Guidelines for designing programmes of assessment


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.

Summary

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.