guidelines, probably those that are more difficult to enunciate or less certain in their utility. For example topics like *combining information* remain still highly debated, and no complete and final answers can be provided at this time.

In trying to be as comprehensive as possible we acknowledge the risk of being over-inclusive. We would like to stress that when designing a programme of assessment, these guidelines should be applied with caution. We recognise and indeed stress that contexts differ and not all guidelines may be relevant in all circumstances. Hence, designing an assessment programme implies making deliberate choices and compromises, including the choice of which guidelines should take precedence over others. Nevertheless, we feel this set combined with the framework of programmes of assessment enables designers to keep an overview of the complex dynamics of a programme of assessment. An interrelated set of guidelines aids designers in foreseeing problematic areas, which otherwise would remain implicit until real problems arise.

We must stress that the guidelines do not replace the need for assessment expertise. Hence, given our fitness-for-purpose perspective on quality, putting the challenge in applying these general guidelines to a local context. Such a translation from theory into practice is not easy and we see the possibility of providing a universally applicable prescriptive design plan for assessment programmes to be slim. Only, if a specific purpose or set of purposes could be decided upon, one could argue that a set of guidelines could be prescriptive. However, thus far it has been the experience that one similar purpose across contexts is extremely rarely found, let alone a similar set of purposes.

What our guidelines do not support is how to make decisions, but they stress the need for decisions to be underpinned and preferably based on solid evidence. This challenge also provides an opportunity to learn from practice. Different ways of applying the guidelines will likely result in more sophisticated guidelines, and provide a clearer picture of the relations in the framework. Thus, it is probably inevitable that some guidelines are not self-evident and need more explanation. Real-life examples from different domains or educational levels will be required to provide additional clarity and understanding. This is a longer term endeavour beyond the scope of this study. Also, it will involve more data gathering and examples from various domains.

Although validation by the opinions of experts is susceptible to biases, it was suitable in our study for generating a first concrete set of guidelines. The validation at this stage is divergent in nature and therefore inclusive and, as such, the guidelines might be over-inclusive. This is only one form of validation and not all guidelines can be substantiated with scientific evidence or best practice. Therefore further validation through specific research is necessary, especially in the area of implementation and translation to practice. Different programmes of assessment will have to be analysed in order to determine whether the guidelines are useful in practice and are generally applicable in different contexts. A practical validation study is now needed. It is encouraging to have already encountered descriptions of programmes of assessment in which to some extent the guidelines are intuitively or implicitly appreciated and taken into account. Of course this is to be expected since not all guidelines are new. However, we think that the merit of this study is the attempt to provide a comprehensive and coherent listing of such guidelines.

References

AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing.* AERA, Washington.

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

Cilliers, F., Schuwirth, L., Adendorff, H., Herman, N., & van der Vleuten, C. (2010). The mechanism of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education, 15*(5), 695-715.

Cilliers, F., Schuwirth, L., Herman, N., Adendorff, H., & van der Vleuten, C. (2011). A model of the pre-assessment learning effects of summative assessment in medical education. *Advances in Health Sciences Education, 17*, 39-53.

Dannefer, E. F., & Henson, L. C. (2007). The Portfolio Approach to Competency-Based Assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine, 82*(5), 493-502.

Davies, H., Archer, J., Southgate, L., & Norcini, J. (2009). Initial evaluation of the first year of the Foundation Assessment Programme. *Medical Education, 43*(1), 74-81.

Dijkstra, J., Van der Vleuten, C., & Schuwirth, L. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education, 15*(3), 379-393.

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

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

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.

Ricketts, C., & Bligh, J. (2011). Developing a Frequent Look and Rapid Remediation Assessment System for a New Medical School. *Academic Medicine, 86*(1), 67-71.

Savage, J. K. (2006). In-training assessment (ITA): designing the whole to be greater than the sum of the parts. *Medical Education, 40*(1), 13-16.

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. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher, 33*(6), 478-485.

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

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.

Chapter 4

Practicality and explanatory power of design guidelines for programmes of assessment: A case study

Submitted for publication

Joost Dijkstra, Marjan J.B. Govaerts,
Pauline A. McAvoy, Karlijn Overeem,
Cees P.M. Van der Vleuten,
Lambert W.T. Schuwirth

Abstract

Assessment in medical education serves various purposes, which clearly cannot be achieved by one single instrument. Therefore, assessment in medical education requires a carefully designed assessment programme, consisting of a purposeful mix of assessment components. In previous studies a framework defining what constitutes a programme of assessment was developed. Subsequently a set of **g**uide**l**ines for **a**ssessment **d**esign (GLAD) was formulated, and during development these GLAD were evaluated with respect to *clarity*, *consistency*, and *parsimony.*

The current study focuses on evaluating GLAD in context. A case study and multiple qualitative inquiry two-step approach are used to evaluate practicality and explanatory power of GLAD. In step 1 *practicality (i.e.* if and how GLAD were taken into account during the assessment design) was investigated through deductive content analysis of assessment documents and semi-structured interviews. We distinguished 4 levels of use of GLAD in assessment design: Well-addressed, Partly-addressed, Not addressed, Missing GLAD. In Step 2 the *explanatory power* was evaluated, by using GLAD to explain statements of perceived strengths and issues that were raised in interviews with key stakeholders. The logic argument informed us about the relevance of GLAD in terms of validity.

Results suggest that The GLAD are comprehensive and logically applicable in practice and thus meet the practicality criterion. One design-element could not be coded with GLAD and led an additional GLAD. Designing an assessment programme is a balancing act, where compromises are required to optimally contribute to the assessment purpose. The GLAD offer a meaningful vocabulary to organisations and stakeholders to *describe and explain* the quality assessment programmes; GLAD thus also meet the explanatory-power criterion.

Introduction

Assessment in medical education usually serves various – and sometimes divergent – purposes, including determining (minimal) competence, influencing learning behaviour, as well as evaluating instruction. Clearly, one single instrument cannot achieve this diversity in purposes, or even a single purpose in full. Therefore, assessment in medical education requires a carefully designed programme of assessment, consisting of a purposeful mix of various assessment components. When designing programmes of assessment, a broad range of factors such as scheduling of assessments, combining different assessment outcomes, and exam regulations have to be taken into account in order to optimally achieve assessment purposes (Dijkstra et al., 2012). In the past decades, a wealth of research has resulted in specifications of strengths and weaknesses of single assessment instruments used in medical education and a resulting view that any selection of assessment instruments requires trade-offs between various quality aspects, such as reliability, validity, and educational impact (Newble et al., 1994; Schuwirth and Van der Vleuten 2004; Van der Vleuten, 1996). However, in contrast to the literature on quality of single instruments, literature about what constitutes quality of programmes of assessment, or how to balance various factors in programmatic approaches to assessment design is scarce.

The available standards for educational and psychological testing (AERA, APA, NCME, 1999) contain generic descriptions of quality in testing, covering a wide range of assessment instruments and activities that are relevant in assessment of professional competence. However, guidance provided by these standards mainly focuses on the quality of individual tests (i.e. measurement instruments), with far less attention to embedding individual instruments in a specific assessment programme. Furthermore, studies on quality criteria used to evaluate combinations of different assessments often apply psychometric approaches only, e.g. combined reliability estimates (Moonen-van Loon et al., 2013; et al., 2001; Verhoeven et al., 2000). In addition, their applicability is often limited to specific well-described educational philosophies or contexts, such as competency-based assessment programmes (e.g. Baartman, 2008). Although several recent papers describe programmes of assessment that illustrate the added value of a programmatic approach (e.g. triangulation of information from multiple assessments), these examples are hard to replicate in other settings because these programmes are designed for a specific local context (e.g. Dannefer and Henson, 2007; Ricketts and Bligh, 2011). Quality criteria or guidelines produced in one setting may not necessarily be relevant or useful in another. Hence, there is a need for guidelines that are applicable (or easy to adapt for use) in a broad range of assessment contexts.

We have aimed at developing guidelines, which are widely applicable across different contexts to support assessment developers in achieving a high-quality programme of assessment. In two previous studies (Dijkstra et al., 2010; Dijkstra et al., 2012) we developed a framework that defines the assessment components that constitute a programme of assessment; and from this framework we subsequently derived a set of **g**uide**l**ines for **a**ssessment **d**esign (GLAD). In order to make the GLAD applicable independent of specific assessment context or philosophy, guidelines were formulated from a fitness-for-purpose quality perspective (Dijkstra et al., 2010; Harvey and Green, 1993). Hence, quality, design decisions, and application of GLAD depend on the purpose of the assessment programme.

Successful implementation of (educational) guidelines requires evaluation and evidence, equal to the call for evidence-based clinical practice guidelines (Basinski, 1995). In this evaluation process three phases can be distinguished: (1) evaluation during the development; (2) evaluation of programmes in which guidelines play a role; (3) scientific evaluation (Basinski, 1995). In order to validate GLAD the first phase has been described in Dijkstra et al. (2010, 2012) and consisted of expert validation focussed on achieving *clarity*, *consistency*, and *parsimony* (Prochaska et al., 2008). Terminology was made explicit and was carefully defined, care was taken that guidelines would not contradict each other, and complexity as well as redundancy in guidelines was minimized.

The present study focuses on evaluation of GLAD in context. We used an instrumental case study and a multiple qualitative inquiry two-step approach to evaluate *practicality* and *explanatory power* (Prochaska et al., 2008). To evaluate practicality of GLAD (step 1) we investigated if GLAD are found in practice, if they are complete, and if they are taken into account during the design of an assessment programme. Document analysis and interviews with multiple stakeholders in the assessment process were conducted to gain in-depth information about decisions and considerations made during the design process. Based on the results from step 1, we investigated the explanatory power of GLAD (step2). Statements about quality of the assessment programme, as perceived by relevant stakeholders, are evaluated in terms of GLAD and explained by the practicality analysis.

Methods

Context of the case study

NCAS (National Clinical Assessment Service) was selected as case in this study, because it is widely regarded as having a high-quality programme of assessment (best practice) and is well documented (www.ncas.nhs.uk/publications). NCAS provides a national service in giving confidential advice and support in the resolution of concerns about professional practice of doctors, dentists and pharmacists in the UK, with the aim to resolve these issues. Usually, employers refer a professional for assessment. Each referral will be evaluated and if accepted for further assessment a specific, individual assessment plan will be constructed.

This implies that NCAS does not uniformly use a standardized set of assessment methods in each case, but that tailor-made programmes are constructed for each individual referral, according to standard procedures (see www.ncas.nhs.uk). The approach towards assessment and procedures are extensively described and underpinned. NCAS has to function in a high-stakes environment with a high risk of appeal and legal challenges, that all have been withstood thus far. This legal robustness, the large numbers of referral requests, as well as the appreciation for the programme of assessment confirm that NCAS is a best practice example of a programme of assessment.

Position of the research team

The members of the research team (JD, LS, KO, MG) who analysed the data were not affiliated to the assessment institute and had no conflict of interest in the assessment programme. There were no hierarchical relations between these members of the research team and participants in the study.

Step 1: Practicality

Data collection strategies

Document analysis and semi-structured interviews were used to explore practicality of guidelines. The assessment institute provided 10 documents describing the assessment programme and underpinning design decisions (e.g. about purpose and services, methods and principles, training assessors, tailoring assessment, and quality assurance). In some instances these documents referred to other resources, which were used by the research team for clarification, but were not included in the document analysis. A semi-structured group-interview was held with five participants, selected by the institute based on their involvement with and knowledge of the assessment programme. All participants were employed by the assessment institute and did not receive any compensation for participating in the interview. They fulfilled roles in management, assessment development, and/or as assessment advisors. The semi-structured interview was held at the assessment institute and moderated by JD. The discussion was structured according to the dimensions of the framework for programmes of assessment (Dijkstra et al., 2010, 2012), and the following topics were subsequently addressed: (1) Purpose of the assessment programme (2) the way assessment information was collected, combined, and valued in making decisions (3) quality assurance and staff development (4) rules and regulations, learning environment, and domain map (5) quality improvement and change management (6) efficiency, effectiveness, and acceptability (for an overview see headings of Table 4.1 and see Dijkstra et al., 2010 for a more elaborate description). For each topic, the central question was: 'Are the corresponding guidelines considered during the design (either explicit or implicit)?' Participants were asked to illustrate their statements with examples and to elaborate on the rationale behind the use (or not) of the guideline. In addition to the semi-structured group-interview, two other assessment designers were interviewed by telephone with similar procedure and structure, because they were not able to take part in the semi-structured interview. All interviews were audio-recorded and summarized. In a member check procedure the summary was sent to participants to allow them to correct mistakes or misinterpretations. Only textual comments were made and used to improve the clarity of the summary.

Step 1: Practicality - Data analysis and procedures

We used deductive content analysis (Elo and Kyngas, 2005; Hsieh and Shannon, 2008) to investigate if and how GLAD were taken into account during the assessment design. First, two coders (JD, LS) independently analysed the documents and interviews using the 72 guidelines (Dijkstra et al., 2012) as a coding scheme. In order to assess conformability, a third coder (KO) was added, who has expertise in assessment and is not affiliated to the assessment programme nor involved in the construction of GLAD. To reach one measure of use, the codes of the three researchers were compared. Since frequency of use does not necessarily reflect the extent to which guidelines are taken into account, we additionally used coding categories indicating the level of use:

1. Well-addressed (W)
   GLAD are explicitly mentioned as taken into account (e.g. decisions made), and a concrete description of assessment components that resulted from using GLAD was provided. GLAD were applied to the complete assessment programme.
2. Partly-addressed (P)
   A) GLAD were mentioned (or implied) as taken into account, however, not supported by concrete descriptions of assessment components.
   B) GLAD were not mentioned, but descriptions of assessment components of the programme imply that GLAD are taken into account (implicitly).
   C) GLAD were not addressed in the programme as a whole, but parts only (incomplete).
3. Not addressed (N)
   Guidelines were not described in terms of process or outcome.

To indicate when GLAD did not cover assessment descriptions, we included a fourth additional category:

4. Missing guideline (M)
   Elements (descriptions) that could not be coded by guidelines.

Researchers met regularly to discuss coding results and any discrepancies were discussed between them until consensus was reached. The interviews were used to support and supplement (triangulate) the document analysis.

Results - Step I: Practicality

Table 4.1 summarizes the use of GLAD in the NCAS design process (the GLAD have a specific character and number for reference purposes). The table is divided in 6 dimensions (A to F), in which guidelines are grouped in broader elements. The table shows that all guidelines have been found in the documents or mentioned in the interviews, although addressed with varying frequency and level of appropriateness.

Table 4.1: Overview of results

W = Well-addressed; **P** = Partly-addressed; **X** = Not addressed	Practicality of GLAD				Expl. Power	
GUIDELINE - a full description of the guidelines can be found in the addendum	Documents		Interview	**Overall**	Strong	Issue
	Freq.	Addressed				
PURPOSE OF THE PROGRAMME		**W**	**W**	**W**		
A1 One principal purpose of the assessment programme should be formulated.	2	W	W	W	5	1
A2 Long-term and short-term purposes should be formulated. But the number of purposes should be limited.	5	W	W	W		1
A3 An overarching structure which projects the domain onto the assessment programme should be constructed.	7	W	W	W	4,5	
INFRASTRUCTURE		**P**	**P**	**P**		
A4 Opportunities as well as restrictions for the assessment programme should be identified at an early stage and taken into account in the design process.	8	P	P	P		6
A5 Design decisions should be checked against consequences for the infrastructure. If necessary compromises should be made, either adjusting the purpose(s) of the assessment programme or adapting the infrastructure.	4	P	P	P		3,6
STAKEHOLDERS		**P**	**P**	**P**		
A6 Stakeholders of the assessment programme should be identified and a rationale provided for including the expertise of different stakeholders and the specific role(s) which they should fulfil.	12	P	P	P	1	
A7 The level at which various stakeholders participate in the design process should be based on the purpose of the programme as well as the needs of the stakeholders themselves.	7	P	P	P	1	4

W = Well-addressed; **P** = Partly-addressed; **X** = Not addressed	Practicality of GLAD				Expl. Power	
GUIDELINE - a full description of the guidelines can be found in the addendum	Documents		Interview	**Overall**	Strong	Issue
	Freq.	Addressed				
PROGRAMME IN ACTION		**W**	**W**	**W**		
Collecting Information		**W**	**W**	**W**		
B1 When selecting an assessment component for the programme, the extent to which it contributes to the purpose(s) of the assessment programme should be the guiding principle.	17	W	W	W	3,4,5	
B2 When selecting an assessment (component or combination), consideration of the content (stimulus) should take precedence over the response format.	13	W	W	W	3,4,5	
B3 The assessment should sample the intended cognitive, behavioural or affective processes at the intended level.	15	W	X	W	3,4,5	
B4 The information collected should be sufficiently informative (enough detail) to contribute to the purpose of the assessment programme.	15	W	X	W	3,4,5	
B5 The assessment should be able to provide sufficient information to reach the desired level of certainty about the contingent action.	13	W	X	W	3,4,5	
B6 The effect of the instruments on assessee behaviour should be taken into account.	7	P	P	P		2
B7 The relation between different assessment components should be taken into account	6	W	X	W		
B8 The overt and covert costs of the assessment components should be taken into account and compared to alternatives.	0	X	P	P		6
B9 Assessment approaches that work well in a specific context (setting) should first be re-evaluated before use in another context (setting) before implementation.	0	X	W	W		
B10 A programme of assessment should deal with error and bias in the collection of information. Error (random) is unpredictable and should be reduced by sampling (strategies). Bias (Systematic) should be analysed and its influence should be reduced by appropriate measures.	10	P	W	W	3	
B11 Any performance categorisation system should be as simple as possible.	0	X	W	W		
B12 When administering an assessment (component), the conditions (time, place, etc.) and the tasks (difficulty, complexity, authenticity, etc) should support the purpose of the specific assessment component.	10	W	W	W	3	3

W = Well-addressed; **P** = Partly-addressed; **X** = Not addressed	Practicality of GLAD				Expl. Power	
GUIDELINE - a full description of the guidelines can be found in the addendum	Documents		Interview	**Overall**	Strong	Issue
	Freq.	Addressed				
B13 When scheduling assessment, the planning should support instruction and provide sufficient opportunity for learning.	6	W	X	W	3	3
Combining Information		**P**	**P**	**P**		
B14 Combination of the information obtained by different assessment components should be justified based on meaningful entities either defined by purpose, content, or data patterns.	6	P	P	P	5	
B15 The measurement level of the information should not be changed.	2	P	P	P	5	
B16 The consequences of combining information obtained by different assessment components, for all stakeholders, should be checked.	2	P	P	P	5,6	2,4
Valuing Information		**W**	**W**	**W**		
B17 The amount and quality of information on which a decision is based should be in proportion to the stakes.	5	P	P	P	4,5	1
B18 A rationale should be provided for the standard setting procedures.	3	W	W	W	6	
Taking Action		**W**	**W**	**W**		
B19 Consequences should be proportionally and conceptually related to the purpose of the assessment and justification for the consequences should be provided.	8	W	W	W	5,6	2
B20 The accessibility of information (feedback) to stakeholders involved should be defined.	8	W	W	W	5	
B21 Information should be provided optimally in relation to the purpose of the assessment to the relevant stakeholders.	10	W	W	W	4,5	
SUPPORTING THE PROGRAMME		**W**	**W**	**W**		
Construction Support		**W**	**W**	**W**		
C1 Appropriate central governance of the programme of assessment should be in place to align different assessment components and activities.	13	W	W	W	1,6	3,4
C2 Assessment development should be supported by quality review to optimise the current situation (Programme in Action), appropriate to the importance of the assessment.	15	W	W	W	1,6	3

W = Well-addressed; **P** = Partly-addressed; **X** = Not addressed	Practicality of GLAD				Expl. Power	
GUIDELINE - a full description of the guidelines can be found in the addendum	Documents		Interview	**Overall**	Strong	Issue
	Freq.	Addressed				
C3 The current assessment (Programme in Action) should be routinely monitored on quality criteria.	5	P	P	P	1,6	3
C4 Support for constructing the assessment components requires domain expertise and assessment expertise.	17	W	X	W	1,4, 5,6	3
C5 Support tasks should be well-defined and responsibilities should lie with the right persons.	14	W	X	W	1,4	3
Political and Legal Support		**W**	**W**	**W**		
C6 The higher the stakes, the more robust the procedures should be.	14	W	W	W	1,4,6	1,3,4
C7 Procedures should be made transparent to all stakeholders.	18	W	X	W	1,2	3
C8 Acceptance of the programme should be widely sought.	10	W	X	W	1,2	3,4
C9 Protocols and procedures should be in place to support appeal and second opinion.	5	W	W	W	1,3,6	3
C10 A body of appeal should be in place	0	X	W	W	1,3,6	3
C11 Safety net procedures should be in place to protect both assessor and assessee.	12	W	W	W	1,3,6	3
C12 Protocols should be in place to check (the programme in action) on proportionality of actions taken and carefulness of assessment activities.	15	W	X	W	1,3,6	2,3
DOCUMENTING THE PROGRAMME		**P**	**P**	**P**		
Rules and Regulations (R&R)		**P**	**P**	**P**		
D1 Rules and regulations should be documented.	8	W	W	W		
D2 Rules and regulations should support the purposes of the programme of assessment.	6	P	X	P		
D3 The impact of rules and regulations should be checked against managerial, educational, and legal consequences.	3	X	X	X		1
D4 In drawing up rules and regulations one should be pragmatic and concise, to keep them manageable and avoid complexity.	3	X	P	P		
D5 R&R should be based on routine practices and not on incidents or occasional problems.	4	P	P	P		
D6 There should be an organisational body in place to uphold the rules and regulations and take decisions in unforeseen circumstances.	3	P	W	W		

W = Well-addressed; **P** = Partly-addressed; **X** = Not addressed	Practicality of GLAD				Expl. Power	
GUIDELINE - a full description of the guidelines can be found in the addendum	Documents		Interview	**Overall**	Strong	Issue
	Freq.	Addressed				
Learning Environment		**P**	**W**	**W**		
D7 The environment or context in which the assessment programme has to function should be described.	10	P	W	W	3	
D8 The relation between educational system and assessment programme should be specified.	5	P	W	W	3	
Domain Mapping		**P**	**X**	**P**		
D9 A domain map should be the optimal representation of the domain in the programme of assessment.	5	P	X	P	4	
D10 A domain map should not be too detailed.	3	P	X	P	4	
D11 Starting point for a domain map should be the domain or content and not the assessment component.	3	P	X	P	4	
D12 A domain map should be a dynamic tool, and as a result should be revised periodically.	2	P	X	P	4	
IMPROVING THE PROGRAMME		**W**	**W**	**W**		
R&D		**W**	**W**	**W**		
E1 A regular and recurrent process of evaluation and improvement should be in place, closing the feedback loop.	14	W	W	W		5
E2 If there is uncertainty about the evaluation, more information about the programme should be collected.	5	P	X	P		5
E3 In developing the programme (re-design) again improvements should be supported by scientific evidence or evidence of best practice.	5	W	X	W		5
Change Management		**W**	**W**	**W**		
E4 Momentum for change has to be seized or has to be created by providing the necessary priority or external pressure.	3	W	W	W		
E5 Underlying needs of stakeholders should be made explicit.	1	W	X	W		
E6 Sufficient expertise about change management and about the local context should be sought.	5	W	W	W		
E7 Faculty should be supported to cope with the change by providing adequate training	2	W	W	W		

W = Well-addressed; **P** = Partly-addressed; **X** = Not addressed	Practicality of GLAD				Expl. Power	
GUIDELINE - a full description of the guidelines can be found in the addendum	Documents		Interview	**Overall**	Strong	Issue
	Freq.	Addressed				
JUSTIFYING THE PROGRAMME		**W**	**W**	**W**		
Effectiveness - Scientific Research		**X**	**P**	**P**		
F1 Before the programme of assessment is designed, evidence should to be reviewed.	5	X	P	P	6	
F2 New initiatives (developments) should be accompanied by evaluation, preferably scientific research.	1	X	X	X		5
Effectiveness - External Review		**W**	**W**	**W**		
F3 The programme of assessment should be reviewed periodically by a panel of experts.	2	W	W	W	6	
F4 Benchmarking against similar assessment programmes (or institutes with similar purposes) should be conducted to judge the quality of the programme.	2	P	X	P	6	
Efficiency: cost-effectiveness		**X**	**W**	**W**		
F5 In order to be able to justify the resources used for the assessment programme, all costs (in terms of resources) should be made explicit.	2	X	W	W		6
F6 A cost-benefit analysis should be made regularly in light of the purposes of the programme. In the long term, a proactive approach to search for more resource-efficient alternatives should be adopted.	2	X	W	W		6
Acceptability: political-legal justification		**W**	**W**	**W**		
F7 Open and transparent governance of the assessment programme should be in place and can be held accountable	2	W	W	W	2	
F8 In order to establish a defensible programme of assessment there should be one vision (on assessment) communicated to external parties.	1	W	X	W	2	
F9 The assessment programme should take into account superseding legal frameworks.	3	P	W	W	2	
F10 Confidentiality and security of information should be guaranteed at an appropriate level.	8	P	W	W	2	

Well-addressed guidelines

About two-thirds of GLAD were well addressed (See Table 4.1). Specifically guidelines regarding the purpose of the assessment (A1,A2,A3), selection of an assessment instrument (B3,B4,B5) and procedures and acceptance (C7,C8) are very well addressed. One document specifically addresses the purpose and all interviewees had a clear and similar view on the main purpose of the assessment programme.

Partly-addressed guidelines

About one-third of GLAD were partly-addressed (See Table 4.1). For example *Infrastructure* (A4,A5) was not explicitly mentioned as taken into account, but clearly the implications were visible in the assessment programme. The other way around: *Combining information* (B14,B15,B16) was addressed at several instances in the documents and during the interview, however, these GLAD were applied to specific components of the assessment programme, rather that to the assessment programme as a whole.

Not (sufficiently) addressed

Two subsets of guidelines appeared to be insufficiently addressed. In general, GLAD about rules and regulations (R&R) were stated in documentation, but not addressed in much detail. Especially *D3* 'The impact of R&R should be checked against managerial, educational, and legal consequences.' seemed to be neglected since it was not described in the documents nor mentioned by interviewees. The R&R support the main assessment purpose and some implications are mentioned, but it is not clear *how* the R&R were drawn up.

GLAD F2 'New initiatives (developments) should be accompanied by evaluation, preferably scientific research.' was also insufficiently addressed. Although the assessment programme is well underpinned, using evidence-based principles for assessment design, F2 was mentioned only once anecdotally.

Missing guidelines

In the documents there were two sections that were relevant for the design of the programme of assessment, but could not be labelled to a specific guideline.

> ***Equality***
> *"[The assessment institute] is committed to meeting the requirements of those who have a disability. If a practitioner undergoing the [...] assessment has any particular requirements that they wish [...] to take into*

> *consideration so that we can make reasonable adjustments, they are asked to let […] know as soon as possible."*

There are GLAD that are aimed at protecting the assessee and to assure carefulness, but these safety net procedures (C11) and protocols for carefulness of activities (C12) are in effect at hindsight. No GLAD deals with equality or fairness for different groups of assessees.

> ***Gaining the agreement of the practitioner and referring body***
> *"[…] the referring body and practitioner will have the opportunity to consider whether they want to agree to participate in an assessment. […] By signing the agreement to assessment document all parties agree to commit to the procedures set out in the document and to taking forward recommendations arising from the assessment report […]"*

There is no GLAD addressing the issue of agreement or contracts. However the fact that there is a contract for each assessment can explain why less attention is paid to the GLAD about Rules and Regulations.

Concluding Step I: Practicality

Based on the results above we conclude that the GLAD meet the practicality criteria. Frequency of coding (during data analysis) is not a valid indicator of the degree to which the GLAD are actually applied during the design process. For instance B11 'Any performance categorisation system should be as simple as possible.' was well addressed, but not coded in the documents. One description provided during the interview was sufficient evidence that this was taken into account during all assessments. Also, it may very well be that some guidelines are so self-evident, that paying explicit attention to related assessment components in documentation seems superfluous. Overall, the results in step I provide a sufficient basis to evaluate the *explanatory power* of GLAD.

Step II: Explanatory power

Data collection strategies

To gather statements about quality of the assessment programme JD conducted telephone interviews with various stakeholders to explore their perceptions of the quality of the programme. We used purposeful sampling to ensure maximal diversity. The assessment institute invited 17 stakeholders to participate in the study. In total 15 stakeholders participated of which 7 internal staff members (management, developers, and advisors) and 8 external stakeholders (Clinical Advisors, Assessors, Referring organisation, Assessee, Lay person), constituting a representative sample of various interest groups involved in the NCAS programme

The interviews started with clarifying the participants' role in the assessment and their view on the purpose of the assessment programme. Next, the participants were invited to speak their own mind about strengths and weaknesses of the assessment programme. The participants were then asked to elaborate more about the statements they made and provide reasons or examples. It was stressed that no right or wrong answer could be given. The interviews were recorded and took between 45 and 90 minutes.

During the analysis of the interviews it became clear that weaknesses result from trade-off decisions and compromises linked to strengths built into the NCAS assessment programme. Therefore it might be more correct to talk about issues that NCAS has to deal with. JD summarized the interviews and sent these to all individual participants for verification. Except for a few clarifying remarks, all interviewees indicated that the summaries accurately represented the interview.

Step II: explanatory power - Data analysis and procedures

JD and MG independently analysed the summaries of the interviews using inductive content analysis to identify quality statements: i.e. strengths (practices supporting the purpose) and issues (situations and circumstances NCAS has to deal with). Both readers came to a similar result and after discussion quality statements were grouped in broader themes.

First the GLAD were used to describe the quality statements. This description allowed us to check whether all statements could be addressed and covered by GLAD. Next, the results from step 1 (practicality analysis) were used to check whether these statements follow logically from the application of GLAD. This logical argument informed us about the relevance of GLAD in terms of validity. As the final step in the analysis we checked which GLADs were not used to describe or explain quality statements.

Results - Step II: Explanatory power

Based on analysis of the interview transcripts, 6 major strengths and 6 major issues could be identified and are described and explained with GLAD below. Characters and numbers refer to the GLAD in Table 4.1. Quotes are used to illustrate quality statements and it is indicated whether the quote comes from an internal

(IN1 to IN7) or an external (EX1 to EX8) participant. The final section describes the GLAD not used in the description and explanation of quality statements.

Strengths

1) Quality Assurance of the assessment - Many different stakeholders (roles) provide input in the report writing and a wide range of procedures are in place to ensure high quality. The assessors are carefully selected and all participate in mandatory training sessions on a regular basis.

> *"Not just anyone who puts himself forward, but actually we do attempt to be rigorous about recruitment and selections of assessors. High quality!" [IN3]*

This strength can be described using GLAD C1 to C12, which concern supporting the assessment programme (i.e. quality assurance); and GLAD A6,A7 about stakeholders and their roles. The practicality analysis can explain this strength as GLAD C1 to C12 are well-addressed. However, rather surprisingly, GLAD A6,A7 are only partly-addressed. How stakeholders are taken into account is not mentioned explicitly. More attention for procedures around quality assurance might compensate for less attention paid to defining stakeholders. These two groups of GLAD are interrelated.

2) Acceptance of the assessment - Stakeholders (e.g. referrers, employers, public) generally accept the assessment. The assessment is set up in a respectful, developmental, supportive way and *"As it is a national body it has legitimacy and authority. It also shows the public that concerns are taken serious." [EX1]* That the assessment programme deals with concerns – *"Making things move forward" [EX6]* - contributes to acceptance. These high levels of acceptance can be explained by the fact that attention is paid to GLAD C8 'Acceptance of the programme should be widely sought', but also C7 (transparency) and F7 to F10 (political-legal justification) are well-addressed in the design process.

3) Fairness of the assessment - NCAS is an independent party and possible conflicts of interest are avoided. The assessment is tailored to the individual to match the assessees' own work environment. Cultural biases are minimized and procedures are well documented and transparent. Assessees are stimulated to actively engage in the assessment and provide their view on the issues raised. Fairness is described in the first place by GLAD D7,D8 (environment) as understanding the working environment is essential in this case, followed by the selection of instruments (B1 to B5):

> *"it is designed in context, not just an academic exercise using what available techniques there are" [IN1].*

More explicitly the GLAD about procedures for carefulness (C9 to C12), dealing with bias (B10), and administering and planning the assessment (B12,B13) address this strength. By paying attention to GLAD D7, D8 and B1 to B5 (ensuring contextualized, tailor-made assessment) as well as through measures to avoid bias (B10) and ensure equality, stakeholders place their trust in the assessment instruments. Combined with transparent procedures in place to assure carefulness explain this strength.

4) Comprehensiveness of the assessment - The assessment is generally perceived as comprehensive and holistic, including all competency domains. Performance is evaluated on a broad range of tasks, and multiple assessment instruments are utilized. Feedback includes comments on strengths and weaknesses, resulting in a balanced and nuanced assessment outcome (report).

> *"The beauty is the whole thing and they take every thing into account when they make a recommendation." [EX3]*

This particular strength follows from extensive attention being paid to the overarching structure (A3) and selection of instruments (B1 to B5). Comprehensiveness of assessment is also related to combining information (B17) and providing feedback (B21), however these two are discussed further in the next strength. Surprisingly, the GLAD about domain mapping (D9 to D12) are only partly addressed. Although GLAD D9 to D12 can be considered a requisite to assessment comprehensiveness, the process of tailoring every assessment to individual needs, supported by robust procedures, is likely to reduce the need for a standard domain map.

5) Meaningfulness and usefulness of the assessment - The report (outcome) is a well-written, report, containing detailed descriptions of assessees performance evaluations. This is regarded as meaningful, insightful, and useful, as the assessment takes place in context and identifies reasons underlying poor performance; to gain in-depth insight in the problem and to support the assessee. Many GLAD are relevant in relation to this strength. Well-addressed GLAD around taking action (B19 to B21) explain this strength. However, the action (report) has no meaning without addressing the GLAD about good information (B17) from individual instruments (B1 to B5). Although less explicit in the documentation the attention given to GLAD about combining information (B14 to 16) also contributes to understanding this strength.

6) Robustness and Defensibility - The report (outcome) is used in high-stakes decisions and therefore robust evidence is strived for. Reports have regularly been scrutinized in legal settings and proven defensible. Central to robustness and defensibility of programmes are GLAD C9 to C12 (appeal and carefulness), which are well-addressed. However, additional guidelines need to be taken into account to ensure a clear and consistent way of assessing. For instance, combining info (B16) can have consequences, which should support the action to be taken, so that it can be proportionally and conceptually related to the purpose of the assessment. GLAD C6 summarizes this: ‘The higher the stakes, the more robust the procedures should be.’ Finally, other factors contributing to Robustness and Defensibility are strong central governance (C1), extensive quality assurance (C2,C3), on-going training of assessors (C4), a clear rational for standard setting (B18), use of scientifically validated instruments (F1), and external review (F3,F4).

Issues

Inherent to design of assessment programmes is making compromises. As a consequence, strengthening particular features of the assessment programme may result in accepting certain negative side-effects, which is very similar to the reliability-validity dilemma in test design.

1) Need for defensibility – The developmental purpose of the NCAS assessment is shared among stakeholders. However, due to the high-stakes environment and strong emphasis on defensibility (which is a strength) compromises have to be made regarding the developmental purposes.. Information about an assessees’ performance that is considered less trustworthy (robust) is excluded from the report, although it might very meaningful and useful for the assessee to improve his/her performance.

“Sometimes difficult to prove your professional judgement, because you have to be evidence based.” [EX1]

Relevant in this issue are, first of all, the description of the purpose (A1,A2) and given the context B17: ‘The amount and quality of information should be in proportion to the stakes’. Although only partly-addressed it does appear to have had a big influence in the design of the assessment. This issue seems inevitable when combining developmental and administrative purposes. Within this high-stakes context, NCAS succeeds in optimally supporting the developmental purpose of the assessment programme.

2) Effect on assessee - Acceptance by the assessee is not obvious; although they acknowledge personal benefits and feel they are assessed properly. All parties sign an agreement, but in reality it would be difficult for an assesse not to engage with an assessment if the employer had requested it – they could be in breach of their contract. Undertaking an assessment could have an effect on reputation even if the assessment found that the concerns were not substantiated. NCAS acknowledges the impact of health, mental and physical on performance and all assessees have a health and behavioural (psychological) assessment to determine this impact and their fitness to undergo an assessment. Concerns identified may result in the assessment of clinical performance being cancelled or of adjustments being made to the conduct of the assessment. In all cases advice on appropriate support is given to the practitioner and their employer.

Effect on assessee is literally described by GLAD B6, and also by GLAD about consequences of (B16) combining information or (B19) the action taken. From the design perspective the only thing that can be done is to minimize known negative effects by making sure that due process is carefully followed (C12). Although the relevant GLAD are addressed, the issue of negative effects on assessees will be inevitable.

> *"It is always stressful, … there is no solution, other then manage their case in a fair way." [IN2]*

3) Duration - The assessment process takes a long time and some time-consuming processes are not visible for stakeholders.

> *"Delivering the assessment is time consuming, some of the things we have control over and others we don't." [IN4]*

It takes time to tailor a full performance assessment that meets quality criteria. (Un)availability of assessors makes the planning difficult. Risk averseness leads to safe procedures but sometimes with some unavoidable inefficiencies. This issue is described by (A5) effect of design decisions on infrastructure; and by (B12,B13) about administering and scheduling assessment. Also, the choice for extensive attention to supporting (C1 to C12) the programme (e.g. procedures for quality assurance) explains the lengthy process. Although ways of reducing time are being evaluated, the issue of defensibility and robustness remain top priority.

4) Structure leads to formalistic assessment - The processes are standardized and structured, which makes judgements very consistent, but can limit flexibility. As a consequence, the way in which assessor judgements are recorded using structured forms, may make the information less meaningful and useful to the assessee. The fact that the report writing is led by and finalized by a case manager could also diminish the feeling of ownership of assessors, although no report is issued until the assessors are in agreement that it accurately reflects their judgements and conclusions. The strict procedures are probably due to strong central governance (C1) and relate to the extensive quality assurance procedures (C6). The trade-off decisions that have to be made in this sense relate to a number of GLAD e.g. the level at which various stakeholders participate (A7); consequences of combining information (B16).

5) Feedback loop: improving the programme Although quality assurance procedures are in place, quality assurance data are not systematically used for continuous improvement of the programme. In 10% of the assessments there is an observer added to the assessors for Quality Assurance purposes. Reports from this are reviewed and recommendations for improvement are made to the assessment team. However, implementation can be protracted, due to lack of resources (and time). As a result evaluation is more ad hoc than systematized and improving the assessment beyond the current assessment is rarely the case. This issue is addressed by GLAD (E1 to E3) about research and development. It is addressed in the design and its importance is acknowledged. However, other issues and priorities prevent the programme from reaching its full potential. Not sufficiently addressed is F2 (conducting research). Although there are incidental studies reported, it is not used in a systematic and structural way to improve the programme.

6) Costs - Costs are becoming more of a constraint as funding changes. There is more reason to look at ways to cut costs. The assessment is resource intensive (costly), however, a cost-benefit analysis is hard to make because for instance the added value of getting a professional back on track is hard to objectively quantify. Costs are described by the GLAD in terms of specific costs related to instruments (B8); costs related to the programme as a whole (F5,F6), but also resources (A4,A5) not directly expressed in terms of money. That the issue was less addressed until now is explained by the availability of funding in the recent past. However, due to changes in circumstances this becomes a very explicit issue.

GLAD not used to describe or explain strengths and issues

The GLAD not used in describing and explaining the quality statements are not per definition irrelevant. For instance B7, about the relation between different assessment components, was not explicitly mentioned in explaining quality statements as described above. However, the guideline was discussed in interviews as something that might be worthwhile investigating again to see if other choices might strengthen the programme even more or tackle cost issues in a different way.

> *"The assessment can get a bit compartmentalized. Work has been done looking at individual components, but not on all components. We haven't looked at relationships." [EX7]*

Similar B9 (re-evaluating in new context) and B11 (performance categorisation) are addressed in the design, but not explicitly mentioned in the description of the quality statements.

The set of GLAD concerning the Rules and Regulations (D1 to D6) was not well addressed in the documentation and interviews; it could be expected that this would lead to issues in the assessment programme. However, it appears that this is not the case as they are not used to describe the quality statements. Given the specific context, it might be simply more logical to address the GLAD about procedures (C1 to C12), which compensates for less attention given to R&R. Although the GLAD about change management were not used to explain quality statements, these are well-addressed in the design. However, the implementation of improvements in the programme does not stand out in the programme as a strength or issue.

Discussion

Our goal was to evaluate the GLAD in a best-practice case, comparable to the second phase of evaluation of evidence-based clinical practice guidelines (Basinski, 1995). With this evaluation we can support the validity of the GLAD. Our evaluation focussed around two steps: *practicality* and *explanatory power* (Prochaska et al., 2008).

The GLAD meet the practicality criterion in the sense that they are comprehensive and logically applicable in practice. Results furthermore show that GLAD are a quality evaluation framework and not merely a tick list. The quality in which a GLAD is addressed in a programme of assessment does not depend on the frequency of a certain GLAD being mentioned, but on the level of detail of the application and on the outcomes based on adhering to this GLAD. Therefore, the practicality analysis is a qualitative analysis by default.

In the analysis two elements could not be coded by the GLAD. The element on *equality* could not be fully described with the GLAD. We did not formulate any GLAD addressing this area, which is an important issue well beyond the context of this case-study. Therefore, an addition to the GLAD can be: *C13 "Protocols should be in place to assure assessment activities are equally accessible and fair for different (relevant) groups of stakeholders."* The element in relation to *Gaining the agreement of the practitioner and referring body* was also not be covered by specific GLAD. However, the issue of contracts and consent is not common in assessment practice, or is generally dealt with in the form of published rules and regulations. Related GLAD can be used to cover the issues, e.g. by GLAD C5 (responsibilities should be well-defined), C8 (acceptance), C12 (carefulness), and F9 (superseding legal frameworks). Therefore in this case it is more likely a specific contextual issue and thus an extension to the GLAD is not required.

The GLAD also meet the explanatory-power criterion. The GLAD offer a framework and terminology to *describe and explain* the quality of the programme of assessment as perceived and expressed by the stakeholders during the interviews. However, not all GLAD were used to describe and explain the quality statements. This is probably due to case-specificity and the selection of the most notable themes in quality statements. GLAD that were well addressed did explain strengths. However, explaining the issues NCAS has to face using the GLAD is in some occasions more difficult. The GLAD are descriptive and not predictive in nature. Hence, if a GLAD is not sufficiently addressed, it does not automatically mean that there will be an issue or even a weakness in the programme. Furthermore, we found during the analysis that specific GLAD might compensate for lack of effort spent on other GLAD. This is likely to occur on more occasions as the GLAD are interrelated and the ordering of GLAD in the framework and dimensions (Dijkstra et al., 2010, 2012) is not the only possible structure. Hence, it is likely that in different assessment programmes the ordering or sequence of addressing the GLAD is different.

In addition, the analysis shows that designing a programme of assessment is a balancing act, in which trade-off decisions have to be made as well as compromises to optimally contribute to the purpose of the assessment. This is illustrated by the issue on the negative effects on the assessee - as discussed above - which cannot be completely eliminated, despite the attention given to the GLAD (B6) on this topic. Hence, taking GLAD into account does not mean that all problems and issues will be resolved. There will always be issues that cannot be influenced (yet) and have to be dealt with by addressing the GLAD making compromises. The value of the GLAD in this sense is that it can serve as an assessment framework to provide insights in the strengths and weaknesses of the programme of assessment as a whole. Important to realize in this respect is the dynamic and context-specific nature of applying the GLAD: i.e. what is conceptualized as a high-quality assessment design may change over time, due to changes in the assessment context. What is relevant at this moment in time might not have been relevant before. A good example is the funding (costs): When there is more funding the GLAD concerning costs will be less urgent. Dealing with less financial resources, however, will make the GLAD about costs and efficiency more important, with likely a different outcome of the assessment design.

As such GLAD provide a framework and vocabulary to organisations and stakeholders to describe their programme of assessment, and enabling them to evaluate and improve the assessment.

References

AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing.* AERA, Washington.

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

Basinski, A. S. (1995). Evaluation of clinical practice guidelines. *Canadian Medical Association Journal, 153*(11), 1575-1581.

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.

Dijkstra, J., Galbraith, R., Hodges, B., McAvoy, P., McCrorie, P., Southgate, L. et al. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*, 20.

Dijkstra, J., Van der Vleuten, C., & Schuwirth, L. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education, 15*(3), 379-393.

Elo, S., & Kyngas, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107-115.

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

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288.

Moonen-van Loon, J. M. W., Overeem, K., Donkers, H. H. L. M., Vleuten, C. P. M., & Driessen, E. W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*, 1087–1102.

Newble, D., Dawson, B., Dauphinee, D., Page, G., Macdonald, M., Swanson, D. et al. (1994). Guidelines for assessing clinical competence. *Teaching and Learning in Medicine: An International Journal, 6*(3), 213 - 220.

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model. *Applied Psychology: An International Review, 57*(4), 561-588.

Ricketts, C., & Bligh, J. (2011). Developing a Frequent Look and Rapid Remediation Assessment System for a New Medical School. *Academic Medicine, 86*(1), 67-71.

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.

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.

Verhoeven, B. H., Hamers, J. G. H. C., Scherpbier, A. J. J. A., Hoogenboom, R. J. I., & Vleuten, C. P. M. v. d. (2000). The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. *Medical Education, 34*(7), 525-529.

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.

Chapter 5

Calibrating guidelines for assessment design against quality criteria for competence assessment programmes

Submitted for publication

Joost Dijkstra, Angelique A. Timmerman
Liesbeth K.J. Baartman, Marjan J.B. Govaerts
Paul Ram, Cees P.M. Van der Vleuten
Lambert W.T. Schuwirth

Abstract

Designing assessment programmes in medical education settings is a complex process. Designers have to deal with multiple assessment purposes and environmental influences at the same time. A programmatic approach, which is holistic in nature, is advocated to achieve assessment programmes that are fit for purpose. **G**uide**l**ines for **a**ssessment **d**esign (GLAD) are developed in earlier studies and systematic evaluation and research of the GLAD is required for successful implementation. After validating the GLAD during the development and in practice, the evaluation in the present study is aimed at investigating the effects of application of GLAD within specific assessment contexts.

In a case study the GLAD are calibrated against criteria for competence assessment programmes (CAP). Based on interviews, document analysis, and a self-assessment tool a competency-based assessment programme was evaluated. Outcomes of both quality evaluations are analysed to determine whether the GLAD meet the criteria of *utility* (useful and meaningful outcomes) and *productivity* (build on research) compared to the validated CAP.

Generally both evaluations covered similar issues in assessment. Use of GLAD lead to useful recommendations for the competency-based assessment program, which are corroborated by the outcome of the validated CAP and we conclude that GLAD meet the *utility* criterion. The GLAD also meet the *productivity* criterion because it extends the CAP criteria with new areas for evaluation of programmes of assessment within the competence-based assessment context. Limitations and further implications are discussed.

Introduction

Designing assessment programmes in medical education settings is a complex process. It requires thorough preparation and perseverance, as well as sufficient expertise in various areas (Bok et al., 2013). Drawing a conclusion as to an individual student's competence is not easy; competence is a complex phenomenon with multiple interacting facets. Therefore, simply adding up results from separate tests on knowledge, skills, and attitude does not suffice. Even more complex is the combination of different purposes of assessment and the attempt to meet them at the same time with the same assessment programme (Bok et al., 2013; Van der Vleuten et al., 2012). To combine for example both selective and developmental assessment functions by merely accumulating together separate assessment instruments will not work and lead to contradicting messages to students. All this is further complicated by political influences and other limiting conditions - typically infrastructural and resource-related, It is for this reason that a programmatic approach to assessment is suggested to better deal with the complexity of assessment design and combining multiple assessment purposes (Bok et al., 2013; Van der Vleuten and Schuwirth, 2005; Van der Vleuten et al., 2012).

An all-encompassing single assessment instrument does not exist and due to content specificity many measurements are required (Van der Vleuten et al., 2010). A programmatic assessment approach is more holistic in nature. Information richness of multiple assessment instruments is purposefully combined in order to acquire a complete impression of an individual's qualities and to determine subsequent action (e.g. remediation or pass/fail decision).

Until recently, available guidance to support design of assessment programmes often focused only on separate instruments used for a specific purpose within a specific context. As a result, that guidance did not provide support for optimally combining assessment instruments nor were they easily transferable for application in contexts.

In our previous studies (Dijkstra et al., 2010, 2012, and Chapter 4) we developed **g**uide**l**ines for **a**ssessment **d**esign (GLAD), which are structured in an overarching framework for designing programmes of assessment. The framework is divided into several dimensions and is placed in the context of *stakeholders* and *infrastructure* (represented as the outer layer). The key element in the framework, the starting point is the *purpose of the programme* (key element in the framework). Five dimensions surrounding the purpose were distinguished. (1) '*Programme in action'* describes the core activities of a programme, i.e. collecting information (such as on student performance or competence development), combining and valuing this information (i.e. drawing conclusions), and taking subsequent action. (2) '*Supporting the programme'* describes activities that are aimed at optimizing the current programme of assessment, such as improving test construction and faculty development, as well as gaining stakeholder acceptability and possibilities for appeal. (3) '*Documenting the programme'* describes the activities necessary to achieve a defensible programme and to capture organizational learning. Elements of this are: rules and regulations, learning environment, and domain mapping. (4) '*Improving the programme'* includes dimensions aimed at the re-design of the programme of assessment, after the programme is administered. Activities are R&D and change management. (5) The final dimension '*justifying the programme*' describes activities that are aimed at providing evidence that the purpose of the programme is achieved by taking effectiveness, efficiency, and acceptability into account.

These GLAD were designed from a utilitarian - fitness-for-purpose - perspective and are therefore intended to be context-independent. Therefore, they are formulated to support the design of high-quality programmes of assessment regardless of the assessment purposes. Thus we expect them to be able to support programmatic assessment design decisions across a wide variety of contexts and educational philosophies. However, the successful development, implementation and application of these guidelines call for systematic evaluation and research. To establish sufficient underpinning for GLAD we adopted the evaluation process of evidence-based clinical practice guidelines (Babinski, 1995), which defines three phases of evaluation. In the first phase - evaluation during development - we used expert validation focused on achieving a set of GLAD that is carefully defined, consistent, and non-redundant (Dijkstra et al., 2010, 2012). In the second phase - evaluation in real practice - we used an instrumental case study to determine the practicality and explanatory power of GLAD (Prochaska et al., 2008). GLAD were found to be sufficiently relevant for application in assessment practice, and we were able to explain almost all strengths and weaknesses in the assessment design (Dijkstra et al., 2013). GLAD serve as a useful and sufficient framework to describe, evaluate and assessment programmes. They provide the expert with a helpful vocabulary to support decisions made in assessment design.

In the present study we evaluate the GLAD in the third phase - evaluation of the effects of guidelines within defined environments. We evaluated a competence-based assessment programme using the GLAD and investigated whether the outcome of this evaluation is comparable to the outcome of an evaluation method specifically aimed at competence-based environments. The same case was also evaluated with a well-researched and validated, method for evaluating quality of Competence Assessment Programmes (CAP) (Baartman et al., 2007; 2007a; 2007b; Baartman et al., 2011; Jonsson et al., 2009). The framework as described by Baartman et al, (2007) consists of 12 quality criteria for CAP, which is applicable to any assessment programme aimed at measuring competencies. CAP criteria have been evaluated by investigating opinions of teachers who work with competence assessments (Baartman et al., 2007a), by comparing the criteria against Messick's (1984, 1994, 1995) framework of construct validity, and in a number of practical evaluations of assessment programmes in (higher) competence-based education (Jonsson et al., 2009;

Baartman et al., 2011; Baartman et al., 2013). See Table 5.1 for an overview of CAP criteria and a brief description, based on Baartman et al. (2007).

Table 5.1: Quality criteria for competence assessment programmes (based on Baartman et al., 2007)

Criterion	Brief description
1. Fitness-for-purpose	Alignment between curriculum goals and what and how is assessed. Criteria and standards should address all competences and the mix of methods should be fit to assess competence
2. Reproducibility of decisions	Decisions about students should be based on multiple assessors, multiple tasks and multiple situations
3. Transparency	CAP should be clear and understandable for all stakeholders
4. Acceptability	All stakeholders should approve of the assessment criteria and methods
5. Comparability	Assessment tasks, criteria, working conditions and procedures should be consistent with respect to key features of interest
6. Fairness	Students should get a fair chance to demonstrate their competences, for example by letting them express themselves in different ways and making sure the assessors are trained and do not show biases
7. Self-assessment	CAPs should stimulate self-regulated learning, for example by using self-assessments, and letting students formulate their own learning goals
8. Meaningfulness	CAPs should be learning opportunities in themselves and generate rich and useful feedback
9. Cognitive complexity	CAPs should enable the judgment of thinking process, besides assessing the product or outcome
10. Authenticity	The degree of resemblance of a CAP to the future workplace
11. Educational consequences	The degree to which the CAP yields positive effects on learning and teaching
12. Costs and efficiency	The feasibility of carrying out the CAP for assessors and students

The CAP criteria incorporate the idea that assessments not only serve a summative purpose, but also a formative one: providing information-rich and valuable feedback to students while stimulating students' learning process towards competence development. Therefore, CAP criteria not only focus on quality issues such as validity and reliability, but also on acceptability, meaningfulness, educational consequences, and fitness for self-assessment. The application of CAP criteria to evaluate programmes of assessment leads to valuable insights in strengths and weaknesses, as well as areas for improvement of the assessment programme.

In this case study we calibrate our GLAD against the CAP criteria. We compared the outcome of both quality evaluations to determine the *utility* and *productivity* (Prochaska et al., 2008) of the GLAD in relation to the CAP criteria. The utility criterion establishes whether using GLAD in a systematic and meaningful way leads to useful outcomes compared to the outcomes of the evaluation with the CAP criteria. CAP criteria are validated in a competence-based educational environment. Thus, it is expected that this evaluation yields useful results, to test the outcome of the GLAD on the *utility* criterion.

The productivity criterion is determined by checking whether GLAD can build on previous research, i.e. the CAP criteria, and can offer a more comprehensive framework and generate new areas for research. CAP criteria are well researched, which defines a sound baseline for evaluating GLAD against the *productivity* criterion.

To evaluate the quality of a competence-based assessment programme using GLAD, interviews and document analysis were conducted. The case was also evaluated by the CAP criteria using a self-evaluation tool followed by a group interview (see: Baartman, et al., 2007). Both evaluations have resulted in an in-depth qualitative analysis of the assessment programme.

Method

Context of the case

The case selected in this study is the general practitioners (GP) residency programme at Maastricht University, the Netherlands (www.huisartsopleiding.nl & www.huisartsgeneeskundemaastricht.nl). The assessment programme was purposefully selected on the basis of a well-defined educational environment (i.e. competence-based assessment) for which there is a well-researched evaluation method (i.e. the CAP criteria). The institute was interested in participating, because they wanted a quality evaluation of their competence-based assessment programme.

A national body governs the Dutch GP residency programmes by describing the assessment principles, regulations and instruments to be used, which underpins the competence-based approach. Within the boundaries of the national protocols, the local departments have certain degrees of freedom to implement the assessment in practice, taking local circumstances into account, such as logistics and patient mix.

The purposes of the assessment are to determine if a resident is progressing at the expected level (selective function) as well as to support their development/learning (educational function). The seven competencies (CanMeds) are further explicated in 22 sub-competencies.

The three-year residency programme consists of a work-based learning programme in the authentic setting (4 days per week) i.e. in a GP practice (years 1 and 3), a hospital, a nursing home, and a psychiatric clinic. In addition, the residents attend learning formal education settings at the educational institute one day per week.

The assessment programme consists of work-based assessments aimed at observation in actual practice, and it includes instruments such as written assessment for knowledge testing. All assessment data are aggregated by a specifically designed instrument, which combines the various test results and observations, to gain in-depth information about the trainees' achievement in each of the 22 sub-competencies. The programme director makes a Go-No-go decision about promotion to the next training year or graduation.

Research procedure and data collection

The research procedure contains three steps. First, an evaluation of the competence-based assessment programme is undertaken using the GLAD, determining its strengths and weaknesses, and areas for improvement and further development (recommendations). Second, an evaluation of the same case using the CAP criteria takes place to determine the degree to which the current assessment programme meets the criteria (i.e. strengths or weaknesses) and to identify if improvements are necessary. The third step consists of comparing both evaluations from Step 1 (the use of GLAD) and Step 2 (the use of CAP). The similarities and differences in between the two evaluations are investigated in coverage of elements in assessment programmes as well as evaluation outcomes regarding strengths, weaknesses, and recommendations for improvement.

Step 1: GLAD evaluation
Data for the evaluation with GLAD were gathered by conducting interviews, supplemented with document analysis. The interviews focused on the quality of the Maastricht GP residency programme and were held with the former national assessment coordinator (PR), who was also highly involved in implementing the assessment in the Maastricht context. In multiple sessions moderated by JD, the GLAD were applied to the current assessment programme of the GP residency programme in detail (per guideline) and to determine whether and how GLAD were addressed during the design process. Specific documents describing the assessment programme - including documents provided by the national organization - protocol describing assessment procedures and assessment plan describing a vision on assessment and instruments (www.huisartsopleiding.nl) were used to support the interview process and to check specific details. A written interpretative summary of the analysis was made by JD and AT and checked with the assessment coordinator, who clarified and nuanced specific statements. The strengths and weaknesses, and the associated

recommendations were divided into 12 themes based on context-specific groups of guidelines (see Appendix A) that were applied in accordance with this particular context.

Step 2: CAP evaluation
Data for the evaluation with CAP criteria were gathered according to the procedure used and described by Baartman et al. (2007). The research team (JD and AT), jointly with the programme director, selected 12 stakeholders based on their involvement and knowledge of the assessment programme: 10 agreed to participate. Their positions were: program director, associate programme director, curriculum coordinator, assessment coordinator, GP-supervisor, resident/assessee (see also Gulikers et al., 2010 for a justification of the selection of stakeholders in the evaluation).

First, all participants discussed the CAP that should be evaluated in order to develop shared understanding of the components and purposes of the assessment program. They also received instruction on how to interpret and use the 12 evaluation criteria and indicators (see Table 5.1 for an overview) in order to guarantee uniformity among all participants regarding the CAP used and the evaluation starting point. Second, nine participants (one could not participate due to personal circumstances) filled out a self-evaluation tool on the CAP. The self-evaluation tool consisted of the operationalization of the 12 CAP criteria in the form of 4 to 6 indicators per criterion. For each criterion, the participants indicated to what extent their CAP concurs with the criterion (by means of a qualitative slide bar) and they were asked to provide the rationale or some form of evidence to substantiate their judgement.

The participants were subsequently asked to reach consensus on each criterion ('Does our CAP meet the criterion and do we consider this sufficient?') by discussing the individual evaluations in a group session led by one of the researchers (JD). The aggregate of individual evaluations and anonymous comments were used to structure the discussion. The qualitative remarks were summarized into meaningful statements per CAP criterion. The conclusions were sent to all participants as a member-check procedure. Only textual improvements and one minor clarifying addition were made.

Step 3: Outcome comparison of GLAD and CAP criteria
In step 3, the outcomes of both evaluations were systematically compared (matched). First, the GLAD and CAP criteria were matched on terminology and assessment components described. AT and JD independently matched the GLAD to the outcome of the CAP evaluation, and subsequently discussed the matching until consensus was reached. Using a similar procedure, LB and JD matched CAP criteria to the outcome of the GLAD-evaluation. Next, both matching procedures were discussed by the research team and summarized into an overall matching outcome (see Table 5.2). Subsequently, this was analysed on coverage of the content,

level of detail, and nature of the evaluation. JD provided a preliminary analysis of these items, which was discussed by the research team until consensus was reached.

Results

Both evaluations covered similar issues in assessment. The results of the matching process are summarized in Table 5.2. The division of issues in assessment in 12 CAP criteria and 72 GLAD grouped under 12 themes complicated the comparison, especially because every division in larger groups has some sort or arbitrariness to it. A more in-depth analysis is required to compare the outcome of both evaluations, i.e. recommendations and conclusions.

When matching certain GLAD to CAP, at a first glance these GLAD could be applied to all CAP criteria. For instance, GLAD B2 ('content over format') could be matched to all CAP criteria that deal with an assessment instrument. Similarly, the theme about stakeholders could be matched to multiple CAP criteria because in many criteria stakeholders are mentioned.

It is remarkable is that there is no match between Theme 1: purpose of the programme and CAP1: Fitness-for-purpose. This may be due to the nature of the related CAP criteria and GLAD differ in the sense that the GLAD have not predefined the purpose of the programme, which could be regarded as the function of the assessment programme. On the other hand the CAP criterion and indicators under fitness-for-purpose focus more on content; hence, the match with Theme 4). Within the competence-based educational philosophy certain choices were made that determined the purpose of the assessment. This philosophy was translated to the CAP criteria in 'Self-assessment/self-directed learning' and 'Meaningfulness'. These criteria are not explicitly addressed in the GLAD because of GLAD's generic nature. The application of GLAD depended on the defined purposes, causing CAP criteria (7, 8 and, 11) to be matched to Theme 1 (Purpose of the programme). An example of a GLAD specifically referring to the purpose related to the CAP criteria 'Meaningfulness' is: '*Consequences should be proportionally and conceptually related to the purpose of the assessment and justification for the consequences should be provided.*'

Furthermore, GLAD Themes 4 and 6 were matched to a CAP criterion more often than any of the other themes. These two themes deal with the 'Programme in action' dimension of the GLAD framework: collecting information (or selecting instruments), combining and valuing information, and taking action. This seems to indicate that the CAP criteria focus more on the programme of assessment as it is run. This is similar to the dimension of our framework: Programme in Action). More specifically, it seems to emphasize the quality of the information collected and its use for learning and decision-making.

Table 5.2 Match between GLAD and CAP

	CAP 1: Fitness-for-purpose	CAP 2: Reproducibility of decisions	CAP 3: Transparency	CAP 4: Acceptability	CAP 5:Comparability	CAP 6: Fairness	CAP 7:Self-Assessment	CAP 8: Meaningfulness	CAP 9:Cognitive complexity	CAP 10:Authenticity	CAP 11: Educational consequences	CAP 12: Costs and efficiency
Theme 1: purpose of the programme							X	X			X	
Theme 2: resources / infrastructure												X
Theme 3: stakeholders				X		X				X		
Theme 4: Content and components	X	X			X				X	X		
Theme 5: Coherence and Effects							X				X	
Theme 6: Decisions, standards, action		X	X		X			X	X	X		
Theme 7: Robustness						X						
Theme 8: Context and implementation												
Theme 9: Procedures; Rules and Regulations			X									
Theme 10: Quality Assurance					X							
Theme 11: Change management												
Theme 12: Justification												

Table 5.2 also shows that the GLAD themes 8, 11 and 12: (context and implementation, change management, and justification), could not be matched with any CAP criterion. This also suggests that the GLAD take a broader perspective on assessment programmes, including organizational issues of implementation and change, as well as justifying the assessment programme to external parties. For instance, in the CAP criterion 3 (transparency) focuses on the CAP in use, e.g. assignments, judgements and procedures. This is covered by the GLAD in a similar fashion. However, GLAD extends the transparency issue to other dimensions in the framework. In the *Infrastructure*, a guideline is formulated around getting transparency about resources; in the dimension *Justifying the programmes* open and transparent governance is advised.

In general, both evaluation outcomes are comparable and similar with respect to the assessment components that are addressed in both evaluations. However, their levels of detail and starting points differ. The CAP criteria tend to be more concrete and targeted towards issues related to competence-based assessment, whereas GLAD are independent of an educational philosophy. When applying the GLAD to a competence-based assessment programme all these specific issues are addressed as well. However, the assessment developer (expert) has to combine several GLAD and take a decision about the issues. For instance CAP-criterion *Authenticity* is not explicitly addressed in the GLAD as it is not a purpose of every programme of assessment. When we compare the outcomes of both evaluations the conclusions are equal. The *Authenticity* criterion is sufficiently met – nearly 100% - according to the CAP, as assessment mainly takes place in real life situations, and written exams take general practice cases as their starting point. The same conclusion can be drawn based on the combination of GLAD in Theme 4 (content and components). Here GLAD B1 to B3 address selection of instruments or assignments; GLAD B12 addresses the circumstances and GLAD D9 the domain map, which result in descriptions of assessment components as part of the assessment programme. The assessment in this specific case is authentic because it takes place at the 'does' level of Miller's pyramid (Miller, 1990); i.e. in real life settings. Here, during the assessment design it was decided to use work-based assessment. Hence, the match between these GLAD and the 'Authenticity' CAP-criterion is not obvious.

Similarity of evaluation outcomes is also illustrated by conclusions regarding assessor expertise: the CAP indicator 'assessors are knowledgeable about the work environment' is evaluated positively. Because assessors are mainly GPs the same conclusions can be drawn using GLAD A6 (role of stakeholders, i.e. assessors) and C4 (required expertise, i.e. knowledge of work environment). However, the GLAD do not prescribe that stakeholders should be knowledgeable about the work environment. Hence, the outcome of the GLAD evaluation states that stakeholders and their roles should be explicated and decided upon. The outcome of CAP resulted in more concrete advice, in contrast to the outcome of GLAD evaluation where the advice to the organization or assessment developer in some occasions is formulated as a suggestion to take a stance in certain aspects or to further evaluate whether GLAD are taken into account.

Discussion

The GLAD meet the *utility* criteria, which is based on the fact that the evaluation of a competence assessment programme with GLAD leads to useful recommendations, which are corroborated by recommendations (or areas for improvement) derived from the well-researched and validated CAP analysis. The CAP approach takes an ideological perspective on the quality of an assessment programme defined by the criteria: quality is equivalent to meeting the criteria.

Compared to the outcome of the CAP evaluation the outcome of the GLAD evaluation also directs attention beyond the defined criteria (e.g. room for improvement, further development of the programme) and enables design choices to be made by assessment developers. GLAD are formulated with an assessment development perspective in mind. In contrast, CAP criteria are aimed at determining the quality of the assessment programme.

The GLAD also meet the *productivity* criterion because it extends the CAP criteria with new areas for evaluation of programmes of assessment within the competence-based assessment context. In addition to the assessment as a measurement issue, programmatic issues such as organization and implementation as well as justifying the assessment to other parties are also addressed as possible recommendations for improvement.

The generic GLAD are applicable to a specific context, in this case a competence assessment programme. Compared to validated quality criteria specifically developed for CAP, we found similarities in the content of evaluation of an assessment programme. The example of Authenticity illustrates the strength of the GLAD being inclusive and taking a programmatic holistic approach, which enables assessment developers to adapt these to specific purposes. However, at the same time it illustrates the weakness of abstract guidelines at a macro level. Choices are open to the assessment developer, where sufficient assessment expertise is required to translate the GLAD to the concrete (educational) context.

Despite all the similarities between GLAD and CAP it has become clear that there is one fundamental difference, which exists by design. CAP clearly starts from the notion that quality of an assessment programme is inherent to it being competence based and is to be used as a measurement instrument for the extent to which a programme actually adheres to the critical aspects of competence-based assessment. GLAD on the other hand are designed from the utilitarian standpoint that the coherence between the stated purpose and values on the one hand and the expertise in the organization, the activities of those experts and the way in which the organization and regulations support the experts in their activities is the key factor in determining the quality of an assessment programme. In other words, GLAD provide the expert user with a helpful vocabulary to describe, evaluate and design or improve and assessment programme. Such 'vocabulary' approaches are also gaining traction in areas such as definition of outcomes and competence and other areas of quality.

This study provides evidence that the GLAD are relevant in practice and lead to valuable areas of improvement of assessment programmes. However, the scope of this study is limited to the application of GLAD to one competence-base assessment environment. Also the application of GLAD is only compared to one evaluation method. Further research is needed to collect more evidence in more diverse range of settings in order to produce more concrete support to apply the abstract GLAD in a broader variety of contexts.

Conflict of interests

Two authors (AT and PR) are faculty members of the GP residency programme, however, given the critical analysis and the outcome of the evaluation a conflict of interest is very unlikely. Furthermore the critical evaluation was done in a safe setting, which enabled subjects to speak freely about the programme. The rest of the team was not affiliated with this institute and had no conflict of interest in their assessment programme.

References

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

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007a). Teachers'opinions on quality for competency assessment programs. *Teaching and Teacher Education, 23*, 857-867.

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 programs. *Studies in Educational Evaluation, 32*, 153-170.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007b). Evaluation assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*, 114 - 127.

Baartman, L. K. J., Gulikers, J. T. M., & Dijkstra, A. (2013). Factors influencing assessment quality in higher education. *Assessment & Evaluation in Higher Education, 38*, 978-997.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Determining the Quality of Competence Assessment Programs: A Self-Evaluation Procedure. *Studies in Educational Evaluation, 33*(3-4), 258-281.

Baartman, L. K., Prins, F., Kirschner, P. A., & Van der Vleuten, C. (2011). Self-evaluation of assessment programs: A cross-case analysis. *Evaluation and Program Planning, 34*, 206-216.

Basinski, A. S. (1995). Evaluation of clinical practice guidelines. *Canadian Medical Association Journal, 153*(11), 1575-1581.

Bok, H. G., Teunissen, P. W., Favier, R. P., Rietbroek, N. J., Theyse, L. F., Brommer, H. et al. (2013). Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Medical Education, 13*(1), 123.

Dijkstra, J., Galbraith, R., Hodges, B., McAvoy, P., McCrorie, P., Southgate, L. et al. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*, 20.

Dijkstra, J., Van der Vleuten, C., & Schuwirth, L. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education, 15*(3), 379-393.

Gulikers, J. T. M., Baartman, L. K. J., & Biemans, H. J. A. (2010). Facilitating evaluations of innovative, competence-based assessments: creating understanding and involving multiple stakeholders. *Evaluation and Program Planning, 33*(120 - 127).

Jonsson, A., Baartman, L. K. J., & Lennung, S. (2009). Estimating the quality of new modes of assessment. The case of an "Interactive Examination" for teacher competency. *Learning Environments Research, 12*, 225-241.

Messick, S. (1984). The Psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63-67.

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model. *Applied Psychology: An International Review, 57*(4), 561-588.

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

Van der Vleuten, C. P., Schuwirth, L. W., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. et al. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher, 34*(3), 205-214.

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 practice & research. Clinical obstetrics & gynaecology, 24*(6), 703-719.

Appendix A: 12 Themes of recommendations - based on context-specific groups of guidelines

The letter-number combinations refer to the guidelines. See the addendum for a more elaborate description and the original division in the dimensions of the framework.

(1) purpose of the programme

A1 one principal purpose of the assessment programme should be formulated.
A2 long-term and short-term purposes should be formulated(limited number).
F8 to establish a defensible programme one vision (on assessment) should be communicated to external parties.

(2) infrastructuur and resources

A4 opportunities and restrictions should be identified early and taken into account in the design process.
A5 decisions should be checked against consequences for the infrastructure.
B8 the overt and covert costs of assessment compared to alternatives.
F5 in order to justify the resources used for the assessment programme, all costs should be made explicit.
F6 a cost-benefit analysis should be made regularly in light of the purposes. In the long term, a proactive approach to search for more resource-efficient alternatives should be adopted.

(3) stakeholder roles

A6 stakeholders should be identified and roles should be assigned.
A7 the level of stakeholder participation should be based on purpose and needs.
C4 support for constructing the assessment components requires domain and assessment expertise.
C5 support tasks and should be well-defined and responsibilities lie with the right persons.
C8 acceptance of the programme should be widely sought.

(4) content and assessment components

A3 an overarching structure which projects the domain onto the assessment programme should be constructed.
B1 the extent to which the assessment component contributes to the purpose(s) should be the guiding principle.
B2 consideration of the content (stimulus) should take precedence over the response format.
B3 sample the intended cognitive, behavioural or affective processes at the intended level.
B12 when administering an assessment component, the conditions and the tasks should support the purpose.
D9 a domain map should be the optimal representation of the domain in the programme of assessment.

D10 a domain map should not be too detailed.
D11 starting point for a domain map should be the domain or content.
D12 a domain map should be a dynamic tool, and as a result should be revised periodically.

(5) coherence and effects of assessment components

B6 the effect of assessment on assessee behaviour should be taken into account.
B7 the relation between different assessment components should be taken into account.
B11 any performance categorization system should be as simple as possible.
B13 when scheduling assessment, planning should support instruction and provide sufficient opportunity for learning.

(6) assessment decisions – cut-off score - actions

B14 combination should be justified based on meaningful entities, either defined by purpose, content, or data patterns.
B15 the measurement level of the information should not be changed.
B16 the consequences of combining information, for all stakeholders, should be checked.
B17 the amount and quality of information should be in proportion to the stakes.
B18 a rationale should be provided for the standard setting procedures.
B19 consequences should be proportionally and conceptually related to the purpose and justified.
B20 the accessibility of information (feedback) to stakeholders involved should be defined.
B21 information should be provided optimally in relation to the purpose to the relevant stakeholders.

(7) robustness

C6 the higher the stakes, the more robust the procedures should be.
C9 protocols and procedures should be in place to support appeal and second opinion.
C10 a body of appeal should be in place.
C11 safety net procedures should be in place to protect both assessor and assessee.
C12 protocols should be in place to check on proportionality of actions taken and carefulness of assessment activities.

(8) context - implementation en programma improvement

B9 assessment approaches should first be re-evaluated for use in another context.
D7 the context in which the assessment programme has to function should be described.
D8 the relation between educational system and assessment programme should be specified.
E1 a regular and recurrent process of evaluation and improvement should be in place, closing the feedback loop.

E2 if there is uncertainty about the evaluation, more information about the programme should be collected.
E3 in developing the program (re-design), improvements should be supported by evidence (scientific or best practice).
F1 before the programme of assessment is designed, evidence should to be reviewed.
F2 new initiatives (developments) should be accompanied by evaluation, preferably scientific research.

(9) procedures, rules and regulations
C1 appropriate central governance of the programme to align assessment components and activities.
C6 the higher the stakes, the more robust the procedures should be.
C7 procedures should be made transparent to all stakeholders.
D1 R&R should be documented.
D2 R&R should support the purposes of the programme of assessment.
D3 the impact of R&R should be checked against consequences.
D4 in drawing up one should be pragmatic and concise and avoid complexity.
D5 R&R should be based on routine practices.
D6 an organisational body should uphold R&R and take decisions in unforeseen circumstances.

(10) quality assurance assessment programme
B4 information collected should be sufficiently informative.
B5 assessment should be able to provide sufficient information to reach certainty about the contingent action.
B10 a programme of assessment should deal with error and bias in the collection of information. Error (random) is unpredictable and should be reduced by sampling (strategies). Bias (systematic) should be analysed and its influence should be reduced by appropriate measures.
B17 the amount and quality of information should be in proportion to the stakes.
C2 assessment development should be supported by quality review to optimise the current situation appropriate to the importance.
C3 current assessment should be routinely monitored on quality criteria.

(11) change management
E4 momentum for change has to be seized or has to be created by providing the necessary priority or external pressure.
E5 underlying needs of stakeholders should be made explicit.
E6 sufficient expertise about change management and about the local context should be sought.
E7 faculty should be supported to cope with the change by providing adequate training.

(12) justification

F3 the programme of assessment should be reviewed periodically by a panel of experts.
F4 benchmarking against similar programmes should be conducted.
F7 open and transparent governance of the assessment programme should be in place and can be held accountable.
F9 the assessment programme should take into account superseding legal frameworks.
F10 confidentiality and security of information should be guaranteed at an appropriate level.

Chapter 6

General discussion

The need for assessment programmes in medical education is eminently clear; however, literature on designing these programmes is scarce, and provides information with limited applicability. To address this need, the general aim of this dissertation has been to develop guidance for design decisions with respect to programmatic assessment and to support assessment developers in achieving high-quality assessment programmes. A utilitarian approach to guidance was taken in order to deal with dilemmas and trade-off decisions inherent with applicability in a broad range of contexts. This approach also provides necessary independence from specific educational philosophies such that the guidelines could be generalized to and made suitable for a variety of assessment purposes.

The studies in this dissertation defined a framework for assessment programmes, from which the **g**uide**l**ines for **a**ssessment **d**esign (GLAD) was developed, validated and evaluated. The first phase of this research was aimed at developing comprehensive and generic guidance (Chapters 2 and 3). This generic guidance was mainly based on expert opinion collected through interviews with international experts in the field of medical education assessment. The next phase of the research was aimed at evaluating this guidance in two purposefully selected case studies, which were analysed in depth (Chapters 4 and 5). The following research questions were principle in this dissertation:

1. What areas or elements can be distinguished in the design of high-quality assessment programmes?
2. What guidelines can be formulated for design support based on the areas of assessment design?
3. What evidence can be provided to substantiate the validity of guidelines based on utilitarian principles in practice?

In this chapter, two main findings will be discussed: 1) a comprehensive framework for assessment programmes and 2) the 73 guidelines for assessment design (GLAD). In this chapter we will further reflect on the evidence for the framework and guidelines, and discuss the limitations of this research. Finally, suggestions for future research and possible implications for practice are discussed.

Framework for assessment programmes

The series of studies in this dissertation sought to gradually build and validate a framework for the design and improvement of assessment programmes. To that end, before any specific support for assessment developers could be provided, it was essential to determine which areas or elements constitute a framework for quality in the design of assessment programmes. In our first study we developed the framework and its dimensions for quality of assessment programmes based on a thematic analysis of experts' opinions regarding best practices in assessment and assessment design. The dimensions defined in this framework are a broad and comprehensive definition of assessment programmes. In addition to the dimension of the

'Purpose of Assessment', this framework includes five principle dimensions: Programme in Action; Supporting the Programme; Documenting the Programme; Improving the Programme; Justifying the Programme. These are embedded in a general context dimension about 'Stakeholders and Infrastructure' (see Chapter 2 for a more detailed description).

'Purpose of Assessment' takes a central role in this framework. During our interviews with the experts about quality of assessment, we found differing opinions; after further exploring these differences it turned out that these were founded in implicitly different purposes of assessment. These differences in opinion also occurred because of different views or philosophies on education and assessment. Although it seems self-evident to define the purpose of assessment first, more often than not, assessment appears to be designed without careful consideration of a clearly defined purpose. An example of that would be the selection of assessment instruments or definition of rules and regulations, which are more often determined by tradition rather than by deliberate choice with a well thought-out assessment purpose. However, in order to avoid confusion and to be able to discuss best practices in the expert focus groups – as mentioned in Chapter 2 – the assessment purpose was explicitly needed to reach an effective construct on quality.

Nonetheless, quality is an elusive, intangible phenomenon. And because assessment purpose was such a central factor in defining quality of assessment, we chose to adopt a utilitarian approach – as opposed to a more deontological approach – by defining quality of assessment in terms of fitness-for-purpose. There are two additional reasons for this:

1. Broad applicability: the fitness-for-purpose principle makes the framework independent of any specific educational philosophy.
2. Long-term applicability: the framework is independent of trends in education and assessment (e.g. portfolios, competency-based education) or paradigms about learning (behaviourism, cognitivism, social constructivism) that are not always consistent over time.

This utilitarian approach also has a downside with regard to assessment development. Although the framework has broad applicability, it requires considerable assessment expertise as well as excellent understanding of assessment purposes and of the context in which the assessment programme has to operate. Before it can be used, the purpose of the assessment has to be defined and – because a framework like this does not function in a vacuum – it has to be applied to a specific case. To address this dynamic, the purpose and five principle dimensions are further embedded into a broader dimension of stakeholders and infrastructure, which are key factors in the assessment context.

Many assessment designs focus exclusively on the 'Programme in action' dimension – the actual running of an assessment programme. In such designs, emphasis is put on the selection of the assessment instruments based on reliability and validity, and on robustness and trustworthiness of decision making (e.g. about student progress). Our comprehensive and inclusive framework, however, stresses the fact that assessment is more than a measurement problem. Moreover, the framework and GLAD developed in the studies in this dissertation show that assessment is not only a design problem, but an organisational problem as well. It is therefore no surprise that most of the guidelines do not address psychometric criteria, but deal with the relation between an assessment programme and the context, logistics, and organisation. The stakeholders and infrastructure dimension that supports all other dimensions illustrates this point.

The framework for assessment programmes (the five principle dimensions embedded within the stakeholders-infrastructure dimension), offers a broad and inclusive definition of assessment. Assessment is not just regarded as a measurement problem, but also organizational issues are incorporated with a stress on continuous improvement. This offers a broader theory on assessment design, comparable to frameworks for instructional curricula. The comprehensiveness and broad scope of the framework is at the same time a pitfall. The interrelatedness and the large amount of relevant assessment components that have to be taken into account make designing an assessment programme a complex endeavour. It requires not only assessment expertise, but also a spectrum of expertise in broad areas, such as change management and organisational knowledge. The design of assessment programmes is inherently a matter of teamwork. This stresses the importance of the stakeholders-infrastructure dimension.

GuideLines for Assessment Design (GLAD)

The GLAD were developed and evaluated in subsequent studies based on the framework. During its development, evaluation of the GLAD focussed on clarity, consistency, and parsimony (Prochaska, 2008). Although the GLAD meet these criteria, the comprehensive and inclusive nature of the framework led to a large set of 73 GLAD. Overall, the GLAD are applicable, useful, and representative in different contexts. The application of the GLAD to two very different cases (in Chapters 4 and 5) provides evidence for transferring and generalising these findings to other contexts. However, the use of such a comprehensive set of guidelines that takes into account many – if not all – components of assessment design makes it a complex process. This is not only due to the number of guidelines and the fitness-for-purpose dimension, but also due to the interrelatedness of the guidelines.

A downside of the utilitarian approach is that it may make the guidelines seem abstract to the assessment developer. On the other hand, it supports the assessment developer by not being prescriptive, which has two main implications for the application of the GLAD. Firstly, as mentioned earlier, application of the GLAD requires a high level of expertise of the assessment developer in multiple areas that vary from assessment measurement to organisational issues. This expertise is often not immediately available within educational and other institutes, which means it has to be developed within an organisation or brought in from external parties. A team of experts (stakeholders) with different background is required. The organisation itself has to be such that it accommodates the teamwork and collective processes as well as the expertise needed for design of high quality assessment programmes. The need to team up various different experts illustrates the complexity of the design and the strong interrelatedness of the dimensions in the framework.

Secondly, application of the GLAD demands that the assessment developer has a high level of reflective ability with a complete overview of the interrelatedness of de GLAD inherent in the dimensions of the assessment programme. An assessment programme is more than the sum of its parts; the design of assessment programmes needs to be approached with a holistic perspective. This has implications for the use of the GLAD. They are unlike quality criteria that are used as algorithms and *if-then* statements to arrive at a decision or determine a subsequent action; the GLAD are not a tick list. As part of an assessment design, they are a set of considerations that should be thought over, regardless of whether or not they are addressed in the final design. Our studies showed that the GLAD meet the *practicality* criterion (Prochaska, 2008) and thus are found and used in practice. However, the relevance of specific GLAD can differ given the purpose and context of the assessment programme.

This is not a trivial deviation from more popular practice in medical education. Often instruments to evaluate elusive concepts – quality, professionalism, or competence – start from a bottom-up notion in which individual items have to be completed and their results have to be aggregated to result in an evaluation of the concept. Multiple-choice items measure knowledge, and questionnaire or Likert scale items measure professionalism, reflection, etc. The GLAD are supportive of a top-down process in which the expert is provided with a comprehensive set of considerations – a vocabulary as it where – that can be used to describe, evaluate and improve the quality of an assessment programme. It may still be too soon, but there are indications that this reversal from bottom-up to top-down is also taking place in different debates. People participating in the medical programme debate are beginning to acknowledge that the definition of outcomes evaluated by ticking them off and adding up the results to determine a student's competence is flawed. These defined outcomes have a more useful function for education providers as a source of jargon to describe, evaluate and improve this student's competence.

The relevance of each guideline has to be judged by a team of experts. However, this means they may be perceived as less readily applicable in practice. Still, the results in Chapter 5 showed that the GLAD offer a framework and terminology to describe and explain the quality of the assessment programme. Herewith, the GLAD meet the *explanatory power* criterion (Prochaska, 2008), which allows an expert to use the GLAD terminology to describe the quality in terms that are relevant, qualitative, and narrative.

The holistic approach towards assessment design also allows assessment developers to shift focus points in the assessment design and combine different GLAD to make trade-off decisions and compromises. In evaluating the case in Chapter 5, the division of the GLAD according to the framework was abandoned because it did not fit the characteristics of the case. Although this study established the usefulness of the GLAD, it was even further optimized for the specific case by the fact that the GLAD where grouped in different themes of specific relevance for the programme of assessment. There are two reasons why it is important to acknowledge that the division of guidelines in the framework is to some extent arbitrary and that choices had to be made to define the key elements in assessment design and get an overview of design processes.

Firstly, application of the GLAD does not follow a specific fixed sequence. While trying to capture the design process in the two case studies it turned out that application of the GLAD is not a linear process. After determining the purpose of the assessment there is no fixed starting point (subsequent step), nor is there a predefined route to take through the GLAD. This is only logical because of differences in relevance of the dimensions of the framework and the GLAD depend on the context in which they are applied. But more important, the design process is an iterative process rather than a linear process. The interrelatedness of the GLAD means that changing one aspect in the design will influence others. Thus, decisions made at some point might have implications for the previous decisions.

Secondly, although the division of the GLAD in the framework can differ, and the order of addressing the GLAD is not fixed, there is a need for underpinning of use of the GLAD. This underpinning is provided by the utilitarian approach to quality of assessment. This approach puts the definition of the purpose central to the design process and provides a direction for the assessment design. It can also help in setting priorities in variations of GLAD to address their specific purposes. What already became clear during the first expert interviews when developing the framework, became even more evident when the GLAD were applied to the cases: Without a clear purpose of assessment the outcome of the design process and the relevance of the GLAD remains undetermined and remains stuck on the 'ifs' and the 'maybes'. With the exception of clarity of assessment purposes, no GLAD is inherently better or more important than another.

The iterative character of the design process based on the GLAD can be seen as a recursive effect in which quality assurance and improvement are part of the definition of quality. Within the design of the programme of assessment the redesign is already foreseen in the *Improving the Programme* dimension. The design of an assessment programme is not a one-off exercise, but requires continuous improvement or redesign. The redesign and improvement in itself need to be based on the GLAD, illustrating a never-ending cycle – comparable to quality assurance cycles in a quality culture. This connects to the definition of assessment as more than a measurement problem and more than a design problem. It is also obvious that no ultimate programme can be defined as a panacea for all assessment purposes. Programmes of assessment do not function in a vacuum and are the result of the interplay with developments in context. Hence, the outcome of a programmatic design process is not fixed. Both cases in Chapters 4 and 5 illustrated that changing circumstances (e.g. finance structures and regulations) require the assessment programmes to adjust to these new demands; i.e. programmatic design of high-quality assessment programmes is a dynamic process. The framework and GLAD can support the developers to keep an eye on the whole, when responses to specific changes are required.

The studies in this dissertation show that the GLAD meet the criteria for theory building according to Prochaska, 2008. The framework and GLAD can therefore be regarded as a theoretical framework to guide design of assessment. As mentioned earlier, the framework and GLAD are not tick-list instruments, they must be regarded more as an expert-support-system that provides a vocabulary to assessment developers and experts to describe, evaluate, justify, and improve an assessment programme. In applying the framework and GLAD to a specific setting, experts need to provide meaning to the GLAD within the intended context. Similar to a language, the GLAD can be regarded as the glossary and the expertise required to apply the guidelines can be seen as the grammar. Just as a glossary is not something to be read from A to Z, the 73 GLAD are not to be applied in a set linear sequence from beginning to end. Designing assessment is not a linear stepwise process, but rather a creative process for which the GLAD provide a vocabulary. Assessment programmes are too complex but still we can and must produce evaluative judgements about the quality of a programme, and answer the questions: ‘what should be improved?’, and ‘what are the strengths of the programme?’ The GLAD should be used to describe such judgements and give credibility to them without reverting to a reductionist approach.

Limitations

The studies in this research are inclusive, rather than exclusive, and they resulted in a comprehensive framework and 73 GLAD. Although only one GLAD was added to our framework after in-depth analyses of two assessment cases, there is still a margin of uncertainty about the completeness of the GLAD. This could be explained by the fact that both cases selected for the studies are best practices. Application and evaluation in other cases can provide more insight into the completeness of the GLAD.

The studies are all focussed on verification, rather than falsification of the GLAD. This means that the evaluation is focussed on providing evidence in favour of the GLAD and not against the GLAD. In theory, the context specificity and intended purposes of any assessment programme might be reasons for not finding evidence for GLAD, while at the same time these GLAD can be perfectly relevant in another case, and thus cannot be rejected based on this. Fortunately, all GLAD were supported by evidence in practice (See Chapter 4).

The criteria of Prochaska (2008) were found to be a sound basis to validate the GLAD; however, we did not explicitly check all criteria defined in Prochaska's framework, as this was beyond the scope of this dissertation. *Testability* and *generalizability* were criteria that are inherently investigated throughout the research. Although the framework and GLAD offer a theory for assessment design, the *integration* criterion in which constructs are combined could not be addressed, because mechanisms and laws are not clear currently. The criterion of *Impact* defined as 'efficacy × reach' was also not addressed. The application of the GLAD was studied retrospectively, to determine whether it was taken into account. In order to investigate Integration and Impact an intervention study is required to evaluate the impact in terms of achieving the purpose of assessment. However, the difficulty remains that strict mechanisms and laws are not available in our 'theory'.

Future research

Results from our studies show that design of high-quality assessment programmes is a complex process that requires expertise in various areas. The framework and 73 guidelines are fairly elaborate and therefore not easy to apply in an assessment programme. Therefore, future research should first be directed at studying other aspects of model or framework validity. Transferability seems one of the most logical ones to attempt first. This can be associated with studies into the necessary scaffolding as practical guidance to an expert using GLAD. At a more abstract level it would be important to better understand the narratives or change in narratives that using GLAD instils on its users in order to see which kind of assessment expertise – or to use the same terminology, assessment literacy – is best developed by the users of GLAD. Interesting approaches would then be to determine whether expertise correlates with the user's ability to produce many real life examples of applications and values of the GLAD criteria.

On the other hand, it might be worthwhile to further explore the possibilities of applying the GLAD by providing more concrete support using a specific educational philosophy or a specific assessment purpose. This raises the question about what can be formulated in a more normative and prescribed way, and which situations would benefit from this. Application to more and diverse assessment programmes and evaluating this could lead to common key characteristics which can then be defined as quality criteria. This is beyond the scope of this dissertation, which is to establish generic guidance, but it might be of practical relevance to many institutions that have a specific purpose or philosophy. The focus in that line of research would then be to make the GLAD ready for use in predefined educational environments.

Application in a variety of and a larger number of contexts would not only provide us with more practical and prescribing guidance on how to apply the GLAD, but it could also provide further information about the comprehensiveness of the framework and the GLAD, as well as its relevance in general. This would require more cases that provide more evidence about the generalisation of the GLAD to other contexts.

A very important implication of this work is that the framework is not only suitable to describe and evaluate an assessment programme but at the same time it can be useful to provide an overview of the current research on assessment. The literature seems to show that most research so far has been done in the 'Programme in Action' dimension. Research in the other dimensions seems to be scarcer, especially in relation to achieving the purpose of assessment. The GLAD could therefore be used as the basis of a literature overview to describe the whole body of research on assessment. The evidence provided for the framework and the GLAD is based on expert opinion and two retrospective cohort studies in which the GLAD were evaluated. Research on the design and redesign of assessment programmes in which the GLAD are applied can provide more and diverse evidence, allowing us to fine-tune the framework and the GLAD. The effect of this kind of intervention using the GLAD to redesign a programme of assessment is expected to be an optimal assessment programme.

Implications for practice

The framework and the GLAD developed and evaluated in the studies in this dissertation provide a new perspective on determining quality of assessment programmes. It provides a new theory to look at assessment programmes and a vocabulary that enables assessment experts to describe their holistic judgement of what a sound assessment programme constitutes. This theory places assessment in broader perspective to describe factors that influence the success or failure of assessment in achieving its purpose.

Assessment is not limited to being merely a measurement problem, and defining assessment as an instructional design problem does not suffice either. Assessment is also an organisational problem. Approaching assessment from a programmatic angle has implications for the arrangement of and around the assessment programme. Where traditionally assessment was fitted into the organisation, a programmatic approach also questions the organisational infrastructure, in terms of fitness-for-purpose. Because a programme of assessment does not function in a vacuum, the framework explicitly includes organisational aspects and restrictions, such as financial constraints and resources. Hence, the political developments should be monitored closely and the fitness of the assessment programme should be monitored periodically. In this sense, the design of a fit-for-purpose assessment programme is a continuous process.

A second implication for the organisation is that sufficiently broad expertise has to be available. Although expertise on the content of the assessment is often the most visible one, logistics and legal expertise are also of the utmost importance. For instance, credibility of the assessment cannot be established without sound logistic procedures. Student appeals virtually never address the content of the assessment. However, appeals often address procedures that were not followed properly (e.g. a test being ten minutes late in administering an exam leads to appeals and complaints about lack of fairness as a result of unnecessary extra stress, and thus less credibility). Constructing a programme of assessment has to be a team effort in order to combine sufficient expertise to address all issues (all GLAD). Staff development or even organisational development is required to achieve sufficient and sufficiently broad expertise.

The programmatic approach to assessment and the ideas that are brought forward can also be translated to other areas in which assessment of some sort is involved. For instance selection into medical education is a high-stakes process and becoming more important. For instance, in the Netherlands a recently new law states that medical schools must select students, whereas before selection was based on a lottery system. The research in the field of selection resembles the focus that has been found in assessment research (e.g. Koczwara et al., 2012; Patterson et al., 2012). The selection is broken down in separate criteria and research aims to find the ultimate measurement that accurately measures non-cognitive or non-academic attributes. This also resulted in strengths and weaknesses of various instruments. However, in selecting the right instrument, similar elements play a role, as defined by our framework and the GLAD. It is important to address the purpose first when it comes to selection too (Patterson, 2011, 2012). Selection methods differ when selecting the top or when excluding those who do not meet the minimum requirements. For example, an assessment programme can avoid overreliance on reliability estimates at the cost of validity. As the number of applicants often far exceeds the numbers of available placements, efficiency is an important driver. A 'system' of selection is needed to cope with this complexity and how to deal with multiple outcome measures that are required. The utilitarian perspective that is taken enables us to apply the GLAD to this form of assessment as well.

Another area where a programmatic approach to assessment can be beneficial is in accreditation of medical schools or institutes. For accreditation purposes the framework and the GLAD can be used to describe the assessment programme, but – as indicated before – the GLAD can also be used to define accreditation frameworks and criteria for assessment programmes attuned to assessment contexts, assessment purposes and/or educational philosophies. For instance the ASPIRE initiative taken by the AMEE, aims to reward the best practices (excellence) in medical education (www.aspire-to-excellence.org). Also in this assessment of organisations, we cannot suffice with one measurement or one criterion; however, simply adding up measurements does not reflect an institution's entire array of excellence; the whole is more than the sum of its parts. A programmatic approach can shift the focus in accreditation from the criteria of merely meeting the minimum requirements to that of determining excellence in medical education, while using a vocabulary that describes exactly what the expert implicitly knows regarding the quality of a specific programme.

The application for accreditation also illustrates the inherent dimension in the framework of assessing the assessment. The GLAD are developed for assessment design, but are useful as an evaluation framework as well.

In conclusion

The studies described in this dissertation define the areas or elements that can be distinguished in the design of high-quality assessment programmes. The boundaries of what an assessment programme constitutes are defined. Dimensions of assessment programmes are divided in a framework. From this framework guidelines are formulated for supporting design of assessment programmes. The studies in this dissertation provide evidence to substantiate the application of this guidance formulated from a utilitarian approach, based on a strategy to evaluate clinical guidelines and to evaluate theory development (e.g. relevance, applicability, and usefulness). At the same time we found that defining and determining quality is not a question of meeting criteria, but a question of providing experts with a vocabulary for conveying to others the description, evaluation, and explanation of the quality of an education programme.

References

Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Medical Education, 46*(4), 399-408.

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education, 46*(9), 850-868.

Patterson, F., Zibarras, L., Carr, V., Irish, B., & Gregory, S. (2011). Evaluating candidate reactions to selection practices using organisational justice theory. *Medical Education, 45*(3), 289-297.

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model. *Applied Psychology: An International Review, 57*(4), 561-588.

Addendum

Guidelines for designing programmes of assessment

The authors present these guidelines to be read with the following points in mind.

There is no linear order in the guidelines presented.

When reading the guidelines, you may not immediately come across those guidelines or important topics you would expect to be given priority. There is potentially more than one way of ordering the guidelines. As one example *costs* are important throughout the design process. However, because of the way this framework is constructed, *costs* are addressed near to the end.

There is overlap between guidelines.

It appeared impractical and somewhat artificial to split every assessment activity into separate parts. The guidelines are highly related, and overlap and/or redundancy are almost inevitable. In the example of *costs*, which are primarily addressed as part of *cost-efficiency*, references to *costs* are actually made in several guidelines.

The level of granularity is not equal for all guidelines.

Determining the right level of detail is a difficult endeavour, variable granularity reflects the fact that some issues seem more important than others, and others may have been investigated in depth.

Assessment components and assessment information.

In the guidelines we have sought to find an overarching term that would cover all possible elements of the programme, such as assessments, tests, examinations, feedback, and dossiers. We wanted the guidelines to be broadly applicable, and so we have chosen the term assessment *components*. Similarly for outcomes of assessment components we have chosen assessment *information* (e.g. data about the assessees' competence or ability).

GENERAL GUIDELINES

Three major themes emerged and are set out below. They are general and applicable to the design process as a whole. Although these guidelines are formulated more generally at this point, we will refer to these explicitly again in the separate dimensions as a reminder.

I Decisions (and their consequences) should be proportionate to the quality of the information on which they are based.

This guideline has implications for all aspects of the assessment programme, both at the level of the design of the programme, and at the level of individual decisions about assessees' progress. The higher the stakes, the more robust the information needs to be. In the dimension *Programme in Action* for instance, actions based on (collected) information should be proportionate to the quantity and quality of the information. The more high-stakes an action or decision, the more certainty (justification and accountability) is required, the more the information collection process has to comply with scientific criteria, and usually the more information that is required. If the subsequent action means that a assessee has to retake one examination, it has less impact when the action means the assessee has to retake an entire year of medical school. Therefore, the former can be taken on the basis of less information (e.g. the results of one single test). The latter, however, requires a series of assessments or maybe even a dossier.

II Every decision in the design process should be underpinned preferably supported by scientific evidence or evidence of best practice. If evidence is unavailable to support the choices made when designing the programme of assessment, the decisions should be identified as high priority for research.

This implies that all choices made in the design process should be defensible and can be justified. Even if there is no available scientific evidence, a plausible or reasonable rationale should be proposed. Evidence can be sought through a survey of the existing literature, new research endeavours, collaborative research, or even be outsourced completely. We stress again that the fitness-for-purpose principle should guide design decisions, i.e. which decisions will contribute optimally to achieving the purpose(s).

III Specific expertise should be available (or sought) to perform the activities in the programme of assessment.

This guideline is more specifically aimed at the expertise needed for the assessment activities in the separate dimensions and elements within the assessment programme. The challenge in setting up a programme of assessment is to 'get the right person for the right job'. Expertise is often needed from different fields including assessment expertise. Legal expertise, specific domain or content knowledge, and practical knowledge about the organisation are frequently required. Some types of expertise, such as psychometric expertise for item analysis, and legal expertise for rules and regulations, are obvious. Others are less clear and more context specific. It is useful when designing an assessment programme to articulate the skill set and the body of knowledge that is useful or even necessary to address these issues.

PURPOSE OF THE PROGRAMME

Purpose of the Programme constitutes a central role in the model for programmes of assessment. It is impossible to consider other assessment elements in isolation from the purpose. Regardless of educational approach (e.g. lecture-based education, problem-based learning) or the specific function of **assessment (e.g. learning tool, licensing decisions), the quality of assessment** programmes should be framed in terms of ***fitness-for-purpose***.

A1 One principal purpose of the assessment programme should be formulated.

This principal purpose should contain the function of the assessment programme and the domains to be assessed. The principal purpose should be formulated by high level stakeholders within the organisation, who are able to oversee the big picture and who understand the context which the organisation has to deal with. In many cases a programme of assessment has to take into account (and contribute to) multiple purposes. Defining one principal purpose of the programme might seem too ideal and theoretical, as the real world is messy. However, defining a principal purpose should contribute to coherence and consistency of the programme as a whole. E.g. in case of conflicts of interest a principal purpose should provide guidance for deciding on compromises. The challenge in designing a programme of assessment will be to combine these different purposes in such a way that they are achieved in the optimal way with a clear hierarchy defined in terms of importance.

A2 Long-term and short-term purposes should be formulated. But the number of purposes should be limited.

Although intuitively one would think that defining one principal purpose is ideal, such purposes are often defined to vaguely or too restrictive. Therefore in this guideline we advise to formulate short term (sub) purposes that will define more concretely the road map to achieving the main long term purpose. More than one purpose may be imposed by the dynamics of the environment as well. Not all purposes may come from within the organisation; external stakeholders might also exert influence. Defining these purposes enables their constructive inclusion in planning the programme of assessment.

(A) Purposes should be made concrete and feasible, but also transparent and referable. (B) Purposes should be prioritised based on (among other things) the principal purpose. (C) Purposes and their prioritizing should be justified based on sufficient information (such as: literature and scientific research, stakeholders, educational approach, etcetera). Explanation of these sub purposes should be helpful in managing different stakes.

A3 An overarching structure which projects the domain onto the assessment programme should be constructed.

An overarching structure should provide the big picture of the assessment programme that needs to be designed. *Domain* in this guideline can be interpreted in the broadest sense of the word. The overarching structure should function as a framework to ensure consistency and coherence of the assessment programme. It has to be formulated with high-level descriptors instead of detailed specifications of items on a test. A more detailed description or map of the domain onto the programme has to be documented (see *Documenting the Programme*: *Domain Mapping*). There is no generally applicable overarching structure and existing structures have to be evaluated on a case by case basis.

E.g. the contemporary idea of competency-based education generally uses a series of competencies to structure the domain - e.g. CanMeds (Frank, 2005), Good Medical Practice (General Medical Council, 2006), or ACGME (www.acgme.org) - although a simple list of topics to be covered could work just as well. From a different perspective, Miller's pyramid (Miller, 1990) dividing the domain into aspects of competence (knows, knows how, shows how, does), can also be used to map the domain onto the assessment programme. Another example is the (instructional) curriculum (what is taught, when and where) defining the development of mastery of the domain by learners. When deciding on the overarching framework to use, it might be acceptable to combine different existing structures and select the appropriate aspects from these. However, content is not the overarching structure, it populates it.

INFRASTRUCTURE

The dimension *Infrastructure* deals with the physical and practical systems and structures an organisation needs to have in place to support a functional assessment programme (e.g. an administration office or logistics of assessment). This is in contrast to the element *Learning Environment* in the *Documenting* dimension which describes more intangible aspects such as the culture of the organisation or institution and the educational setting or approach.

A4 Opportunities as well as restrictions for the assessment programme should be identified at an early stage and taken into account in the design process.

This guideline informs decisions regarding compromises on the purpose and/or resources (see A5). On a proactive note it is important to accept that it may not be possible to achieve all of the assessment ideals. Knowing the restrictions at an early stage prevents disappointment during the design phase and potential challenge during implementation.

A5 Design decisions should be checked against consequences for the infrastructure. If necessary compromises should be made, either adjusting the purpose(s) of the assessment programme or adapting the infrastructure.

Depending on the resources, urgency, and need to achieve certain purposes a balance has to be found between investing in infrastructure and making concessions to the purpose. Expertise in administrating tests and in logistics of the organisation is necessary. E.g. deciding to implement computerised examinations might simplify logistics of administrating tests and calculating scores. However, this decision has resource implication and may put a strain upon IT support and computer/network-facilities.

STAKEHOLDERS

Stakeholders are inextricably part of the programme as well as the design process. Although in various elements in the model, stakeholders are mentioned (as in *Acceptability* in the dimension *Supporting the Programme*), it is important to mention some aspects separately.

A6 Stakeholders of the assessment programme should be identified and a rationale provided for including the expertise of different stakeholders (or not) and the specific role(s) which they should fulfil.

This enables an informed decision to be made based on the involvement of stakeholders (in what phase or in what element). The roles they should play in the design and/or assessment programme should be clarified. Different perspectives can be distinguished, in relation to which several subgroups can be defined.

a. **society** (customers): e.g. patients, medical councils, government, tax-payers.
b. **assessee** (product): student, candidate, learner
c. **faculty** (company): management, teachers.

A7 The level at which various stakeholders participate in the design process should be based on the purpose of the programme as well as the needs of the stakeholders themselves.

The purpose of the assessment programme might render participation of some groups unnecessary or require that other groups participate. In each instance, a case should be made to demonstrate how involvement supports the principal purpose.

PROGRAMME IN ACTION

Programme in Action defines the currently running assessment practices. The four core activities of *Programme in Action* are: *Collecting Information*, *Combining Information*, *Valuing Information* and *Taking Action*. This dimension includes the activities minimally required to have a running assessment programme. As such, this is a necessary but not sufficient condition for a high-quality programme. This dimension encompasses activities ranging from collecting information to taking action based on that information.

COLLECTING INFORMATION

B1 When selecting an assessment component for the programme, the extent to which it contributes to the purpose(s) of the assessment programme should be the guiding principle.

In line with general guideline (II), a rationale for the selection of assessment components should be provided, preferably based on scientific research and/or best practice. Equally, the contribution each component makes to achieve the purpose of the assessment programme should be considered.

The guidelines in this section (B2 to B9) should aid in demonstrating the underpinning of the selection choices. Different components have different strengths and weaknesses and these have to be weighed against each other in order to decide the optimal balance to contribute to the purpose of the assessment. The interrelatedness of the guidelines should be taken into account in the design, but feasibility (*Infrastructure*) and acceptability (*Stakeholders*) are also clearly important. This is not as obvious as it seems. Currently design is often focussed almost exclusively on the characteristics of individual assessment components and not on the way in which they contribute to the programme as a whole. Often there is a tendency to evaluate the properties of an assessment component per se and not as a building block in the whole programme.

B2 When selecting an assessment (component or combination), consideration of the content (stimulus) should take precedence over the response format.

The target for assessment should determine the type of assessment components. A common pitfall arises because of the ready availability of assessment components or question formats. This results in fitting the question to the format. From research we know that the stimulus (content of the question) is more important than the format of the assessment. No response format is by definition better than another (form follows content). A similar format e.g. an open-ended question can be used to measure different types of knowledge e.g. Factual knowledge: 'Who is the President of the United States?'; or Clinical reasoning: 'Given the case described, what is the correct diagnosis? (Ward, 2006; Norman et al., 1985; Norman, 1988; Schuwirth et al., 2000)

B3 The assessment should sample the intended cognitive, behavioural or affective processes at the intended level.

In addition to content, the mental, behavioural, and affective processes evoked in assessees should also support the purposes of the assessment. A good doctor has to have the ability to deal with a variety of situations. To assess this ability, a programme of assessment has to be constituted using a range of assessment components. Many, if not all, assessment programmes already use a mix to get a complete picture of assessee performance in the domain of interest.

B4 The information collected should be sufficiently informative (enough detail) to contribute to the purpose of the assessment programme.

The goal of this guideline is to ensure that the information collected in the assessment programme can be used to fulfil (one of) its purpose(s). When selecting assessment components the characteristics of the information should be considered in relation to the purposes. A pass/fail or yes/no could suffice to permit assessees to practise medicine. However, if the (sub) purpose is aimed at improvement of the assessee then further information is necessary to inform the assessee about how to improve. Different characteristics can be of importance under different circumstances; If the priority is to measure incremental change, multiple measurements over time are required. If a purpose is to measure improvement (e.g. of the assessee or the educational programme), the collected information should be comparable to previous measurements. If the aim is to compare results on different topics, the results should not be combined into one score on the assessment, but should provide information on each topic.

B5 The assessment should be able to provide sufficient information to reach the desired level of certainty about the contingent action.

This guideline is a specific instance of general guideline I to ensure that the information gathered is of sufficient quality to ensure that consequent actions are consistent with the strength of the information. The higher the stakes the more certainty that is required to come to a decision and act on it. Sufficient information pertains to the amount of information in relation to its reproducibility. This raises the question 'When is enough enough?' (Schuwirth et al., 2002). It also works the other way around. If sufficient certainty exists that the intended decisions (action) will not change with additional information, collection of more information is not useful. But if the purpose cannot be achieved with sufficient certainty, further collection of information is necessary.

B6 The effect of the instruments on assessee behaviour should be taken into account.

Assessment drives learning. The assessment programme should support the educational principles or perspective on learning (if applicable) and contribute to the instruction (i.e. educational programme). Assessment should not hinder learning (or development), but it can be employed strategically to steer learning behaviour and thus strengthen the instructional value, e.g. Constructive Alignment (Biggs, 1996).

B7 The relation between different assessment components should be taken into account

The goal of this guideline is to avoid competition between instruments, but also to achieve efficiency of the programme. The aim is to achieve an optimal mix of instruments, usually by using strengths of one instrument to compensate for weaknesses of others.

Selection of instruments should result in a balanced compromise. Redundancy of information can be reduced for efficiency reasons; or be fostered in order to triangulate data. In other words, when selecting a mix of instruments, the method of combining information has to be taken into account. Combining information can also influence learning behaviour. If the weighting is unbalanced, assessees tend to study harder for the test that has the most weight.

B8 The overt and covert costs of the assessment components should be taken into account and compared to alternatives.

Costs is a separate element in the *Justifying the Programme* dimension. However, we feel this is an important aspect in the selection of assessment components and also worth mentioning here.

B9 Assessment approaches that work well in a specific context (setting) should first be re-evaluated before use in another context (setting) before implementation.

This guideline refers to that fact that there are not many *assessment activities* that are generally applicable across contexts. Every context has its own issues with feasibility and validity of an assessment component. Although it can be seen as best practice in one situation, it might be less so in another setting. Applicability of the assessment in its own specific context should be considered.

B10 A programme of assessment should deal with error and bias in the collection of information. Error (random) is unpredictable and should be reduced by sampling (strategies). Bias (Systematic) should be analysed and its influence should be reduced by appropriate measures.

Error and bias in assessment are unavoidable and must be taken into account when designing an assessment programme. Error is random: A single measurement or data-point is always flawed as a result of unsystematic error. More samples (as measurements or data-points) are needed to reduce the effect of error on the reliability of the assessment. Depending on the purpose and the stakes influencing the need for reliability, a trade-off decision should be made between efficiency and costs on the one hand and broad sampling on the other.

To deal with systematic bias, awareness of the bias has to be fostered and the source of the bias should be made transparent. Efforts to manage bias should be directed towards the source (e.g. the test material or the user). Bias in the test material (e.g. in a written test) is best tackled by improving the material (structuring, reviewing items, etc.) Bias as a result of (human) judgement, is best tackled by professionalising (training) the assessor. This means that improving the quality of assessment which is based on observation requires effort to be focussed on the user (i.e. assessor) e.g. by training, or fostering acceptability.

B11 Any performance categorisation system should be as simple as possible.

A performance system should be as simple as possible e.g. complex scoring systems do not add value to the assessment information; more often than not they complicate the interpretation. There is a tendency to give more weight to key items (and constructing *killer*-items), in order to increase validity of assessment. However, this contradicts the need for many items to achieve reliability of assessment. Validity should not be addressed by scoring systems, but by constructing high quality items or increasing complexity of items. The decisions based on different scoring systems do not vary much (Swanson et al., 1987). Differences occur around the pass-fail decision.

B12 When administering an assessment (component), the conditions (time, place, etc.) and the tasks (difficulty, complexity, authenticity, etc) should support the purpose of the specific assessment component.

In different stages of the assessment programme (or curriculum) the conditions and tasks may vary in their characteristics. Unnecessary (cognitive) load should be avoided. The level of the assessee should be considered. The context in which competency should be demonstrated should be supported by the selection of the instrument.

B13 When scheduling assessment, the planning should support instruction and provide sufficient opportunity for learning.

Assessment drives learning and as such determines the focus of assessees. Planning an assessment component too close to the time of instruction might undermine the attention given to this instruction, as assessees might already focus on the assessment component. Similarly people need time to study and to learn new things or to remediate deficiencies. Therefore sufficient time should be provided between instruction and assessment, as well as between assessment and re-sitting of the assessment. A distinction can be made between longitudinal development versus a more ad hoc (just in time) development.

COMBINING INFORMATION

B14 Combination of the information obtained by different assessment components should be justified based on meaningful entities either defined by purpose, content, or data patterns.

Different purposes (e.g. decisions regarding assessees versus decisions about the educational programme) require different ways of combining the information from the assessment components. This may involve approaches that do not necessarily combine results merely because they are of the same format (e.g. the results a communication station and a resuscitation station in one OSCE). What meaningful elements are is best defined in the overarching structure of the domain (guideline A3). To illustrate this further, if in patient care we would adopt the same standard procedure of combining information as in many assessment programmes, we would combine the sodium level and the potassium level because they are both of the same 'format'. But this combination is less meaningful than the combination of the sodium level and e.g. complaints of thirst. Unfortunately, available research in this area is minimal. The question remains: How to combine information from various (qualitative and quantitative) sources in a more meaningful way and reach a decision. But the paucity of this research only supports its urgency.

B15 The measurement level of the information should not be changed.

The measurement level of the information should not be changed just to be able to add things up. Qualitative scores should not be converted into quantitative scores. Often qualitative information is more useful than quantitative information. Different kinds of evidence should be juxtaposed in some way or another, but not necessarily numerically just to allow averaging. Hence, combining information is not necessarily about reducing information, but can also be about finding similar messages in the information. Other ways of combining should be explored, such as holistic or global judgments, triangulation, emerging themes and other methods from qualitative research.

B16 The consequences of combining information obtained by different assessment components, for all stakeholders, should be checked.

Combining information can mean loss of information (data reduction). this can be simultaneously useful for some stakeholders and useless for others: knowing whether you passed or failed does not tell you anything about your strengths or weaknesses.

VALUING INFORMATION

B17 The amount and quality of information on which a decision is based should be in proportion to the stakes.

This is a specification of general guideline I and guidelines B4 and B5. Whether enough information is collected depends among other things on the consequences of the actions to be taken based on this information. The higher the stakes, the more information is required to eliminate uncertainty in the outcome of the assessment e.g. failing an assessee from medical school based on one MCQ test is disproportionate.

B18 A rationale should be provided for the standard setting procedures.

This is also the underpinning of the set standard. The standard setting procedure should be chosen in light of to stakes, resources, and acceptability of false positives and false negatives. This is a specification of general guideline I. When stakes are high, justification of the standard setting procedures needs to be stronger to support the defensibility of standards. When the effect of incorrect decisions is severe, more care (effort and evidence) should be put into standard setting. In cases where human judges are assessing, they often use implicit standards. These should be made explicit and justified in order to achieve more defensible standards. Although often a decision has to be made as to what constitutes a pass, in the extreme case of a pure formative assessment, where the stakes are very low, a standard can be implicitly put in narrative feedback from an individual assessor to the assessee. Standard setting is always arbitrary but should never be capricious. Every standard contains more or less arbitrary decisions. To enhance acceptability and defensibility a rationale for these arbitrary decisions should be made explicit e.g. availability of a norm population, number of assessees, availability to provide a judgement, resources available, etc. In addition there is also the perspective on the standard that is set e.g. a statistical, an ethical, or psychological rationale for the standard.

TAKING ACTION

B19 Consequences should be proportionally and conceptually related to the purpose of the assessment and justification for the consequences should be provided.

This is a specification of general guidelines I and II. Severe consequences should be based on extensive and high quality assessment, whereas minor consequences can be justified with less information or information of lower quality.

B20 The accessibility of information (feedback) to stakeholders involved should be defined.

Information should be accessible to the appropriate stakeholders. How much information is provided and to whom, depends on the purpose and context of the assessment.

B21 Information should be provided optimally in relation to the purpose of the assessment to the relevant stakeholders.

In order to have the desired effect, the information (based on the purpose of the assessment programme) should reach the right persons in the right manner. Therefore care has to be taken in determining how to present feedback in order to optimise the intended results (including sub purposes) of the assessment programme. In some cases this means extensive feedback to achieve a change in learning behaviour, whereas in other cases a simple pass-fail notification can be sufficient. Feedback needs to be moderated and annotated so that users or receivers can understand the information and how it was collected, instead of just dumping information on receivers. This also implies that expertise on how to provide information is required e.g. faculty development on giving feedback might be beneficiary.

SUPPORTING THE PROGRAMME

Supporting the Programme includes activities contributing to the quality of the programme of assessment, which more often than not are related to, if not interwoven with, activities categorised under *Programme in Action*. This is about quality support activities, as distinct from infrastructural support. For an activity to support the programme in action and contribute to overall programme quality it should be directed at the purposes 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. The following two support-related themes are congruent with the concept of quality as being fitness-for-purpose. Together with *Programme in Action*, *Supporting the Programme* forms a cyclic process aimed at optimising the internal assessment system.

CONSTRUCTION SUPPORT

C1 Appropriate central governance of the programme of assessment should be in place to align different assessment components and activities.

One of the main problems in a decentralised assessment programme is the level of relatedness of assessment components and assessment activities. A central body of some kind should be in place to avoid sub-optimization and counterproductive initiatives within the programme of assessment.

C2 Assessment development should be supported by quality review to optimise the current situation (Programme in Action), appropriate to the importance of the assessment.

The development of all assessment components should include pre- and post- administration quality procedures. The amount of effort invested in quality review depends on the purposes of the assessment. Pre-administration procedures can include peer review of items (written assessment) or assessor training (observational assessment). Post- administration procedures can include assessor performance evaluation or psychometric analysis of items.

C3 The current assessment (Programme in Action) should be routinely monitored on quality criteria.

Evaluative information should be collected and acted upon. Evaluative information should be fed back to the current assessment programme and fed forward to a redesign to improve the programme and prevent mistakes in the future. Psychometric analysis and user satisfaction can be part of the quality review.

C4 Support for constructing the assessment components requires domain expertise and assessment expertise.

Both types of expertise should be included here. Improvements can then be made on various important aspects e.g. in content, format or assessment design. From this perspective, faculty development is an important quality improvement measure to enhance expertise in assessment issues and, as such, faculty development supports a programme of assessment. It was also seen as an activity to support the development and evaluation of the programme of assessment. Furthermore, it also relates to acceptance of the assessment programme (C8).

C5 Support tasks should be well-defined and responsibilities should lie with the right persons.

Expertise is used in the right place and at the right level (e.g. experts should be involved in constructing items (content) and not ticking boxes on a highly structured form). At the same time tasks should be appointed to specific persons to guarantee that things get done. For the same reason administrative support should be available.

POLITICAL AND LEGAL SUPPORT

C6 The higher the stakes, the more robust the procedures should be.

The procedures around the assessment programme should be robust and should be able to withstand legal challenge. There should be due process, meaning that assessment components and activities should be defensible. This does not mean all procedures should be explicit, standardised, or objective. Rather this implies that procedures should be acceptable to, or defensible for, all stakeholders. E.g. with high stakes examinations, it is more important to have safety net procedures in place as the consequences of the outcome are more severe and have a greater impact on the assessee. In contrast, if the purpose of an exam is only to provide feedback for an individual, the stakes are low and less attention can be given to procedures (more leniency can be permitted). This is a specification of the general guideline (I) that actions should be proportional to the quality of the information. With robust procedures the quality of information and/or the quality of decision-making can be increased.

C7 Procedures should be made transparent to all stakeholders.

Procedures should be made easy to understand. Complexity and exceptions should be avoided as much as possible.

C8 Acceptance of the programme should be widely sought.

As the outcome of the assessment programme often influences stakeholders, it is important that the stakeholders accept the assessment programme. Although accepting the assessment does not necessarily mean liking it, stakeholders should buy into the programme of assessment and the instruments used in it. Especially when the user determines the quality of the instrument (using a scoring form), users' opinions and motivation are critical.

C9 Protocols and procedures should be in place to support appeal and second opinion.

When decisions are made, often some disagreement with the decision arises. Having protocols in place to deal with disagreement makes the defensibility of decisions stronger (if it holds up), contributes to acceptance of stakeholders, and may avoid legal challenge. Such protocols and procedures also constitute a safety net or quality assurance opportunity to identify and address mistakes in the programme of assessment.

C10 A body of appeal should be in place

Filing an appeal should be safe for the applicant. Consideration of the appeal should be sufficiently objective e.g. by establishing a body, which is independent of the organisation, to deal with this.

C11 Safety net procedures should be in place to protect both assessor and assessee.

To avoid perverse actions or decisions both the assessor and assessee should be able to voice their opinion without sanctions. On the one hand assessors should be protected and supported when having to make unfavourable decisions, like failing an assessee, without negative consequences in terms of extra work or litigation. On the other hand, assessees should also be protected from unfair practice in assessment.

C12 Protocols should be in place to check (the programme in action) on proportionality of actions taken and carefulness of assessment activities.

Alongside the robustness of the procedures, processes should be in place to guarantee the appropriateness of the activities conducted in the assessment programme. Where the other guidelines focus on the design of the programme, these protocols are intended to check for appropriateness of the current activities (active after the design process).

C13* Protocols should be in place to assure assessment activities are equally accessible and fair for different (relevant) groups of stakeholders.

*This Guideline is added after the study reported in Chapter 4. No group or individual should be excluded from taking part in the assessment based on the factors other than qualities or requirements that are the same to all groups. The assessment should be fair for different groups of assessees.

DOCUMENTING THE PROGRAMME

Documenting the Programme serves two purposes. Firstly, documentation will facilitate learning of the organisation by allowing the cycle of optimising the programme in action to function properly. Secondly, it enhances the clarity and transparency of the programme. It is about explaining procedures in the programme and what is to be expected. In essence this should be public information.

RULES AND REGULATIONS (R&R)

D1 Rules and regulations should be documented.

Procedures on which decisions are made should be made explicit. Without having these stated (explicitly) decisions might become arbitrary, ambiguous, or *ad hoc*. In order to make decisions defensible, the procedures on which the decisions are based need to be documented. When formulating R&R input from people who are knowledgeable about the assessment programme is required. Legal expertise can contribute to the clarity of the R&R. All stakeholders are impacted by R&R to some extent, although it assesses and assessors are likely to be most affected by the procedures underlying decisions. Although it is expected that

R&R should be documented in virtually any programme of assessment, documentation can differ in degree and is likely to be dependent on the stakes involved.

D2 Rules and regulations should support the purposes of the programme of assessment.

Rules and regulations should not have unintended consequences on the outcomes of the assessment programme or the behaviour of stakeholders. Not only does *assessment drives learning*, assessment also procedures drive learning. For instance when offering many opportunities to pass a test (i.e. a lot of resit possibilities), it can become realistic for assessees to take a test without studying for it, because they might pass by chance. (This is also an opportunity to gain experience of test content before taking a resit.)

D3 The impact of rules and regulations should be checked against managerial, educational, and legal consequences.

The goal of this guideline is to avoid unwanted and unintended consequences of the rules and regulations e.g. overuse of resources of an institution. There should be congruence between R&R and requirements of management, educationalists, and legal staff. Therefore it is important to have expertise not only in drawing up rules and regulations, but also in awareness of the higher level implications in the organisation.

D4 In drawing up rules and regulations one should be pragmatic and concise, to keep them manageable and avoid complexity.

Rules and regulations should be formulated and made transparent, available, unambiguous, fair, and simple. It is important that R&R can be understood by all relevant stakeholders and proportional effort should be spent in making sure this is achieved. Involvement of stakeholders in the review of the R&R does contribute to achieving this guideline and has the benefit of fostering acceptance of the R&R at the same time.

D5 R&R should be based on routine practices and not on incidents or occasional problems.

The more rules and regulations, the more time is needed to maintain them. While trying to be comprehensive in all instances one can spend a disproportionate amount of effort in covering rare cases or cases that will never occur.

In practice often R&R slowly increase in scope and complexity in response to one off incidents. This will tend to decrease transparency and increase complexity. If the R&R are too specific or detailed the programme can become inflexible and incapable of dealing with unforeseen circumstances. Although incidents can be a trigger to review R&R, these rare instances should normally be covered by a general clause (e.g. 'In circumstances that these Rules and Regulations do not foresee, the certifying committee has the final say').

D6 There should be an organisational body in place to uphold the rules and regulations and take decisions in unforeseen circumstances.

The responsibility of upholding the R&R should be clearly defined and an appropriate mandate should be given to a separate body.

LEARNING ENVIRONMENT

Guideline D7 is about the context of the assessment programme and guideline D8 refers to the educational approach underlying an assessment programme.

D7 The environment or context in which the assessment programme has to function should be described.

The goal of this guideline is to contribute to the feasibility of implementing a programme of assessment with long term sustainability. Knowing the specifics of your own context and making them explicit supports the transfer of scientific research and best practice to one's own practice.

D8 The relation between educational system and assessment programme should be specified.

The goal of this guideline is to achieve a match between the assessment programme and the underlying assumptions regarding learning, instruction, and assessment. In order to have an assessment programme that can fulfil its purpose, it needs to be in line with the educational approach. In this sense it can be also applied to a non-educational organisation (e.g. a certifying body), as these organisations (consciously or unconsciously) also use educational paradigms. Knowing the educational approach can contribute to a more consistent and coherent programme of assessment.

DOMAIN MAPPING

The term *blueprinting* is deliberately not used here, because this term is often used to denote a specific tool using a matrix format to map the domain (content) to the programme and the instruments to be used in the programme. With *Domain Mapping*, a more generalised approach is implied. Not only should content match with components, but the focus should be on the assessment programme as a whole in relation to the overarching structure (e.g. the educational curriculum) and the purpose.

D9 A domain map should be the optimal representation of the domain in the programme of assessment.

First of all, a domain map relates to the overarching structure (guideline A3). This domain map entails a more detailed specification of the overarching structure, including assessment components and content elements. A domain map is closely tied to the sampling of content and the strategies used (*Collecting Information*), and to combining information on specific content from different sources. This is related to the fact that a single instrument is never sufficient to claim that a particular domain is completely and validly assessed. A variety of assessment components is required, and these have to be mapped onto the domain. As such, *Domain Mapping* is part of the validation process and, in accordance to guideline B2, content prevails over format. Aspects to consider in describing the domain are the content (knowledge, skills, attitude), and the level of authenticity (simulated versus real). The programme should sample purposefully through content and levels of authenticity.

D10 A domain map should not be too detailed.

This guideline is formulated in order to avoid the pitfall of atomizing a programme of assessment into the smallest possible units of analysis. This would harm the integrative nature of a programmatic approach towards assessment. As such, a domain map should not contain too much detail, or too many dimensions (axes). Too much detail would diminish the degrees of freedom in assessment and frustrate the process of designing as well as administering the assessment programme.

D11 Starting point for a domain map should be the domain or content and not the assessment component.

In congruence with guideline B2 the assessment component (e.g. type or format) is a tool not a goal. The domain or content that should be measured is part of the goal. Often an assessment component is available or familiar to the user or designer and therefore becomes the first choice when designing an assessment programme. In some cases there will be a sound match between content and instrument, however, this is not guaranteed. Starting from the purpose and the nature of the domain to be assessed will focus the design process on achieving the purpose.

D12 A domain map should be a dynamic tool, and as a result should be revised periodically.

There is a risk that the domain map quickly becomes outdated, as virtually every field develops at a rapid pace. Also priorities and ideas change over time. Therefore the domain map should be updated periodically. The frequency with which updating is needed, depends on the context of the assessment programme and the pace of developments in the domain.

IMPROVING THE PROGRAMME

Activities related to *Improving the Programme* generally have no immediate effect on the currently running programme, but impact in the (re)design of (parts of) the programme, usually at a later date.

R&D

Scientific research is dealt with in the next dimension (*Justifying the Programme*). *Research* in R&D is defined as the systematic collection of all necessary information to establish a careful evaluation (critical appraisal) of the programme with the intent of revealing areas of strengths and areas for improvement. *Development* should then be interpreted as re-design and therefore all other guidelines apply.

E1 A regular and recurrent process of evaluation and improvement should be in place, closing the feedback loop.

Not only should information be collected about the functioning of the programme, it should be acted upon as well (e.g. plan-do-check-act).

E2 If there is uncertainty about the evaluation, more information about the programme should be collected.

This is a specification of general guideline I. Actions based on evaluation of the programme (development) more often than not have large implications for the organisation. Therefore before changes are implemented based on the evaluation information, a high level of certainty about the information is required.

E3 In developing the programme (re-design) again improvements should be supported by scientific evidence or evidence of best practice.

This is a specification of general guideline II. In a sense all guidelines are applicable in the case of re-design.

CHANGE MANAGEMENT

Apart from measures to solve problems in a programme, political change or new scientific insights can also trigger improvement. *Change Management* refers to activities designed to cope with potential resistance to change. (Political) acceptance of changes refers to changes in (parts of) the programme.

E4 Momentum for change has to be seized or has to be created by providing the necessary priority or external pressure.

It is likely that many stakeholders do not perceive or experience the same imperative for change. The need for change has to be communicated or awareness should be raised to diminish possible resistance to change. The sense of urgency can be influenced by e.g. leadership, external pressure, and time, but also by making sure that stakeholders understand the reason(s) for change and how they can benefit from it. Often resistance to change stems from uncertainty and anxiety. Making the reasons for change explicit and communicating the benefits to stakeholders often decreases resistance. A change which at first sight is unpopular, may be accepted if it contributes sufficiently to the needs of the stakeholders.

E5 Underlying needs of stakeholders should be made explicit.

Because the *wants* of different stakeholders may seem to compete at first sight, the underlying needs of stakeholders should be made explicit. Needs may not be clear to the stakeholders themselves as they are often less noticeable than wants. Wants can concern a specific lay-out of a form, whereas the need might be that the form can be filled out quickly.

E6 Sufficient expertise about change management and about the local context should be sought.

Similar to *Construction Support*, there is a need to translate general concepts to the local situation. In assessment practices such as item writing the same issue occurs. One cannot write a test item without content expertise, nor without expertise on writing an item. Both expertises have to be combined in order to develop good items.

E7 Faculty should be supported to cope with the change by providing adequate training

Uncertainty and low efficacy lead to resistance. Clear explanations and introductions as well as training or faculty development, can assist in overcoming resistance.

JUSTIFYING THE PROGRAMME

Justifying the Programme relates to the increasing demand for public accountability. The aim of activities in this dimension is to defend the current practices of the programme in action and demonstrate that purposes are met. *Justifying the Programme* deals with the rationale behind it based on the leading purpose. Three elements can be distinguished in justifying the assessment programme. First the *Effectiveness of the Programme* deals with the question of whether the purposes of the programme can be achieved, by providing evidence of due practices. The second element, *Efficiency*, is concerned with the realities of limited resources and providing evidence of cost-effectiveness. The third element, *Acceptability*, relates to the dimension of *Stakeholders*. The focus of this element is on the broader framework of legislation and external stakeholder groups.

EFFECTIVENESS

Scientific Research

Scientific research on assessment components and activities is needed to support practices with sound evidence, which is in line with the prominence in medicine of the drive for evidence-based practice. Although this is a general principle which should guide the design of the programme as a whole, it comes into effect when one has to account for choices made in the programme.

F1 Before the programme of assessment is designed, evidence should to be reviewed.
This is a specification of general guideline II. This way the design can be informed by and based on scientific evidence and/or best practices. The relevant literature has to be reviewed in order to make state of the art decisions regarding the design of a programme of assessment or even elements of the programme.

F2 New initiatives (developments) should be accompanied by evaluation, preferably scientific research.
Scientific research which supports the activities in the programme of assessment is but one form of justifying the effectiveness. The domains or areas of research may be diverse as education and assessment are based on various scientific domains in humanities, social science, psychometrics and cognitive (neuro)sciences.

External Review

Justification also requires *external review* of programmes of assessment. Assessment programmes are also shaped by the needs and wishes of external stakeholders.

F3 The programme of assessment should be reviewed periodically by a panel of experts.
Stakeholders within the organisation have two disadvantages when justifying the programme. First, they are not independent and may have conflicts of interest. Second, they may have developed blind spots over time.

F4 Benchmarking against similar assessment programmes (or institutes with similar purposes) should be conducted to judge the quality of the programme.
It is possible to determine to what degree the purposes of the programme are met. However, without comparison between programmes, it is impossible to judge whether this is the best possible programme given the circumstances.

EFFICIENCY: cost-effectiveness

In every institution or organisation, resources - including those for assessment programmes - are limited. *Cost-effectiveness* is regarded as a desirable endeavour.

F5 In order to be able to justify the resources used for the assessment programme, all costs (in terms of resources) should be made explicit.

Decisions are made based on cost-effectiveness. In order to be able to make these decisions the required resources must be made explicit. Resources can be: time, money, materials, expertise, etc.

F6 A cost-benefit analysis should be made regularly in light of the purposes of the programme. In the long term, a proactive approach to search for more resource-efficient alternatives should be adopted.

If the programme of assessment can be made more efficient, resources can be freed up for other activities.

ACCEPTABILITY: political-legal justification

As an assessment programmes does not exist in a vacuum, political and legal requirements often determine how (part of) the programme of assessment has to be (re)designed and justified.

F7 Open and transparent governance of the assessment programme should be in place and can be held accountable

With every design step one has to ask oneself: 'Can I defend this, if it ends up in the media?'; 'Can I explain and rationalise the actions taken?'

F8 In order to establish a defensible programme of assessment there should be one vision (on assessment) communicated to external parties.

Choices need to supported or at least accepted by all internal stakeholders. If it is not supported inside the organisation, it is hard to sell to the outside world.

F9 The assessment programme should take into account superseding legal frameworks.

When designing a programme of assessment it is important to know which laws apply e.g. university regulations, national law, international law. It might even be necessary to involve a legal department in the design.

F10 Confidentiality and security of information should be guaranteed at an appropriate level.

An issue, strongly related to guideline C12 about proportionality of actions based on combined data, is the use of information by third parties. Information should be stored with appropriate security measures and procedures should be in place to protect the information from being used inappropriately. Here, the proportionality principle should be heeded again. The more personal and sensitive the information, the more extensive the safety measures should be e.g. disclaimers about acquiring consent before the use of the combined data for purposes that are not specified is required at the outset.

References

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

Frank, J. (Ed.). (2005). *The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care.* Ottawa: The Royal College of Physicians and Surgeons of Canada.

General Medical Council. (2006, 13 november 2006). *Good Medical Practice.* Retrieved 14 April 2011 from http://www.gmc-uk.org/guidance/good_medical_practice/GMC_GMP.pdf

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63-67.

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

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22*, 270 - 286.

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., Verheggen, M. M., Van der Vleuten, C. P. M., Boshuizen, H. P. A., & Dinant, G. J. (2000). Validation of short case-based testing using a cognitive psychological methodology. *Medical Education, 35*, 348-356.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment Of Clinical Competence: Written And Computer-Based Simulations. *Assessment & Evaluation in Higher Education, 12*(3), 220 - 246.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6*(1), 1-11.

Summary

For long, research on assessment in medical education has been mainly focussed on single assessment instruments. The aim of most of these studies was to achieve the best possible instrument to measure separate elements of student abilities. This research led to a toolbox of instruments and valuable insights into the strengths and weaknesses of these single assessment instruments (Van der Vleuten et al., 2010).

With this assessment approach, decisions about student achievement are typically based on the collection and combination of the separate outcomes of each of the examinations without taking into account if and how these building blocks represent a complete and integrated picture of professional competence. Simply adding up or lumping together individual and independent exams does not comprehensively capture competence. Competence is to be regarded as a whole task, which cannot be broken down into separate parts. Competence does not consist of one-dimensional traits, but is a complex integrated construct (Schuwirth and Van der Vleuten, 2006).

It is only logical to conclude that no single instrument will ever be able to provide all the information for a comprehensive evaluation of competence in a domain as broad as medicine. Furthermore, while acknowledging the importance of psychometrics, it is clear that exclusively focussing on psychometrics is an insufficient basis for selecting assessment instruments. Not only should reliability and validity be taken into account, but educational impact, acceptation, and costs need to be considered too (Newble et al., 1994; Schuwirth and Van der Vleuten 2004; Van der Vleuten, 1996).

More important, assessment in medical education entails more than just determining competence (assessment of learning). Multiple and divergent goals also need to be addressed by assessment, such as facilitating or influencing development (assessment for learning) as well as evaluating instruction (quality improvement). Any single instrument only has a certain value and therefore cannot completely meet all or even one assessment purpose(s). Thus, assessment in medical education requires a carefully designed assessment programme, consisting of a purposeful mix of various assessment components that correspond with the goals of assessment (and/or the curriculum at large) in the best possible way.

A programmatic approach to assessment design is advocated in order to help assessment developers in dealing with the complexity of the design process and combining multiple assessment purposes (Lew et al., 2002; Schuwirth et al., 2002; Van der Vleuten and Schuwirth, 2005). Assessment design must take into account more than just the strengths and weaknesses of separate assessment components. It must also include the interrelatedness of these components and the implementations of assessment in practice. Inevitably, such an approach does not only consider assessment as a measurement problem, but also as an educational design problem in which trade-off decisions have to be made.

Designing assessment programmes in medical education settings is a complex process influenced by a broad range of factors that have to be taken into account in order to optimally achieve assessment purposes (Dijkstra et al., 2012) Serving multiple purposes makes assessment design complex and challenging. Assessment programmes that are perceived to be of high quality in one particular context may not be suitable in other contexts. We need, therefore, guidelines that not only provide a framework for design of an integrated assessment of professional competence, but are also applicable (or easily adaptable) in a broad range of settings.

In **Chapter 1** the scarcity of literature addressing criteria and guidance for assessment design is reviewed and highlighted. This leads to the overall aim of this research to provide generic support for achieving high-quality assessment programmes.

We take a utilitarian approach, whereby quality is defined as fitness-for-purpose (Harvey and Green, 1993). The advantage of this perspective is that it makes the quality framework more broadly applicable and less reliant solely on current ideas on education and assessment. From a fitness-for-purpose view, weaknesses of assessment components can be perfectly acceptable if the strengths contribute optimally or sufficiently to the purpose of assessment.

In Chapter 1 the research questions are described, which were leading for the studies in this dissertation.

- What areas or elements can be distinguished in the design of high-quality assessment programmes?
- What guidelines can be formulated for design support based on the areas of assessment design?
- What evidence can be provided to substantiate the validity of guidelines based on utilitarian principles in practice?

The validation process of the support for assessment design is similar to the development of theories or frameworks and evaluation of clinical guidelines. Therefore Basinski's (1995) work on evaluation of guidelines and criteria for theory building described by Prochaska et al., 2008) are used to validate guidelines for assessment design.

Chapter 2 describes the development of our framework for assessment programmes and specification of areas and elements that have to be covered, when formulating design guidelines. Because of the absence of a common vocabulary for programmatic assessment, we used focussed group interviews as an exploratory, qualitative method to probe the experiences, views and ideas of nine experts in assessment in medical education, concerning good practices and new ideas about theoretical and practical issues in assessment programmes. 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. This resulted in an overarching framework for programmatic assessment, which defines the scope of what constitutes an assessment programme, and should be covered by our guidelines. The overarching framework for designing programmes of assessment consists of six assessment programme dimensions: Purpose of the programme, Programme in Action, Support, Documenting, Improving and Justification (previously named Accounting). Embedded in a seventh stakeholder-infrastructure dimensions describing the context. The framework provides a shared construct of how to define assessment programmes, but also a comprehensive picture of the dimensions to be covered when formulating guidelines for assessment design. It helps identifying areas concerning assessment in which ample research and development has been done. But, more important, it also helps to detect underserved areas. One of the main conclusions in this study was the importance of the assessment purpose. A guiding principle in design of assessment programmes is fitness-for-purpose. High quality assessment can only be determined in terms of achieving the purpose(s).

Chapter 3 describes how a set of **g**uide**l**ines for **a**ssessment **d**esign (GLAD) was derived from this framework. A fitness-for-purpose approach defining quality was adopted to develop and validate guidelines, since the aim of this study was to formulate guidelines that are general enough to be applicable in a variety of contexts, and yet at the same time meaningful and concrete enough to support assessment designers. We started with a brainstorm, to generate ideas for guidelines based on our framework for programmes of assessment using the input of nine international experts in the field of assessment in medical education. This was followed by structured interviews and afterwards fine-tuning of the guidelines through analysing the interviews. Finally, validation was based on expert consensus via member checking. In this first phase of gathering validity evidence *during* the development of guidelines, the expert consensus procedure focussed on achieving *clarity*, *consistency*, and *parsimony* (Prochaska et al., 2008) of the guidelines. More specifically, attention was given to creating explicit terminology and defining the guidelines carefully. The guidelines were grouped logically to avoid any contradiction with each other. Some guidelines were found to be clear and concrete, others were less straightforward and were phrased more as issues for contemplation. Finally, complexity as well as redundancy of the guidelines was minimized. This led to a comprehensive set of guidelines (See the Addendum for a complete overview and description). In total 72 guidelines were developed and in Chapter 3 the most salient guidelines are discussed. The guidelines are related and grouped per dimension of the framework. Some guidelines were so generic that these are applicable in any design consideration. These are: the principle of proportionality, rationales should underpin each decisions, and requirement of expertise.

Chapters 4 and 5 describe the next steps in the validation process. In **Chapter 4** the evaluation of GLAD was done in a real life setting. An instrumental case study and a multiple qualitative inquiry two-step approach were used to evaluate the *practicality* and *explanatory power* of GLAD (Prochaska et al., 2008). The practicality of GLAD was investigated through document analysis and interviews with multiple stakeholders in the assessment process. More specifically, we used a deductive content analysis on documents and semi-structured interviews to investigate if GLAD could be found in actual practice and if they were taken into account during the process of design. Results yielded in-depth information about decisions and considerations made during the design process. We distinguished 4 levels of use: Well-addressed, Partly-addressed, Not addressed, Missing GLAD. In Chapter 4 the practicality of specific GLAD is described and discussed. Overall, the GLAD are comprehensive and logically applicable in practice and thus meet the practicality criterion. One design-element could not be coded with GLAD and led an additional GLAD. Based on the results from the practicality evaluation, the explanatory power of GLAD was investigated in Step 2. The *explanatory power* was evaluated, by the ability of GLAD to describe and evaluate statements of perceived strengths and issues that were identified through analysis of interviews with relevant stakeholders. In total 6 major strengths and major issues were derived from the interviews. All could be explained by GLAD (and its Practicality), how the GLAD were used to describe the strengths and issues is described in Chapter 4. The GLAD offer a vocabulary to organisations and stakeholders to *describe and explain* the quality assessment programmes and thus the GLAD meet the explanatory-power criterion.

The second case study as described in **Chapter 5** aims to investigate the *utility* and *productivity* of GLAD (Prochaska et al., 2008). The utility of GLAD in the evaluation of assessment programmes was investigated by comparing evaluation outcomes and processes to a well-researched and validated set of quality criteria for Competence Assessment Programmes (CAP). The *productivity* of GLAD is determined by investigating whether GLAD contributes to existing research. More specifically, the productivity in this study looks at whether GLAD adds to the established and validated CAP criteria. A competence based assessment programme was purposefully selected and was evaluated based on interviews, document analysis, and a self-assessment tool. Firstly, we evaluated the programme using GLAD by conducting interviews and document analysis. Secondly, the programme was evaluated by the CAP criteria using a self-evaluation tool followed by a group interview (see: Baartman et al., 2007). Both evaluations are an interpretation of an in-depth qualitative analysis of the assessment programme. Outcomes of both quality evaluations are analysed to determine whether the GLAD meet the criteria of *utility* (useful and meaningful outcomes) and *productivity* (build on research) compared to the validated CAP. Generally both evaluations covered similar issues in assessment. Differences in the outcome of the evaluations are discussed, as levels of detail and starting points differ. Application of the GLAD resulted in useful recommendations, which are corroborated by the outcome of the validated CAP. We therefore concluded that GLAD meet the *utility* criteria. The GLAD also meet the *productivity* criterion because it extends the CAP criteria with new areas for evaluation of programmes of assessment within the competence-based assessment context.

Finally, in **Chapter 6**, the main findings are summarized and discussed. Further reflection on the evidence for the framework and guidelines is provided. Limitations of this research are discussed and suggestions are presented for future development and evaluation of support for designing programmes of assessment.

Finally, possible implications for practice are explored. The general aim of this dissertation has been to develop guidance for design decisions with respect to programmatic assessment and to support assessment developers in achieving high-quality assessment programmes. The first phase of this research was aimed at developing comprehensive and generic guidance (Chapters 2 and 3). The second phase of the research was aimed at evaluating this guidance (Chapters 4 and 5). The studies in this disseratation defined a framework for assessment programmes, from which the GLAD was developed, validated and evaluated. The two main findings were: 1) a comprehensive framework for assessment programmes and 2) the 73 guidelines for assessment design (GLAD). The criteria of Prochaska et al. (2008) are addressed as well as the downsides of the comprehensiveness and abstract level of the GLAD.

Although the studies in this research are inclusive, rather than exclusive, still there is a margin of uncertainty about the completeness of the GLAD. The studies all focussed on verification, rather than falsification of the GLAD. Fortunately, all GLAD were supported by evidence in practice (See Chapter 4). The criteria of Prochaska et al. (2008) were found to be a sound basis to validate the GLAD. However, not all criteria could be explicitly evaluated. This is beyond the scope of this dissertation and will be the domain of future studies. We therefore feel that future research should first be directed at transferability of GLAD, by studies into the necessary scaffolding as practical guidance to an expert using GLAD, and exploring the possibilities of providing more concrete support using a specific educational setting. Application of the GLAD in a variety of contexts would provide further information about the comprehensiveness of the framework and the GLAD, as well as its relevance in general.

The framework and the GLAD developed and evaluated in the studies in this dissertation provide a new perspective on determining quality of assessment programmes. It provides a new theory to look at assessment programmes and a vocabulary that enables assessment experts to describe their holistic judgement of what a sound assessment programme constitutes. The programmatic approach to assessment and the ideas that are brought forward can also be translated to other areas, in which assessment of some sort is involved, for instance selection into medical education and accreditation of schools. The application for accreditation also illustrates the inherent dimension in the framework of assessing the assessment. The GLAD are developed for assessment design, but are useful as an evaluation framework as well.

The studies in this dissertation provide evidence to substantiate the application of this guidance formulated from a utilitarian approach. At the same time we found that defining and determining quality is not a question of meeting criteria, but a question of providing experts with a vocabulary for conveying to others the description, evaluation, and explanation of the quality of an education programme.

References

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Determining the Quality of Competence Assessment Programs: A Self-Evaluation Procedure. *Studies in Educational Evaluation, 33*(3-4), 258-281.

Basinski, A. S. (1995). Evaluation of clinical practice guidelines. *Canadian Medical Association Journal, 153*(11), 1575-1581.

Dijkstra, J., Galbraith, R., Hodges, B., McAvoy, P., McCrorie, P., Southgate, L. et al. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*, 20.

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

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.

Newble, D., Dawson, B., Dauphinee, D., Page, G., Macdonald, M., Swanson, D. et al. (1994). Guidelines for assessing clinical competence. *Teaching and Learning in Medicine: An International Journal, 6*(3), 213 - 220.

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model. *Applied Psychology: An International Review, 57*(4), 561-588.

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 psychometric models in educational assessment. *Medical Education, 40*, 296-300.

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

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 practice & research. Clinical obstetrics & gynaecology, 24*(6), 703-719.

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.

Samenvatting

Dutch Summary

Lange tijd is onderzoek naar toetsing in medisch onderwijs voornamelijk gericht geweest op afzonderlijke toetsinstrumenten. Het doel van veel van deze studies was om het best mogelijke instrument te ontwikkelen voor het meten van afzonderlijke onderdelen van bekwaamheid van studenten. Dit onderzoek heeft geleid tot de ontwikkeling van een toolbox van instrumenten en waardevolle inzichten in de sterktes en zwaktes van deze afzonderlijke toetsinstrumenten (Van der Vleuten et al., 2010).

Dit perspectief op toetsing leidt tot beslissingen over de bekwaamheid van studenten, dat typisch gebaseerd is op de verzameling en combinatie van verschillende resultaten van elke toets, zonder rekening te houden met de vraag of en hoe deze toetsen bijdragen aan een volledig en geïntegreerd beeld van professionele competentie. Het simpelweg optellen of combineren van afzonderlijke en onafhankelijke examens vat competentie niet in zijn geheel. Competentie moet gezien worden als een 'complete taak', die niet kan worden ontleed in afzonderlijke delen. Competentie bestaat niet uit eendimensionale kenmerken, maar is een complex en geïntegreerd construct (Schuwirth & Van der Vleuten, 2006).

Het is logisch om te concluderen dat geen enkel toetsinstrument ooit in staat zal zijn om alle informatie te verschaffen voor een allesomvattende evaluatie van competentie in een domein zo breed als geneeskunde. Het wordt daarnaast duidelijk, dat ondanks de erkenning van het belang van de psychometrie, een exclusieve focus hierop onvoldoende basis biedt voor de selectie van een toetsinstrument. Niet alleen de betrouwbaarheid en de validiteit van een toets moeten worden meegenomen in de keuze, ook impact op het onderwijs, acceptatie en kosten moeten worden afgewogen (Newble et al., 1994; Schuwirth & Van der Vleuten 2004; Van der Vleuten, 1996).

Misschien nog belangrijker is het feit dat toetsing in medisch onderwijs meer inhoud dan alleen het bepalen of iemand competent is (toetsen van leren). Meerdere en divergente doelen moeten worden bereikt met toetsing, zoals het faciliteren en beïnvloeden van ontwikkeling (toetsing voor leren), maar ook de evaluatie van instructie en onderwijs (kwaliteitsverbetering). Elk afzonderlijk toetsinstrument heeft een bepaalde specifieke waarde en kan niet aan een enkel doel voldoen, laat staan aan meerdere doelen tegelijk. Daarom is er een sterke behoefte in medisch onderwijs voor een met zorg ontwikkeld programma van toetsing. Deze dient te bestaan uit een doelmatige mix van verscheidene toetscomponenten die bijdrage aan de doelen van toetsing (en de doelen van het curriculum als geheel) op een zo effectief mogelijke manier.

Een programmatische aanpak van toetsing wordt geadviseerd om toets-ontwikkelaars te ondersteunen in het omgaan met de complexiteit van het ontwerpproces en het combineren van meerdere toetsdoelen (Lew et al., 2002; Schuwirth et al., 2002; Van der Vleuten & Schuwirth, 2005). Toetsontwerp moet rekening houden met meer dan alleen de sterktes en zwaktes van afzonderlijke toetscomponenten. Ook de relaties tussen deze componenten en de implementatie van toetsing in de praktijk moet daarbij worden betrokken. Het is onvermijdelijk dat een dergelijke benadering van toetsing niet alleen als een meetprobleem gedefinieerd kan worden, maar ook als een onderwijskundig ontwerp probleem, waarbij keuzes en compromissen gemaakt moeten worden op basis van voor- en nadelen.

Het ontwikkelen van toetsprogramma's in een medisch onderwijskundige setting is een complex proces dat beïnvloed wordt door een breed spectrum aan factoren, waarmee rekening gehouden moet worden om optimaal aan de doelen van toetsing te kunnen voldoen (Dijkstra et al., 2012). Het moeten voldoen aan meerdere doelen van toetsing tegelijkertijd. Dit maakt toetsontwerp complex en uitdagend. Toetsprogramma's die worden beschouwd als zijnde van hoge kwaliteit in de een specifieke setting, kunnen ongeschikt zijn in een andere setting. Er is daarom behoefte aan richtlijnen die een raamwerk bieden voor het ontwerp van geïntegreerde toetsing van professionele competentie en tegelijkertijd toepasbaar zijn in (of eenvoudig aan te passen aan) een breed spectrum van contexten.

In **Hoofdstuk 1** wordt de schaarsheid van literatuur over criteria en ondersteuning voor toetsontwerp aangehaald en belicht. Dit leidt tot het algemene doel van dit onderzoek om te komen tot generieke ondersteuning voor het ontwikkelen van hoogkwalitatieve toetsprogramma's.

We kiezen hiervoor een utilistische benadering, waarbij kwaliteit wordt gedefinieerd als fitness-for-purpose (Harvey & Green, 1993) – geschiktheid om het doel te bereiken. Het voordeel van deze benadering is een bredere toepasbaarheid van het kwaliteitsraamwerk en de verminderde afhankelijkheid van de huidige ideeën en trends over onderwijs en toetsing. Vanuit een fitness-for-purpose perspectief is een zwakte van een specifiek toetsonderdeel acceptabel zolang de sterktes ervan optimaal of voldoende bijdragen aan het doel van toetsing.

In Hoofdstuk 1 worden de onderzoeksvragen beschreven, welke leidend zijn geweest voor de studies in dit proefschrift.

- Welke gebieden of elementen kunnen worden onderscheiden in het ontwerp van hoogkwalitatieve toetsprogramma's?
- Welke richtlijnen kunnen geformuleerd worden ter ondersteuning van toetsontwerp, op basis van deze gebieden en elementen?
- Welk bewijs kan worden geleverd om de validiteit te onderbouwen van de richtlijnen gebaseerd op utilitaristische principes in de praktijk?

Het valideringsproces voor de ondersteuning van toetsontwerp is vergelijkbaar met de ontwikkeling van theorie en de evaluatie van klinische richtlijnen. Daarom is Basinski's (1995) werk over evaluatie van richtlijnen en de criteria voor theorie-ontwikkeling beschreven door Prochaska et al., 2008) gebruikt om de *guidelines for assessment design* – richtlijnen voor toetsontwerp – te valideren.

Hoofdstuk 2 beschrijft de ontwikkeling van ons raamwerk (model) voor toetsprogramma's en de specificatie van de verschillende gebieden en elementen die meegenomen dienen te worden tijdens de formulering van ontwerprichtlijnen. Vanwege het gebrek aan een gemeenschappelijk vocabulaire voor programma's van toetsing, hebben we *focussed* groepsinterviews gehouden. Op een exploratieve, kwalitatieve manier zijn negen experts op het gebied van toetsing in medisch onderwijs bevraagd naar hun ervaringen, perspectieven en ideeën over *good practices* en nieuwe ideeën theoretische en praktische aandachtspunten in toetsprogramma's. De discussie is geanalyseerd door alle relevante aspecten in een raamwerk te plaatsen, dat door een iteratief proces is aangepast en bijgesteld om de data zo goed mogelijk te beschrijven, totdat saturatie was bereikt en geen aanpassingen meer nodig waren. Dit resulteerde in een overkoepelend raamwerk voor programma's van toetsing, welke de omvang definieert van waaruit een toetsprogramma bestaat en welke gedekt moeten worden door de te ontwikkelen richtlijnen. Het overkoepelende raamwerk voor toetsprogramma' s bestaat uit zes dimensies: (1) Doel van het programma, (2) Programma in actie, (3) Ondersteuning, (4) Documentatie, (5) Verbetering en (6) Verantwoording Ingebed in een zevende stakeholder-infrastructuur dimensie, die de context beschrijft. Het raamwerk biedt een gedeeld construct van hoe een toetsprogramma gedefinieerd kan worden, maar ook een veelomvattend beeld van de dimensies die gedekt moeten worden door de te ontwikkelen richtlijnen. Het raamwerk helpt bij het identificeren van toetsgebieden waarin weinig onderzoek en ontwikkeling heeft plaatsgevonden. Maar bovendien helpt het bij het detecteren van gebieden die te weinig aandacht hebben gekregen. Eén van de hoofdconclusies in deze studie is het belang van het doel van toetsing. Een richtinggevend principe in het ontwerp van toetsprogramma's is fitness-for-purpose. Hoogkwalitatieve toetsprogramma's kunnen alleen worden geduid in termen van het bereiken van de doelen.

Hoofdstuk 3 beschrijft hoe een set van richtlijnen voor het ontwerpen van toetsing - **g**uide**l**ines for **a**ssessment **d**esign (GLAD) – is ontwikkeld aan de hand van dit raamwerk. Het doel van de studie was om richtlijnen te formuleren die generiek genoeg zijn om toegepast te kunnen worden in verscheidene contexten, maar tegelijkertijd betekenisvol en concreet genoeg zijn om toets-ontwikkelaars te ondersteunen. Daarom is een fitness-for-purpose benadering gekozen voor de definitie van kwaliteit voor het ontwikkelen en valideren van richtlijnen. We startte met een brainstorm om ideeën te genereren voor richtlijnen gebaseerd op ons raamwerk voor programma's van toetsen en gebruik makend van de input van negen internationale experts in het veld van toetsing in medisch onderwijs. Vervolgens zijn er gestructureerde interviews gehouden waarna *fine-tuning* van de richtlijnen heeft plaatsgevonden op basis van de analyse van deze interviews. Tot slot is de validatie gebaseerd op expert consensus via een *member check* procedure. In de eerste fase van het verzamelen van bewijs voor de validiteit *tijdens* de ontwikkeling van de richtlijnen lag de focus op het bereiken van *duidelijkheid*, *consistentie* en *spaarzaamheid* van de richtlijnen (Prochaska et al., 2008). In het bijzonder is aandacht geschonken aan het expliciteren van terminologie en de zorgvuldige formulering van de richtlijnen. De richtlijnen zijn logisch gegroepeerd om tegenstrijdigheden te voorkomen. Een aantal richtlijnen zijn rechtlijnig en concreet, waar andere minder vanzelfsprekend waren en meer geformuleerd werden als een onderwerp waaraan aandacht besteed moet worden. Uiteindelijk is de complexiteit en de overlap tussen richtlijnen geminimaliseerd. Dit leidde tot een veelomvattende lijst van richtlijnen (zie het addendum voor een compleet overzicht). In totaal 72 richtlijnen werden ontwikkeld en in hoofdstuk 3 worden de meest opvallende richtlijnen bediscussieerd. De richtlijnen zijn gerelateerd aan elkaar en gegroepeerd per dimensie van het raamwerk. Enkele richtlijnen waren dusdanig generiek dat deze van toepassing zijn op elke ontwerpbeslissing. Dit zijn: het principe van proportionaliteit, onderbouwing van beslissingen, en de noodzaak van expertise.

De hoofdstukken 4 en 5 beschrijven de volgende stappen in het validatie proces. In **Hoofdstuk 4** de evaluatie van de GLAD vond plaats in de daadwerkelijke praktijk. Een instrumentele case study en een meervoudige kwalitatieve onderzoek aanpak is gebruikt in een tweetraps aanpak om de *practicality* – gebruik in praktijk - en *explanatory power* – verklarende kracht - van de GLAD (Prochaska et al., 2008) vast te stellen. De practicality van de GLAD is bepaald op basis van document analyse en interviews met meerdere stakeholders betrokken bij het toets proces. Meer specifiek hebben we een deductieve inhoudsanalyse gebruikt om de documenten en de semi-gestructureerde interviews te analyseren en vast te stellen of de GLAD terug te vinden zijn in de daadwerkelijke praktijk en of deze in overweging genomen zijn tijdens het ontwerp proces. De resultaten leverden gedetailleerde informatie over de genomen beslissingen en overwegingen tijdens het ontwerp proces. We onderscheidde 4 niveaus van gebruik: goed overwogen, deels overwogen, niet overwogen, ontbrekende GLAD. In hoofdstuk 4 de practicality van specifieke GLAD is geschreven en bediscussieerd. In het algemeen zijn de GLAD veelomvattend en logisch toepasbaar in de praktijk, waarmee aan het practicality-criterium wordt voldaan. Eén onderdeel in het ontwerp kon niet worden gecodeerd met behulp van de GLAD en leidde tot een additionele richtlijn. Op basis van de resultaten van de practicality-evaluatie is de explanatory power van de GLAD is onderzocht in stap 2. De explanatory power is geëvalueerd aan de hand van de mogelijkheid om met de GLAD uitspraken over de gepercipieerde sterktes en aandachtspunten van het toetsprogramma te beschrijven en te evalueren. Deze uitspraken zijn verkregen door een analyse van interviews met relevante interne en externe belanghebbenden. In totaal 6 sterktes en 6 aandachtspunten zijn gedestilleerd uit de interviews. Alle uitspraken konden worden verklaard met behulp van de GLAD (en de practicality ervan). Hoe de GLAD gebruikt zijn om de sterktes en aandachtpunten te beschrijven, is te vinden in hoofdstuk 4. De GLAD bieden een vocabulaire aan organisaties en stakeholders om de kwaliteit van toetsprogramma's te *beschrijven en te verklaren*. Daarmee voldoen de GLAD aan het explanatory-power criterium.

De tweede case study beschreven in **Hoofdstuk 5** was gericht op het onderzoeken van de *utility* – bruikbaarheid – en *productivity* – productiviteit – van de GLAD (Prochaska et al., 2008). De utility van de GLAD bij de evaluatie van toetsprogramma's werd geëvalueerd door de uitkomsten van de evaluatie te vergelijken met de uitkomsten van een grondig onderzochte en gevalideerde set van kwaliteitscriteria voor Competence Assessment Programmes (CAP) – Competentie Toets Programma's. De *productivity* van de GLAD is bepaald door te onderzoeken in hoeverre de GLAD bijdragen aan bestaand onderzoek. Meer specifiek, de *productivity* in deze studie is beoordeeld in het licht van de toevoeging aan de bestaande en gevalideerde CAP criteria. Een competentie gebaseerd toetsprogramma was doelgericht geselecteerd en geëvalueerd op basis van interviews, document analyse en met behulp van een zelfbeoordelingsinstrument. Ten eerste hebben we het programma geëvalueerd met behulp van de GLAD door interviews en documentanalyse. Ten tweede is het programma geëvalueerd met behulp van de CAP criteria door gebruik te maken van een zelfbeoordelingsinstrument gevolgd door een groepsinterview (zie: Baartman et al., 2007). Beide evaluaties zijn een interpretatie van een kwalitatieve diepte-analyse van het assessment programma. De uitkomsten van beide kwalitatieve evaluaties zijn geanalyseerd om te bepalen of de GLAD voldoen aan het *utility* criterium (bruikbare en betekenisvolle uitkomsten) en het *productivity* criterium (voortbouwen op onderzoek) in vergelijking met de gevalideerde CAP criteria. In het algemeen dekken de beide evaluaties dezelfde onderwerpen in toetsing. Verschillen in de uitkomsten van de evaluatie worden bediscussieerd, want het niveau van detaillering en de uitgangspunten verschillen. De toepassing van de GLAD resulteerde in bruikbare aanbevelingen welke ondersteund worden door de uitkomsten van de gevalideerde CAP. Op basis daarvan concluderen we dat de GLAD voldoen aan het *utility* criterium. De GLAD voldoen ook aan het *productivity* criterium. Omdat het verder gaat dan de CAP criteria en aandachtsgebieden van toetsprogramma's in competentie gebaseerde toetsing context toevoegt in de evaluatie.

Tot slot, in **Hoofdstuk 6**, worden de belangrijkste bevindingen samengevat en bediscussieerd. Verder wordt er een reflectie gegeven op het bewijs dat gegeven is voor het raamwerk en de richtlijnen. Beperkingen van dit onderzoek worden bediscussieerd en suggesties voor verdere ontwikkeling en evaluatie van ondersteuning voor het ontwerpen van programma's van toetsen worden gepresenteerd. Mogelijke implicaties voor de praktijk worden verkend. Het algemene doel van deze dissertatie was het ontwikkelen van ondersteuning bij ontwerpbeslissingen met betrekking tot programmatisch toetsen en het ondersteunen van toetsontwikkelaars om hoogkwalitatieve programma's van toetsing te bereiken. De eerste fase van dit onderzoek was gericht om het ontwikkelen van veelomvattende en generieke richtlijnen (hoofdstuk 2 en 3). De tweede fase van het onderzoek was gericht op het evalueren van deze ondersteuning (hoofdstuk 4 en 5). De studies in deze dissertatie definiëren een raamwerk, waaruit de GLAD ontwikkeld zijn en vervolgens geëvalueerd en gevalideerd zijn. De twee belangrijkste resultaten zijn: 1) een veelomvattend raamwerk voor toetsprogramma's en 2) de 73 richtlijnen voor toetsontwerp (GLAD). De criteria van Prochaska et al. (2008) zijn daarin meegenomen, maar ook de nadelen van veelomvattende en abstract niveau van de GLAD.

Ondanks dat de studies in dit onderzoeksproject inclusief van aard zijn, (in plaats van excluderend met betrekking tot richtlijnen) is er toch een bepaalde onzekerheid ten aanzien van de volledigheid van de GLAD. De studies waren allemaal gericht op verificatie en minder op falsificatie van de GLAD. Gelukkig werden alle GLAD ondersteund door bewijs uit de praktijk (zie hoofdstuk 4). De criteria van Prochaska et al. (2008) zijn een solide basis om de GLAD te valideren. Echter, niet alle criteria konden expliciet geëvalueerd worden. Deze criteria gaan verder dan het bereik van de studies in deze dissertatie en zal in toekomstige studies aan bod moeten komen. Daarom zal naar onze mening verder onderzoek zich op de eerste plaats dienen te richten op het gebruik van GLAD in andere settings/ contexten, met behulp van studies naar de benodigde *scaffolding* als praktische ondersteuning van een expert die de GLAD gebruikt. De mogelijkheden moeten onderzocht worden om meer concrete handvaten te kunnen geven in specifieke onderwijskundige settings. Toepassing van de GLAD in verscheidene contexten zou verdere informatie over de compleetheid van het raamwerk en de GLAD kunnen opleveren, alsmede de relevantie in het algemeen.

Het raamwerk en de GLAD die in de studies in deze dissertatie zijn ontwikkeld, bieden een nieuw perspectief op het bepalen van kwaliteit van toetsprogramma's. Het biedt een nieuwe manier om naar toetsprogramma's te kijken en een vocabulaire dat experts in staat stelt om een holistisch oordeel over deugdelijke toetsprogramma's te verwoorden. De programmatische aanpak en de ideeën die naar voren zijn gebracht kunnen ook vertaald worden naar andere gebieden waarin toetsing een rol speelt, bijvoorbeeld selectie voor toelating tot (medisch) onderwijs of accreditatie van opleidingen. De toepassing bij accreditatie illustreert ook de inherente dimensie in het raamwerk van de toetsing getoetst. De GLAD zijn ontwikkeld voor toetsontwerp, maar zeker bruikbaar en nuttig als evaluatie-raamwerk.

De studies in deze dissertatie bieden het bewijs om de toepassing van deze ondersteuning geformuleerd vanuit een utilitaristisch perspectief te staven. Tegelijkertijd is het definiëren van kwaliteit niet een kwestie van het behalen van criteria, maar een kwestie van experts een vocabulaire aanreiken om hun beschrijving, evaluatie en verklaring over de kwaliteit van een toetsprogramma, over te brengen aan anderen.

Referenties

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Determining the Quality of Competence Assessment Programs: A Self-Evaluation Procedure. *Studies in Educational Evaluation, 33*(3-4), 258-281.

Basinski, A. S. (1995). Evaluation of clinical practice guidelines. *Canadian Medical Association Journal, 153*(11), 1575-1581.

Dijkstra, J., Galbraith, R., Hodges, B., McAvoy, P., McCrorie, P., Southgate, L. et al. (2012). Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education, 12*, 20.

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

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.

Newble, D., Dawson, B., Dauphinee, D., Page, G., Macdonald, M., Swanson, D. et al. (1994). Guidelines for assessing clinical competence. *Teaching and Learning in Medicine: An International Journal, 6*(3), 213 - 220.

Prochaska, J. O., Wright, J. A., & Velicer, W. F. (2008). Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model. *Applied Psychology: An International Review, 57*(4), 561-588.

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 psychometric models in educational assessment. *Medical Education, 40*, 296-300.

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

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 practice & research. Clinical obstetrics & gynaecology, 24*(6), 703-719.

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.

Dankwoord

Acknowledgements

Dankwoord

Toen ik met de bijna definitieve versie van het manuscript thuiskwam, met daarin de laatste tekstuele opmerkingen en verbeteringen, zei Ian: “Nu snap ik waarom het 7 jaar moest duren.” Ian was op dat moment ook 7, dus ik heb zijn hele leven tot dan toe eraan gewerkt. Werken aan een proefschrift heeft invloed op je hele gezin en andersom, zeker als het gezin groeit tijdens het promotietraject. Ian en Lars hebben voor voldoende afleiding gezorgd tussendoor, afschakelen geen probleem. Jeremy op zijn beurt heeft me goed wakker gehouden tijdens de laatste eindsprint. En met Patricia, wetend hoe het is om aan een proefschrift te schrijven, als mijn steun en toeverlaat, zonder jou was het proefschrift er niet gekomen. Patricia, Ian, Lars en Jeremy… dank voor jullie liefde.

Maar het begon nu bijna 9 jaar geleden met de vraag van Lambert of ik aan een of ander toetsproject voor buitenslands gediplomeerden wilde komen werken. Na een project over onderwijskwaliteit indicatoren wist ik zeker dat daar mijn promotie-onderzoek *niet* over moest gaan en in een discussie met Cees en Lambert, kwam een uit een lade een A4-tje met een onderzoeksvoorstelletje… met als doel *“The identification of a model for assessment programmes. Using the model as a tool for the actual design of assessment programmes.”* Het bleek iets complexer dan aanvankelijk gedacht van 6 onderdelen in het begin naar 73 guidelines aan het einde. Cees en Lambert, jullie gaven mij van af het begin het vertrouwen dat ik nodig had. Misschien hebben jullie tijdens de rit jullie twijfels gehad, maar daar heb ik niets van gemerkt. Lambert, Cees… dank voor deze kans en vooral jullie vertrouwen.

Het project vorderde langzaam maar gestaag en kreeg een extra kwaliteitsimpuls toen Lambert naar Australië vertrok. Nee, niet omdat hij wegging, maar omdat Marjan aan het begeleidingsteam werd toegevoegd. Dat hadden we veel eerder moeten doen. De vanzelfsprekendheden in de uitwerking van het project werden kritisch tegen het licht gehouden en aangescherpt waar nodig. De impliciete aannames die Cees, Lambert en ik voor vanzelfsprekend hielden, werden blootgelegd. Maar vooral belangrijk de discussies aan het einde van het traject en je meedenken en nog een keer opnieuw meedenken hebben het proefschrift scherper gemaakt. Marjan… dank voor je scherpe en kritische blik.

Wat ik in de afgelopen jaren heb geleerd is dat je bovenal met plezier naar je werk moet gaan. In mijn geval fluitend. Dat plezier hangt voor het grootste deel af van mijn collega’s bij O&O en de mensen met wie ik nauw samenwerk. Aan allen mijn dank voor een prettige werkomgeving, maar een aantal mensen hebben meer invloed gehad dan anderen op dit proefschrift. Ron, hoe heb je het met me volgehouden al die tijd op één kamer? Germiek, Robert, Judith, Arnout, Guus, Paul, we blijven nog wel even samenwerken als het aan mij ligt en heel fijn dat jullie mij bij de laatste afronding van het proefschrift uit de wind hebben gehouden. Mascha, heel kort dan maar op een kamer, maar wel heel krachtig, altijd klaarstaand met wijze raad. Herman, zonder de zondagavondskype zou het werkoverleggen veel ingewikkelder worden. Het is erg verfrissend om met je samen te werken… dank allen en hopelijk zetten we dit nog even voort.

Je bent bij O&O of SHE nooit alleen als je promoveert en verschillende promovendi in verschillende fases van hun promotie traject vormen een troost als het even niet zo loopt als verwacht (een afgewezen artikel of feedback die verschillende kanten opgaat). Maar het is vooral ook gezellig. Als je dan ook nog samen op congres gaat en dan met stel Denen in een kroeg in Trondheim zit, wordt Floris en Joost al snel *Joris and Floost*. Een nieuwe laptop laten staan op het terras in Malaga is ook niet echt handig, maar gelukkig is Jeantine voorstander van goede communicatie ook in het Spaans.

Renée… Bij een van de vele bijeenkomsten / vergaderingen (of misschien was het wel een proefpromotie van ik weet niet meer wie), maakte ik een of andere te lollige, overbodige, droge opmerking. Waarop jij zei: "Met Joost heb je voor elke situatie een tegeltjeswijsheid bij de hand."… met deze vierkante vormgeving als resultaat. De 73 guidelines zou je heel goed als tegeltjeswijsheden kunnen gebruiken. Maar behalve dit ook vooral dank voor je eerlijkheid en oprechtheid.

Het onderzoek had niet kunnen plaatsvinden zonder de experts die meer dan welwillend bereid waren om mee te denken en te discussiëren over wat nu een programma van toetsing inhoud. Maar voor case-studies heb je casus nodig.

A special thanks to NCAS and the people there, who freed time in their busy diaries to help me conduct the in-depth analysis and speaking freely about their assessment programme. Especially Pauline McAvoy for enabeling a look behind the scenes and Marie Bunby for facilitating my NCAS-study and finding interviewees.

Ook bij Huisartsopleiding in Maastricht kreeg ik een kijkje in de keuken met medewerking van verschillende stakeholders. Met name Paul Ram, Jean Muris en Bas Maiburg hebben dit mogelijk gemaakt, maar het zou niet geslaagd zijn zonder urenlange discussies met Angelique Timmerman.

Angelique was ook een van de mede-onderzoekers net als Karlijn Overeem, Liesbeth Baartman en Angelique Timmermans. Het was heel prettig om met mensen die van een afstandje naar het onderzoek kunnen kijken te sparren over de guidelines... Dank voor jullie inzet.

Roy… wie had dat in de zomer van 1997 gedacht? Op de dag dat jij je proefschrift verdedigt, stuur ik het naar de leescommissie. Je krijgt nog een tientje van me.

Last but not least Lilian en Nicky. Waar LOL en hard werken hand in hand gaan. Met veel plezier de OOK samenstellend (als die verschijnt: dat moeten we OOK nog doen!). Maar het probleemoplossend vermogen en meedenken over *vanalles* en *nogwat* is van onschatbare waarde... Dank voor de gezelligheid.

Curriculum Vitae

Joost Dijkstra was born on 22 November 1978 in Sittard, the Netherlands. From 2001 to 2005 he studied Educational Sciences at the Faculty of Psychology of Maastricht University. He specialized in 'management of learning' and wrote his Master thesis on expertise development in 'job one'.

Soon thereafter he started working as a researcher at the department of Educational Development and Research at the faculty of Medicine (now faculty of Health, Medicine and Life Sciences - FHML) at Maastricht University. He first started working on a research project about quality indicators of education, before he switched to the research project about programmes of assessment in January 2007.

Next to the writing his dissertation Joost has been working as assessment coordinator of different test methods and assessment programmes within the FHML; e.g. Case-based Key-feature Test for clinical Reasoning, the Assessment for Foreign Graduate Doctors (under the authority of the Department of Health, Welfare and Cultural Affairs, the Netherlands), Master of Medicine, Bachelor of Medicine) and as a member of the taskforce assessment he has developed as an assessment generalist. Furthermore he has been involved in educational development projects, course coordination and teaching responsibilities for both the FHML and the School of Health Professions Education.

He has done several consultancies and given invited lectures on topics related to assessment. His work has been published in national and international peer-reviewed journals. In 2011 he received the 'Best Scientific Paper Award' at the yearly congress of The Netherlands Association for Medical Education (NVMO), for his paper on the development and validations of guidelines for designing programmes of assessment.

Currently Joost still works at the department of Educational Development and Research as an assistant professor where he will continue doing research, coordinating and developing assessment, and teaching activities

Joost is married to Patricia Jaspers and has three children: Ian, Lars, and Jeremy.

SHE Dissertation Series

The SHE Dissertation Series publishes dissertations of PhD candidates from the School of Health Professions Education (SHE) who defended their PhD theses at Maastricht University. The most recent ones are listed below. For more information go to: www.maastrichtuniversity.nl/she.

Frambach, J.M. (26-03-2014) The cultural complexity of problem-based learning across the world

Hommes, J.E. (26-02-2014) How relations, time & size matter in medical education

Van der Zwet, J. (30-01-2014) Identity, Interaction and Power. Explaining the affordances of doctor-student interaction during clerkships

Watling, C.J. (22-01-2014) Cognition, Culture, and Credibility. Deconstructing Feedback in Medical Education

Winston, K. (12-12-2013) Remediation Theory and Practice: Transforming At-Risk Medical Students

Kamp, R.J.A. (28-11-2013) Peer Feedback to Enhance Learning in Problem-Based Tutorial Groups

Junod Perron, N. (24-10-2013) Towards a learner-centered approach to postgraduate communications skills teaching

Pratidina Susilo, A. (24-10-2013) Learning to be the Patient Advocate
The Development of a Communication Skills Course to Enhance Nurses' Contribution to the Informed Consent Process

Alves de Lima, A. (23-10-2013) Assessment of clinical competence: Reliability, Validity, Feasibility and Educational Impact of the mini-CEX

Sibbald, M. (09-10-2013) Is that your final answer? How doctors should check decisions

Ladhani, Z. (05-07-2013) Competency based education and professional competencies: a study of institutional structures, perspectives and practices in Pakistan

Jippes, M. (01-02-2013) Culture matters in medical schools: How values shape a successful curriculum change

Duvivier, R. J. (12-12-2012) Teaching and Learning Clinical Skills. Mastering the Art of Medicine

De Feijter, J.M. (09-11-2012) Learning from error to improve patient safety

Prescott, L. (09-11-2012) Ensuring the Competence of Dental Practitioners through the Development of a Workplace-Based System of Assessment

Cilliers, F.J. (05-09-2012) The Pre-assessment Learning Effects of Consequential Assessment: Modelling how the Examination Game is Played

Spanjers, I. A.E. (05-07-2012) Segmentation of Animations: Explaining the Effects on the Learning Process and Learning Outcomes

Al-Kadri, H.M.F. (28-06-2012) Does Assessment Drive Students' Learning?

Leppink, J. (20-06-2012) Propositional manipulation for conceptual understanding of statistics

Van Zundert, M.J. (04-05-2012) Conditions of Peer Assessment for Complex Learning

Claramita, M. (30-03-2012) Doctor-patient communication in a culturally hierarchical context of Southeast Asia: A partnership approach

Kleijnen, J.C.B.M. (21-03-2012) Internal quality management and organizational values in higher education

Persoon, M.C. (19-01-2012) Learning in Urology; The influence of simulators and human factors

Pawlikowska, T.R.B. (21-12-2011) Patient Enablement; A Living Dialogue

Sok Ying Liaw, (14-12-2011) Rescuing A Patient In Deteriorating Situations (RAPIDS): A programmatic approach in developing and evaluating a simulation-based educational program

Singaram, V.S. (7-12-2011) Exploring the Impact of Diversity Factors on Problem-Based Collaborative Learning

Balslev, T. (24-11-2011) Learning to diagnose using patient video cases in paediatrics: Perceptive and cognitive processes

Widyandana, D. (19-10-2011) Integrating Pre-clinical skills training in skills laboratory and primary health care centers to prepare medical students for their clerkships

Durning, S.J. (09-09-2011) Exploring the Influence of Contextual Factors of the Clinical Encounter on Clinical Reasoning Success (Unraveling context specificity)

Govaerts, M.J.B. (08-09-2011) Climbing the Pyramid;Towards Understanding Performance Assessment

Stalmeijer, R. E. (07-07-2011) Evaluating Clinical Teaching through Cognitive Apprenticeship

Malling, B.V.G. (01-07-2011) Managing word-based postgraduate medical education in clinical departments

Veldhuijzen, J.W. (17-06-2011) Challenging the patient-centred paradigm: designing feasible guidelines for doctor patient communication

Van Blankenstein, F. (18-05-2011) Elaboration during problem-based, small group discussion: A new approach to study collaborative learning

Van Mook, W. (13-05-2011) Teaching and assessment of professional behavior: Rhetoric and reality

De Leng, B. (8-12-2009). Wired for learning. How computers can support interaction in small group learning in higher education

Maiorova, T. (29-05-2009). The role of gender in medical specialty choice and general practice preferences

Bokken, L. (04-03-2009). Innovative use of simulated patients for educational purposes

Wagenaar, A. (18-09-2008). Learning in internships. What and how students learn from experience

Driessen, E. (25-06-2008). Educating the self-critical doctor. Using portfolio to stimulate and assess medical students' reflection

Derkx, H. (18-06-2008). For your ears only. Quality of telephone triage at out-of-hours centres in the Netherlands

Niessen, Th. (30-11-2007). Emerging epistemologies: making sense of teaching practice

Budé, L. (05-10-2007). On the improvement of students' conceptual understanding in statistics education

Niemantsverdriet, S. (26-07-2007). Learning from international internships: A reconstruction in the medical domain

Marambe, K. (20-06-2007). Patterns of student learning in medical education – A Sri Lankan study in traditional curriculum

Pleijers, A. (19-01-2007). Tutorial group discussion in problem-based learning

Sargeant, J. (21-09-2006). Multi-source feedback for physician learning and change

Dornan, T. (12-06-2006). Experience-based learning

Wass, V. (12-05-2006). The assessment of clinical competence in high stakes examinations

Prince, K. (21-04-2006). Problem-based learning as a preparation for professional practice

Section
1.
2.
3.
4.
a
b
c
d
klet.